IBM eServer z900 I/O subsystem

by D. J. Stigliani, Jr.
T. E. Bubb
D. F. Casper
J. H. Chin
S. G. Glassen
J. M. Hoke
V. A. Minassian
J. H. Quick
C. H. Whitehead

The IBM eServer z900 is the first in a generation of future eServers that continues its leadership via a new I/O subsystem with enhancements in capability, performance, configuration management, and qualities of service. Significant features of the I/O subsystem are included to support the 64-bit z/Architecture™ and configuration-management enhancements [e.g., assignable channel path identifiers (CHPIDs) and dynamic channel path management (DCM)]. A 1GB/s self-timed interface (STI), I/O infrastructure, and I/O card cage are described which support the high-bandwidth I/O (e.g., FICON™, Ethernet). These improvements yield enhanced configuration flexibility and connectivity, as well as reliability, availability, and serviceability (RAS), while providing for future bandwidth growth. The various types of I/O ports supported by the IBM eServer z900 platform are also discussed. A common I/O platform is discussed which has been used to provide a uniform, high-bandwidth attachment of industry-standard peripheral computer interface (PCI) cards, while maintaining the leadership functionality and RAS of the eServer zSeries™. A high-density (16-port) ESCON® I/O card has been designed by exploiting IBM advanced CMOS and state-of-the-art fiber optic technologies. Finally, a high-performance, high-density intersystem channel (ISC-3) coupling link I/O card has been developed for the IBM eServer z900 by leveraging advanced technology and packaging techniques.

1. Introduction

The information technology (IT) industry is evolving at a rapid rate to provide customers with enhanced data processing capabilities. This evolution (which some might call a revolution) is being driven by the rapid acceptance of the Internet and Intranet (with its high-bandwidth infrastructure) as required business tools, and by a fundamental need to reduce costs and improve time-to-market and competitiveness. The new business paradigm, along with classical (e.g., transaction and batch processing) and new business applications, has demanded extraordinary price/performance improvements in server technology to meet these business needs. For example, server performance has typically improved by 20–25% per year, with prices remaining relatively flat.

This environment, in turn, requires the input/output (I/O) subsystem to scale in price/performance with the processor capacity. A system balanced between processing capacity and I/O capacity must be maintained in order to effectively support the growing new business initiative applications along with the traditional IT workloads. These new applications (e.g., Internet, data warehousing, rich text, and multimedia) are creating unprecedented demands for storage, which is growing in excess of 50% per year. The requirement for low-latency, high-bandwidth access to storage is fostering enhanced server I/O and new interconnection infrastructures [e.g., storage area

0018-8646/02/$5.00 © 2002 IBM

network (SAN)] far beyond the normal improvements commensurate with higher-performance processors. The IBM eServer z900 contains a new generation of I/O infrastructure which is intended to meet these demands and provide a foundation for future enhancements. Also, the scaling of the IT infrastructure beyond standalone units with Parallel Sysplex* and Geographically Dispersed Parallel Sysplex* technology is crucial to IT environments for both large and medium-sized customers. Not only do they need to scale by using parallel processing technology, but they also require a high-availability processing complex that affords a continuous (24/7) operation and disaster-recovery capability. The interconnection links between central processing units (CPUs) must also scale commensurately with processing improvements.

The IBM eServer z900 I/O subsystem design also focuses on many other areas that are important to our customers. Improvements in cost competitiveness are achieved by leveraging IBM advanced, high-density CMOS technologies and increasing the performance of the individual I/O port. Leadership in connectivity is enhanced by improved configuration management, which enables full utilization of allowed I/O ports, independent of I/O type and card port density, and movement of selected I/O resources to logical partitions (LPARs) as needed. Design enhancements for superior quality of service (e.g., reliability, availability, and serviceability) continue unabated.

The IBM eServer z900 is the first in a generation of future IBM eServers that continues its leadership via a new I/O subsystem with enhancements in capability, performance, configuration management, and qualities of service. The requirements, design, and key features of the I/O subsystem are discussed in this paper. Both I/O subsystem Licensed Internal Code (LIC) and hardware are discussed.

2. Licensed Internal Code enhancements

The new z/Architecture [1] is the latest version of a long series of evolutionary system architectures, providing support for 64-bit addressing, 64-bit operands, and a number of other enhancements. Both the I/O subsystem Licensed Internal Code (LIC) and hardware are enhanced to support the new z/Architecture*. The major I/O subsystem LIC enhancements include 1) support for data buffers anywhere in a 64-bit address space and 2) I/O support for 64-bit logical partitions. Significant enhancements are also made in configuration flexibility and management, as well as availability and serviceability.

z/Architecture support

The IBM eServer z900 I/O subsystem incorporates a number of new features provided by the z/Architecture. A full description can be found in Reference [1]; two of the most significant I/O subsystem changes are discussed in the following sections.

Data buffers within a 64-bit address space

A key requirement of the I/O subsystem is to provide support for data transfer to or from any absolute memory address within the 64-bit address space provided on the IBM eServer z900. I/O data transfers are performed in one of two ways:

1. For a traditional channel-attached device, the I/O operation is executed according to a channel program that is constructed by software and initiated by the execution of the *start subchannel* (SSCH) instruction. An operand of the SSCH is the *operation request block* (ORB) instruction, which describes the starting address of the channel program in memory and other dynamic parameters that are in effect for the I/O operation.
2. For internal networking adapters, a shared-queue structure, implemented in the main storage of each LPAR, is used to pass packets to and from the adapter.

The main difference between the two methods is that for trusted adapter platforms, implemented within the machine boundary, efficiency can be gained by sharing memory and eliminating the need to perform a SSCH and I/O interrupt for every I/O operation. The queued direct I/O (QDIO) interface, initially designed to meet the 64-bit requirements, has been enabled in the z/Architecture and IBM eServer z900. Its queue structure supports the extended addressing with complete 64-bit structures.

For channel-attached I/O, the IBM eServer z900 has a requirement to retain support for older existing channel programs as well as supporting the new z/Architecture. A channel program consists of one or more channel command words (CCWs) chained together in main storage. Each eight-byte CCW, aligned on an integral boundary, contains a command, count, address, and flags. CCWs typically reside in consecutive memory locations, with limited support for skipping and unconditional branching provided by special commands. Prior to the IBM eServer z900, two CCW formats were provided: the original S/360 format limited to 24-bit addresses, and the 31-bit format introduced with S370/XA. Support for both variations of this 8-byte construct is retained. In order to provide support for noncontiguous real memory, an additional construct called an indirect address word (IDAW) list is also provided. The IDAW list allows scattering of data in memory for noncontiguous real pages.

To provide channel program addressing for a 64-bit address space while not disturbing compatibility and adding unnecessary cost associated with transition to a larger CCW, a new Format 2 IDAW (the prior version of IDAW was called Format 1 IDAW) is introduced.

The Format 2 IDAW is a 64-bit construct that provides support for page boundary crossings on either 2KB or 4KB boundaries. This allows support for data buffers anywhere in a 64-bit space, while retaining support for all other channel program modes. A single channel program is limited to either Format 1 or Format 2 IDAWs. The operation request block (ORB) has also been extended to include a number of new features, including indications describing the IDAW format and the page boundary crossing that are in effect for a specified channel program.

64-bit logical partitions

The z/Architecture provides a facility (the zone relocation facility) that allows the I/O subsystem to relocate storage addresses for one or more logical partitions. This facility allows each logical partition to be mapped into IBM eServer z900 storage such that it appears to have its own unique memory space starting at absolute address zero. Each logical partition is assigned a storage relocation zone, which is associated with a zone identifier, a zone origin, and a zone limit specifying the machine addresses into which the LPAR memory space is mapped. I/O operations which are associated with a zone identifier and all main storage accesses made by the I/O subsystem for a logical partition are relocated to the absolute machine address. The capability exists to dynamically assign or change these parameters for support of dynamic storage reassignment. The IBM eServer z900 provides support for storage zones up to the capability of installed memory, with the zone located anywhere within a 64-bit absolute address space.

Configuration management

The IBM eServer z900 has incorporated several enhancements that provide improved configuration management and flexibility, and the ability to dynamically manage I/O resources and workloads, as discussed below.

Assignable channel path identifier

A channel path identifier (CHPID) is a one-byte identifier that is both an external (customer) and internal (software) name used to identify a channel path. For S/390* systems prior to the IBM eServer z900, CHPIDs were assigned to I/O card slots in a static and predetermined uniform manner. This had the advantage of simplicity in that the physical locations of the channel paths were clearly defined and static, based on system configuration. In addition, labels identifying the CHPIDs could be placed close to the I/O cable attachment point. This approach was successful for many years, but as machines became larger, more efficient use of CHPIDs was required. Two items were particularly important:

1. The I/O card types increased from only ESCON* and parallel channel to also include FICON* (the IBM eServer z900 version of Fibre Channel [2]), networking, cryptographic processors, and intersystem coupling (ISC) card types with varied port densities. The method of static assignment is no longer acceptable, since the range changed from 2 to 16 port increments with this complement of IBM eServer z900 I/O cards.
2. For some machine upgrade options, it was preferred for IBM to reassign the CHPID numbers that were previously assigned to a channel. Also, for some customers it would ease migration to newer IBM eServers if CHPID assignments were capable of being ported to the replacement machine. For the IBM eServer z900 system, these issues are addressed by providing the customer an option to selectively assign CHPIDs to physical channels as desired. This is the user's choice, however; IBM still provides appropriate default mappings for customer use if desired. In both cases, the definition in effect is preserved during system upgrades.

The assignable channel identifier is provided by the IBM eServer z900 service subsystem, which maintains a mapping table that relates physical location to CHPID. This mapping table is nonvolatile and is preserved during the life of the installation. The mapping table is updated as required by the default algorithms during channel adds/deletes or by user actions. User changes to the mapping table are made via facilities provided on the service subsystem or the hardware master console. Changes are supported concurrently with system operation, with the targeted channels configured in service mode. A historical record of changes is kept by the service subsystem for user convenience and to allow correlation of events that may have occurred prior to the reassignment. A new display panel is provided on the service display to present the mappings in various ways and to assist the customer in determining the I/O cabling location in lieu of CHPID labels.

Channel paths activated by Licensed Internal Code

As channel card densities increase, it is desirable to allow a customer to purchase channels in smaller increments than are packaged on a card. In particular, the introduction of the IBM eServer z900 16-channel ESCON card provides a hardware port increment of 16, whereas the desired customer offering is a four-channel increment. This would require a larger purchase than desired. The situation is accentuated if the customer chooses to configure the channels in different failure domains, requiring the use of multiple cards. Having channel paths activated by Licensed Internal Code (LIC) [3] addresses this issue; it maintains the four-channel purchase

increment and allows IBM to fill the order for availability (different domains), while reserving a number of channels that are immediately available for future activation when required by the customer.

Each I/O card contains a secure globally unique identifier stored on a card in vital product data (VPD). The individual channels on a card are made available to the customer according to a secure profile associated with the identifier. The identifier, profile, and LIC work in conjunction with the customer order process to fulfill orders without the actual delay associated with adding parts. If the order can be satisfied in this manner, a new profile is provided for the machine that allows the channels to be activated concurrently with machine operation. This process is currently performed by IBM service personnel who perform other installation duties required to complete the customer request, including I/O cabling and diagnostic checkout.

Dynamic channel path management

Dynamic channel path management (DCM) [4] enables the system to respond to dynamically changing channel requirements by reconfiguring channel paths from less heavily used I/O attachments to those more heavily used. It is one of two important new I/O features of the intelligent resource director (IRD). Channel paths may be dynamically assigned to match the workload priorities with the I/O load. For example, in an environment in which an installation normally requires four channels to several attachments, but occasionally needs as many as six, the user must currently define all six channels to each control unit that may require them. DCM allows the customer to define four channels to the attachments and indicate that DCM may add an additional two. As the attachment becomes more heavily used, DCM may assign channels from a pool of managed channels, identified by the user, to the attachment. If the work shifts to other attachments, DCM will reconfigure them from the less heavily used attachments to what are now the more heavily used ones. Overall, this allows for more efficient use of channel paths, reducing customer cost and providing relief for the 256-channel path limit. The IBM eServer z900 I/O subsystem supports this process via the dynamic I/O reconfiguration, and via measurement facilities provided in z/Architecture and its supporting elements.

Channel subsystem priority queueing

Channel subsystem priority queueing [4] (CSSPQ) is the other important I/O feature of IRD that further extends the IBM eServer z900 lead in managing multiple heterogeneous workloads with various business priorities to achieve their goals. While DCM was designed to cope with shifting heavy workloads, CSSPQ adds support for smaller variations in workload by assigning priorities to I/O requests which are then used by the IBM eServer z900 to schedule the work to resources within the I/O subsystem. Higher-priority workloads which are running in a logical partition can be given higher I/O priority; the I/O request will be sent to the attached I/O devices ahead of lower-priority workloads. In this way customers may identify workloads that are the most critical, and z/OS* can work with the processor subsystem to allow the critical work greater access to I/O subsystem resources.

ESCON channel path sparing

Channel path sparing is an availability and serviceability enhancement of the z/900 I/O system that allows a failing channel to be logically replaced by one or more channels that are candidate spares. This feature allows the actual replacement of the failing card to be deferred indefinitely or until a convenient time that has minimal customer impact. As card port densities increase, both the requirements and the economics change to allow this capability. Currently, this feature is provided only for the ESCON channel card.

Each 16-channel ESCON card reserves at least one channel as a spare candidate to be used for substitution of any customer-available channel on the card. If other channels are available on the card that have not been purchased by the customer, these too may be used as spare candidates. When an ESCON channel failure occurs, a service call is scheduled. If a spare candidate is not available, service personnel will replace the card; otherwise, they are directed to perform only the sparing action. Repair procedures are provided via the service subsystem to move the CHPID assignment from the failed channel to the new channel and validate the new channel. The I/O attachment cable must be moved from the failed channel port to the replacement channel port. LEDs are located next to the channel ports to aid in identifying the two physical ports involved in the service action. After sparing is complete, the spare candidate has been assigned the CHPID value of the failed channel, and this physical channel remains as the permanent replacement for this CHPID unless other events cause it to be relocated.

3. Hardware overview

The new I/O subsystem and infrastructure is the foundation for the high-performance I/O required to support the IBM eServer z900 applications. It is the foundation which enables the IBM eServer zSeries* to maintain its I/O performance leadership capability at a significantly improved cost, and it is the basis for future I/O subsystem enhancements.

The IBM eServer z900 hardware I/O subsystem consists of several LIC enhancements (discussed in the previous section), an improved I/O signal-distribution infrastructure from the processor cage to the I/O card, a new mechanical

cage, and dense I/O cards and book packages. The IBM eServer z900 hardware I/O subsystem takes advantage of the latest IBM CMOS technology, state-of-the-art fiber optic technology, high-speed connectors, high-speed copper coaxial technology, and improved packaging. The hardware I/O subsystem, with I/O cards, has the following significant attributes:

- *Bandwidth*
  The I/O subsystem contains an I/O infrastructure bandwidth three times larger than that of the prior machine. This infrastructure will support two 1Gb/s or two 2Gb/s channels per I/O card (e.g., FICON and Ethernet).
- *Package density*
  In terms of I/O ports, one IBM eServer z900 I/O cage, with the new-generation I/O cards discussed in this paper, is equivalent to three prior machine I/O cages. The effect of this improvement in density is to allow customers the same or improved capacity with fewer I/O cards.
- *Cost*
  The I/O subsystem costs significantly less than the prior-generation S/390 platform because of the use of advanced technology, performance improvements, and increased package density.
- *Growth*
  The high-speed self-timed interface (STI) between the I/O card and the host is capable of speed selection based on the I/O card requirement. This flexibility enables the bandwidth offered to the I/O card to match the I/O card requirements and enables the bandwidth to be increased for future eServer zSeries platforms. The I/O infrastructure and I/O have been designed to allow future I/O cards to exploit this enhancement capability.
- *Reliability, availability, and serviceability (RAS)*
  All cards in the I/O subsystem, including the cage power and cooling structure, allow for concurrent repair or upgrade. The LIC enhancements were discussed earlier.

I/O subsystem hardware structure overview

The I/O subsystem infrastructure begins with a high-speed data bus emanating from the processor multichip module (MCM) [5]. The high-speed data bus is fanned out to the I/O cards in the IBM eServer z900. The infrastructure and its attributes are discussed later in this paper.

I/O cage description

The I/O cage provides high-bandwidth I/O slots to enable a higher number of I/O ports per card, as well as high-bandwidth ports (e.g., Gb/s) at a substantially reduced cost. The IBM eServer z900 I/O cage offers six additional I/O slots (receptacles for I/O cards), with an effective slot bandwidth improvement of approximately 20× compared to the previous I/O cage.

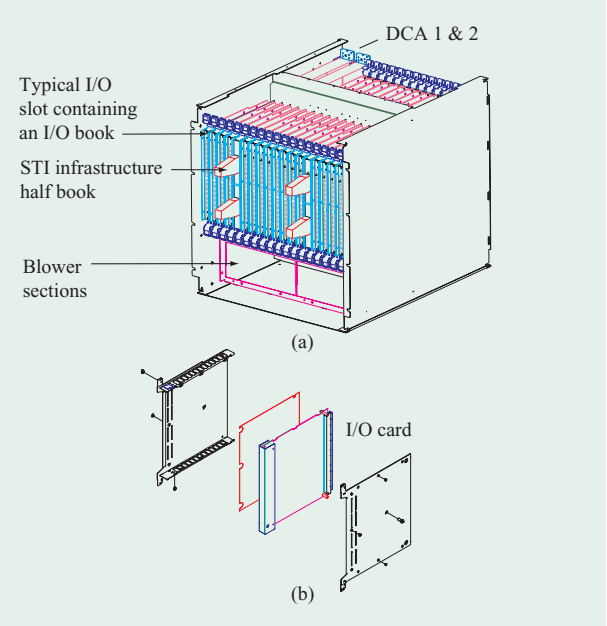


Figure 1

IBM eServer z900 I/O cage (a) and I/O book package (b).

The I/O cage is a double-sided unit which contains 35 slots [see **Figure 1(a)**]. Four slots are predefined for STI infrastructure cards, three slots are predefined for the cage power subsystem, and 28 slots are available for I/O cards. Each I/O slot accommodates a book package [**Figure 1(b)**] compatible with a card size of 11-in. width and 13.42-in. depth. There are 16 I/O book slots in the front and 12 in the rear. All I/O cards (including multiplexor/demultiplexor infrastructure cards) are encased in a metallic book to provide shielding for electromagnetic compatibility (EMC). All bezel plates of the I/O books are flush when plugged into the I/O cage, enabling EMC shielding at the cage level and access to the fiber optic connectors on the card edge. However, the STI infrastructure cards are 12.42 in. deep to allow the STI top card connector to be recessed approximately an inch. The recess allows improved fiber optic connector space for I/O books adjacent to the multiplexor/demultiplexor slots. The unique PSC-24 card indicated in the bottom of slot 28 is used to turn on/off various attached devices from the IBM eServer z900 system. It is not widely used and is not discussed in this paper.

The I/O cage uses a "light strip" of light-emitting diodes (LEDs) to indicate a slot position and the field-replaceable unit (FRU). For most I/O cards, the FRU is the I/O subsystem book in the slot. This is used to aid in

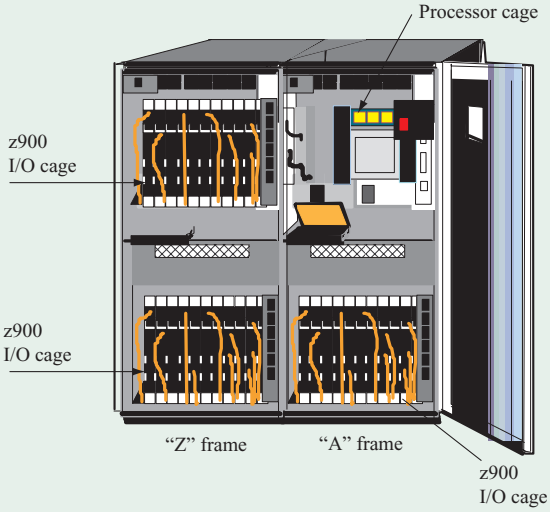


Figure 2

IBM eServer z900 I/O system (front view).

serviceability and dynamic installation of additional I/O cards. The light strip is controlled by the service subsystem. In some instances, however, there is more than one FRU within a slot (e.g., two multiplexor/demultiplexor cards in a single slot and two ISC-3 adapter books within a base book). In this case, the FRUs will have LED indicators to direct the service personnel to the correct FRU for replacement. The FRU LEDs are also controlled by the service subsystem.

The power subsystem for the cage consists of two fully redundant distributed converter assemblies (DCAs), which convert the bulk power to the appropriate logic-level voltages and power controllers. The power subsystem structure allows concurrent maintenance of each DCA within the I/O cage without compromising the operation of the I/O cards. The voltages available for I/O cards are 1.8, 2.5, 3.3, 5, and 24 V.

System configuration

The IBM eServer z900 maximum I/O configuration is illustrated in **Figure 2**. The cage below the processor cage, in the "A" frame, is always present, even in the smaller models. For larger IBM eServer z900 models, a maximum I/O configured system may contain two additional I/O cages in the expansion ("Z") frame, for a maximum total of three I/O cages. The Z frame can contain two cages, with any mix of the IBM eServer z900 cage and an I/O cage from a prior machine generation (prior I/O cards that are not compatible with the new I/O cards), depending on the quantity and type of I/O cards used. A customer planning to install previously owned I/O cards in the eServer z900, for example, will require prior-machine-generation cages in the Z frame. This enables a customer to protect his existing I/O investment. The implementation details of this attachment are not discussed in this paper.

I/O port types supported and migration

I/O book types supported in the IBM eServer z900 machine are illustrated in **Table 1**. ESCON channels are supported in both the new and the prior I/O cage. The new IBM eServer z900 I/O cage supports the new 16-channel ESCON card (ESCON-16), whereas the prior I/O cage supports only the four-channel ESCON card (ESCON-4). This allows a customer to move any existing ESCON-4 cards to IBM eServer z900 while adding any new ESCON requirements in the new technology (ESCON-16).

Parallel channels are supported only in the prior-generation I/O cage for eServer z900. The parallel channel was originally manufactured for the S/360 machine; it is rapidly being replaced by serial [e.g., Enterprise Systems Connection (ESCON), Fiber Connection (FICON)] I/O channels. Both the three-channel and four-channel parallel cards will be accommodated. Customers who already have parallel channels must retain at least one prior-generation cage.

The IBM eServer z900 will also support Fiber Distributed Data Interface (FDDI) and Ethernet (10 Mb/s)/Token Ring I/O cards in the prior-generation I/O cage for existing cards on upgrades only. Both Token Ring and Ethernet ports are offered in the IBM eServer z900 I/O subsystem and cage, using a new common technology platform. All other I/O cards [the cryptographic peripheral computer interface (PCI) is considered an I/O card for the purpose of this paper] will be shipped in the new IBM eServer z900 I/O subsystem cage (shown in Table 1). For upgrades, old I/O cards (Parallel Channel, Ethernet/Token Ring, and FDDI), except for ESCON, may be exchanged with equivalent new I/O subsystem cards. This will allow customers to immediately take advantage of the new I/O structure and cost/performance improvements, and will enable future growth. The third-generation intersystem channel card (ISC-3) is available only in the new IBM eServer z900 I/O subsystem. ISC-3 is backward-compatible with the prior-generation ISC ports and can be used to attach S/390 platforms to the IBM eServer z900 in a Parallel Sysplex. A new common I/O platform card has been designed to enable attachment of PCI I/O adapters, as listed in Table 1.

New I/O cards

The IBM eServer z900 I/O subsystem includes the development of three new low-cost, high-density I/O cards

Table 1 I/O cards supported by IBM eServer z900.

I/O card type	*IBM eServer z900 I/O cage*	*Prior platform I/O cage*
ESCON-4 (existing)	No	Yes
ESCON-16	Yes	No
Parallel Channel (existing)	No	Yes
Fiber Distributed Data Interface (existing)	No	Yes
Ethernet/Token Ring (existing)	No (replaced with common platform version)	Yes
ISC-3 (1Gb and 2Gb)	Yes	No
Common I/O platform		
FICON (1Gb and 2Gb) SX/LX laser*	Yes	No
Fast Ethernet (10/100 Mb)	Yes	No
1Gb Ethernet SX/LX laser	Yes	No
ATM (155Mb) SX laser only	Yes	No
Token Ring (4/16/100 Mb)	Yes	No
Cryptographic PCI	Yes	No

*Supports FICON or FCP for Linux**. SX = short-wavelength version; LX = long-wavelength version.

to support IBM eServer z900 and future zSeries machines. The new I/O cards (for which the technical design is discussed later in this paper) for the IBM eServer z900 are the following:

- *Common I/O platform*
  The common I/O platform (CIOP) card is intended to provide a unified logical and physical interface to the IBM eServer z900 for industry-standard PCI I/O adapters. All high-speed storage area network (SAN) attachments (e.g., FICON, Fibre Channel), networking attachments (e.g., Ethernet, Token Ring, Asynchronous Transfer Mode (ATM), and cryptographic PCI applications are designed to utilize the common platform. The common platform also provides the essential functions of a firewall for I/O attachment networks, logical host partitioning support, access authorization, and error recovery in conjunction with the IBM eServer z900 software.
- *ESCON-16*
  A new high-density ESCON-16 card is designed with 16 ESCON ports (15 active, one spare). The ESCON port is fully interoperational and compatible with existing ESCON ports. The fiber optic technology is a state-of-the-art small-form-factor (SFF) application-specific integrated circuit (ASIC) and connector that enables 16 ports on the card edge. The fiber optic transceiver and connector technology is an industry standard that is common with other low-speed ASICs (e.g., ATM). Cable connector conversion harnesses are available to allow connection to the IBM Fiber Transport Services (FTS) cabling system. In addition, a connector conversion cable is available for customers who prefer to connect to the existing ESCON fiber optic cable plant.
- *ISC-3*
  An enhanced four-port ISC-3 card is included in the IBM eServer z900 I/O subsystem. It consists of a base card with two types of adapter card. Each adapter card contains two ISC-3 ports. The 1Gb/s link will be fully compatible with prior 1Gb/s ISC versions. The ISC-3 card provides improved cost, link utilization, and performance, particularly at long distance (e.g., 10 km). The ISC-3 links use state-of-the-art small-form-factor (same optics package as ESCON) fiber optic technology to maintain S/390 technical leadership in the application of fiber optics to interprocessor communication.

4. I/O interconnection infrastructure

The I/O interconnection infrastructure comprises an interconnection network beginning with the self-timed interface (STI) on the memory bus adapter (MBA) chip in the multichip module (MCM) [5]. There are a total of 24 primary STIs (six per MBA). The STI is full-duplex and operates at a baud rate of 1 GB/s in each direction. The I/O interconnection infrastructure distributes the high-bandwidth I/O capability at the processor (24×2 GB/s, full-duplex) MCM to the multiplicity of I/O cards associated with the IBM eServer z900 I/O cage. The various I/O cards have differing bandwidth requirements based on the type (e.g., ESCON, Ethernet) and the number of ports per card. The I/O infrastructure provides the bandwidth and connectivity needed by each I/O card by using a multiplexor/demultiplexor (mux/demux) ASIC as the central hub and fan-out mechanism of the network. The primary STIs emanating from the processor complex are directly connected to the mux/demux ASIC, which in turn generates four lower-speed secondary STI data interfaces to attached

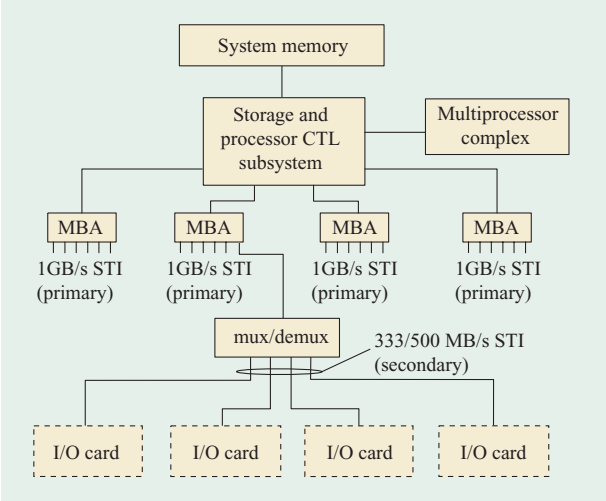


Figure 3

I/O interconnection infrastructure.

I/O cards. Therefore, the 24 primary STIs can be fanned out into 96 secondary STIs. **Figure 3** illustrates the logical interconnection infrastructure of one of the primary STI paths from the processor MCM. The other STIs are treated in the same manner for I/O. This approach allows I/O configuration granularity (individual I/O card types and quantity determined by customer), uniform capability per I/O slot, and enhanced RAS through I/O cards having different paths to the processor complex.

Self-timed interface (STI)

A new STI was invented to meet the configuration, distance flexibility, and speed requirements of the IBM eServer z900, and to reduce cost and complexity compared to prior-generation infrastructure. The STI link is point-to-point full-duplex, with an operational length from 0.01 to 10 meters. The link supports up to 15 transactions at a time and operates at three different link speeds: 333, 500, and 1000 MB/s independent of the host cycle time. Error checking and retry are provided at the STI link level. Source and destination addressing allows switching and routing capability. The physical interface is 10 bits wide in each direction and is self-timing, with the capability of realigning up to three bit-times between any two wires of the interface. In general, a channel or an I/O card is the requester making a request or multiple requests to an element (responder) in the IBM eServer z900 via the STI. Most of the time the request is to a storage element in the system to either store data or fetch data. The responder then responds with the data for a fetch request or status indicating that the store request was successful or that the fetch or store request was not successful. The request and response travel end to end through one or two mux/demux ASICs plus the MBA. The STI protocol is a point-to-point protocol that is confined to the two adjacent ports connected by an STI link. In addition to this protocol, there is the end-to-end requester–responder protocol, which is mostly unchanged from prior CMOS machines.

STI link protocol overview

The STI link protocol is a credit-based protocol in which the credit is established at link initialization time. The send protocol logic unit selects a buffer ID to send a packet and keeps track of the acknowledgments (ACKs) for each buffer ID. Each buffer ID can have only one ACK outstanding at a time. The packet or the returning ACK may be lost because of some error condition on the STI link. If one or the other gets lost, the sending port times out and asks the ACK for the last good packet received for that buffer ID. The receive protocol logic unit responds with an ACK for the last packet that was successfully received for that buffer ID. The sending port then determines whether the ACK was for the last packet sent or the previous packet for that buffer ID. If the ACK was for the last packet sent for that buffer ID, the link is recovered. If the ACK was for the previous packet sent for that buffer ID, the sending logic unit re-sends the last packet sent for that buffer ID.

STI logic units

The STI link port functions are shown in **Figure 4**. The STI port at each end of the link contains an STI transmit logic unit, a buffer unit, an STI outbound physical logic unit, an STI inbound physical logic unit, and an STI receive logic unit.

The STI transmit logic unit provides the interface between the host logic and the STI link. This logic unit controls the loading of the outbound buffers from the host logic using the host clock. On the other side of the buffer, this logic unit contains all of the controls for forming an information packet and transmitting it from the buffer onto the STI link. This section of the logic is synchronized to the STI transmit clock.

The buffer unit buffers the STI packets and also provides the synchronizing interface between the host clock section and the STI clock section. The buffers are two-port arrays; one port is the write port in one clock section and the other is the read port in the other clock section.

The STI outbound physical logic unit serializes a word-wide dataflow from the transmit buffer under control of the STI transmit logic unit into a byte-wide dataflow. This data is transmitted along with the transmit clock on the STI link [6, 7].

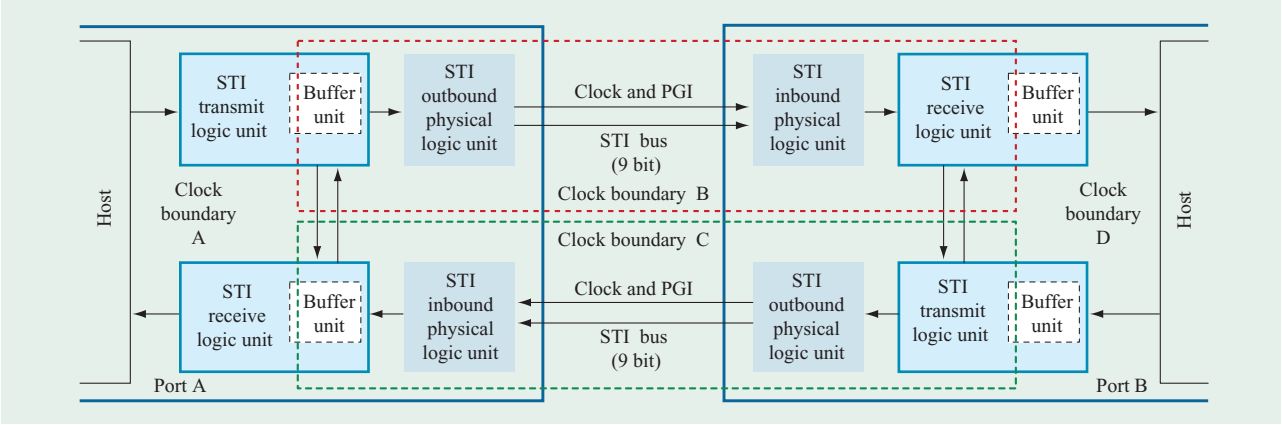


Figure 4

STI link port functions.

The STI inbound physical logic unit dynamically aligns each of the nine data bits (8 + 1) with the received STI clock, de-skews the data bits into bytes, and deserializes the bytes into words [6, 7]. These words are then loaded into the receive buffer under control of the STI receive logic unit.

The STI receive logic unit controls the loading of the inbound buffer from the STI inbound physical logic unit using the clock received on the STI link. On the other side of the buffer, this logic unit has the controls for the host logic to read the data out of the buffer using the host clock. This logic unit, in conjunction with the STI transmit logic unit, controls all of the STI link protocols.

Format of the transmitted data

Information is passed across the link as information packets separated by at least one link control word (packet separator). Information packets are used to transmit commands and data. The length of the information packet is measured in units of 8 bytes (two words), and is composed of two parts, the header block and an optional data block (**Figure 5**).

The header block is further subdivided into a link header, a data header, and an extended data header, as shown in Figure 5. The link header contains a source address, a designation address, a data count, and a control field. The data header contains requester-to-responder or responder-to-requester control information. The extended data header, if present, contains information transported by the STI protocol from the requester to the responder. At the end of the header block is the header block longitudinal redundancy check (LRC) word. This is a 4-byte field generated by the STI transmit logic unit and checked by the receive logic unit to make sure that the header block was delivered error-free.

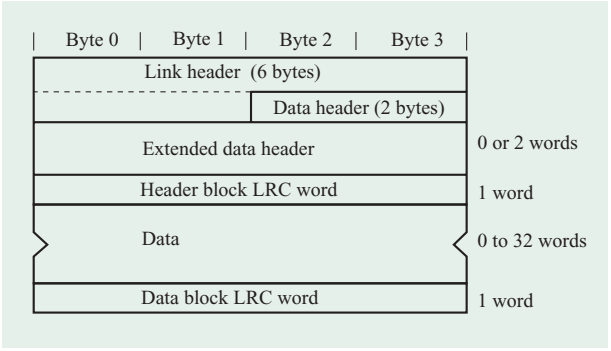


Figure 5

Fields of the STI information packet.

The data block (see Figure 5) is transmitted after the header LRC. The data block length is indicated in the data count field in the link header. The data block can be from 0 to 128 bytes, in 8-byte increments. The information in the data block is transported by the STI protocol from the requester to the responder or from the responder to the requester. The data block also contains an LRC word that is a 4-byte field generated by the STI transmit logic unit and checked by the receive logic unit to make sure that the data block was delivered error-free.

STI multiplexor/demultiplexor ASIC

The STI multiplexor/demultiplexor (mux/demux) ASIC is designed to fan out a single-input STI link (primary link) into four independent output STI links (secondary links) while providing the speed matching necessary to deliver

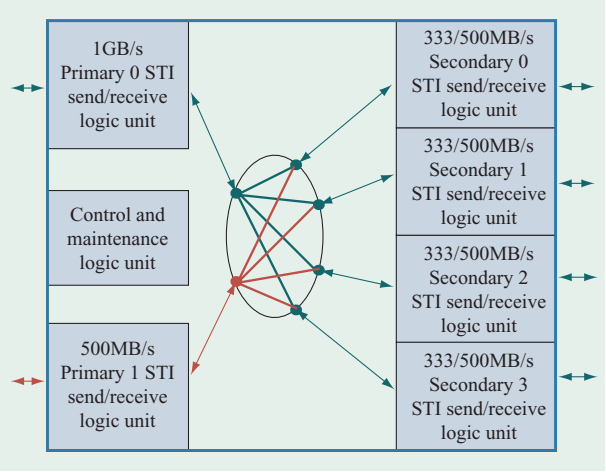


Figure 6

Mux/demux ASIC high-level block diagram.

the optimum bandwidth to each I/O card. **Figure 6** illustrates the high-level block diagram of the mux/demux ASIC. The mux/demux ASIC contains six STI ports, two primary ports (0, 1) and four secondary ports (0, 1, 2, 3). The primary 0 port operates at a data rate of 1 GB/s and the primary 1 port operates at 500 MB/s. The four secondary ports operate at either 333 MB/s or 500 MB/s. Only one of the primary interfaces may be configured as active, while all secondary interfaces may be active simultaneously. Data from the active primary port is multiplexed across the secondary ports as determined by the addressing in the STI protocol. The higher-speed primary 0 port is used for connection to the processor complex via the memory bus adapter chip in the MCM. The speed of the secondary port is configured by the IBM eServer z900 service subsystem on the basis of the connected I/O card type. Each secondary port is configured independently.

Another important aspect of the mux/demux ASIC is the ability to cascade two mux/demux ASICs to provide additional fan-out capability. This is accomplished by attaching one of the secondary ports of the first mux/demux ASIC to the primary 1 port (Figure 6) of the second ASIC. This feature enables a single primary STI from the processor complex to directly connect up 16 (4×4) I/O ports. For example, this capability is exploited for the ISC card (discussed later in this paper) by placing a mux/demux ASIC on the ISC card to provide direct STI connection to four ISC ports.

I/O cage interconnection

The connection from the processor complex to the I/O ports in the I/O cage is accomplished via the mux/demux ASICs. The mux/demux ASICs are contained on mux/demux half books (special books which are half the height of a standard I/O book, enabling two special books to fit in special book slots in the I/O cage (see Figure 1). A mux/demux card pair is illustrated in **Figure 7(a)**, and the I/O cage and board configuration structure is illustrated in **Figure 7(b)**. The primary 0 port is connected to the processor complex via a top card connector and a parallel copper cable. The secondary ports are wired from the mux/demux card to the I/O slots (point to point), through the I/O board, as shown in Figure 7(b). Each mux/demux card connects to four I/O slots. The I/O cage contains seven domains (one domain per 1GB/s STI connection). There are seven mux/demux cards (corresponding to the seven STI domains) per I/O cage. Each mux/demux slot and the corresponding four I/O slots are indicated by slots labeled A through G in Figure 7(b). The unique PSC-24 card indicated in the bottom of slot 28 is used to turn on/off various attached devices from the z900. It is not widely used and is not discussed in this paper. The I/O bandwidth capability into the cage is 14 GB/s, full-duplex (7 GB/s in each direction). All slots have equal bandwidth capability, and I/O books are generally plugged to yield maximum availability (spread across STI domains).

Performance extendibility is a fundamental aspect of the I/O infrastructure. The design parameters have been chosen such that the four secondary STIs from the mux/demux ASIC have a maximum aggregate data rate (4×500 MB/s) which is twice the input primary 0 STI data rate. Consequently, the secondary links, in general, run at 50% utilization, which is more than adequate for the IBM eServer z900 I/O cards. As future I/O cards require higher data throughput, it can be provided by increasing the primary and secondary STI data rates while maintaining data-rate compatibility with older I/O cards. For example, a 3+ GB/s STI link from the MBA to the primary 0 port and a secondary STI capability of 333, 500, or 1000 MB/s would allow for higher-speed I/O cards (e.g., a 10Gb/s Ethernet card) to be utilized in the I/O cage by enhancing the mux/demux ASIC and card. New I/O card designs may immediately take advantage of the higher bandwidth while maintaining backward compatibility with all existing cargo I/O cards. The I/O cage board wiring has been designed for the higher speeds.

The 1GB/s STIs in the IBM eServer z900 are also used to couple IBM eServer z900 to IBM eServer z900 for Parallel Sysplex using an integrated cluster bus (ICB-3) protocol. Also, support of coupling to prior-generation S/390 machines and attachment of a S/390 I/O cage are done via the mux/demux housed on special cards in the processor cage. These features are not discussed in this paper.

5. Common I/O card platform

The IBM eServer z900 platform has evolved from relatively few IBM unique I/O channels to a platform that provides support for key industry-standard and emerging attachments. It is expected that the IBM zSeries will continue to support these technologies as the market demands. Most of these technologies are typically supported by standards and vendors that react quickly with solutions offered as peripheral computer interface (PCI) adapter cards. The use of industry-standard PCI-based solutions in eServer zSeries machines enables improvements in development efficiency, allowing timely support for a broader range of industry attachments (e.g., Ethernet, Fibre Channel, Token Ring) directly benefiting the customer.

The direct use of PCI cards while maintaining the zSeries level of functionality and quality of service presents several technical challenges, a few of which are described here:

- *Adapter bus attachment characteristics*
  The STI, which is used to attach z/900 I/O cards, provides a point-to-point, high-bandwidth, independently clocked, split-transaction (packet-switched), and limited-transfer-size bus. These characteristics are well suited to support the IBM eServer z900 requirements for concurrent installation, replacement of adapters, effective fault isolation, and excellent bandwidth sharing between adapters. This is in contrast to the PCI bus, which is arbitrated, multi-drop, and synchronously clocked, with potentially long and unpredictable data-transfer sizes. In addition to the differences stated above, the architectural memory model on the IBM eServer z900 includes checking and reporting for program exceptions, storage errors, and machine errors that PCI does not address. Bus adaptation is required to incorporate PCI adapters into the IBM eServer z900 I/O structure.
- *zSeries architectural adaptation*
  The IBM eServer z900 I/O programming model, introduced with S/360, has resulted in significant software investment that must be preserved. For channel-based applications, such as FICON, this architecture includes channel programs consisting of channel command words (CCWs), indirect address words (IDAWs), and I/O devices that are represented by objects called subchannels, as well as all of the necessary instructions and other mechanisms used to initiate, terminate, report status, and present I/O interruptions to the program. PCI cards have no knowledge of these requirements. The PCI-based architecture is centered around a memory-mapped I/O model. PCI cards are available that provide the FC-0 to FC-2 [8] layer protocol (e.g., Fibre Channel adapters used for FICON-based applications), required to communicate with a device, but these cards are not directly compatible with the z/Architecture. For networking-based applications, the channel model has been largely replaced by a new, highly efficient shared memory interface called queued direct I/O (QDIO); however, this again is not directly compatible with PCI cards. An adaptation of these architectures to PCI is required.
- *Support for logical partitioning and adapter sharing*
  The IBM eServer z900 supports the Enterprise Multiple Image Facility (EMIF), which enables all I/O ports, starting with ESCON, to support adapter sharing by two or more logical partitions (LPARs). This also requires a unique adaptation layer to provide a high degree of protection and separation between partitions

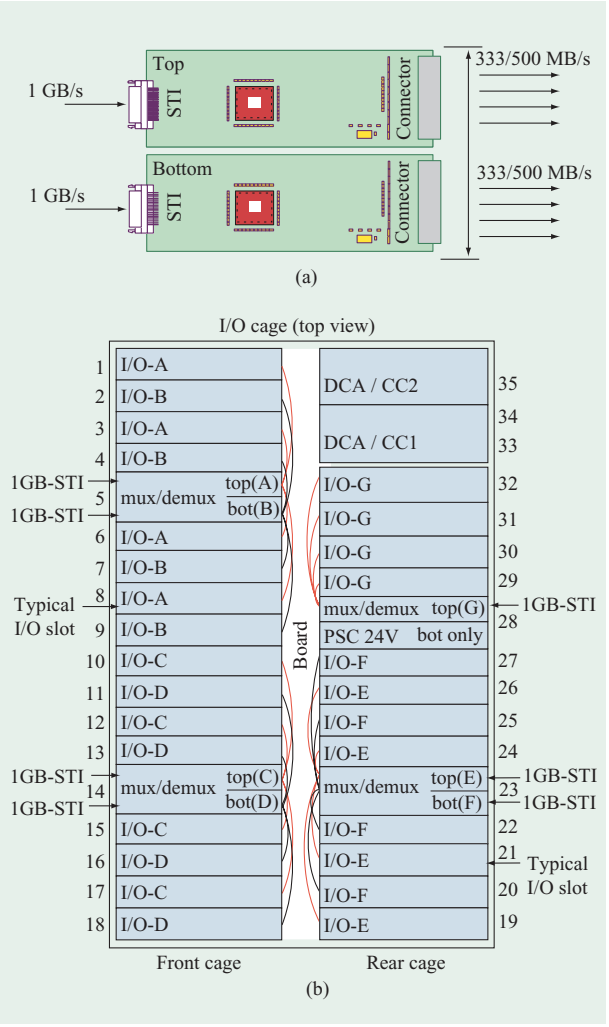


Figure 7

IBM eServer z900 I/O cage configuration: (a) mux/demux card pair; (b) I/O cage and board structure (top view).

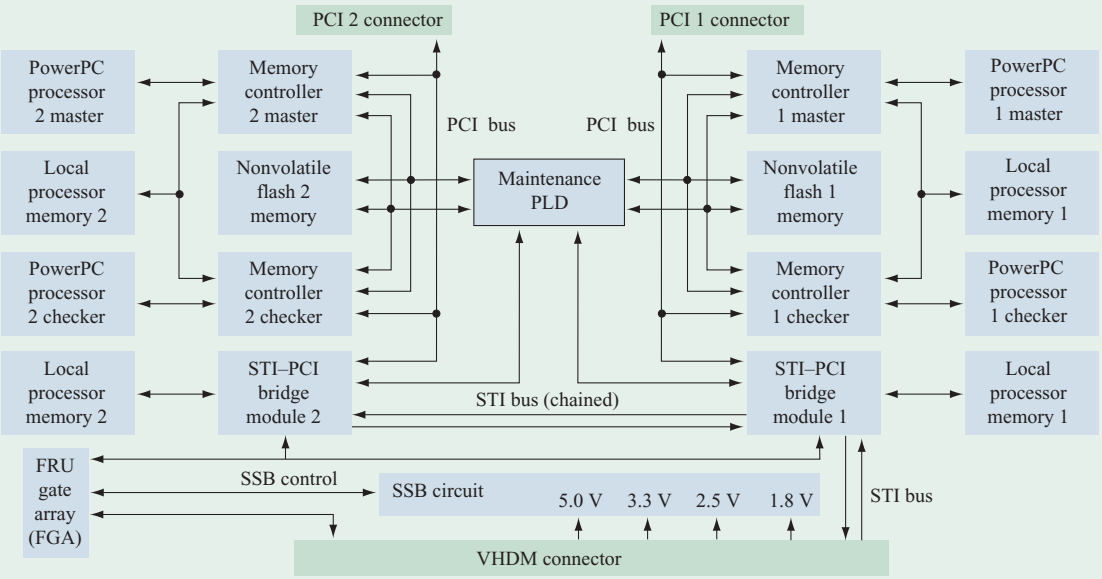


Figure 8

Common I/O platform block diagram. (SSB = soft switch breaker; VHDM = very high density metric.)

that cannot be provided directly by PCI cards. Key requirements include the appearance of entirely separate memory spaces for each LPAR provided by the zone relocation facility, and complete replication of architectural I/O interruption facilities, provided by the zonal interruption facility. Through the use of these and other facilities, each logical partition is given the appearance that it has its own I/O adapter. In order to achieve this, the I/O adapter must have knowledge of logical partitions. For example, an inbound Internet protocol (IP) frame must be routed to the correct OS IP stack by mapping the destination IP address to the correct logical partition. The PCI architecture and I/O adapters do not provide these functions.

The solution of these issues and others across a broad range of applications, in a manner that allows scaling to hundreds of adapters without placing undue burden on the processor complex, requires a high-performance, programmable adaptation layer. The common I/O platform was designed to fulfill these needs. Other design criteria for this adapter include the need to exploit, where possible, other high-volume technologies available within and outside IBM while maintaining enterprise-class characteristics.

The common I/O platform (CIOP) is a processor-based PCI-to-STI adaptation bridge with extensive RAS features. It consists logically of four major elements:

1. A processor complex consisting of dual cross-checked PowerPC* microprocessors.
2. Dual cross-checked PowerPC bridges with a combined L2 cache and L3 memory controller supporting up to 128 MB of error-correction-code (ECC)-protected program memory.
3. A uniquely designed STI–PCI bridge ASIC providing a daisy-chain STI system connection, PCI bus support for 32- or 64-bit PCI cards operating at either 33 or 66 MHz, a data buffer controller supporting up to 128 MB of ECC-protected data storage, and other special features including traditional error checking to assist in meeting IBM eServer z900 RAS requirements.
4. An industry-standard server-class PCI adapter.

The CIOP supports two independent PCI connectors, to which adapters (not shown) are attached; a flow diagram depicting two sets of the above elements is shown in **Figure 8**. The hardware is personalized with Licensed Internal Code (LIC) for each application. It is structured in two major parts:

1. Common code, which provides low-level hardware functions that are required for all applications, including hardware initialization, hardware logging, recovery, diagnostic support, and concurrent LIC patch support. Critical sections of this code are contained in flash memory on the CIOP card.

2. Application-unique code, which provides the individual characteristics for each of the CIOP applications (e.g., Fibre Channel, Ethernet). This code is loaded from IBM eServer z900 system memory during initialization or recovery. A list of applications (port types) offered with the common I/O platform card is shown in Table 1.

Collectively, these elements, along with the associated LIC, implement the features required to support the z/Architecture. In most cases, the CIOP applications implement a z/Architecture channel path, designated by its channel path identifier (CHPID). A channel path provides the necessary facilities to communicate with one or more logical partitions, processors, and system memory in order to effect transfer of data and control information to/from one or more I/O devices. Depending on the function required (e.g., FICON, Gigabit Ethernet), the channel path may have unique characteristics defined by the channel path type. Each I/O adapter and its supporting CIOP logic (see Figure 8) is considered to be a completely separate channel path/CHPID and is operated, configured, initialized, and recovered independently of the other.

In all cases, each of the four major elements of a common I/O platform collectively perform the following tasks:

- The dual IBM PowerPC 740 microprocessor complex accepts and executes z/Architecture work requests signaled from a logical partition. In the process, STI signals are passed through the STI–PCI bridge to alert the microprocessor. The microprocessor, using facilities within the STI–PCI bridge, fetches the required information describing the request (a channel program, consisting of CCWs, IDAWs, etc.). The I/O requests are then passed to the PCI adapter for execution. A number of hardware assists are available within the STI–PCI bridge to aid the microprocessor in achieving high-performance data movement and data integrity verification. This includes data mover queues (DMQs), high-priority storage requesters (HPSRs), cyclic redundancy check (CRC) logic, and longitudinal redundancy check (LRC) logic.
- The microprocessor signals completion or progress of an I/O operation by initiating a request for a z/Architecture I/O interruption. Conditions detected at the PCI card indicating these events are intercepted by the processor and transformed accordingly.
- The processor handles various other z/Architecture requests, including the ability to terminate or cancel I/O operations in progress when the initiating jobs are canceled or terminated for recovery reasons.
- Depending on the application, the processor may have to inspect data arriving from the PCI card in order to route it to the appropriate LPAR memory space. For example, an Internet protocol (IP) frame arriving from the PCI card is first transferred to the data buffer memory, allowing routing decisions to be made by highly checked logic prior to routing to the destination address.
- I/O performance measurement data is collected and reported to the appropriate logical partition. This information is used by the z/OS resource measurement facility for capacity planning and problem analysis. It is also used by DCM to dynamically allocate I/O resources.
- All channel path I/O reset functions are performed by this logic.
- Extensive recovery actions are provided to attempt recovery of intermittent errors and to isolate the failing components for repair and verification.
- The microprocessor collects both hardware and LIC log-outs, reporting them to the IBM eServer z900 I/O subsystem for effective engineering and field problem analysis. This includes the generation of field-replaceable-unit (FRU) calls for automated service calls.
- The processor coordinates and applies LIC updates for the common I/O platform code and the PCI adapter card concurrently with ongoing work.

Common I/O channel hardware description

As shown in Figure 8, the common I/O platform channel hardware consists of several elements:

- *PowerPC microprocessors*
  Two PowerPC 740 processors per I/O channel (a master processor and a checker processor) are used. They are clocked at 333 MHz internally and 66 MHz externally to interface with the memory controller. The processors are run in lockstep using a single copy of the LIC. The microprocessor input/output is cross-checked on a cycle-by-cycle basis by supporting logic within the STI–PCI bridge ASIC (discussed later in this section). Any discrepancy in the two processors causes an immediate halt in operation, thus preventing fault propagation and preserving state information for subsequent logging and recovery. The microprocessor subsystem also includes a maintenance programmable logic device (PLD), which provides access to on-card status indicators (LED displays) and to the inter-IC (IIC) interface of the STI–PCI bridge ASIC. The PLD also provides a diagnostic interface (on-card connector) to support test and debug access to microprocessor and flash memory. The diagnostic interface is also used to assist in bringup or manufacturing, as required.
- *Memory controller*
  A standard IBM component is used to provide the dual-memory-controller function. The memory controllers (a master and a checker per channel) serve as a PowerPC-to-PCI bridge, allowing the PowerPC processors to communicate with the rest of the CIOP subsystem. The master controller manages up to 128 MB of SDRAM

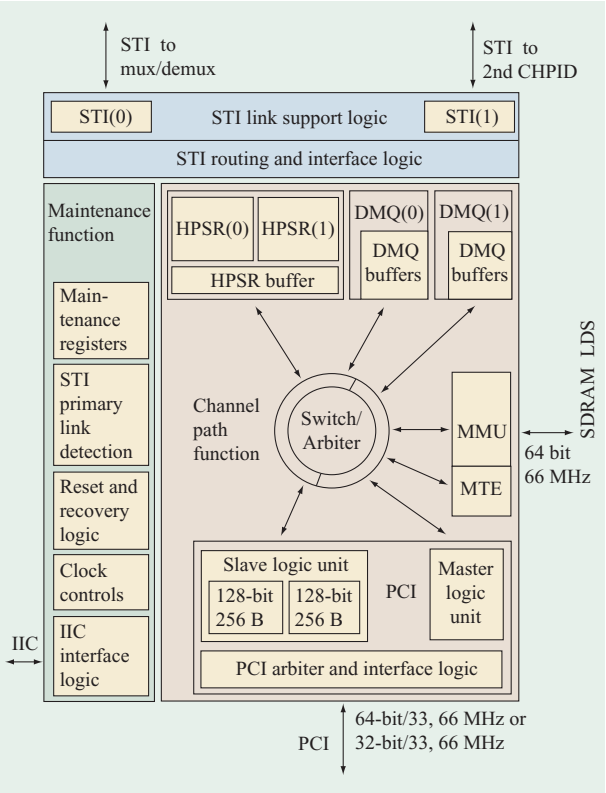


Figure 9

STI–PCI bridge ASIC block diagram.

used as local program storage (LPS) for the PowerPC microprocessors assisted by a large L2 cache. The memory controller also provides an interface to 2 MB of on-card flash memory. The master and checker controllers are also used to implement the cross-checking function. The checker compares the memory and PCI operations generated and driven by the master ASICs with those generated, but not actually driven, by the checker ASICs.

- *STI-to-PCI converter*
  The STI-to-PCI conversion is done by the STI–PCI bridge ASIC (discussed in next section). It is highly programmable and can easily be adapted to attach a variety of PCI-based adapters that provide the I/O interfaces supported by the IBM eServer z900.

STI–PCI bridge ASIC

The STI–PCI bridge ASIC coordinates data and control transfers between the IBM eServer z900 processor complex and the common I/O platform. It also provides the main firewall between the highly reliable eServer zSeries memory interface and the PCI adapter, handling differing RAS requirements, bus protocol mismatches, and differing memory architectures. The major functional blocks of the STI–PCI bridge ASIC are shown in **Figure 9**. The functions are partitioned into three main areas, which are under separate reset and clocking control. They are the STI network, channel path logic, and maintenance.

STI network function

The STI network function, comprising the STI link support logic and the STI routing and interface logic, supports two 333-MHz duplex STI links. The first link, STI(0), is connected to the mux/demux ASIC. The network includes limited routing (chaining) function supporting a second link, STI(1), which may be used for on-card connectivity to another completely separate channel path. Essentially, multiple STI–PCI bridge ASICs can be daisy-chained in a serial fashion. Within each STI–PCI bridge, the STI chaining logic routes packets internally or externally over the second link, STI(1), on the basis of addressing information embedded in each packet. This technique enables bandwidth-balanced sharing of a single STI link by multiple CHPIDs. The interface layer detects and buffers arrival of commands, response, and data packets from the system, routing them to either the maintenance or channel path logic as required, notifying the processor when appropriate. It arbitrates between the channel path and maintenance logic to forward data, control, and signal requests as required. It also generates responses to packets received from the system.

Channel path logic

The channel path logic makes up the bulk of the STI–PCI bridge function. It performs the data-movement and bridging functions that enable channel and networking dataflow. The STI–PCI bridge implements two independent data-mover queue (DMQ) engines. The DMQ engines are designed to efficiently move scattered blocks of data from IBM eServer z900 main storage to any "addressable" location connected to the STI–PCI bridge ASIC, such as the attached synchronous dynamic random-access memory (SDRAM) or any memory-mapped address on the PCI bus. The DMQ engines are controlled by the microprocessor subsystem through the use of firmware (LIC) queues. As part of an I/O operation, the processor builds queue entries for the DMQ engines to execute. These entries define scatter/gather lists consisting of both addresses and counts. The DMQ engine asynchronously executes entries from the queue in a circular fashion. The two DMQ engines available to the microprocessor complex have separate controls and queueing areas.

The high-priority storage requester (HPSR) engines in the STI–PCI bridge are used to execute single high-priority commands over the STI interface. These can be stores or fetches to the IBM eServer z900 system main storage, or hardware accesses (sense/control commands).

Typically, HPSRs are used for accessing small storage-control blocks and providing sense- and control-based interrupt notification to the microprocessors. HPSRs are initiated by LIC using a request block located near the processor. The HPSR and DMQ engines share the STI network. The STI network logic controls arbitration, interleaving data requests on an STI-packet basis between the various DMQ and HPSR facilities.

The PCI interface supports the PCI 2.2 specification, and is used to connect the STI–PCI bridge to the microprocessor complex and the particular adapter card installed. The PCI interface supports both master and slave transactions, and uses two 256-byte buffers (filling one while emptying the other) to speed data transfer. The STI–PCI bridge ASIC serves as the PCI bus arbiter on the common I/O platform and supports a robust token-based arbitration scheme for the various masters on the bus. The PCI interface also allows PCI devices access to memory-mapped control registers inside the STI–PCI bridge ASIC. The control registers are used to manipulate the various hardware engines, perform initialization and reset/recovery functions, and gather logging information. Access to these controls may be restricted to the highly checked processor, thus enhancing overall system integrity. Attempted access by an unauthorized device results in a detected error condition and associated recovery action.

The memory management unit (MMU) logic supports the attachment of up to 128 MB of synchronous DRAM. The storage is used by the processor complex as a buffer memory [local data store (LDS)] and can contain control blocks, channel data, transmission control protocol/Internet protocol (TCP/IP) packets, etc. The MMU supports many features, including bus arbitration with a multilevel priority capability, background scrub support, self-refresh support, 1GB addressability, single-error-correct/dual-error-detect correcting code, bus locking for semaphore operations, full-page burst, and internal memory test engine support. The raw memory bandwidth supported is 528 MB/s. Integrated with the MMU element is a memory test engine (MTE). The MTE is an LIC-controlled element that provides robust test capabilities for the attached SDRAM storage. The MTE tests memory at system speed, and can be programmed to generate and check various patterns, such as checkerboard, walking 0s and 1s, and pseudorandom patterns.

Connecting all of the elements in the STI–PCI bridge is the switch/arbiter block, which is an eight-port, two-way nonblocking switching matrix. The switch uses separate control signals for address and data, and implements a multilevel arbitration scheme.

Maintenance function

The maintenance logic on the STI–PCI bridge provides utility functions such as clock control and distribution, STI link detection (physical layer management), and test/debug interfaces. Much of this function is controlled through an IIC interface. This interface is accessible by the microprocessor complex, and it allows memory-mapped access to much of the low-level functionality of the STI–PCI bridge ASIC. Reset and clocking primitives are initiated through this interface, as well as scan, ASIC self-test, and other level-sensitive scan design (LSSD) functions. The maintenance logic includes dedicated state machines and control logic to manage the STI–PCI bridge recovery functions. The various clock domains can be isolated, logged, and recovered individually with minimal disruption to the remainder of the bridge.

The STI–PCI bridge includes a significant amount of support for traditional IBM eServer z900 reliability. All dataflow is parity-checked. State machines and random control logic are also checked. All internal errors are logged and are available to the processor complex via the maintenance logic. The STI–PCI bridge implements cyclic redundancy check and longitudinal redundancy check generation and checking on data flowing through the ASIC. LIC can use this to provide an additional level of protocol checking on data transfers.

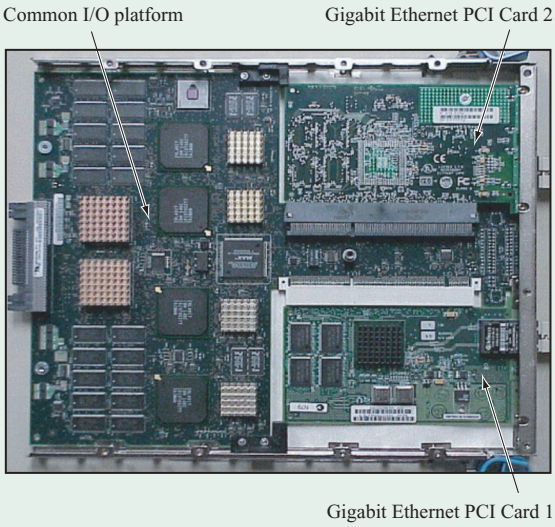


Figure 10

Photograph of Gigabit Ethernet common I/O platform card.

Common I/O platform card physical description

The common I/O platform card is a single wide book package which contains a base card and two PCI adapters of choice. **Figure 10** is a photograph of a Gigabit Ethernet version (two PCI cards) of the common I/O platform card. The single wide book package is standard sheet-metal

construction which has been designed to be mounted directly into specified slots within the IBM eServer z900 I/O cage. The common I/O platform book accommodates a family of PCI cards. The book packages are very similar but not identical, since each PCI I/O adapter type requires a unique bezel and labeling. The PCI adapter card 1 of the common I/O platform card is mounted with its components facing in the same direction as the base card, while the PCI adapter card 2 is mounted with its components facing in the opposite direction from the base card. This approach enables the various sizes of standard PCI card to be accommodated in the common I/O platform book. The overall book dimensions are height, 12.0 in., width, 1.25 in., and depth, 13.5 in. Each common I/O platform card bottom edge connector receives the STI interface directly from the I/O board, and the top edge of the book contains the LED status indicator and the network interface connectors.

Base card

All versions of the common I/O platform (CIOP) use a 20-layer, 50-Ω printed wiring board (PWB) to provide the wiring connectivity. The base card also provides the required PCI voltages (+5 V and +3.3 V) for the PCI adapter. There are four soft switch breaker (SSB) circuits that provide controls to switch power on and off to the hardware on the base card. Each of the system voltage inputs (5 V, 3.3 V, 2.5 V, and 1.8 V) have SSB protection and power control used for concurrent maintenance and secondary fault isolation. The 24-V and 3.4-V standby inputs are protected by a current-limiting resistor and are used for continuous powering of the control and vital product data logic. The SSB circuits limit current draw to their rated amount for about 7 ms. If load current exceeds the rated amount for longer than this time, the breaker will "trip" latching power off until the turn-on line or power is cycled. Current limiting will occur within 8 ms, with transient currents not exceeding twice the breaker rating during this initial period. The SSB components are located along the bottom card edge on either side of the board connector, as shown in Figure 10. All PCI adapters that plug into the common I/O platform are compliant with the PCI specification for both standard or short-form-factor cards, and are able to work in the 3.3-V signaling environment.

The CIOP card also includes maintenance functions that provide for installation, removal, verification, and sensing if the PCI adapter is plugged and tracking the adapter hardware. These maintenance functions for the overall package are supported by power/controller functions in the DCAs plugged into the IBM eServer z900 I/O cage. The maintenance interface is connected through the FRU gate array (FGA) ASIC located on the base card. The FGA ASIC provides an interface to the CIOP service electrically erasable programmable read-only memory (EEPROM), soft switch breaker, I/O port status LEDs, and channel reset lines.

Common I/O platform integrated design

A version of the common I/O platform card, used for FICON and Fibre Channel ports, integrates the PCI adapter function on the base card instead of having a pluggable PCI card. This enables the elimination of duplicate components and PCI connectors, simplifies the base PWB, and lowers the overall cost structure. IBM purchases the custom Fibre Channel ASIC and firmware from a business partner. The remaining parts are supplied by IBM and integrated onto the base card. The logic flow and CIOP are the same as the PCI card version; however, the CIOP uses a new small-form-factor (SFF) optical transceiver.

6. ESCON 16-channel I/O card

The ESCON 16-channel (ESCON-16) serial I/O card is a lower-cost, higher-port-density follow-on design to the four-channel ESCON (ESCON-4) serial channel I/O card used in S/390 machines [9]. The ESCON-16 I/O card plugged into the new IBM eServer z900 I/O cage can provide up to 15 active ESCON channel ports; one channel is reserved as a spare port. The number of active channels on an ESCON-16 card is controlled by LIC activated channel paths (discussed earlier in this paper).

Greater channel density and lower cost were achieved on the ESCON-16 I/O card by integrating the function and content of six cards (four ESCON-4 I/O and two infrastructure) in the prior-generation S/390 machine. The new I/O card design was made possible by two key changes in packaging technology. First, the 16-port ESCON adapter uses the new industry-standard SFF optical transceiver that supports a smaller multimode fiber optic cable connector (type MT-RJ). The second key change was the replacement of the single-channel ESCON ASIC with the ESCON-4C ASIC, which uses advanced CMOS technology and supports four channels per ASIC plus the link serializer, deserializer, and retiming function.

ESCON-4C ASIC description

The ESCON-4C ASIC is the latest in a series of ESCON channel designs. While objectives for the design include cost reduction and new function, the main thrust is to help maintain a balanced system by achieving a packaging consolidation while maintaining the same performance and improved RAS characteristics per channel. This allows more overall I/O bandwidth to fit within the I/O cage configurations, using fewer I/O slots and thereby allowing increased usage of newer high-performance channel and networking cards (e.g., common I/O platform cards).

ESCON channel function

The ESCON channel has provided IBM servers with a serial optical interface for I/O connectivity since 1990, operating at speeds up to 17 MB/s over a distance of three kilometers [9]. The ESCON architecture, functionality, and hardware structure have been described in considerable detail in previous papers [10–12].

The ESCON ASIC contains two main functions for each channel: the I/O interface link dataflow logic and the channel processor with LIC storage. Full-duplex dataflow to and from the optics consists of a 200Mb/s serial bit stream utilizing 10-bit data characters assembled into frames. A balanced 8-bit/10-bit code is used to help minimize link errors, in addition to allowing for special control characters used for link idles and frame delimiters. Outbound and inbound data respectively pass through a digital serializer and deserializer (SERDES). On the channel side of the SERDES, the balanced 10-bit code passes through synchronizing buffers in both directions, since the internal system channel clock runs asynchronously with the link clock. The channel clock frequency can vary depending on how the ESCON-4C ASIC is used (e.g., channel or control unit), but the link speed is always 200 Mb/s, driven from an independent external link oscillator and a PLL internal to the ASIC. On the channel clock side of the synchronizing buffers, the inbound 10-bit code is translated to standard 8-bit data plus parity. In a similar fashion, outbound 8-bit data is translated to the 10-bit code used by the SERDES. Transmissions consist of data frames or control frames, and a CRC is included on all frames for link error detection. Outbound and inbound state machines, along with an assortment of data and header buffers, control the assembly and disassembly of frames.

The channel processor executes channel programs and performs link initialization and recovery. The processor is a horizontally coded engine using a 38-bit control word, allowing for a high degree of control of the I/O interface dataflow hardware and maximizing throughput performance. The control word consists of several encoded fields to allow for control of many hardware functions in a single cycle of execution. Four levels of subroutine calls are supported. The LIC is loaded into static random-access memory (SRAM) on the ASIC during channel reset. By loading different LIC into the SRAM, the highly flexible design allows the hardware to be used for different functions such as channel, control unit, channel-to-channel (CTC), or diagnostics.

ASIC physical design and packaging

Achieving the level of ESCON I/O card density requires a similar consolidation of the ESCON channel logic. The ESCON-4C ASIC leverages the advanced IBM CMOS technology to integrate four channels into a single ASIC.

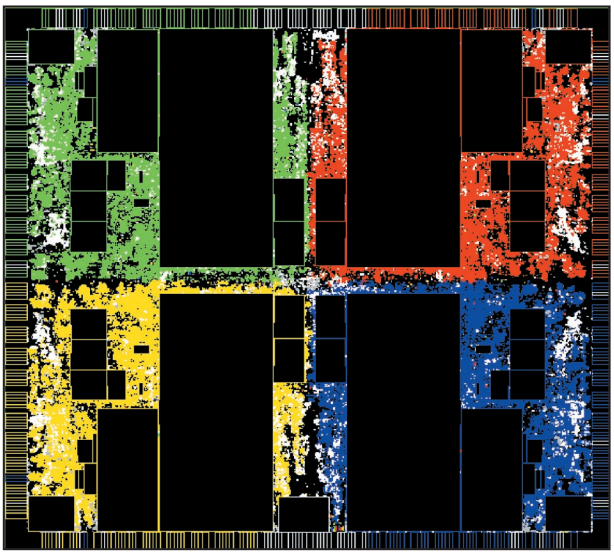


Figure 11

ESCON-4C chip layout.

The 10-mm^2 die size used is able to accommodate nearly six times as much logic as the previous single ESCON channel ASIC, which is also a 10-mm^2 die. Thus, in addition to being able to accommodate four channel engines, it allows space for doubling the amount of LIC storage for each channel. This is a new capability added to allow for growth of code to support future enhancements to the eServer architecture. Standard growable SRAM arrays are used in the ESCON-4C ASIC instead of custom arrays in the prior ESCON ASIC. The amount of module I/O required for four independent ESCON channels requires the use of a 32-mm^2 ceramic ball grid array (CBGA) module. A significant benefit of the combined consolidation and latest technology is also a reduction in power required.

The physical layout of the ESCON-4C chip (**Figure 11**) begins by placing one ESCON channel in each quadrant of the chip (designated by the different colors). The associated SERDES logic unit is replicated four times on the silicon chip, and the units are placed in the four corners of the ASIC to maintain the least noise interference between them. Fitting the large LIC storage arrays toward the center of the ASIC helps to define the remaining layout, although the PLL for the SERDES clock generation takes a center spot along the south edge of the die such that the arrays are not placed back to back. Smaller data buffers, control arrays, and I/O cells are then floorplanned to match the dataflow for each of the four channel engines such that the general channel logic can fall into place. The compact layout, SERDES

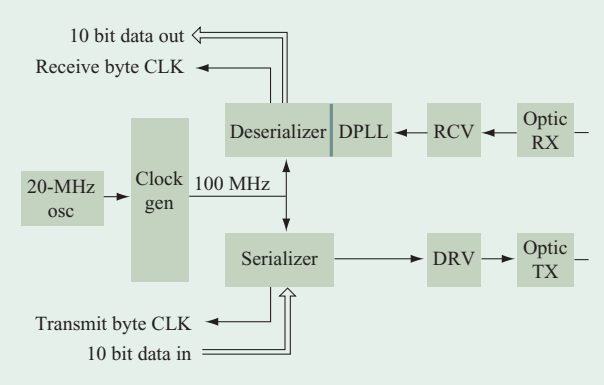


Figure 12

ESCON channel DPLL/SERDES used in a fiber optic link application.

logic, and limitations in the porosity of the arrays required the use of five levels of metal for global chip wiring.

Reliability, availability, and serviceability (RAS)

The ESCON-4C ASIC was designed, from the beginning, with RAS in mind. While there are four channels on the chip, and there is no interaction between them such that a failure in one would not affect the operation of the others, they each have independent channel clocks, so that one channel can be stopped and recovered while the others remain running. Single points of failure are minimized. A low-alpha-particle solder is used for the ASIC solder-ball connections to minimize soft-error fails of the arrays. In addition, as in previous ESCON channels, the LIC code arrays are scrubbed for soft errors. When invoked in card test and system environments, a diagnostics package with extensive control hooks into the channel logic ensures good hardware.

Digital SERDES

The integration of the serializer/deserializer (SERDES) function with the main ESCON channel logic was a key factor, along with advances in the IBM CMOS technology, that allowed the cost savings and high-density packaging achieved with the ESCON-4C ASIC. Prior to development of the digital SERDES, the SERDES function was accomplished using an analog phase-locked loop and SERDES packaged in a separate and expensive bipolar ASIC.

A typical serial data link, such as the IBM ESCON channel, comprises a serializing transmitter and a deserializing receiver on each end of the link. The serializer converts a 10-bit parallel data byte into a stream of serial data bits; the deserializer, along with a digital phase-locked loop (DPLL), samples the serialized data stream using a local reference clock and reassembles the data back into a 10-bit-wide parallel data byte referenced to the local clock. **Figure 12** shows one end of a typical ESCON fiber optic link utilizing a SERDES and a DPLL.

The digital SERDES uses digital circuit techniques in the PLL design, wherein traditional clock extraction from the serial bit stream is replaced by phase-shifting, or delaying, the incoming serial data stream so that it may be reliably sampled by a fixed local clock. The DPLL phase-shifts the data to allow the local clock to optimally sample the data. A selected phase of the data can be sampled reliably with the local clock until the frequency drift between the transmit clock and the local clock creates additional phase error that must be corrected. The clock drift between the transmit and receive clocks is specified as a small fraction of the data rate of the link, of the order of a few hundred parts per million (ppm) or less.

The DPLL samples the incoming serial data stream and captures data on both the rising edge and the falling edge of the local reference clock. This partitions the data stream into two-bit samples, which the DPLL presents to the deserializer. The deserializer completes the serial-to-parallel conversion and assembles the received data back into 10-bit-wide bytes. The deserializer also generates the receive byte clock (RBC) used for presenting the parallel data to the ESCON channel logic.

The DPLL tracks the frequency drift between the transmit clock and the local receive clock by feeding the received serial data into a long chain of delay elements which are actually identical inverters. This creates many phases of the data, each phase separated by one inverter delay. The DPLL identifies the locations of the data bit edges and tracks them with an edge detector [6, 13]. Knowing the location, or inverter address, of the data edges allows the optimum data sample point to be determined by simple arithmetic. It is helpful to picture a standing-wave pattern inside the delay chain, as shown in **Figure 13**. For example, given a data bit time of 5 ns and an inverter delay of 500 ps, the data bit would span ten delay elements and have edges at inverter outputs 0 and 10. The optimum data sample point is calculated to be at inverter 5. The standing-wave pattern inside the delay chain will drift either to the left or to the right, as time passes, depending upon whether the local receiver clock is faster or slower than the transmit clock that launched the serial data.

The individual delay element is designed in such a way that there will always be at least two bit times worth of data in the delay chain under all conditions (i.e., temperature, power-supply, and CMOS manufacturing process variations). The requirement of two simultaneous data valid windows is essential for tracking the frequency drift; however, only one data window is considered "current" (data being read) at a time.

Figure 14 illustrates two data windows, window 1 and window 2, in relation to the delay chain. Assume that window 1 is the "current" data window: In the presence of a faster transmit clock, the "current" data window will drift to the right until a certain point at which a *transfer-up* procedure will occur, and window 2 will become the "current" window. As time passes, window 2 continues to drift to the right until it reaches the upper end of the delay chain. At this point, a *wrap-back* procedure occurs, and window 1 again becomes the "current" window. **Figure 15** illustrates the process when the transmit clock is slower than the local reference clock. Again assume that window 1 is the "current" data window; in the presence of a slower transmit clock, the "current" window will drift to the left. However, in Figure 15 the "current" window is already at the lower end of the delay chain. Consequently, an immediate *wrap-forward* procedure will occur, and window 2 will become the "current" window. As time passes, window 2 will continue to drift to the left until a certain point at which a *transfer-down* procedure will occur, and window 1 will become the "current" window. These procedures continue repeatedly and indefinitely as the DPLL tracks the incoming serial data stream.

The deserializer receives 2 bits of retimed data from the DPLL each 100-MHz local clock cycle. The deserializer produces a 20-MHz receive byte clock (RBC) from the local clock, which yields one 10-bit byte every 50-ns RBC period. In the process of tracking the incoming data stream, the deserializer periodically adds or subtracts one bit time from the RBC cycle time. Whenever the DPLL performs a *wrap-back* procedure, the deserializer shortens the RBC cycle by one bit time, and whenever the DPLL performs a *wrap-forward* procedure, the deserializer lengthens the RBC cycle by one bit time. Thus, during a wrap procedure the deserializer presents 10 bits of data to the ESCON channel logic in either 45 ns or 55 ns, depending on whether the incoming serial data is running faster or slower than the local clock.

The wrap procedures tightly couple the DPLL and deserializer. The deserializer acknowledges a *wrap* request only at the end of an RBC cycle to synchronize the *wrap* events and maintain data integrity. In the *wrap-back* case, in which the RBC cycle is shortened by one bit time, an acknowledgment is sent from the deserializer to the DPLL, where the second data window must be sampled, allowing the deserializer to acquire an extra bit during the local clock cycle. For that particular local clock cycle, three data bits are acquired from the DPLL. The *wrap-back* case is illustrated in Figure 14. In the *wrap-forward* case, in which one bit time is added to the RBC cycle, the DPLL twice presents a particular bit, which must be discarded by the deserializer. The *wrap-forward* case is illustrated in Figure 15.

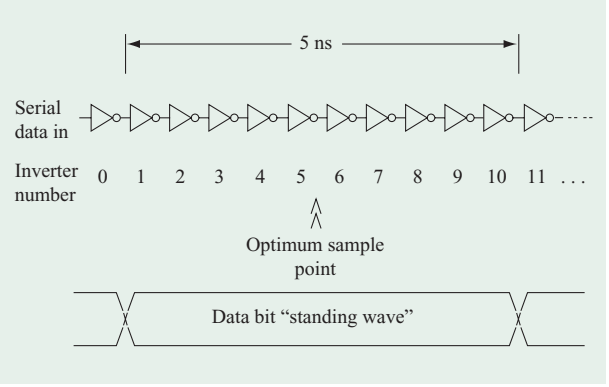
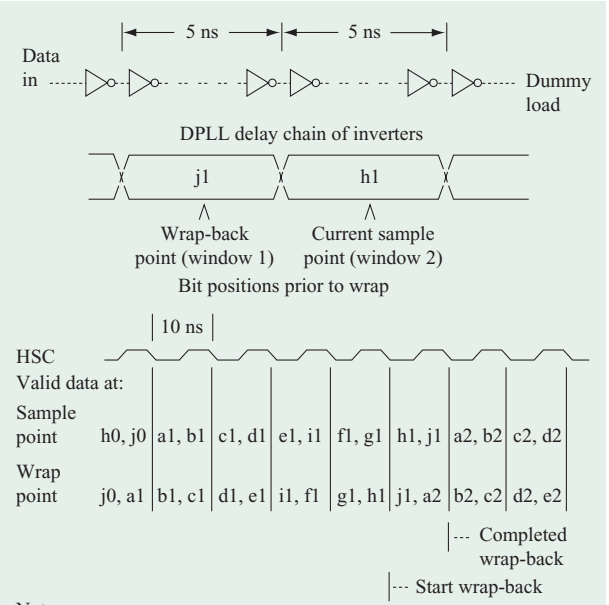


Figure 13

Delay chain sampling window.

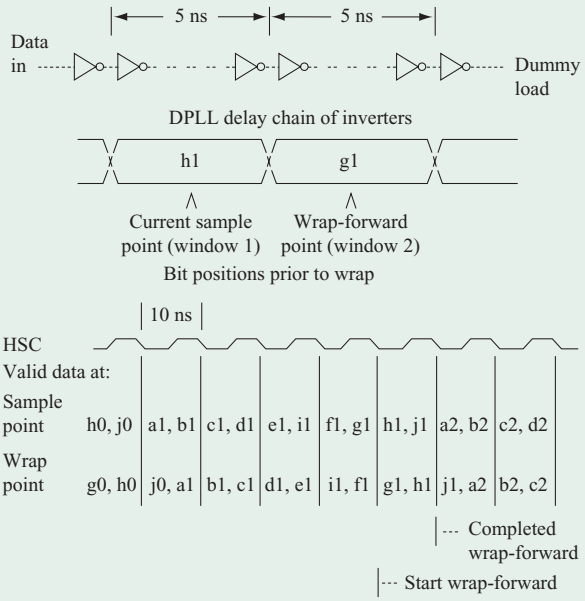


Notes:
1. After wrapping backward, the "wrap-back point" becomes the new sample point.
2. Notice the cycle following the wrap-back; the a2 bit has been missed.
3. Therefore, the a2 bit must be sampled simultaneously with bits h1 and j1.
4. The a2 bit must be inserted back into the data stream and the (external) receive byte clock is shortened by one bit time (5 ns).

Figure 14

Data-valid window configuration for wrap-back procedure.

A novel feature of the DPLL design is the use of three edge detectors. Two edge detectors are used to track the "current" data window, and the third is used to locate the



Notes:
1. After wrapping forward, the "wrap-forward point" becomes the new sample point.
2. Notice the cycle following the wrap-forward; the j1 bit has been sampled twice.
3. Therefore, by extending the (external) receive byte clock by one bit time (5 ns), the extra j1 bit can be dropped from the data stream.

Figure 15

Data-valid window configuration for wrap-forward procedure.

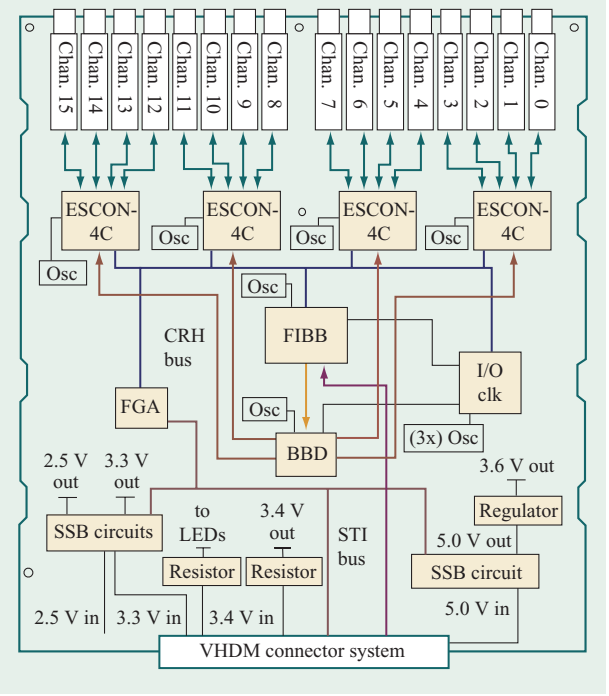


Figure 16

ESCON-16 serial I/O card logical dataflow.

second data window. During a *transfer-up* or *transfer-down* procedure, the contents of the registers containing the address of the "current" data window (two consecutive data edges) are simply transferred to the appropriate registers for the new data window. Similarly, prior to a *wrap-back* or *wrap-forward* procedure, the third edge detector is employed to locate the second data window. This data window becomes the "current" data window once the *wrap* procedure is completed. This allows for the second window to be located and "locked" onto well in advance of the time at which the window must actually be used.

ESCON-16 I/O card description

The ESCON-16 I/O card is a field-replaceable unit (FRU) when fully assembled in its sheet metal enclosure to ensure electromagnetic compatibility, and it can be installed or replaced while the system is operating. The I/O card comprises 16 ESCON SFF transceivers mounted on the top side along the top edge of the 50-Ω, 14-layer, 79-mil-thick printed wiring board (PWB). There are also 16 port indicators (orange-colored LEDs) attached to the rear side of the PWB at each optical transceiver location to improve RAS functions when servicing or reconfiguring fiber optic cables. These indicators are visible through the enclosure when they are installed in the IBM eServer z900 I/O cage. There are four ESCON-4C ASICs, and each supports four channels. The other active devices on the card are unique logic ASICs, an I/O clock ASIC, FGA, oscillators, other miscellaneous industry ASICs, and passive components. The ESCON-16 I/O card is designed for 100 000 power-on hours of operation.

The I/O card signal and voltage requirements are supplied at the Very High Density Metric (VHDM**) signal/power connector. The voltage into the card power planes is controlled via power MOSFETs residing in the soft switch breaker (SSB) circuit, using discrete components mounted directly on the adapter. Card power sequencing and resets are controlled through the service element and FGA interface included on each adapter. The VHDM connector incorporates a strip-line shielding scheme that effectively allows 100% of the pins to be used for signals while minimizing crosstalk and reflections. The connector features a solderless compliant pin press-fit termination.

The ESCON-16 I/O card logical dataflow is shown in **Figure 16**. The fast internal bus buffer (FIBB) ASIC is connected to the STI. This ASIC, in conjunction with the

bidirectional bus distribution (BBD) ASIC, supports the attachment of 16 ESCON channels via the four ESCON-4C ASICs over the channel request handler (CRH) bus. Support of up to 16 LPAR partitions is done in these two ASICs. All initialization, command, and data transfers from and to the ESCON-4C are done under control of the ASICs. The I/O clock provides all clocking signals, self-test, and scan support for the I/O card. Each ESCON-4C ASIC connects to four optical transceivers that convert electrical signals to light and vice versa. The ESCON-16 serial I/O channel design has timing and electrical properties (except for power requirements) that are compatible with those of S/390 CMOS ESCON channels. The functional and diagnostic LIC as well as design data have been updated to support the 64-bit z/Architecture and other enhancements. **Figure 17** is a photograph of the ESCON-16 I/O card.

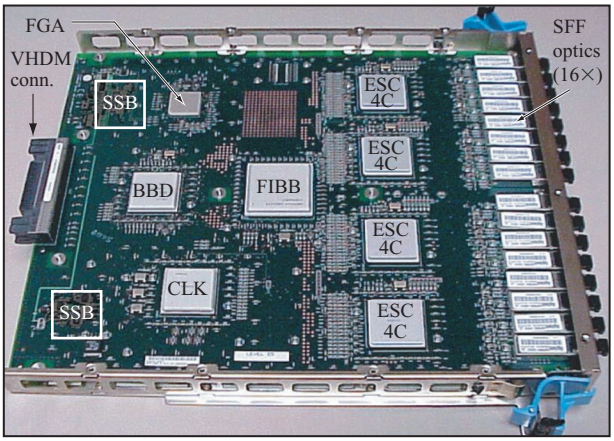


Figure 17

ESCON-16 serial I/O card.

7. Intersystem channel (ISC-3) card

A Parallel Sysplex environment is formed when multiple systems are tightly coupled to cooperate and work in parallel by using IBM eServer z900 message architecture to pass information back and forth over high-speed paths or links. These message paths can be internal within a central processing complex (CPC), over high-speed parallel copper links (ICB-3), or over high-speed optical fiber links using intersystem channel third-generation (ISC-3) links [14]. Multiple CECs can be placed close together for higher performance, or positioned farther apart for increased availability. Parallel copper links [integrated cluster bus 3 (ICB-3)] are generally used for short distances (about 10 m) and have a peak data rate of 1 GB/s. ISC-3s are serial optical links using long-wavelength (1300-nm) lasers over single-mode fiber optic cables as the transport. ISC-3 links have a peak data rate of 2.125 Gb/s and can interconnect IBM eServer z900 systems for distances up to 10 km. The distance can be extended for special circumstances.

Functional description

In prior IBM S/390 systems, this optical serial interface was driven by a large variety of unique ASICs. The ISC-3 design combines the functions of multiple precursor ASICs into a single advanced CMOS ASIC, appropriately referred to as the ISC-3 ASIC, thereby substantially reducing the cost and at the same time improving the performance and link utilization. The IBM 801 processor used by prior-generation ISC cards (ISC-1 and ISC-2) is no longer required on the ISC-3 card. All of the real-time functions performed by the IBM 801 processor have been converted to hardware state machines in the new ISC-3 ASIC. The remaining non-performance-critical functions (e.g., initialization, recovery, time-outs, logging) are done by the system assist processor (SAP).

The ISC-3 and ICB-3 share a common design base. The main line commands sent by the processors are the same for both designs, and the numbers and sizes of the buffers are also the same. Furthermore, both ISC-3 and ICB-3 use the same command set and data mover design. Sharing the command set with ISC-3 greatly reduces the microcode effort in supporting both ISC-3 and ICB-3. In support of the 64-bit architecture, both ISC-3 and ICB-3 designs utilize 48-bit absolute addressing to access main storage. The IBM eServer z900 design improves the path and link performance while greatly reducing the cost and the number of interfaces/channels required.

The ISC-3 functions utilize a base-adapter-card design concept for the IBM eServer z900 machine, in which each ISC-3 base card can accept up to two ISC-3 adapter cards. **Figure 18** shows data paths through major components of both base and adapter cards. A 500MB/s secondary STI from mux/demux ASIC interfaces is connected to a cascaded (second-level) mux/demux ASIC located on the ISC-3 base card. The four secondary STI interfaces (running at 333 MB/s) from the second-level mux/demux ASIC are connected to the four ISC-3 ASICs located on the ISC-3 adapter cards. The generated ISC frame is then passed from the ISC-3 ASIC to a serializer/deserializer ASIC, which sends the differential serial data to fiber optic laser transceivers at a data rate of either 2.125 Gb/s or 1.0625 Gb/s depending on the operating mode of the ISC-3 adapter card. The optical transceivers are designed to comply with Class 1 laser safety standards, and no additional safety interlocks are required.

Link operating modes

The ISC-3 port can be configured to operate in two distinct modes: peer mode or compatibility mode. The

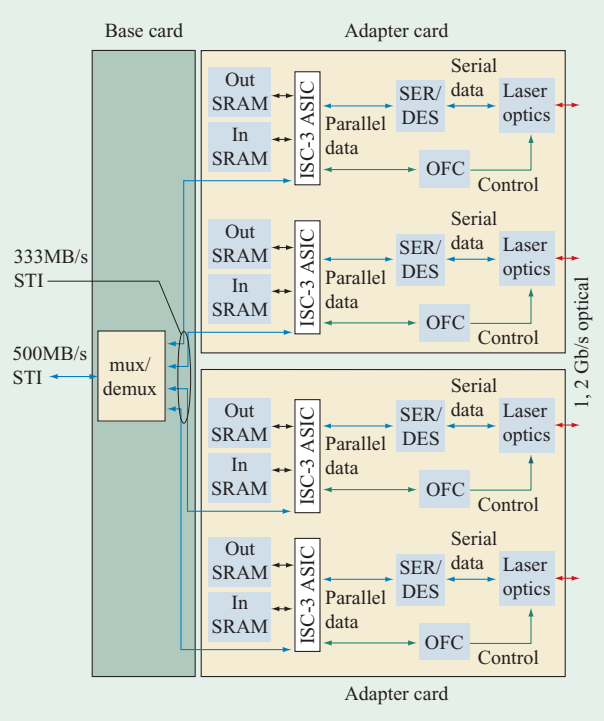


Figure 18

Logical dataflow of ISC-3 card.

peer mode is the new ISC-3 mode; it is not supported on the precursor ISC-1 or ISC-2 links. Peer-mode operation allows ISC-3 cards to be connected to other ISC-3 cards with the fiber link speed running at 2.125 Gb/s. This is a symmetrical mode of operation in which each end of the ISC-3 link can perform both sender and receiver functions at the same time. The number of buffer sets for each ISC-3 channel operating in peer mode increases to 28 (seven primary originator, seven primary recipient, seven secondary originator, and seven secondary recipient), and these buffers can be shared between a common sender and receiver. With a higher number of buffer sets, more messages can be active on a given link at a given time, thereby increasing link utilization. ISC-3 peer mode supports up to 1KB message command blocks (MCBs) and message response blocks (MRBs). This larger MCB/MRB size allows for future coupling control/response exploitation. The data buffer size is increased to 64 KB for peer mode. With larger data buffer sizes, the efficiency and distance are increased. Peer mode allows for much more efficient use of the links and reduces the number of links needed.

The compatibility mode of operation is provided so that ISC-3 cards can connect to prior-generation coupling links. In this operating mode, the fiber optic link speed must be set to run at 1.0625 Gb/s, and the open fiber control (OFC) protocol is used to initially establish and activate the link. OFC is a laser safety interlock scheme implemented on the ISC-3 adapter card to allow connection to predecessor ISC cards that comply with OFC protocol. The OFC safety interlock system ensures that a pair of laser transceivers connected by a point-to-point fiber optic link must first perform a specific handshake sequence in order to initialize the link before data transmission can begin. In compatibility mode, the ISC-3 channels use two originator and two recipient buffer sets, operating in sender or receiver modes, but not both at the same time. Furthermore, the MCB and MRB sizes are limited to 256 bytes, and the data buffer size is limited to 4 KB.

Card packaging description

The ISC-3 card design uses a base-adapter concept for higher granularity. The base card assembly can accept a maximum of two ISC-3 adapter cards, where each card assembly is an independent field-replaceable unit (FRU). Base and adapter card assemblies are enclosed in their individual sheet metal shells or book packages. This book packaging enclosure provides physical support to the ISC-3 cards, as well as positive latching and air baffling, and it ensures electromagnetic compatibility. The adapter card book assembly slides into the base card book package on guide rails, and it is securely seated in place by two finger-tab latches located on the adapter card bezel. **Figure 19** is a photograph of the ISC-3 card showing the mechanical layout and the components.

Base card

At the left side of the base card (Figures 18 and 19), a single right-angle female connector is used for attachment into the IBM eServer z900 I/O cage board. Two right-angle male connectors on the right side of the base card accommodate the attachment of two separate ISC-3 adapter cards. All right-angle connectors used on both the ISC-3 base and adapter card utilize the VHDM connector system, incorporating a strip-line shielding scheme that effectively allows 100% of the pins to be used for signals while minimizing crosstalk and reflections. This VHDM connector system features solderless eye-of-the-needle press-fit terminations. The STI interface, the system support interface (SSI), and the required voltages (+24 V, +5 V, +3.4 V standby voltage, +3.3 V, +1.8 V, and ground) are supplied to the base card via the bottom VHDM connector.

Concurrent maintenance capability of both ISC-3 base and adapter cards is achieved by the implementation of three independent soft switch breaker (SSB) designs, using discrete components mounted directly on the base card. Two SSB circuits are dedicated to the adapter cards, and

the third SSB circuit is dedicated to the base card. The +3.3-V and +1.8-V power voltages supplied to the base card are switched onto the card power planes via the power MOSFETs residing in the SSB circuit. The primary purpose of the SSB circuit is to provide control (i.e., soft turn-on and turn-off) and protection for the two power levels (+3.3 V and +1.8 V). Having two separate and independent SSB circuits (one SSB for each adapter card) allows turn-on, turn-off, and replacement of one ISC-3 adapter card FRU without disrupting the operation of the second ISC-3 adapter card. In the event of an over-current condition occurring on one of the adapter cards, the SSB will latch off (turn off) the output voltages to the failing adapter card in less than one microsecond and will activate a "power fail" signal back to the controlling module [FRU gate array (FGA)], while the other ISC-3 adapter card continues to function undisturbed.

The service subsystem controls the power, reset, and maintenance functions of both ISC-3 base and adapter cards via the path starting at the cage controller (CC), located in the distributed converter assembly (DCA). The control messages are passed from the CC via the system support interface (SSI) lines to the master FGA ASIC, located on the DMA card. The master FGA then communicates these commands to the slave FGA ASIC, residing on the ISC-3 base card. The slave FGA senses and controls all three SSBs, controls the LEDs on both adapter cards, supplies power-on and reset control signals to the STI-mux and four ISC-3 ASICs, supports the SSI interface lines from the master FGA, provides IIC interface lines to read and write the vital product data (VPD) ASIC from/to the service EEPROM on the base and adapter cards, and supports the universal service interface (USI) lines to the mux/demux ASIC and four ISC-3 ASICs. The mux/demux ASIC, located on the ISC-3 base card, provides the STI fan-out (cascading) required by the ISC-3 card.

Adapter card

The ISC-3 card assembly contains two adapter cards. Each adapter card contains two independently operating ISC-3 links (Figure 18). A single right-angle female VHDM connector, located at the left side, is used for attachment into the ISC-3 base card book package. All required voltages (i.e., +5.0 V, +3.4 V standby voltage, +3.3 V, +1.8 V, and ground), the 333MB/s STI, the USI interface, the IIC serial interface, and the LED control lines are provided by the ISC-3 base card via the VHDM connector. Mounted at the right edge of the adapter card are two pin-through-hole SFF fiber optic transceivers that transmit and receive serial optical data. The laser transceiver contains a duplex optical connector (type LC) and is capable of a maximum data rate of 2.125 Gb/s, up

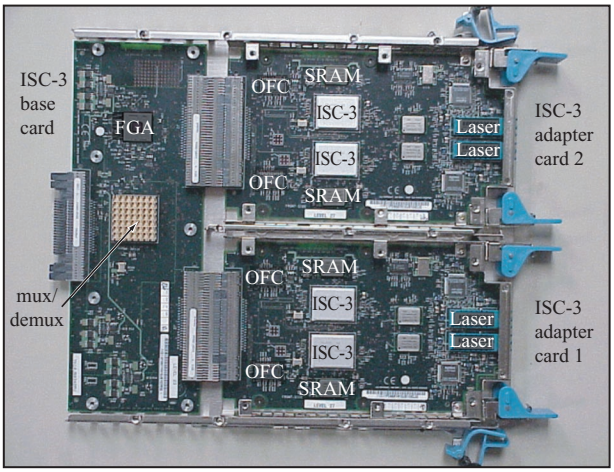


Figure 19

ISC-3 I/O card.

to a distance of 10 km. Also at the right edge of the adapter card, one amber LED is mounted on the component side (i.e., FRU indicator) to easily identify the correct ISC-3 adapter card assembly for repair and verification. Two additional orange LEDs (i.e., ISC port availability indicators) are mounted on the solder side of the adapter card (one LED under each FOSA transceiver port) to identify the available ISC-3 port. The control signals to turn these three LEDs on and off are driven by the FGA ASIC residing on the ISC-3 base card.

Two ISC-3 ASICs are mounted on each adapter card. They are in a 25-mm by 32-mm ceramic ball grid array package and were designed with advanced IBM CMOS technology. On the inbound side, the ISC-3 ASIC connects to the 333MB/s STI interface provided by the second-level mux/demux ASIC. On the outbound side, the ISC-3 ASIC connects to a serializer/deserializer ASIC which interfaces with the transceiver. The outbound parallel data is 8-bit/10-bit-encoded. The SERDES accepts 10-bit-wide encoded parallel data characters from the ISC-3 ASIC and serializes them into a high-speed serial differential data stream, delivered to the fiber optic laser transceiver. On the inbound side, the SERDES accepts a serial differential data stream from the fiber optic transceiver and converts it back into 10-bit-wide encoded parallel data, presented to the ISC-3 ASIC.

Two independent 8Mb SRAM ASICs are used for each ISC-3 ASIC to buffer incoming and outgoing data streams. Both inbound and outbound SRAM ASICs have the same 8-ns cycle time, clocked at a frequency of 125 MHz from the respective ISC-3 ASIC, and the same basic read/write operations.

8. Conclusions

The IBM eServer z900 platform is a new generation of machine that has evolved to meet the requirements of the information technology industry. The new IBM eServer z900 I/O subsystem extends the capability of the IBM eServer zSeries to provide the bandwidth, connectivity, and RAS requirements to sustain the I/O technology leadership of the platform. The new STI infrastructure provides the enhanced flexibility and bandwidth to support the high bandwidth and varied I/O port types required by many new applications. Significant improvements have been made in LIC to aid the customer to achieve maximum performance (e.g., dynamic channel path management) from the platform as well as providing new flexible configuration tools (e.g., assignable CHPIDs). The new IBM eServer z900 I/O subsystem supports the enhanced addressing and other features of the ESAME 64-bit architecture [1]. A new common I/O platform has been incorporated in the IBM eServer z900 to enable a common attachment of industry-standard I/O ports. The IBM eServer z900 hardware enhancements have exploited IBM advanced CMOS technology (e.g., ESCON-16 and ISC-3 I/O cards) to achieve higher I/O port density while reducing I/O cost.

These improvements were made while protecting the customer's I/O investment in prior S/390 platforms by also enabling the prior I/O infrastructure to be included in the IBM eServer z900. This feature enables an easier customer migration to the IBM eServer z900 from earlier S/390 platforms so that the customer may take advantage of the new features of the leadership eServer zSeries platform as soon as possible.

Acknowledgments

The authors thank H. Bagheri and F. Ferraiolo for their careful reading of the manuscript and suggested improvements. We also thank L. Greenberg for supplying the ESCON-4C layout figure. Finally, we wish to thank the many individuals (too numerous to list here), both within and outside IBM, whose hard work and dedication have helped to make the IBM eServer z900 I/O subsystem a reality.

References

1. IBM Corporation, *z/Architecture Principles of Operation*, Order No. SA22-7832, 2001; available through IBM branch offices.
2. *Single-Byte Command Code Sets-2 Mapping Protocol (FC-SB-2)*, T11/Project 1357-D/Rev. 2.1, American National Standards Institute, Washington, D.C., December 2000.
3. J. Probst, B. D. Valentine, C. Axnix, and K. Kuehl, "Flexible Configuration and Concurrent Upgrade for the IBM eServer z900," *IBM J. Res. & Dev.* **46,** No. 4/5, 551–558 (2002, this issue).
4. W. J. Rooney, J. P. Kubala, J. Maergner, and P. B. Yokom, "Intelligent Resource Director," *IBM J. Res. & Dev.* **46,** No. 4/5, 567–586 (2002, this issue).
5. H. Harrer, H. Pross, T.-M. Winkel, W. D. Becker, H. I. Stoller, M. Yamamoto, S. Abe, B. J. Chamberlin, and G. A. Katopis, "First- and Second-Level Packaging for the IBM eServer z900," *IBM J. Res. & Dev.* **46,** No. 4/5, 397–420 (2002, this issue).
6. J. M. Hoke, P. W. Bond, T. Lo, F. S. Pidala, and G. Steinbrueck, "Self-Timed Interface for S/390 I/O Subsystem Interconnection," *IBM J. Res. & Dev.* **43,** No. 5/6, 829–846 (September/November 1999).
7. J. M. Hoke, P. W. Bond, R. R. Livolsi, T. C. Lo, F. S. Pidala, and G. Steinbrueck, "Self-Timed Interface of the Input/Output Subsystem of the IBM eServer z900," *IBM J. Res. & Dev.* **46,** No. 4/5, 447–460 (2002, this issue).
8. *Fibre Channel—Physical and Signaling Interface (FC-PH)*, X3T9.3/Project 755D/Rev. 4.2, American National Standards Institute, Washington, D.C., October 1993.
9. C. DeCusatis, E. Maass, D. Clement, and R. Lasky, *Handbook of Fiber Optic Data Communication*, Chapter 13, Academic Press, Inc., New York, 1998, pp. 439–495.
10. T. A. Gregg, "S/390 CMOS Server I/O: The Continuing Evolution," *IBM J. Res. & Dev.* **41,** No. 4/5, 449–462 (1997).
11. J. C. Elliott and M. W. Sachs, "The IBM Enterprise Systems Connection (ESCON) Architecture," *IBM J. Res. & Dev.* **36,** No. 4, 577–591 (1992).
12. J. R. Flanagan, T. A. Gregg, and D. F. Casper, "The IBM Enterprise Systems Connection (ESCON) Channel: A Versatile Building Block," *IBM J. Res. & Dev.* **36,** No. 4, 617–632 (1992).
13. R. C. Jordan, R. S. Capowski, D. F. Casper, F. D. Ferraiolo, W. C. Laviola, and P. R. Tomaszewski, "Edge Detector," U.S. Patent 5,577,078, November 19, 1996.
14. T. A. Gregg and R. K. Errickson, "Coupling I/O Channels for the IBM eServer z900: Reengineering Required," *IBM J. Res. & Dev.* **46,** No. 4/5, 461–474 (2002, this issue).

Received October 16, 2001; accepted for publication April 3, 2002

Daniel J. Stigliani, Jr. *IBM Server Group, 2455 South Road, Poughkeepsie, New York 12601 (danst@us.ibm.com).* Dr. Stigliani received the B.Engr. degree in general engineering from Stevens Institute of Technology, and the M.S. and Ph.D. degrees in electrical engineering from the University of Illinois. In 1969, he joined the IBM Federal Systems Division in Owego, New York, working on a broad range of projects in optical signal processing and communications. In 1974, Dr. Stigliani transferred to the System/390 Division, where he was responsible for the development of optical fiber communications for data processing applications. He has received three Division Awards, three IBM Outstanding Technical Achievement Awards, a Publication Award, and two IBM Invention Achievement Awards. He has published numerous technical papers and has co-authored two books in the field of optical communications and computers. Dr. Stigliani is currently a Senior Technical Staff Member in the Server Group Hardware System Design organization, responsible for future processor data communications infrastructure, I/O, and fiber optic applications.

Tim E. Bubb *IBM Server Group, 2455 South Road, Poughkeepsie, New York 12601 (bubb@us.ibm.com).* Mr. Bubb is an Advisory Engineer in the IBM eServer I/O Hardware Development group. He received the B.S. degree in electrical engineering from Virginia Polytechnic Institute in 1988, and the M.S. degree from Purdue University in 1989. He joined IBM at Poughkeepsie, New York, in 1990 and has held various technical and management positions in the eServer I/O design area. Mr. Bubb has received an IBM Outstanding Innovation Award for his work on the Hydra I/O subsystem, and he has received two IBM Outstanding Technical Achievement Awards for his work on the Multiprise 3000 and IBM eServer z900 I/O subsystems.

Daniel F. Casper *IBM Server Group, 2455 South Road, Poughkeepsie, New York 12601 (casper@us.ibm.com).* Mr. Casper is a Senior Technical Staff Member in the System Design group. He received a B.S. degree in electrical engineering from the University of Wisconsin at Madison in 1970, joining IBM at the Kingston Laboratory that same year. Mr. Casper has held various technical positions in the area of channels and system control element development. He holds numerous patents relating to channel design and has received ten IBM Invention Achievement Awards. He has received seven other formal awards, including two IBM Outstanding Innovation Awards, an IBM Outstanding Technical Achievement Award, and an IBM Technical Excellence Award. Mr. Casper is a member of the Institute of Electrical and Electronics Engineers.

James H. Chin *IBM Server Group, 2455 South Road, Poughkeepsie, New York 12601 (jchin@us.ibm.com).* Mr. Chin is an Advisory Engineer in the eServer I/O Hardware Development group. He received the B.S. degree in electrical engineering from New York Institute of Technology in 1981, joining IBM at Poughkeepsie, New York, that same year. Mr. Chin has held various technical and leadership positions in the eServer I/O design area. He has received numerous awards, including an IBM Outstanding Technical Achievement Award for his contributions to the zSeries I/O microcode development and an IBM Outstanding Technical Achievement Award for FICON microcode development. Mr. Chin is currently involved in the I/O microcode design for the next-generation eServer.

S. G. Glassen *IBM Server Group, 2455 South Road, Poughkeepsie, New York 12601 (glassen@us.ibm.com).*

Joseph M. Hoke *IBM Server Group, 2455 South Road, Poughkeepsie, New York 12601 (jmhoke@us.ibm.com).* Mr. Hoke is an Advisory Engineer in the eServer I/O Hardware Development group. He received the B.S. degree in electrical engineering from the University of Illinois at Chicago in 1987 and continued his studies under a university fellowship, receiving the M.S. degree in electrical engineering from Northwestern University in 1989. He joined IBM at Poughkeepsie, New York, in 1989 and has held various technical positions in the eServer I/O area. Mr. Hoke holds several patents used in the IBM ESCON and Sysplex products, and he has received two IBM Invention Achievement Awards. He has received IBM Outstanding Technical Achievement Awards for his work on ESCON, his work on the G5 server, and his contributions to the zSeries eServer.

Vahe A. Minassian *IBM Server Group, 2455 South Road, Poughkeepsie, New York 12601 (vahe@us.ibm.com).* Mr. Minassian is an Advisory Engineer in the eServer I/O Hardware Development group. He received the B.S. degree in electrical engineering from Stevens Institute of Technology in 1979, joining IBM at the Kingston laboratory that same year. Mr. Minassian has held various technical and leadership positions in the eServer I/O design area. He holds a patent used in the IBM ESCON converter products and received an IBM Invention Achievement Award. He received an IBM Outstanding Technical Achievement Award for his contributions to the zSeries ISC-3 hardware development. Mr. Minassian is currently involved in the I/O hardware design for the next-generation eServer.

John H. Quick *IBM Server Group, 2455 South Road, Poughkeepsie, New York 12601 (jhquick@us.ibm.com).* Mr. Quick is a Senior Software Programmer responsible for high-end eServer I/O logic card development and has more than 30 years in the electronic packaging field. Prior to joining IBM Kingston in 1981, he developed electronic packaging designs for military and commercial products. His previous IBM assignments in the Kingston and Poughkeepsie laboratories include various technical and leadership positions for designing ESCON directors and I/O channel card and backplane designs. He holds a B.S. degree from Empire State College at New Paltz, S.U.N.Y. Mr. Quick has published several packaging papers; he has received various IBM awards and, more recently, an IBM Outstanding Technical Achievement Award for his design contributions to the IBM eServer z900.

Carl H. Whitehead *IBM Server Group, 2455 South Road, Poughkeepsie, New York 12601 (whitehea@us.ibm.com).* Mr. Whitehead is a Senior Engineer in the eServer I/O Hardware Development group. He received the B.S. degree in electrical engineering from Manhattan College. He joined IBM at Poughkeepsie, New York, in 1979 and has held various technical positions in the eServer processor and I/O areas. He has received an IBM Outstanding Technical Achievement Award for his work on ESCON, in addition to two others for his I/O development work on the G5 and G6 servers.