The IBM Enterprise Systems Connection (ESCON) Director: A dynamic switch for 200Mb/s fiber optic links

by C. J. Georgiou
T. A. Larsen
P. W. Oakhill
B. Salimi

This paper describes the function and hardware structure of the Enterprise Systems Connection (ESCON™) Director™, an I/O switch capable of providing dynamic, nonblocking, any-to-any connectivity for up to 60 fiber optic links operating at 200 Mb/s. Optoelectronic conversion at the switch ports allows the switching of the fiber optic links to be done electronically. The establishment of paths in the switching matrix is done by means of a hard-wired, pipelined controller at a maximum rate of five million connections/disconnections per second. Routing information is provided in the header of data frames. The switch-port function, switching matrix, and matrix controller were implemented in the IBM 1-μm CMOS "standard cell" technology. The paper discusses the system interconnection philosophy, details of the data flow, the switch hardware architecture, the design methodology, and the approach to technology implementation.

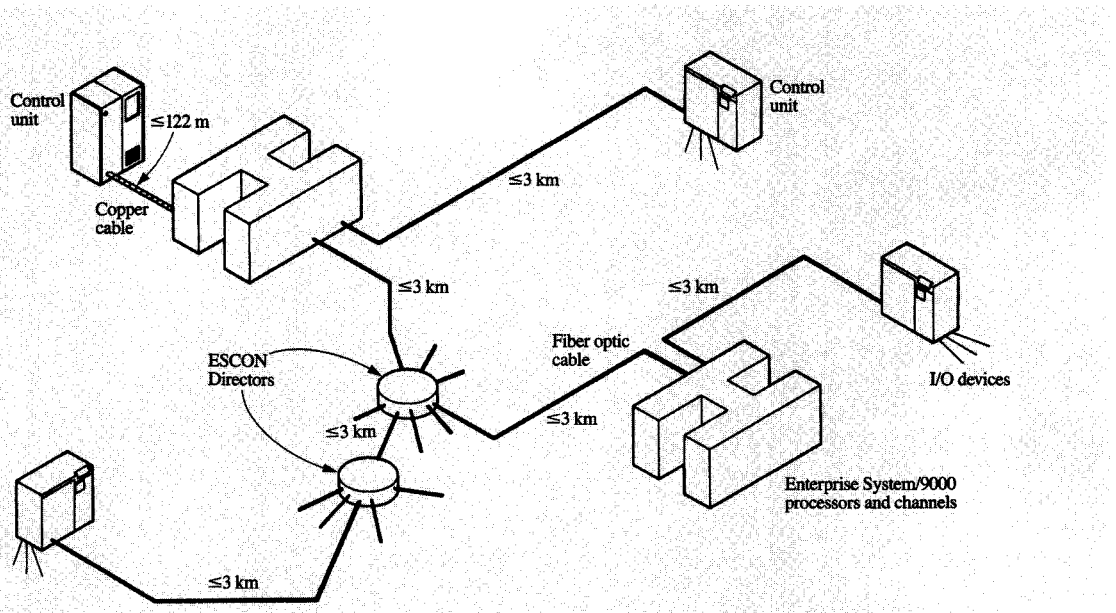

Figure 1

System connectivity via the ESCON Director.

1. Introduction

The ESCON Director™ provides the primary interconnection mechanism of the IBM Enterprise Systems Connection (ESCON™) Architecture™ for System/390® machines (see [1]). It is an implementation of a new switched point-to-point topology interconnecting serial System/390 I/O channels and control units. **Figure 1** shows a typical computer system configuration. The ESCON Director provides nonblocking, dynamic, any-to-any connectivity for up to 60 channels and control units that are attached to its ports via fiber optic links. Optical link distances of up to three kilometers are provided*, thus permitting channel-to-control-unit or channel-to-channel distances of up to 6 kilometers with a single ESCON Director, or up to 9 kilometers with two chained Directors, as illustrated in Figure 1. The ESCON Director is offered in two models: the 9032, ranging in size from 28 to 60 ports, and the 9033, ranging in size from 8 to 16 connection ports.

The switched point-to-point topology is central to the design of the ESCON architecture and ESCON Director. To appreciate the advantages of the switched point-to-point topology, we can compare it to point-to-point and multidrop topologies, those prevalent through the 1980s for computer I/O.

- *Topology comparison*

As **Figure 2(a)** shows, point-to-point topology means simply that a separate physical path is required between any two points that communicate. In the figure, each channel and control unit pair requires a separate communication path. A control unit that communicates with more than one channel needs a completely separate path to each channel. The point-to-point topology has the advantage of simplicity but the considerable disadvantage that channels and control units are often not used very efficiently. In addition, the number of physical paths and, consequently, the amount of cabling can be very large in even a relatively small computer center. It should be pointed out that the ESCON architecture also supports point-to-point topology for those cases in which the advantages of switched point-to-point connection are not needed.

In order to reduce the number of paths and increase channel utilization, multidrop topology can be used, because it allows a number of control units to share a path from a channel. In **Figure 2(b)**, the three control units can be accessed from any of the three channels. In this case, each channel has a separate path from which the control

*The 3-km limit has been extended to 20 km, the 6-km limit extended to 40 km, and the 9-km limit extended to 60 km with the introduction of the laser link product called the ESCON Extended Distance Feature (ESCON XDF), which utilizes single-mode fiber. The ESCON XDF was introduced as an IBM product in the fall of 1991.

units are "dropped," and each control unit requires a separate interface for each channel.

Switched point-to-point topology, shown in **Figure 2**(c), overcomes the disadvantages of point-to-point and multidrop. It provides the ability to switch connections via a switching unit, so that only one link is needed from each control unit or channel to the switching unit to realize any channel-to-control-unit or channel-to-channel connection. When a channel or control unit is added, it needs a path only to the switching unit rather than to all the nodes with which it communicates. With this arrangement, a potentially large number of point-to-point connections are made possible among a group of nodes. In some cases, this reduces the number of channels and control unit interfaces required and the total amount of cabling. To further reduce cabling requirements, the switching capability can be put where it is most useful. For example, if there are many more control units than channels, the switching capability can be put closer to the control units.

The ESCON Director is designed to operate as a nonblocking circuit switch. This means that in a system with N ports, $N/2$ simultaneous full-duplex connections can be established between any combination of port pairs. Once a connection is established, a direct path between the two attached end points remains in place until a command to disconnect is received by the ESCON Director. The total throughput capacity for full-duplex connections made available by the switched point-to-point topology provided by the ESCON Director is $2(N/2)BW_{link}$, where BW_{link}, the bandwidth of a simplex link, is 200 Mb/s. The network latency for initiating a connection between a source and destination, if the destination is not busy, is dependent only on the switch controller performance characteristics and not on the existing network traffic between other source–destination pairs. These are significant performance advantages over network topologies that share the transmission medium, such as bus topologies (e.g., Ethernet® [2]) and ring topologies (e.g., FDDI [3]).

• *System connectivity and configuration management*
The switched point-to-point topology implemented by the ESCON Director not only offers performance characteristics (i.e., aggregate network bandwidth) superior to those of other topologies, but simplifies the physical planning and reconfiguration normally associated with data processing installations.

The ESCON Director ports are versatile in that they allow attachment of either channels or control units. This flexibility allows the ESCON Director to adapt to the changing characteristics of a data processing installation. Link attachment changes are less disruptive in switched point-to-point topology because the addition or deletion of a link does not affect the connectivity of the other links in the network. With on-line applications continuing to grow and with 24-hour-per-day/7-day-per-week operation becoming increasingly common, system disruption cannot be tolerated. The ESCON architecture permits the modification of an installation while it remains operational. ESCON control units and channels can be added or disconnected from the ESCON Director without affecting the operation of the rest of the system.

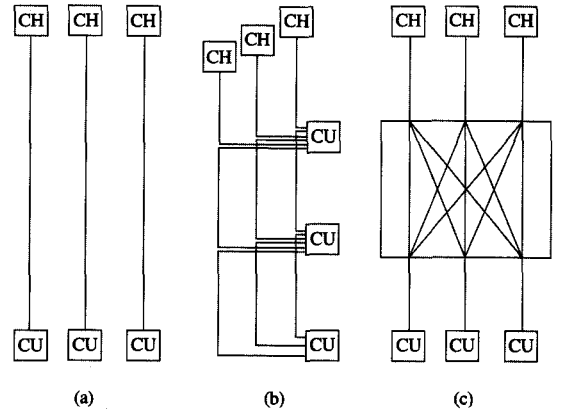

Figure 2

Comparison of network topologies: (a) point-to-point; (b) multidrop; (c) switched point-to-point.

Through the use of two cascaded ESCON Directors, I/O devices may be located up to 9 kilometers from the host processor complex. The greater distances provide increased flexibility in the physical planning of data processing installations, as well as new alternatives for disaster backup and recovery planning.

The any-to-any connectivity characteristics of the ESCON Director at an installation can be customized. The customization can be done either by means of an easy-to-use operator interface provided by a standard PS/2® computer or by means of an interface to a host running the ESCON Manager™ application program. The ESCON Director can be adapted to varying system needs; e.g., a single physical ESCON Director can be logically partitioned into multiple subDirectors, thereby ensuring data security/integrity protection among multiple applications that may be sharing it. The isolation of test and production systems is a case in which this capability may be valuable. The ESCON Director includes a password capability to protect against unauthorized use of selected operator console functions. It also includes a local audit-trail capability for time-stamping and logging

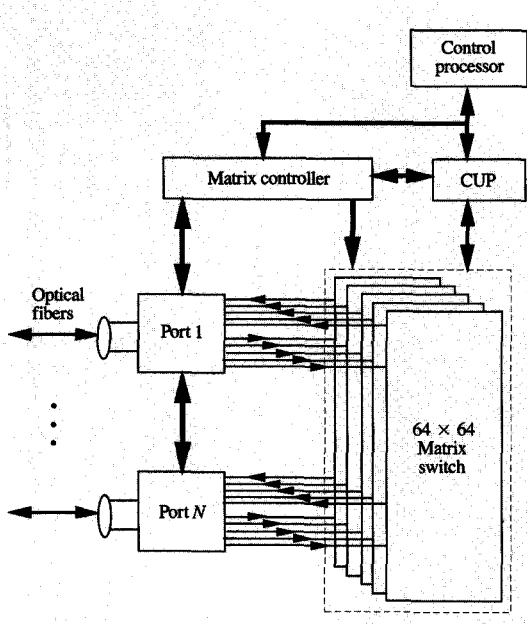

Figure 3

ESCON Director block diagram.

configuration changes, password changes, and access to certain utilities.

System availability can be enhanced by the appropriate planning of Director connections. A computer facility can have more than one Director, with each control unit connected to ports on at least two Directors. Also, hosts can be connected to at least two Directors by separate channels. This ensures that communication can continue if a Director or a link to or from a Director should become unavailable.

The ESCON Director port granularity (ports are packaged in groups of four) provides the necessary flexibility to accommodate the growth requirements of an installation. An ESCON Director with the minimum number of ports may satisfy the initial connectivity requirements of a new installation. As the number of processors, channels, and control units in the installation grows, more ports can be added. The ESCON Director can also be installed with more ports than are initially required. As configurations grow and additional channels and control units are installed, they can be attached to the unused ESCON Director ports with no disruption.

- *Outline of paper*

This paper is organized as follows: Section 2 gives a functional description of the ESCON Director, Section 3 describes its hardware structure, Section 4 discusses the packaging, Section 5 the performance analysis, and Section 6 the design system. In Section 7, we present the considerations that led to the selection of the technology, and in Section 8, we provide a summary of the technology analysis that proves the validity of the design. Finally, Section 9 presents the physical design. A summary is provided at the end.

2. Functional description

Figure 3 shows the basic components of the ESCON Director in block diagram form. The ESCON Director consists of ports that interface to fiber links, a matrix switch that provides connectivity between the ports, a matrix controller that communicates with the ports to establish or remove connections via the matrix switch, a control processor, and an internal port (the CUP or control unit port). The CUP does not have a fiber link attached to it, but can be connected to a host channel via the matrix switch and another Director port. When the host and CUP are connected in this manner, the host channel communicates with the Director as with any other control unit.

The initial concepts of the Director structure can be found in [4–6]. The Director was designed to be a crosspoint, nonblocking circuit switch. Such a switch combines simplicity of structure with significant performance advantages over other types of switches that have been used in telephony or parallel processing (e.g., multistage networks [7]). The relatively small size of the Director (i.e., 60 ports) and the availability of low-power, high-density VLSI technologies (e.g., CMOS) made a crosspoint matrix switch feasible and attractive. A single high-performance, centralized matrix controller that has sufficient capacity for handling the connect/disconnect requests required for system I/O operations was developed [5]. Performance analysis (see Section 5) verified this point. Some of the other functional characteristics of the Director are discussed in the following sections.

- *Port partitioning and static connections*

A Director allows any port to be connected with any other port through the matrix. However, to satisfy specific application requirements, these port connections may be *dynamic* or *dedicated*. A dynamic connection is automatically established and broken for each communication exchange [1], whereas a dedicated connection joins two ports continuously, to the exclusion of all other ports.

Unrestricted *dynamic* connectivity allows all devices attached to a Director's fiber optic ports to communicate with each other dynamically. This means that any attached channel or control unit may communicate with any other attached channel or control unit that is not currently

connected. A dynamic connection occurs at the time it is needed rather than at a predetermined or fixed time. A temporary communication path, determined from routing information contained in the data, is established between two ports for the duration of data transmission only. The connectivity of each port may be restricted to specified ports.

A Director can also support *dedicated* connections between some or all of its ports. A dedicated connection provides a continuous communication path between two Director ports and restricts those ports from communicating with any other ports. The data stream is passed directly through the Director, from one port to the other, without regard to routing information in the data. The Director contains a configuration table—a static RAM—that specifies the status of dynamic and dedicated connections.

• *In-band signaling*

As previously discussed, the ESCON Director conforms to the ESCON architecture [1]. Data transmissions occur in the form of frames, each frame containing a header, data field, and transmission-error-detection field (i.e., cyclic redundancy check, CRC). The transmitted bits are encoded in 8B/10B code [8]. A frame is bounded by special 10-bit characters, called delimiters.

An *in-band signaling* technique was chosen as the method of passing connection information to the ESCON Director. With this technique, the frame-routing information for dynamic connections is included in the frame header, whereas with out-band signaling, the frame-routing information is sent separately, on another link. The advantages of in-band signaling over out-band signaling are that it requires only one link per node, instead of two, and it minimizes the latency in making and breaking connections in the switch by limiting the number of "handshakes." This is an important protocol design parameter in view of the long distances supported by the architecture and the propagation delays of the fiber link (5 μs/km).

The adoption of in-band signaling meant that a connect frame (a frame specifying the establishment of a new connection) had to be buffered in the port until its addressing information could be processed by the matrix controller. Thus, the size of the buffer became an important design parameter. A large buffer would allow large connect frames to be buffered, at the expense of port complexity and additional cost. Because of the CMOS functional density at the time (see Section 7), the buffer size was set to 80 bytes, limiting the size of connect frames.

• *Handling destination-busy conditions*

Another important ESCON Director feature is the way it handles destination-busy conditions (connect requests to busy ports). Two approaches were considered. In the first approach, also referred to as "camp-on," the connect request would be stored in the port buffer and remain pending until the destination port was freed. Then, the matrix controller would make the connection in the matrix switch, and the connect frame would be allowed to reach its destination. In the second approach, the source port would respond immediately to the sender with a "destination-busy" frame, i.e., a negative acknowledgment. The sender could retransmit the connect request to the same destination after a time-out. The latter approach was chosen, since it was determined that the camp-on would significantly increase complexity without a compensating increase in performance. Furthermore, camp-on was not required, because a channel normally attempts to connect to a different control unit (for the sake of efficient utilization of resources) instead of retrying on the busy path.

• *Handling error conditions*

Errors (unusual conditions detected by the Director) and Director hardware failures are reported with an "unsolicited alert" message over previously designated primary or alternate links. These events are logged on a diskette in the Director PS/2 operator interface. Director hardware failures are also indicated to the local operator. The identity of a failing field-replaceable unit (FRU) is included in the reporting. Under certain conditions, an indicator on the failing FRU is also turned on.

3. Hardware structure description

• *ESCON Director port*

ESCON Director ports (**Figure 4**) provide the interfaces to which the fiber optic channels or control units are attached, and contain the following major components: a fiber optic transmitter/receiver, a serializer/deserializer, and a port adapter.

The fiber optic receiver [9] contains a PIN-type photodiode for converting serial optical signals from the incoming multimode fiber into serial electrical signals. Receiver electronics amplify the electrical signal from the photodiode and provide the automatic gain control that is needed to accommodate the large fluctuations in the received optical power. The serial signals are received at a rate of 200 Mb/s and, as previously mentioned, are encoded in 8B/10B code [8]. The optical receiver also includes an error-detection mechanism that indicates when optical power has been lost on the incoming fiber. The fiber optic transmitter consists of a 1.3-μm "surface" light-emitting diode (LED) and a driver chip, which converts incoming logic levels to currents capable of modulating the LED output power. Details on this fiber optic link are provided in another paper in this issue [9].

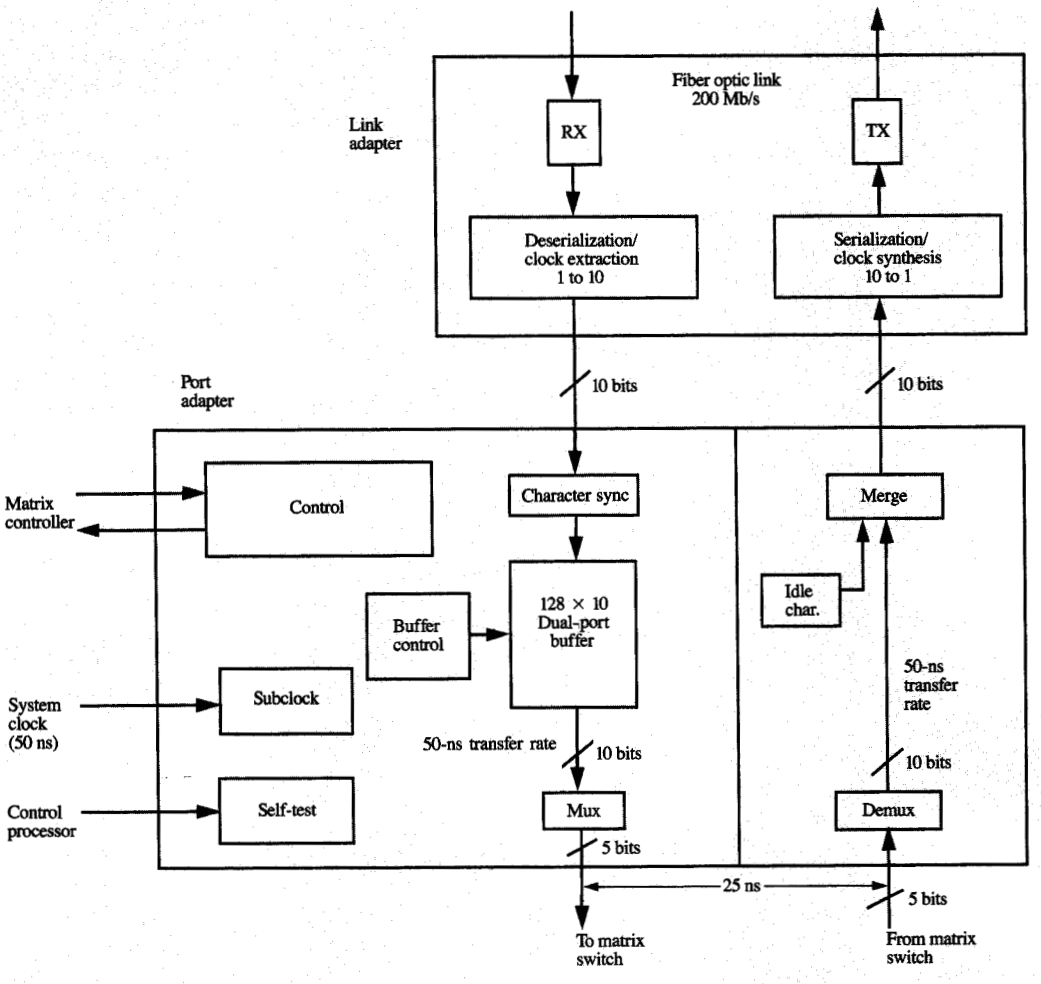

Figure 4

ESCON Director port.

The serializer/deserializer module performs several functions. It contains a phase-lock loop (PLL), whose function is to reshape the incoming data stream and then extract a 5-ns clock from it. This clock, called the bit clock, is used to transfer the incoming data into a 10-bit deserializer. A 50-ns clock, called the character clock, is derived from the bit clock and is used to transfer, in parallel, the 10-bit characters into the port adapter. Similarly, data from the port adapter, 10 bits wide, are transferred to the serializer/deserializer module by use of the 20-MHz (50-ns) Director system clock. This system clock is also provided to the transmitting section of the serializer/deserializer, where it is multiplied to a frequency of 200 MHz by means of a PLL. The resulting 200-MHz clock is used to serialize the parallel data to be transmitted and transfer the data to the 1.3-μm LED driver. The serializer/deserializer module can perform a "serial electrical wrap" function, for diagnostic purposes, in which the outgoing data are sent back to the incoming data path.

The port adapter module utilizes the IBM 1-μm CMOS technology [10] and contains the complete function of two ports, a single set of self-test circuitry, and one Director system-clock-generation function, implemented with approximately 28 000 cells on a 9.4-mm chip. This chip is packaged on one 36-mm MCP (metallized ceramic with polyimide) module with 183 signal pins.

Besides communicating with the attached device or channel over the fiber link, the port has internal interfaces to the matrix controller and the matrix switch. The

primary functions of the port adapter module are to buffer the data that will be sent to another port through the matrix switch; perform code checking on the incoming data in order to detect special characters or character sequences that may cause a dynamic switch connection to be established or removed, identify frame boundaries, or signify a link condition; determine the state of character synchronization and detect transmission code violations; initiate connect or disconnect requests to the matrix controller; transmit reject/busy messages to the attached device ports when appropriate; and perform diagnostics functions.

A monolog/dialog scheme is implemented, which allows maximum switch transparency when two ports are actively communicating with each other. The monolog state resolves potential races that may arise during the establishment of connections, such as the case of "ships passing in the night." For example, if port B requests a connection to port C at the same time that port A is requesting a connection to port B, port A receives a busy response. The dialog state exists when two ports are connected. Data in both directions are monitored for disconnect data-frame delimiters, so that the connection can be broken either by the source port (the port that initiated the connection) or by the destination port. Finally, a port is in the inactive state when no connection has been either initiated or established.

The 60 ports of the ESCON Director are arranged in eight groups of eight ports each, except for group one. Group one contains only six ports, two of which are spares, and the control unit port. These groupings play a role in the port priority scheme for port requests to the matrix controller. Installed ports are recognized at IPL (initial program load) time when the address-check table is loaded with all zeros.

• *Matrix controller*

The matrix controller (MC) services the connect and disconnect requests from all of the ports and controls the port-to-port interconnection paths through the matrix switch. As part of this process, it maintains tables that contain the current status of the ports, such as which ports are available, connected, operational, etc. The requests serviced by the MC fall into four categories: a successful request by a port to be connected to another port, a request by a port to be connected to another port that is then rejected for any of a number of reasons, a request to break an existing connection, and a request by a port to present its status to the control processor.

The matrix controller, shown in **Figure 5**, achieves its performance objective by means of a parallel pipeline design [5]. The MC pipeline consists of four 200-ns stages, each with four 50-ns clock subcycles. The processing of a connect request is described here as an example of the operations that take place in the four stages of the pipeline. In the first stage, when a connect request occurs, the port presents the connect request and the destination port address to the MC. In the second stage, the port presents the source port address to the MC. At the same time, the pipeline might begin processing the first stage of another connect request (presenting the request and destination address to the MC). In the third stage, the source and destination addresses are transferred to the matrix switch, and the connection command is issued in order to establish the forward and return paths. Finally, in the fourth stage, all error conditions, if any, are processed.

The MC interface to the ports includes one connect request line and one disconnect request line from each of the eight groups of ports. In response to a request, the MC activates either the connect select or the disconnect select line for that group. A port in the selected group (chosen as described below) responds by sending source and destination port addresses to the MC, which uses these addresses to make or break the desired connection. For a connect operation, the involved ports are notified by the MC that they are being connected, and the source port is sent a signal by the destination port to begin transmitting through the matrix switch.

Disconnect requests have higher priority in the MC than connect requests, in order that connections that are no longer needed can be broken as quickly as possible. This helps reduce the number of times a connect request encounters a busy destination port. More specifically, every 200 ns the MC samples the port request lines for active disconnect or connect requests. If there are disconnect requests, it selects one, in round-robin fashion, and activates the disconnect select line to that group. If there is a connect request and no active disconnect request, the MC chooses a connect request, in round-robin fashion, and activates the connect select line to its group.

The select lines of each group pass serially, in effect, through each port in that group, so that the priority of a port within its group is determined by its position on the select line. This arrangement provides a defined order of priorities within each group of ports which can be taken advantage of, if necessary, to optimize the performance of the ESCON Director in fully configured, high-traffic environments. This can be accomplished by assigning the busier nodes to the higher-priority ports.

The matrix controller logic is implemented on two CMOS chips, MC1 (9.4 mm) and MC2 (7.5 mm), as shown in Figure 5. MC1 contains host/operator command execution (pipeline), error reporting, interfaces to the matrix switch and ports, the activity and partition/check table arrays, ten-to-eight decode, and the subclock and self-test logic. MC2 contains the round-robin port-polling circuitry, the subclock and self-test logic, and the interface

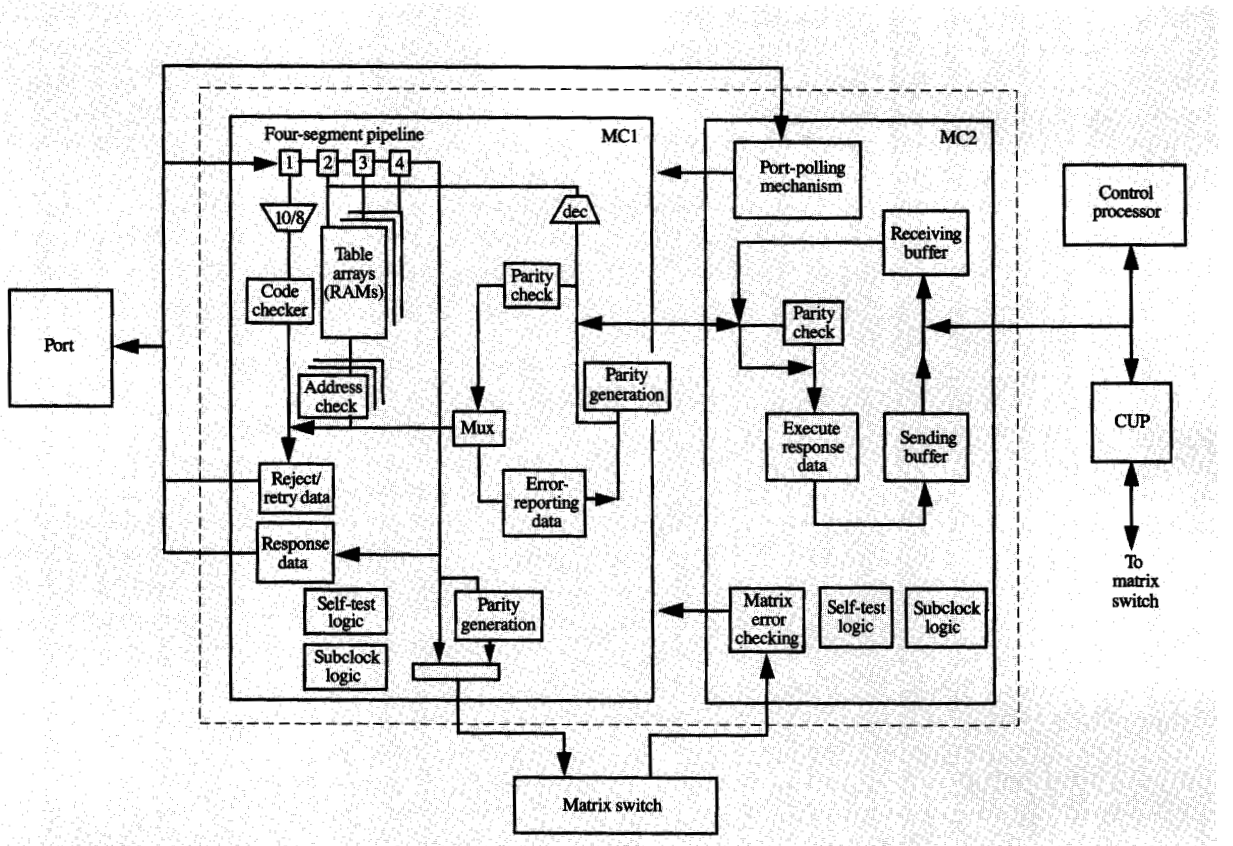

Figure 5

Matrix controller.

to an Intel® 8255 chip, via which MC2 can be accessed by the control processor and CUP.

Dynamic operations between the MC and the ports are handled by two functional islands: "round robin" in the MC2 chip and "port interface" in MC1. The round-robin functional island contains the control logic for prioritizing and selecting the service requests from the ports and the control processor. The port interface logic processes the selected request. As previously discussed, service request processing is pipelined, with the ability to start a new service request every 200 ns and complete the operation in 800 ns; thus, both of these functions, i.e., prioritization and selection of service requests and request processing, can take place at the same time but for different service requests.

• *Matrix switch*

The matrix switch provides nonblocking, any-to-any, full-duplex connection for up to 60 ports. Once a connection is established by the matrix switch at the request of the matrix controller, a stream of characters is transferred. The matrix switch maintains the connection for the data stream until the matrix controller requests a disconnection. For each data stream, each character is transferred in two 25-ns slots, five parallel bits at a time. The matrix switch consists of five identical planes, each plane handling one bit of data for all 60 possible ports in a fully configured ESCON Director. A multiple-stage, clocked design enables the matrix switch to meet the required data-transfer rate of one 10-bit character every 50 ns for each established connection.

The internal structure of the matrix switch plane is shown in **Figure 6**. The plane performs a logical crosspoint function, since it provides nonblocking connectivity between any pair of attached Director ports. The actual implementation of the internal plane connections is done via multiplexers. The plane contains sixty-four 64:1 multiplexers (60 for data ports, one for the CUP, and three unused) plus a spare, input and output latches, and control logic. Each multiplexer selects which one of the 64 plane

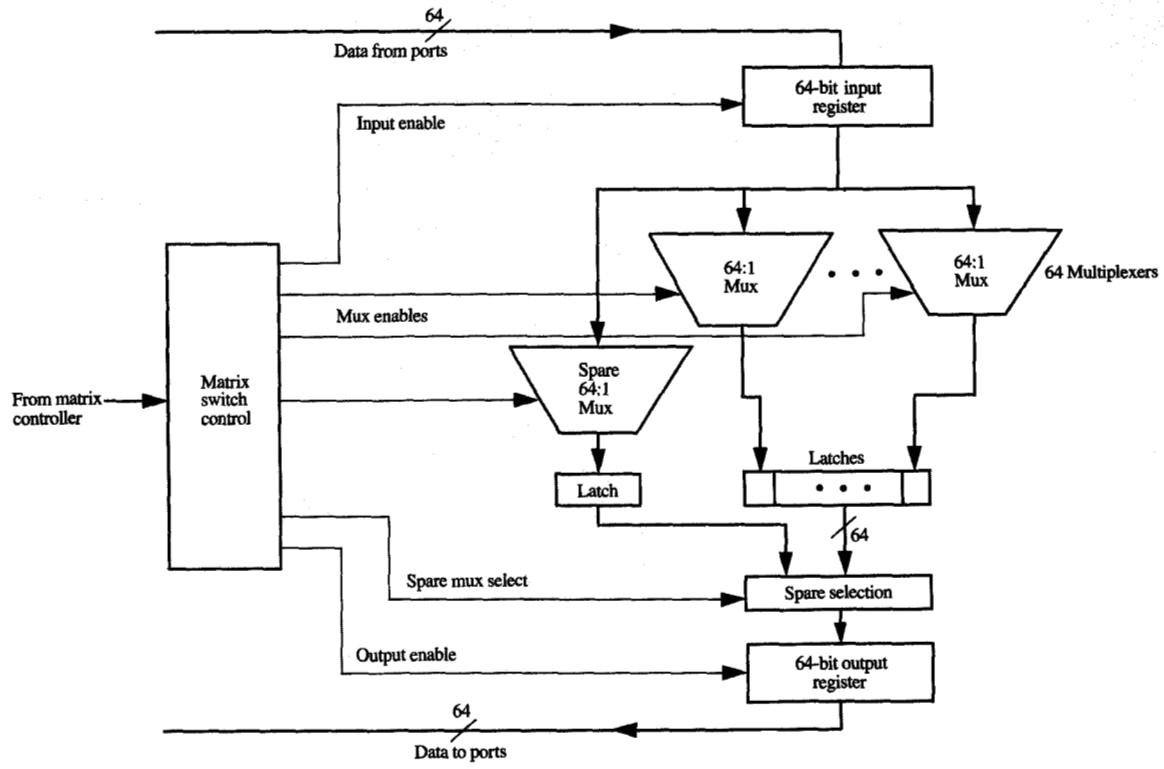

Figure 6

Matrix switch chip.

inputs is to be connected to its output. Thus, a logical 64 × 64 crosspoint configuration is realized. The transmission of each bit of the data streams through the matrix switch takes place in the following stages: transferring the bit from the sending port to the switch input register, passing it through the multiplexer that is selecting the sender of the bit, setting the output latch of the multiplexer, transferring the latch contents to the switch output register, and sending the output register bit to the corresponding receiving port. The multiple-stage transfer (completed in three 25-ns cycles) is pipelined, with new character bits entering each cycle. The multiple-stage transfer is done simultaneously for all established connections. Each matrix switch plane is implemented on a 7.5-mm chip containing 15.5K circuit cells. The multiplexer macro was already available in the CMOS technology (see Section 7), so it could readily be incorporated in the design.

The matrix switch is under the control of the matrix controller, which, as previously discussed, provides it with control signals required to establish and break connections, perform diagnostics, and use backup logic in the case of internal hardware errors. Parity is checked on all of the control signals sent from the matrix controller to the matrix switch, and all matrix switch errors are reported to the matrix controller.

The interface between the matrix switch and the port adapters created two performance problems that had to be addressed. First, the CMOS technology did not allow all 60 matrix off-chip drivers to be switched at the same time. Therefore, a switching "window" was created to skew, in time, the switching of 30 drivers with respect to the remaining 30 drivers. Second, because of large differences in printed circuit wiring length and variations in chip performance, there was the possibility of data being captured in the wrong cycle. A detailed discussion of the above problems and the solutions found is presented in Section 8.

The matrix switch can maintain its nonblocking characteristics even in the case of an internal hardware failure. This is accomplished in a different way for each of the two ESCON Director models. The matrix switch for

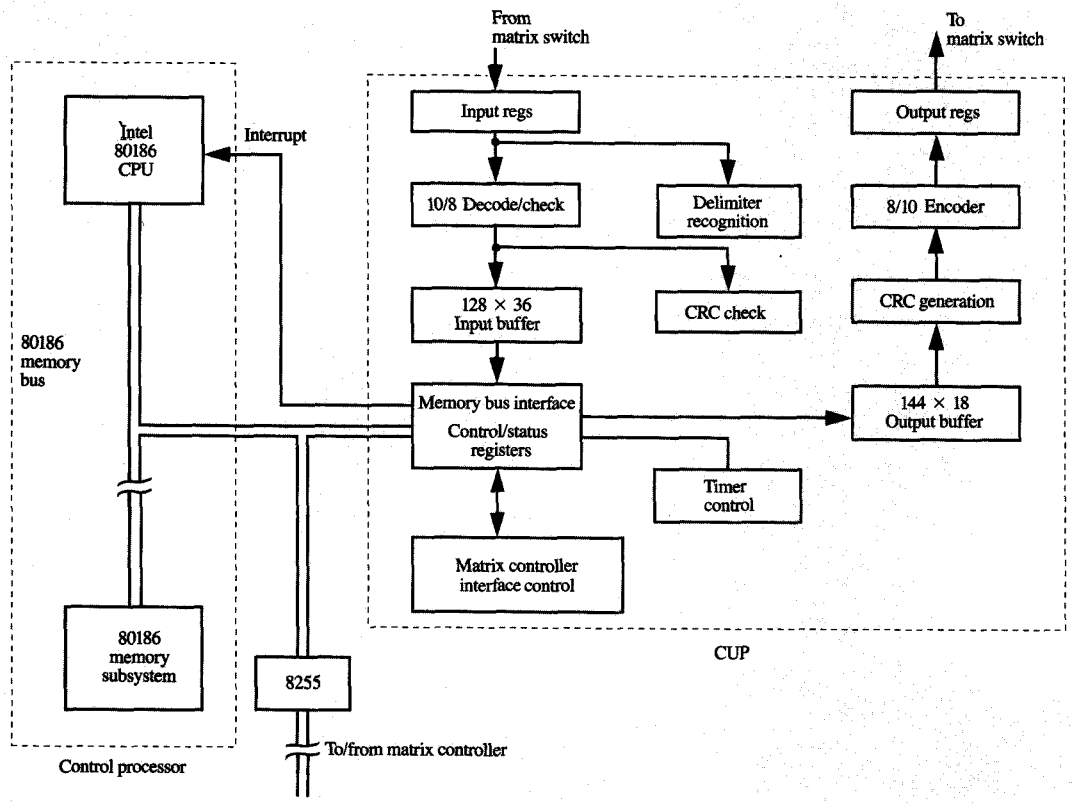

Figure 7

Internal port (CUP).

the Model 9033 Director (16 ports) has full duplication of all internal logic. In the case of a detected internal hardware fault, the backup internal logic becomes the primary logic. Since all crosspoint connections are replicated in the backup logic, there is no disruption of connections when switching over to the backup logic. The matrix switch for the Model 9032 Director (60 ports) provides duplication of most of the data paths. In case of an internal hardware fault, a spare multiplexer replaces the failing one. This provides fault tolerance for roughly 50% of all internal logic.

• *Internal port (CUP) and control processor*

The internal port (**Figure 7**) functions as the control unit port (CUP) for the ESCON Director, in accordance with the ESCON architecture. As all other ports, the CUP receives and interprets link-level frames, which contain the control information for the ESCON Director. In the case of the CUP, these include requests to establish and remove connections between the CUP and a host processor via a channel. The CUP also receives device-level frames, which are passed on to the control processor. These frames contain user data and control information, which are used to partition the Director, to transmit or collect status and error information, and to perform similar functions.

Unlike externally attached channels and control units, the CUP is directly attached to the matrix switch (i.e., there is no fiber optic link and port adapter between the CUP and the matrix switch). As a result, the CUP requests via its MC interface that connections and disconnections be made. The CUP is physically located on the control processor (CP) card and consists of one 9.4-mm CMOS chip. The CP is an Intel 80186® microprocessor. All communication between the CUP chip and the CP takes place through memory-mapped status (to the 80186) and control registers (from the 80186). Two memory-mapped arrays are resident on the CUP chip for use in transmission and reception of frames. These arrays are called the output buffer and the input buffer.

• *ESCON Director clocking*

The clocking function generates and controls distribution of clock signals throughout the Director. It includes a master clock, a master clock distribution tree, subclocks, and subclock distribution trees. Under control of the CP, the clock function also manages special operating modes, such as level-sensitive scan design (LSSD) scanning [11] and self-test, by generating and distributing appropriate clock signals.

The Director clock function was designed to keep all ports synchronized, thus eliminating the need for transaction buffering and development of asynchronous protocols between the individual ports. To meet the strict synchronization requirements, a distributed architecture was chosen. The clock designs for both the Model 9032 and Model 9033 Directors feature this distributed architecture. From a controlling central function called the master clock, a bus carrying carefully timed and matched oscillation pulses is distributed to clock subassemblies throughout the system. The clock distribution network also controls and distributes test clock sequences during system initialization and diagnostic operations.

Had certain clock-distribution paths failed, the entire Director might have failed; therefore, a redundancy scheme was created. There are two crystal oscillators. If the primary oscillator fails, the backup oscillator is switched in. Also, there is similar redundancy for the on-chip clock function within the master clock itself. In both cases, the redundancy provides fault tolerance for most of the key clocking circuits.

The primary and backup oscillators are both 40-MHz crystal oscillators, mounted next to the master-clock chip. The 40-MHz oscillator is the common clock reference for all Director timing, both on the master-clock assembly and on the clock subassemblies. The base clock and its subclocks create the 25-, 50-, and 100-ns cycle times that are used throughout the design.

At each functional logic partition, there is a resident subclock that transforms the pulses into latch and trigger clock pairs. Most of the partitions use multiple cycle times to meet their critical timing requirements. The subclock generates the array clock for those partitions that use on-chip arrays.

A major advantage of this distributed architecture is the simplification of clock distribution throughout the Director. The approach greatly reduces the board wiring complexity, which otherwise might have been prohibitive. Second, it places clock logic into the functional islands for added flexibility. Although each subclock is fundamentally the same, the use of the clock was optimized for each area. With a centralized clock design, this flexibility would have been more difficult to achieve.

Individual subclock functions can be switched between system mode and idle mode in response to a control signal from the master-clock chip. When a partition is idle, it receives and responds to test clock signals and diagnostic procedures. The subclock is designed to cooperate with the self-test macro, also resident on the chip.

• *Self-test*

Self-test, in which the tests are randomly generated, is a special case of LSSD scan. LSSD has been the IBM chip and card test approach for many years [11]. Instead of applying LSSD patterns to a chip and card from an external source, the self-test function creates the test patterns by means of on-chip logic. The result of each test is a "signature." This signature can be compared to a previously defined "good" signature in a go/no-go test.

Built-in self-test is implemented on the 9032 and 9033 models for use by card test and unit test. At the card-test level, a "tester" (test equipment) is customized to initiate the card self-test function and to collect and interpret the results. At the unit-test level, self-test is initiated when the machine is turned on.

Self-test is implemented as follows: Two linear-feedback shift registers (LFSRs) are embedded into each CMOS chip. One LFSR is configured as a pseudorandom-pattern generator (PRPG), and the other is configured as a multiple-input signature register (MISR). The PRPG supplies pseudorandom test vectors to the LSSD scan paths, and the vectors are subsequently applied to the logic under test. The MISR compresses the output test results into a signature. After a predetermined number of scan and system clock sequences (controlled by microcode), the compressed signature is scanned out of the chip and compared to a "known good" signature (stored bit pattern) in the test equipment. A mismatch indicates a chip or card fault.

Drivers, receivers, and wiring between chips or cards are tested by means of a set of predetermined patterns. These tests are run at unit power-up as part of the self-test function. The self-test sequences are embedded in a PROM on the control processor card and controlled by microcode. The control structure of the microcode allows the self-test function to be run at power-up with no need for external stimuli.

4. Packaging

The ESCON Director utilizes "card-on-board" technology [12] to meet its cost and performance requirements. Several different board and card technologies were analyzed in detail for the CMOS environment by means of ASTAP (the Advanced Statistical Analysis Program) [13]. The following summarizes the first-, second-, and third-level packaging as well as connector systems that were chosen for the ESCON Director. All the packages described have been developed and manufactured by the IBM Technology Products Division.

- *Module*

All modules used in the Director are single-chip PIH (pin-in-hole) modules [14]. The following highlights some of their characteristics:

- *MC/FL (metallized ceramic/fine line)* This is a 36-mm pin-grid array (PGA) module with 150- and 166-signal I/Os for 7.5- and 9.4-mm chips, respectively.
- *MCP/RP (metallized ceramic with polyimide/reformed pins)* This is a 36-mm high-performance module with power and ground planes to reduce on-module coupled noise and ΔI noise [15]. This SCM (single-chip module) was used extensively throughout the Director to carry critical logic chips with many simultaneously switching drivers. This module supports 179- and 183-signal I/Os for 7.5-mm and 9.4-mm chips, respectively.

- *Card*

The printed circuit card technology used, known as 4S3P [12], contains four signal and three power planes. This card technology was selected because preliminary wiring analysis at the card level indicated that multiple signal planes with multiple wiring channels were necessary to meet wiring and performance requirements. This card has a sufficient number of power planes to reduce on-card wiring crosstalk. The 4S3P card has three-line-per-channel (3 L/C) capability (i.e., three parallel conductors between the card vias), allowing more circuit nets to be wired in the same channel with shorter lengths. However, off-card critical nets were wired utilizing a 2 L/C strategy with center line omitted. This approach reduced the coupled noise significantly, as the drivers used in the critical paths had very large dv/dt (up to 5.0 V/ns).

- *Board*

The printed circuit board technology chosen for the Director design uses 3 L/C on the S1 and S6 signal planes and 4 L/C on the S2, S3, S4, and S5 signal planes. The number of lines allowed on a signal plane depends on the distance between the signal plane and its nearest power plane and the way the signal plane is sandwiched between the power planes. The close proximity of the signal planes to power planes results in reduced line-to-line coupling coefficients and, consequently, reduced coupled noise. The center line on S1 and S6 and one of the center lines on S2, S3, S4, and S5 were eliminated to reduce the coupled noise for critical paths.

- *Connector system*

The card connector used contains 268 pins in four rows, satisfying the I/O density requirements. The coupled noise through this connector is reduced significantly as a result of power and ground planes sandwiched between the middle pins (rows 2 and 3). This structure reduces the effective mutual inductance between the connector pins, and, as a result, the fast voltage transitions of CMOS drivers do not generate significant coupled noise.

5. Performance simulation

Because the ESCON Director is involved in every I/O transfer between the channels and control units attached to it, the time it requires to make connections and propagate frames is an important factor in overall system performance. The use of pipelining in the matrix controller enables it to begin servicing a new connect or disconnect request every 200 ns. This means that the Director can make and break connections at a maximum rate of five million per second. Early in the design of the Director, performance modeling was used to determine the effectiveness of the priority scheme used in servicing the requests for connects and disconnects. The objective of the modeling was to determine what effect this five-million-per-second service rate capability would have on the delay a channel or control unit would experience when initiating a connection through the Director with up to 60 ports requesting service concurrently (the theoretical limit). Through performance modeling, approximations of the distributions of these queuing delays were determined for various configurations and traffic loads.

- *Performance model*

The priority scheme that was described in the subsection on the matrix controller can be summarized as follows: Disconnect requests have priority over connect requests; requests from port groups are serviced with equal priority in round-robin fashion; there is a fixed priority of ports within each group. The model used to evaluate the performance of the Director essentially emulates this priority scheme.

The model was implemented using RESQ (RESearch Queueing Package), a software package for constructing and analyzing queuing network models [16]. Because RESQ simulations are driven by pseudorandom numbers, the results are actually those of a statistical experiment and are subject to the inaccuracies inherent in such random trials. To specify the statistical accuracy of the results, one can provide confidence levels that will, in turn, determine the length of a simulation or the number of independent replications required to achieve the desired level of confidence.

The hierarchical modeling capability of RESQ was utilized to construct the performance model of the ESCON Director. A submodel was defined to represent one group of eight ports and was then replicated eight times to model all the ports in the switch. Each of these replications generates jobs, according to specified interarrival time distributions for each port in its group, that represent connect requests. Associated with each of these jobs is

such information as port number, connection duration, destination port number, group number, and priority. Each job proceeds through the model to the connect queue for its group (the part of the model that represents the ports in the group waiting for a connect select from the matrix controller). The appropriate group connect requests are sent to the matrix controller model. Statistics are collected about the queue lengths and waiting times.

A submodel representing the round-robin, intergroup priority scheme found in the matrix controller periodically polls the group connect and disconnect requests and, every 200 ns, sends a select to one of the group submodels that has an active disconnect or connect request. The connect request with the highest priority in the connect queue of the selected group is allowed to proceed through the model. If the destination port chosen by this job is already connected to some other port, the requested connection cannot be made, and the job is routed to a specified delay block that represents the time taken by a channel or control unit to turn around a "destination busy" response and retransmit the connect frame back to the switch. After this delay expires, the job re-enters the connect queue for its group, where it again waits for service with the other active requests.

When a connect request is selected and its destination port is available, it proceeds through a series of fixed delays that represent the time required for the matrix controller to establish the connection and propagate the connect frame through the switch. It then enters a user-specified delay that represents the total duration of the connection, including fiber propagation delays and overhead at the channel and control units for sending frames back and forth. After this delay, the job is turned into a disconnect request and enters the appropriate disconnect priority queue, where it waits for service from the matrix controller model to break the connection. Some time after the connection has been broken, a new connect request is generated for that port. (The "interarrival time" from request to request for a port is a random number, generated according to a negative exponential distribution with a specified mean. If, however, the interarrival time is less than the sum of all the times to process the request and then break the connection, the interarrival time is not used. Instead, the new connect request is generated after a fixed post-disconnect delay.) This circulation of connect and disconnect jobs continues for the duration of the simulation.

As implied in the previous description of model operation, there is a group of input parameters for each port in the model that allows one to specify certain characteristics of the attached channel or control unit, the configuration, and the traffic generated by each attached entity. Specifically, the input parameters available for each port include

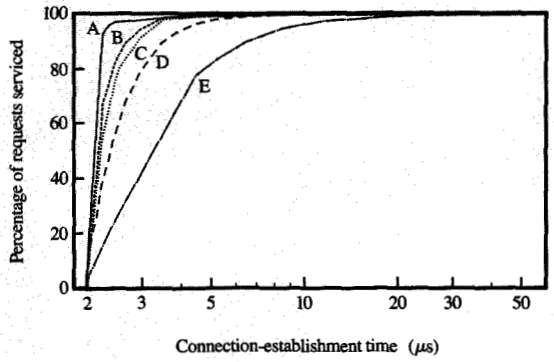

Figure 8

Cumulative distribution of connection-establishment times.

Table 1 Request rates for simulations of Figure 8.

Simulation	*Aggregate request rate*	*Connect request rate*	*Disconnect request rate*	*Percent destination busy*
A	0.444	0.414	0.030	92.7
B	0.908	0.878	0.030	96.6
C	2.938	2.880	0.058	98.0
D	3.979	2.330	1.649	29.3
E	4.929	2.980	1.949	34.6

Note: Rates are in millions of requests per second.

- The mean interarrival time of connect frames to the source port.
- The allowable destination ports for each source port. The actual destination assigned to each connect frame is chosen with equal probability from the allowable destinations specified.
- The busy-response turnaround time. The busy-response time specifies the delay from the time the busy-response frame is sent from the switch until the connect frame to retry the connection is received at the switch.
- Total connection duration. This parameter specifies how long a connection will stay in place before a disconnect request is generated to break it. It includes frame transfer times, fiber delays, interframe gaps, and frame turnaround time at the end points.

- *Performance simulation results*

Various traffic loads were simulated for the ESCON Director by using different combinations of the above parameters, and connection-establishment-time distributions were collected (**Figure 8** and **Table 1**).

Connection-establishment time is defined here as the delay from the time the beginning of the connect frame enters the source port until the connection has been established and the beginning of the connect frame exits the destination port of the switch—that is, the time it takes for a connect frame to get through the ESCON Director. Each curve shows the cumulative distribution of connection-establishment times for all of the 60 ports involved in the corresponding simulation. The simulations represent aggregate request rates (sum of connect and disconnect requests) ranging from less than 500 thousand per second to over 4.9 million per second, which is 98% of the maximum rate the ESCON Director can achieve. Table 1 presents the request rates and the percentage of requests to busy destinations.

Figure 8 shows the cumulative percentage of requests serviced as a function of connection-establishment time for various traffic patterns. The configuration represented by curves A and B has 30 channels and 30 control units; any channel can request a connection to any control unit, and vice versa. The mean interarrival time of initial connect requests (i.e., requests that are not the result of retrying a busy destination) from each channel and control unit is 2 ms. The duration of connections initiated by channels is 72 μs; the duration of control-unit-initiated connections is 385 μs. (Although the configuration and input parameters are the same for curves A and B, the results are different because more of the randomly chosen destinations for curve B turned out to be busy when connections were attempted. This resulted in a higher connect request rate because of repetitive retries of connect requests to the busy destinations.) Configuration C has 20 channels and 40 control units; the mean interarrival time of initial connect requests is 600 μs from the channel and 1.2 ms from each control unit. Connection durations and allowable destinations are the same as for the simulations resulting in curves A and B.

Curves A through C represent situations in which the probability of a connect request encountering a busy destination is quite high, as shown by the "percent destination busy" column in Table 1. A relatively short busy response turnaround time of 2 μs was used for these three simulations. This has the effect of increasing the connect request rates because channels and control units frequently retry their connect requests to the busy destinations. The other factor that contributes to the high destination-busy percentage is the duration of the connection (i.e., the time two ports remain connected); it was chosen to be large relative to the busy-response turnaround time. Because the ratio of connect requests to disconnect requests is large in these situations, the effect of the disconnect-request processing on the time required to establish a connection is negligible compared to the effect of contention among the connect requests.

In configurations D and E, each of the 60 switch ports has a mean interarrival time for initial connect requests of 30 μs. Connection duration is 4 μs. Any port can request a connection to any other port. The busy-response turnaround times for configurations D and E are 10 μs and 2 μs, respectively.

Curves D and E represent configurations with much shorter connection durations and shorter time gaps between initial connect requests than curves A through C. As a result, the matrix controller is much busier. The destination is busy a lower percentage of the time, and the ratio of connect to disconnect requests is much closer to unity. Aggregate request rates are higher, however, because of the much shorter mean interarrival time of initial connect requests from each port.

The aggregate request rates for the configurations represented by curves C, D, and E are unrealistically high for today's typical configurations and applications. The parameters were chosen to given an indication of the connect performance of the ESCON Director over a range of request rates approaching its maximum capability. It must be emphasized that the performance represented by these curves is only an approximation, based on the simulation model previously discussed. There are many combinations of variables in real configurations that can affect the connect times actually realized.

6. Design system

Adequate design automation tools are essential elements in the development of a complex electronic product. The design methodology utilized in developing the ESCON Director is hierarchical, requiring the use of design tools with the capability of supporting both a high-level design-description language and a technology-dependent design-description language, for both IBM and vendor technologies. The design tools must also provide general documentation—e.g., charts that can consist of boxes, circles, and straight lines; schematic and wiring diagrams; and timing diagrams and flowcharts.

- *Hierarchical methodology*

We use the term *hierarchical design methodology* to describe treating an agglomeration of units at one level as an entity at the next higher level; e.g., transistors combine to form a gate, a group of gates is a functional island, functional islands form a chip, chips form a module, modules comprise a card, a group of cards is a board. Each successive level of the design process is simply built upon the previous level. To the card-level designer, for example, a chip module is a simple entity, even though it actually comprises thousands of circuits. This is true at all levels of the design. This consistency in treatment allows a single set of tools and a single methodology to be used throughout the entire design, as it evolves from transistor to system.

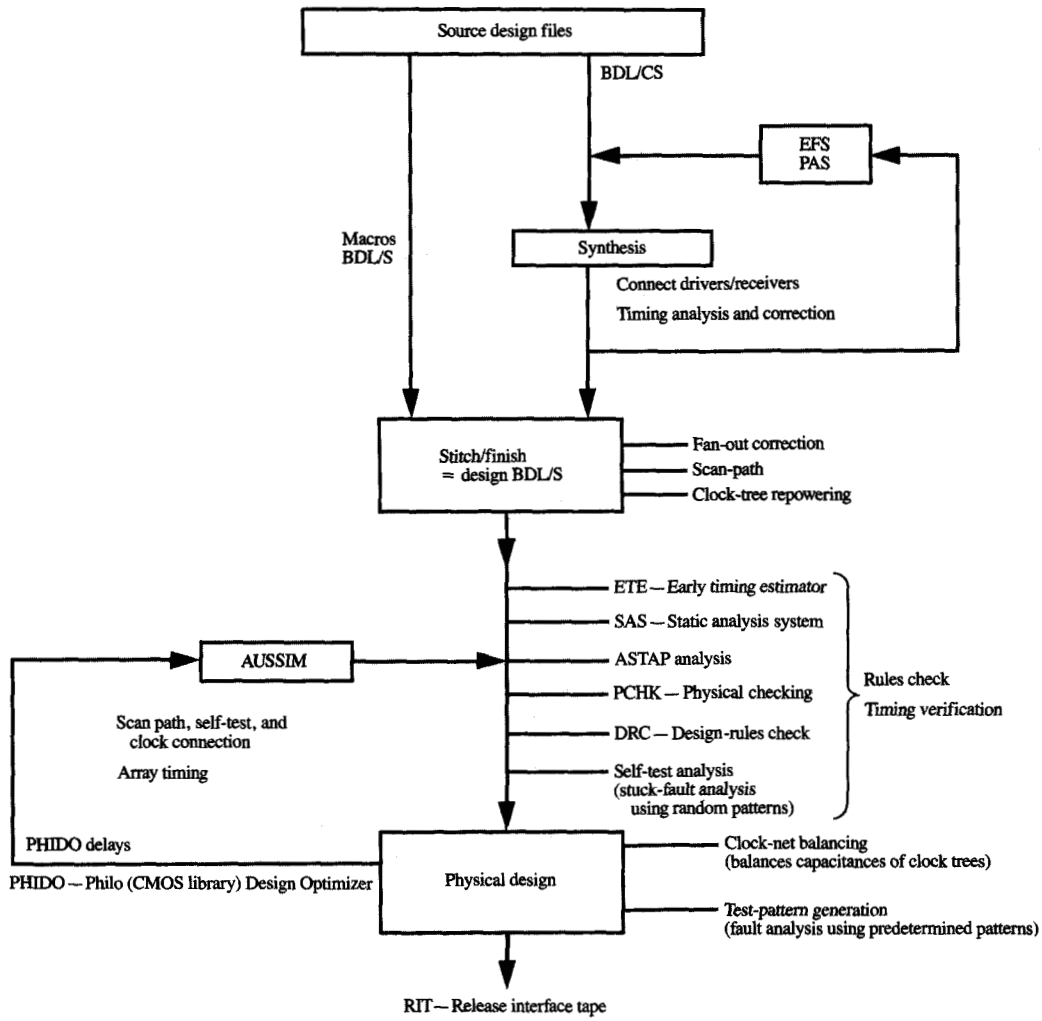

Figure 9

Basic ESCON Director BDL/S methodology.

• *Language representation*

In addition to being hierarchical, our design methodology was also bilingual, in that the same circuit logic was translated into two completely different forms for different uses. Basic Design Language for Cycle Simulation (BDL/CS) [17] was used for model simulation and was the base for synthesizing the physical model. BDL/CS includes no physical information about the design; it is a purely behavioral description. It is usually the first data set generated from the logic description, since design verification precedes physical implementation.

Basic Design Language for Structure (BDL/S) [17] is a physical representation of the logical design. The data that are used to fabricate the hardware are derived from it. For chip-level design, the BDL/S is synthesized from the BDL/CS and is merged with existing circuit macros in the BDL/S format to create a physical model of a chip. **Figure 9** shows the sequence of operations that took place, beginning with BDL/S synthesis from the source design files and ending with the generation of the RIT (release interface tape) for the manufacturing of chips.

All above-chip-level designs, such as cards and boards, were created directly in BDL/S, with the relationship with BDL/CS maintained by the use of common signal names for connected pins.

• *System simulation*

The design of the ESCON Director involved the use of a number of IBM simulation tools, such as the EFS (Enhanced Functional Simulator), PAS (Path Analysis System), and AUSSIM (AUStin SIMulator).

EFS was the centerpiece of functional simulation. Designers provided BDL/CS descriptions and test cases to EFS, which simulated the operation of the logic on a cycle-by-cycle basis, thus providing verification of the logic design. The progress of the EFS simulation was monitored by PAS, which reported the number of logical paths exercised, thus providing feedback about the thoroughness of the test cases. When the logic had passed functional verification, BDL/S generation was performed. Finally, AUSSIM is to BDL/S as EFS is to BDL/CS. AUSSIM was used to verify the functions that were not designed in BDL/CS, such as POR (power-on reset) and scanning logic, clock-generation functions, and the self-test macro. These functions were represented in AUSSIM either by actual block delays or a unit delay-per-block.

• *Timing analysis*

Timing analysis allows the designer to ensure that all of the paths meet best-case and worst-case design constraints. The ESCON Director designers used ETE (the Early Timing Estimator) [18], ASTAP [13], and the delay capability in AUSSIM to verify that the timing requirements were met.

7. Technology considerations

The ESCON Director switch comprises 60 ports, a matrix switch, and a matrix controller, totaling over one million logic circuits (a logic circuit is equivalent to a three-way NOR gate). The Director specifications required the capability to manage up to 30 simultaneous full-duplex 200Mb/s data streams through a switching matrix. This involves reading and interpreting headers and routing the frames, and it imposes severe performance demands on the logic circuits.

To resolve this apparent conflict of density vs. speed, a solution involving parallel processing and routing of data streams was adopted. Parallel processing of data streams allowed the use of low-cost, dense CMOS logic technology. The delay introduced by the serialization and deserialization of the data streams was not a significant design factor, because the signal-propagation delay in the multikilometer fibers dominated the data flow. The introduction of very wide data buses, however, would have put severe demands on the package technology in terms of the required number of module pins and card I/O pins, as well as noise generation, and would have led to an excessive number of modules and cards—hence the need for minimum data bus widths.

• *Data flow*

Figure 4 illustrates the basic functional structure and data flow of the Director ports. ECL circuits are used in the digital sections of the link adapter. (A detailed discussion of the design of the link adapter can be found in [9].) An I/O data width of 10 bits was chosen for the port adapter because it matched the width of one character of data in 8B/10B code. The resultant data rate per bus line (20 Mb/s) was within the capabilities of the CMOS technology with TTL-level off-chip drivers and modest interface wiring lengths (e.g., on the same printed circuit card). The use of a 10-bit-wide bus minimized complexity and power in the link adapter ECL logic; in addition, it optimized the overall number of module I/O pins and related signal noise. The data transfers between the port and link adapters are synchronous.

The data interface between the port and matrix switch was also designed to be synchronous over 5-bit-wide buses, operating at 40 MB/s per bit. Bus-width minimization was especially crucial here, as there are 60 pairs of buses. A bus width of five bits was determined to be optimum within the capabilities of the CMOS technology but demanded very tight timing control. The details of the timing and noise analyses are provided in the section on technology analysis and in the Appendix. Limiting the bus width to five bits was an important consideration, because the use of ten bits would have doubled the interconnect wiring between all the ports and the matrix switch (i.e., $60 \times 20 = 1200$ wires), in addition to doubling the number of matrix chips, thus creating a very difficult wiring problem. This would have resulted in more port cards and a larger and more complex main board, thus significantly increasing the cost of the system. Also, the potential simultaneous switching of 60 matrix switch outputs at 40 Mb/s could result in line-coupled and simultaneous-switching noise conditions that had to be minimized.

• *Logic technology*

The 1-μm CMOS standard cell technology, developed by our fellow laboratory in Burlington, Vermont [10], is an "n-well" process, with one level of polysilicon and two levels of metal, requiring a 5-V power supply. It has a capacity of 17 500 cells on a 7.5-mm chip and 27 700 cells on a 9.4-mm chip. The drivers utilize a series output resistor as the transmission-line termination. The receivers, designed for CMOS voltage levels, appear, effectively, as open circuits and require driver-output series termination to ensure signal quality at the receiving end.

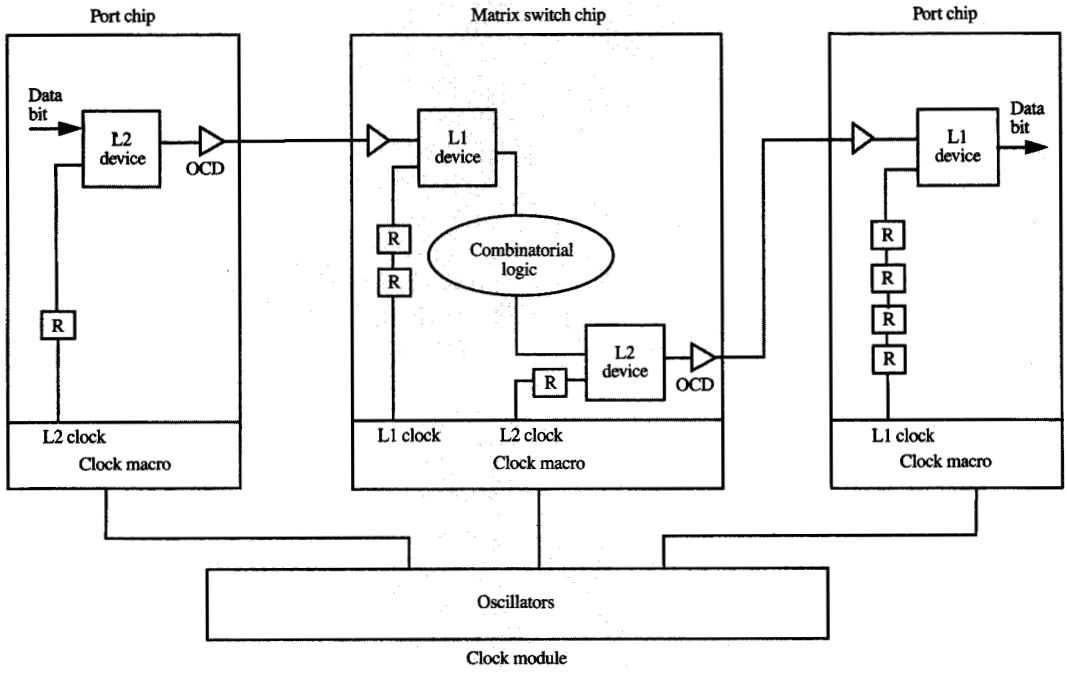

Figure 10

Port-to-matrix data paths (OCD—off-chip driver; R—delay buffer).

The functionally rich CMOS standard-cell library played a critical role in the ability to contain two ports on one 9.4-mm chip and one 64 × 64 matrix switch plane on one 7.5-mm chip. For example, the 128 × 10 dual-port RAM in the port required only 1500 cells, because this function was provided by the macro library. Registers, latches, drivers, and receivers were some of the other custom standard cells used in the ports. Another example is the matrix switch, whose design, illustrated in Figure 6, makes extensive use of multiplexer and decoder macros.

Because the standard-cell CMOS technology was used, the complete 60-port system of approximately one million circuits could be contained on fewer than 45 chips. Because of the power efficiency of the CMOS technology, the power dissipation of the 45 chips was less than 100 W at peak activity. Consequently, the cooling design of the overall system was simplified, with the primary design focus being directed at how to effectively cool the fiber-link electronics, which utilize silicon bipolar technology and contain temperature-sensitive optical components.

To meet the required system performance, the CMOS technology had to satisfy a number of critical performance parameters. Specifically, the buffer arrays in the ports and the table arrays in the matrix controller required a 45-ns cycle time. Standard, three-way-NAND delay had to be under 3 ns, and the maximum delay from port output latch to matrix switch input latch had to be less than 15 ns, since the remainder of the 25-ns cycle had to be allocated to accommodate latch delay and clock skew.

8. Technology analysis

This section presents a detailed analysis of two critical wiring nets in the Director: the data path between the ports and the matrix switch, and the control path between the ports and the matrix controller. The important design issues are the timing required to ensure synchronous data transfers, the noise that may be coupled onto the signal lines from parallel data buses, and noise and reflection affecting signal quality, all of which may cause signal errors.

Figure 10 illustrates the data path and the clock distribution between the ports and the matrix switch. The data transfer time from L2-latch clock rise (data launch) to L1-latch clock fall (data capture) must be accomplished

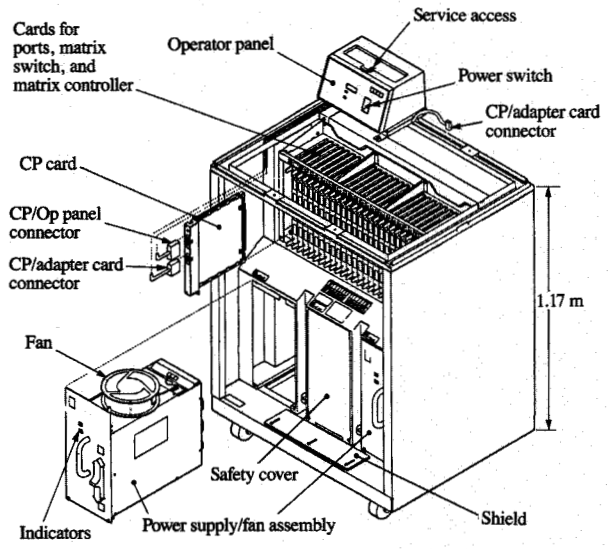

Figure 11

Model 9032 ESCON Director.

Table 2 Noise contributors for TTL-level and CMOS-level drivers and receivers.

Noise contributors	*TTL-level noise* (V)	*CMOS-level noise* (V)
Second-level package	0.60	1.04
Module ΔI	0.50	0.70
Card ΔI	0.20	0.29
Receiving-chip noise	0.20	0.27
Hot pluggability	0.05	0.05
Total noise	1.55	2.35
Total dc noise budget	1.65	2.45
Safety margin	0.10	0.10

within a 25-ns clock cycle. The contributors to the path delay include circuit delays, wiring-net delays, and clock skew (which reduces the effective clock cycle from 25 ns).

The off-chip wiring-net delay (examined later), consisting of delays in the off-chip driver (OCD), net wiring, and receiver, was planned to be under 15 ns. Since the maximum delays for the worst-case path exceeded 15 ns, the concept of cycle skewing (i.e., extending the clock cycle in one segment of the data flow) was employed to solve the problem. This was done by adjusting the relative time of the L2 and L1 clocks with respect to the reference clock by varying the number of delay buffers (R in Figure 10) in the clock-distribution tree. Normally, the rise of L2 and fall of L1 occur simultaneously. As a result of the delays shown in Figure 10, the interval from data launch to data capture in the path from the port chip to the matrix switch chip is increased by one buffer delay. Similarly, the 25-ns clock cycle between the matrix switch chip and the port is extended by three buffer delays, approximately to 30 ns. Note that the subsequent logic path must be less than 20 ns to satisfy the average 25-ns clock cycle.

The other extreme of this path timing was the possibility of very short delay times due to short wiring nets. This situation occurs in the Model 9033 Director, in which the ports are physically much closer to the matrix switch than they are in the Model 9032 Director. In this situation, it was possible that the L1 clock would capture the data intended for the next cycle. To solve this potential problem, the matrix design was modified to introduce delay padding between L2 and its OCD.

The critical nets between the ports and the matrix controller are of two types: nets consisting of one driver with multiple receivers, and nets with multiple drivers and one receiver. Thus, the consideration was how many drivers or how many receivers could share one bus while keeping the wiring-net propagation time within the allowable bound. The details of the timing analysis are given in the Appendix.

The very large and fast voltage transitions (up to 5.0 V/ns) and corresponding large ΔI of the technology, along with the need for closely spaced components and connector wiring, resulted in significant noise generation, which needed careful analysis and design. Since the quiet-line noise (noise on an inactive signal line) consists of coupled noise and resulting reflection noise due to impedance mismatches at both ends, the worst-case situation was not clear until significant circuit analysis (using the ASTAP program) had been undertaken. The noise-analysis results are summarized below.

Two sets of CMOS drivers and receivers were used in the networks analyzed. The first set, referred to as "CMOS-level," operates at a minimum voltage swing of 3.7 V. The receivers can tolerate a maximum signal-line noise of 2.35 V, but the drivers produce more noise than those of the second set. The second set, referred to as "TTL-level," operates at a minimum voltage swing of 2.0 V. The receivers have significantly lower noise tolerance than the CMOS-level ones (1.55 V), but the drivers, with their low output impedance, are particularly well suited to drive heavily loaded nets.

The total noise budget of the driver/receiver pairs (i.e., the total noise allowed in the net before a false logic state would be caused) with the maximum allowable allocation for each contributor is given in **Table 2**. The various types of generated noise are as follows. Second-level-package

noise is due to noise coupled from active to quiet lines on the cards and at the connectors. Module ΔI is the noise generated on the module power pins due to significant current transients from off-chip drivers resulting in Ldi/dt voltage change that may couple onto quiet signal lines. Card ΔI is the noise generated on the power distribution lines of the card due to similar transient currents. The latter is minimized with on-card capacitors that reduce the effective inductance of the power planes. Receiving-chip noise is the power-pin noise on the receiver module, which couples into the receiver circuit. Hot-pluggability noise is introduced on a line because of signal connections made to the line while power is on. Such a case would occur when port cards are added or removed, thereby increasing the loading on the bus between the ports and the matrix switch. The apportionment to each noise source defined in Table 2 was arrived at after significant analysis and design optimization. Verification analysis is illustrated in the Appendix.

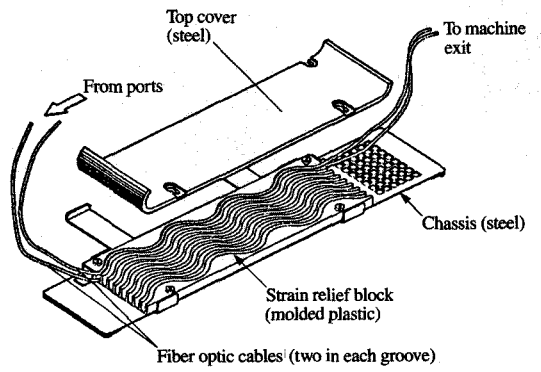

Figure 12

Model 9033 fiber optic cable strain relief (the 9032 version is functionally similar).

9. Physical design

The ESCON Director is offered in two models: Model 1 (9032), which is a floor-standing unit with 28–60 ports, and Model 2 (9033), which is a table-top unit with 8–16 ports. The physical design details of each of these models are discussed in the following sections.

- *Model 9032 Director*

The Model 9032 Director, shown in **Figure 11**, is a floor-standing unit. Power supply/fan assemblies are located at the bottom of the cabinet; the logic assembly is located in the top section. An operator's panel, consisting of the main power switch, a power-on indicator, a power-on-test indicator, and a four-digit hexadecimal display, is located at the top left rear of the machine.

The power supplies convert 220-V, one-phase ac to 5 V dc for logic and 24 V dc for fans. The Model 9032 Director has two supply/fan assemblies, but a redundant third assembly, as well as two spare fiber optic ports, is installed when the Enhanced Availability Feature is ordered. Two such assemblies can power and cool a fully configured Model 9032 Director. With the Enhanced Availability Feature in place, any one of these assemblies can be installed/removed while the machine is running, without causing machine errors.

This logic assembly consists of a sheet-steel cage containing all logic cards, including the control processor (CP) card, an optional spare port card (two ports), from seven to fifteen four-port cards, a clock card, five matrix switch cards, and a matrix controller card. In the event of a port failure, a fiber optic cable can be transferred to a spare port without affecting machine operation. Each card has a fault indicator light (LED) switched on by the CP card when an error condition is detected.

As discussed in the section on packaging, the board has a 6S4P cross section. It was a major achievement to embed the required wiring into the six signal planes of the board for the following reasons: 1) There are 864 wiring nets, making 2064 connections, mostly in a horizontal pattern, in a wiring area already containing many connector pins. 2) Some wirability was lost because most nets are high-speed, and the potential noise exposure limited use to only three of the four lines per channel. 3) Most nets had to be balanced in length, forcing many lines to greater lengths than would otherwise be necessary. 4) Care was taken to prevent certain wire combinations from running next to each other, in order to avoid false switching.

Fiber optic cables connect directly to port cards; there is no "tailgate." Cables enter the bottom of the cabinet through a "radius control boot" and are led up into a strain-relief device between the power supply/fan assemblies (**Figure 12**). The strain-relief device prevents cable pulls external to the machine from being transmitted to, and possibly damaging, the port cards. It consists of a molded plastic block containing serpentine grooves into which the cables are laid. The friction holding the cables into the grooves, combined with the several direction changes provided by the grooves, is effective in resisting cable pulls of at least 267 N (60 lb).

- *Model 9033 Director*

The Model 9033 Director, shown in **Figure 13**, is a table-top unit with steel chassis, a slip-on steel top/side cover, and molded plastic decorative end covers. Its logic assembly consists of a large planar card, the CP card, a

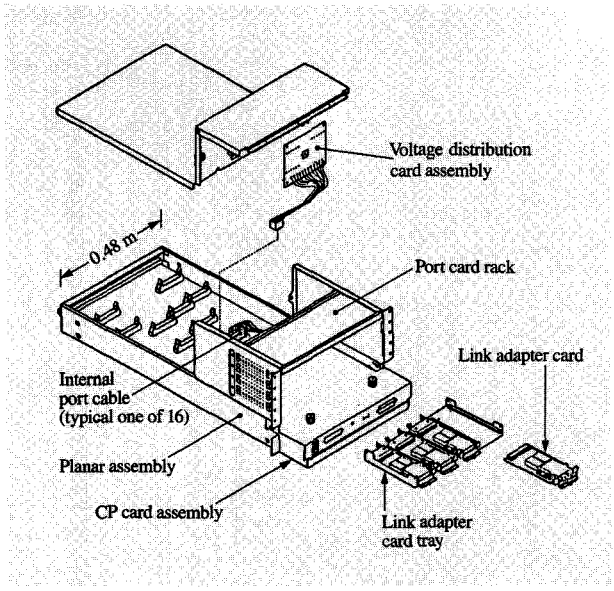

Figure 13

Model 9033 ESCON Director.

rack of eight to sixteen port cards, and eight to sixteen tri-lead cables connecting the port cards to the planar card, all enclosed within a sheet-steel subchassis and EMC (electromagnetic compatibility) cover in front of the port cards. The planar and CP card are sandwiched together by a 268-pin connector carrying low-speed signals only. The fiber optic cables are led through a strain-relief device similar to that described above for the Model 9032 Director.

The use of a large planar card instead of a conventional card-on-board package minimized the number of connectors, thus increasing reliability. It also minimized overall net lengths, thereby reducing noise and improving performance. The CP card is easily replaced without removing the logic assembly from the chassis, by turning two thumbscrews and then sliding the assembly in or out. Each port card is also easily removed/installed individually, with the use of a jack screw. Any port card can be replaced while the Model 9033 Director is operating, without causing errors.

10. Summary

This paper has presented the design philosophy and implementation details of two models of the ESCON Director, the new interconnection element for System/390. The ESCON Director is a dynamic, nonblocking, any-to-any switch for up to either 16 or 60 fiber optic links operating at 200 Mb/s. With 1-μm CMOS technology and innovative packaging techniques, it was possible to implement the Director models in compact, air-cooled boxes. The paper has presented various design challenges and trade-offs in order to give a perspective of the design of these high-performance switching elements.

Appendix: Timing and noise analysis

All paths (critical and noncritical) were categorized and analyzed for best- and worst-case conditions with regard to noise and delay on the basis of their net types (i.e., point-to-point, distributed, etc.). To determine the worst-case delay path, extensive ASTAP circuit modeling was performed on several critical-timing paths with different drivers and receivers. The same paths were also analyzed with regard to noise, as it was often the case that a viable design from the point of view of delay requirements would not satisfy the allowable-noise criteria. Delay of a signal was defined to be from the midpoint of the driver input-signal transition to the midpoint of the receiver output transition. The signal-propagation speed of the printed circuit wiring was 14.5 cm/ns. The work summarized here is based on the CMOS and packaging models used for the product implementation.

For crosstalk analysis of a quiet line, the simulation of the total coupled noise included the contributions of the original signal, as well as its subsequent reflections at both ends of the transmission line. In addition, several wiring patterns in the spaces between vias (also known as the channel) at the card and the board level were examined. In general, one of the potential lines in the channel (located between two utilized lines) was deleted in order to minimize the coupled noise. The inductance and capacitance parameters of the card and board signal lines and of the connector pins were obtained from the manufacturer and were incorporated, without change, in the circuit network models for ASTAP circuit analysis. The module ΔI noise analysis was undertaken by the technology developers using an ASTAP model of the package and considering the specific functional application. After a number of iterations of I/O placements, the resultant module noise was found to be within the limits of Table 2.

A detailed timing and noise analysis was done for the four critical paths in the system: matrix switch to port, port to matrix switch, port to matrix controller, and matrix controller to port. The following sections present a discussion of two of the above paths—matrix switch to port and port to matrix controller.

- *Matrix switch to port*

This is the most critical off-chip delay path in the system (25 ns), exhibiting the largest levels of coupled and ΔI noise. To obtain an adequate noise margin and maintain the quality of the signal, a CMOS-level (tri-state push-pull) driver and a CMOS-level receiver with hysteresis were

Table 3 Best/worst-case noise and timing analysis.

Length			*Far-end noise*		*Near-end noise*		*Delay(R)*		*Delay(F)*		*dv/dt*		*di/dt*	
Matrix card (cm)	*Board* (cm)	*Port* (cm)	*BC* (V)	*WC* (V)	*BC* (V)	*WC* (V)	*BC* (ns)	*WC* (ns)	*BC* (ns)	*WC* (ns)	*BC* (V/ns)	*WC* (V/ns)	*BC* (mA/ns)	*WC* (mA/ns)
4	3	6.62	0.85	0.46	0.56	0.55	4.00	7.61	4.50	8.11	5.35	1.61	148	33
24	36	6.62	0.80	0.75	0.64	0.65	7.60	11.13	8.10	11.63	5.35	1.61	148	33

selected. As previously discussed, CMOS-level receivers have a higher noise margin than TTL-level receivers. The CMOS-level circuits, however, are somewhat slower than the TTL-level circuits, posing a design dilemma. The printed wiring net (card plus board wiring) is point-to-point, as defined above, ranging in length from 13.62 cm to 66.62 cm. The characteristic impedance of the signal lines, as well as the driver output impedance, is 80 Ω. To ensure the viability of the design of this critical net (matrix switch to port), all 60 similar nets in the system had to be investigated, because the noise characteristics do not have a predictable dependence on the interconnection pattern. The worst-case delay, however, would likely be that of the longest net. The following sections summarize the critical sensitivity studies undertaken to find the worst-case noise situation and its corresponding worst-case delay.

Matrix card wire-length sensitivity The coupled noise on a quiet line from the matrix switch to a port was analyzed as a function of the length of common wiring shared with an active line on the matrix switch card. The port cards nearest to and farthest from the matrix switch card were examined. Noise at both ends of the quiet line, i.e., the far-end and near-end noise, was examined. The wire lengths on the board and port card were held constant at 3.0/6.62 cm (nearest port card) and 36/6.62 cm (farthest port card), respectively. The matrix switch card wiring analyzed was from 4 to 24 cm, in 4-cm sections. Although the noise sensitivity as a function of card wire length was influenced by the relative length of board wiring, the general conclusion was that short matrix switch card wiring was most detrimental and required the establishment of a design rule to ensure a minimum of 12 cm for the matrix switch card.

Noise sensitivity to wiring length on the board The far-end coupled noise consistently increased with the length of the board wiring. However, the near-end noise tended to remain constant. This analysis verified that the maximum allowable situation occurred with 12-cm matrix card wiring and 36-cm board wiring, resulting in 1.04 V of noise. Three lines per channel were assumed for this analysis, as previously discussed.

Utilization of all four lines per channel The ability to utilize the fourth wire in the channel on the board was very desirable from the point of view of wirability. However, the introduction of the fourth wire increased the coupled noise on the critical matrix-to-port nets by approximately 20%, resulting in noise levels that exceeded the allowable levels shown in Table 2.

Best-case/worst-case (BC/WC) delay for the shortest/longest nets **Table 3** summarizes the results of the analysis of the most significant paths between the matrix switch and the port. The maximum allowable delay, as indicated in the main body of the text, is 15 ns. Two cases are illustrated with matrix card wiring of 4 and 24 cm, and board wiring of 3 and 36 cm, respectively. Port wiring is fixed at 6.62 cm. Each column has two entries, the first being for fast circuits (i.e., the case in which process variations and operating conditions result in the fastest performance) and the second for slow circuits (i.e., resulting from process variations and operating conditions). The critical net delay is analyzed for input rise (delay R) and fall (delay F) transitions. The coupled noise as well as the parameters affecting the coupled noise [i.e., the rate of change of the driver output voltage (dv/dt) and current (di/dt)] are derived for the two nets.

• *Port to matrix controller*

This net consists of multiple drivers (one per port) and one receiver at the matrix controller. Since each of the 60 ports must communicate with the matrix controller, many ports sharing a common bus would be desirable from the point of view of wiring. But to achieve high performance and low noise, it is most desirable to have few ports sharing a common bus. After analysis, a compromise was reached of a maximum of 17 ports per bus. In general, 16 ports share a bus, except for one bus supporting 17 ports, the extra port being the CUP. It should be pointed out that since there are two ports per chip sharing a common bus driver, the maximum number of common drivers per bus is 9.

Distributed nets are known to cause a considerable degree of reflections because of the stubs on the board and the cards. To reduce signal reflections and resultant ringing due to multiple reflections at the receiving end (i.e., matrix

Table 4 Driver output impedance sensitivity.

Driver output impedance (Ω)	*Far-end noise*		*Near-end noise*		*Delay(R) WC* (ns)	*Delay(F) WC* (ns)	*dv/dt BC* (V/ns)	*di/dt BC* (mA/ns)
	BC (V)	*WC* (V)	*BC* (V)	*WC* (V)				
25	0.79	0.48	0.62	0.48	16.5	18.8	5.99	149.3
35	0.74	0.45	0.42	0.50	17.5	19.7	5.21	127.7
45	0.70	0.44	0.38	0.52	17.6	21.6	4.97	121.5
60	0.60	0.36	0.46	0.50	20.1	24.7	4.21	109.5
80	0.56	0.33	0.46	0.50	25.1	27.7	3.75	104.0

controller card), a discrete series resistor was used in the path before the receiver chip. This resistor significantly reduced the coupled noise by dampening the high-frequency noise at the receiver. Its value was fine-tuned to meet the timing requirements of the path.

As previously discussed, TTL-level drivers were used in the design because they are best suited to drive heavily loaded nets. The output impedance of the bus drivers was adjusted to match that of the wiring nets, because the nets have a considerable amount of capacitance, which decreases the effective characteristic impedance seen by the drivers. These drivers were designed to provide a variety of output impedances in order to match the particular net they were driving.

There are four port-to-matrix-controller nets that are similar; the one that also has a CUP driver is discussed. In addition, there are an equal number of matrix-controller-to-port nets. These nets, however, consist of one driver with multiple receivers and do not have the performance problems associated with the first set of nets.

The most complex wiring net consists of 16 ports with two ports per driver plus the CUP driver (total of nine drivers) sharing the bus. To determine the worst combination of delay and noise, a number of sensitivity studies were performed.

Maximum far-end/near-end noise as a function of driver position This study examined the effect of port card position with respect to noise generated at the near end and far end of the quiet line. The far-end noise was the largest when the switching driver was farthest from the quiet line receiver. This was not really unexpected, since this situation maximized the line coupling. On the other hand, the reflection effects from the driver closest to the matrix controller, which are due to the maximum stub length, resulted in the maximum delay to the matrix controller receiver output.

High-frequency noise-damping resistor (RSER) sensitivity This study examined the sensitivity of far-end noise and worst-case overall net delay with respect to the value of the series resistor used at the far-end input to the matrix controller receiver. This resistor, being part of an *RC* network, dampens the high-frequency noise resulting from multiple reflections on the net, but the signal transitions deteriorate with increasing resistor values. A value of 270 Ω proved adequate to ensure that the noise did not exceed the limit of 0.60 V when the line was quiet, and the worst-case net delay (24.6 ns) was within the design limit.

Utilization of 4 L/C vs. 3 L/C on the GRAD board with R = 270 Ω This study examined the noise impact of utilizing all four lines per channel vs. three lines. With four lines per channel, the coupled noise was increased by 50%, unlike the higher-impedance port-to-matrix-switch wiring nets where the noise increase was considerably lower. This led to the decision to limit the implementation to three lines per channel.

Driver output impedance sensitivity This study examined the appropriate output impedance for the driver. It is more difficult to calculate the effective impedance of a multidrop net than that of a point-to-point net. A low-output-impedance driver was most desirable from the point of view of performance, since it provided more current to charge and discharge the load capacitance. In other words, it provided a better impedance match to the relatively low impedance of the "stubbed" network. However, the faster transitions not only aggravated the line-to-line coupling noise but also aggravated the module noise because of simultaneous switching of the off-chip drivers. **Table 4** summarizes the results of best- and worst-case circuit simulation runs for different driver output impedances.

Acknowledgments

The authors thank John Coons, Carl Corriere, Ray Eustace, John D. Flanagan, Gary Goodstal, Tom Irene, Laura McGoogan, John Nakoski, Leon Skarshinski, Jordan Taylor, and Lowell Thing for their contributions to this paper. In addition, thanks go to Paul Case and John DeVeer for their critical review of the paper.

ESCON, ESCON Director, Enterprise Systems Connection Architecture, Enterprise System/9000, and ESCON Manager

References

1. J. C. Elliott and M. W. Sachs, "The IBM Enterprise Systems Connection (ESCON) Architecture," *IBM J. Res. Develop.* **36**, 577–591 (1992, this issue).
2. *Carrier Sensing Multiple Access with Collision Detection (CSMA/CD), Access Method and Physical Layer Specification,* ANSI/IEEE Standard 802.3, Institute of Electrical and Electronics Engineers, New York, 1985.
3. *Fiber Data Distributed Interface (FDDI) Token Ring Media Access Control*, American National Standard X3.139, American National Standards Institute, New York, 1987.
4. C. J. Georgiou, "Fault-Tolerant Crosspoint Switching Networks," *Proceedings of the 14th International Conference on Fault-Tolerant Computing (FTCS 14)*, June 1984, pp. 240–245.
5. C. J. Georgiou, "Controller for a Cross-Point Switching Matrix," U.S. Patent 4,630,045, December 1986.
6. C. J. Georgiou, "Full Duplex, One-Sided, Cross-Point Switch," U.S. Patent 4,635,250, January 1987.
7. L. R. Goke and G. J. Lipovski, "Banyan Networks for Partitioning Multiprocessor Systems," *Proceedings of the 1st Annual Symposium in Computer Architecture*, 1973, pp. 21–28.
8. A. X. Widmer and P. A. Franaszek, "A DC-Balanced, Partitioned-Block, 8B/10B Transmission Code," *IBM J. Res. Develop.* **27**, 440–451 (1983).
9. N. R. Aulet, D. W. Boerstler, G. DeMario, F. D. Ferraiolo, C. E. Hayward, C. D. Heath, A. L. Huffman, W. R. Kelly, G. W. Peterson, and D. J. Stigliani, Jr., "IBM Enterprise Systems Multimode Fiber Optic Technology," *IBM J. Res. Develop.* **36**, 553–576 (1992, this issue).
10. A. Aldridge, R. Keil, J. Panner, G. Pittman, and D. Thomas, "A 40K Equivalent Gate CMOS Standard Cell Chip," *IEEE Custom Integrated Circuits Conference Digest of Technical Papers*, 1987, pp. 248–252.
11. T. W. Williams and K. P. Parker, "Design for Testability—A Survey," *IEEE Trans. Computers* **C-31**, 2–15 (1982).
12. R. Lasky, C. Li, and D. Seraphim, "Principles of Electronic Packaging," McGraw-Hill Book Co., Inc., New York, 1989.
13. *ASTAP User Guide, Revised Edition*, Order No. 5796-PBH, May 1984; available through IBM branch offices.
14. D. J. Bendz, R. W. Gedney, and J. Rasile, "Cost/Performance Single-Chip Module," *IBM J. Res. Develop.* **26**, 278–285 (1982).
15. E. E. Davidson, "Electrical Design of a High Speed Computer Package," *IBM J. Res. Develop.* **26**, 349–361 (1982).
16. C. H. Sauer, E. A. MacNair, and J. F. Kurose, "The Research Queueing Package Version 2: Introduction and Examples," *IBM Research Report RA-138 (#41126)*, April 12, 1982.
17. L. I. Maissel and H. Ofek, "Hardware Design and Description Languages in IBM," *IBM J. Res. Develop.* **28**, 557–563 (1984).
18. W. H. Elder, P. P. Zenewicz, and R. R. Alvarodiaz, "An Interactive System for VLSI Chip Physical Design," *IBM J. Res. Develop.* **28**, 524–536 (1984).

Received January 22, 1991; accepted for publication July 10, 1991

Christos J. Georgiou *IBM Research Division, Thomas J. Watson Research Center, P.O. Box 218, Yorktown Heights, New York 10598 (GEOR at YKTVMH, geor@watson.ibm.com).* Dr. Georgiou received the B.S. degree in electrical engineering and the Ph.D. degree in electrical and computer engineering from the University of California, Santa Barbara, in 1974 and 1981, respectively, and the M.S. degree in electrical engineering and computer science from the University of California, Berkeley, in 1975. From 1976 to 1978, he worked on the design of floppy-disk-based communications subsystems at Scientific Micro Systems, Mountain View, California. From 1978 to 1981, he was the co-founder and engineering manager at Voicetek, Goleta, California, responsible for the development of a line of speech-recognition and voice-response peripherals for personal computers. In 1982, Dr. Georgiou joined the Thomas J. Watson Research Center as a Research Staff Member; in 1984, he became manager of the System Interconnection Structures group, conducting research on high-performance switching systems, interconnection networks for parallel processing, and multiprocessor system architectures. During the academic year 1988–1989, he was on sabbatical leave from the IBM Research Division, as visiting professor at the National Technical University, Athens, Greece, and as visiting research scientist with the Institute of Microelectronics at the Research Center "Demokritos," also in Athens. Dr. Georgiou has received three IBM Outstanding Innovation Awards for his work on fault-tolerant crosspoint switching networks, on hierarchically controlled interconnection networks, and the ESCON Director architecture. He is also the recipient of seven IBM Invention Achievement Awards, and has 20 U.S. patents issued or pending. Dr. Georgiou is a Senior Member of the IEEE.

Thor A. Larsen *IBM Technology Products, East Fishkill facility, Route 52, Hopewell Junction, New York 12533 (THOR at FSHVMFK1, thor@fshvmfk1.vnet.ibm.com).* Mr. Larsen received a B.S. degree in physics in 1960 from Queens College and an M.S. degree in electrical engineering in 1962 from Columbia University. He spent an additional year of study in the field of electrophysics at Rensselaer Polytechnic Institute in 1972, on an IBM Resident Study Fellowship. During most of his career at IBM, he was engaged in technology development for system applications in Kingston, New York. Early in his career, he worked in circuit development as an engineer and manager. He also managed an FET LSI pilot line which provided prototype modules to product developers as well as development of new FET processes. During a two-year assignment at IBM Hursley, England, Mr. Larsen explored silicon-based liquid crystal displays. As a member of the IBM Kingston System Interconnect group, he participated in the early development of the ESCON Director switch system. Mr. Larsen also spent a year and a half as technical assistant to the Director of the GaAs Technology Laboratory at the Thomas J. Watson Research Center. From 1988 to 1992, he was a Research Staff Member exploring fiber-optic-based system interconnections at the Thomas J. Watson Research Center at Hawthorne, New York. Currently, he is a technical assistant to the Assistant General Manager for I/O Subsystems and Subassemblies of Technology Products, Somers, New York. Mr. Larsen has received several IBM awards, including a Research Division Outstanding Contribution Award for his work on fiber optic technology assessment for ESCON. He is also the recipient of three IBM Invention Achievement Awards and has six U.S. patents issued or pending. Mr. Larsen is a member of Phi Beta Kappa and Eta Kappa Nu and a Senior Member of the IEEE.

Peter W. Oakhill *38 Appletree Drive, Rhinebeck, New York 12572 (retired).* Mr. Oakhill was an Advisory Engineer at IBM Kingston, working in the area of switching process development. He received a B.S. degree in physics from

Riverside College in 1954 and also attended the University of California at Riverside from 1954 to 1956. He joined IBM in 1955, in Large Systems Field Engineering in Southern California, where he held various assignments until 1964. From 1964 to 1983, Mr. Oakhill worked on various large-system development projects, including System/360™ Model 65, the FAA Air Traffic Control System, System/370™ Model 135, and 3081 Channels, with development assignments at IBM sites in Hursley, England, and Montpellier, France. From 1983 to his retirement in 1991, he participated in the design and development of the ESCON Directors and earlier switching systems.

Bijan Salimi *IBM Enterprise Systems, P.O. Box 950, Poughkeepsie, New York 12602 (SALIMI at PK705VMA).* Mr. Salimi is currently a Staff Engineer with IBM Enterprise Systems, Poughkeepsie, New York. He received the B.S. and M.S. degrees in electrical engineering from Rutgers University in 1982 and 1984, respectively. He joined IBM Kingston in July 1984 in the Interconnect Products Department, where he was responsible for system technology analysis and implementation. In 1988 he joined the Advanced Bipolar Circuit Design group in IBM Kingston, where he was involved in the design of bipolar VLSI circuits for future mainframe computers. In August 1990, he joined the Advanced Products Technology Department in Poughkeepsie, New York, where he is currently responsible for coordinating a wide range of system technology issues relative to future mainframe computers. From 1987 through 1989, Mr. Salimi taught in the Department of Electrical Engineering at the State University of New York, New Paltz. He has been granted one patent and has another patent under investigation.