Fiber optic interconnects for the IBM S/390 Parallel Enterprise Server G5

by C. M. DeCusatis
D. J. Stigliani, Jr.
W. L. Mostowy
M. E. Lewis
D. B. Petersen
N. R. Dhondy

Fiber optic interconnections ("interconnects") have become an increasingly important part of System/390® over the years. With the introduction of the Generation 5 Parallel Enterprise Server in May 1998, a number of important new fiber optic features and enhancements to existing features were made available for applications in input/output device attachment, networking, and parallel sysplex coupling. In this paper, we describe the main applications of fiber optic data links on System/390, with emphasis on the new features announced with the G5. In particular, we describe the fiber connection (FICON™) links, the Gigabit Ethernet interface for the open systems adapter (OSA Express), the Fiber Transport Services fiber quick connect system featuring a new type of small-form-factor fiber optic connector, and the Geographically Dispersed Parallel Sysplex using the 9729 Optical Wavelength Division Multiplexer. We also describe the use of optical mode-conditioning patch cables in FICON, Gigabit Ethernet, and Parallel Sysplex® data links as a way to reuse an installed multimode cable infrastructure at higher data rates, and to overcome distance-limiting phenomena such as differential mode delay associated with modal power distribution. Technical data is presented for each of these applications, and we discuss how the new fiber optic interconnects fit into the overall IBM strategy for future large systems.

1. Introduction

Enterprise server computer systems have undergone many significant changes in the past ten years. One important trend has been the increasing use of fiber optic communication links as an integral part of the system architecture. In this paper, we describe the main applications for fiber optic data links on the IBM System/390* (S/390*) platform, with particular emphasis on the new fiber optic features for the Generation 5 (G5) Parallel Enterprise Servers announced in May 1998. We can categorize the applications which call for the increased use of optical data links into three groups: input/output (I/O) devices (such as disk, tape storage, or printers); networking on the local-area network (LAN) and wide-area network (WAN); and Parallel Sysplex* links (including both sysplex timer links and coupling links).

0018-8646/99/$5.00 © 1999 IBM

Table 1 Fiber optic channel attachment options.

Channel	*Fiber*	*Connector*	*Bit rate*	*Distance/bandwidth*	*Channel loss*
ESCON (SBCON)	SM	SC duplex	200 Mb/s	20 km	14 dB
	MM 62.5	ESCON duplex	200 Mb/s	2 km/500 MHz-km 3 km/800 MHz-km	8 dB
	MM 50.0	ESCON duplex	200 Mb/s	2 km/800 MHz-km	8 dB
Sysplex timer (ETR/CLO)	MM 62.5	ESCON duplex	8 Mb/s	3 km	8 dB
	MM 50.0	ESCON duplex	16 Mb/s	2 km	8 dB
FICON	SM	SC duplex	1.06 Gb/s	10 km	7 dB
	MM 62.5 (with MCP)	SC duplex	1.06 Gb/s	550 m/500 MHz-km	5 dB
	MM 50.0 (with MCP)	SC duplex	1.06 Gb/s	550 m/400 MHz-km	5 dB
HiPerLinks	SM	SC duplex	1.06 Gb/s	10 km (increased from 3 km May 98) RPQ to 20 km	7 dB
	MM 50.0 (with MCP)	SC duplex	1.06 Gb/s	550 m/500 MHz-km	5 dB
	MM 50.0* withdrawn May 98; RPQ only	SC duplex	531 Mb/s	1 km/500 MHz-km*	8 dB*
ATM 155	SM	SC duplex	155 Mb/s	20 km	15 dB
	MM 62.5	SC duplex	155 Mb/s	2 km/500 MHz-km	11 dB
FDDI	MM 62.5	SC duplex†	100 Mb/s†	2 km/500 MHz-km	9 dB
Gigabit Ethernet	SM 1000BaseLX	SC duplex	1.25 Gb/s	5 km	4.6 dB
	MM 62.5† 1000BaseSX	SC duplex	1.25 Gb/s	220 m/160 MHz-km† 275 m/200 MHz-km†	2.6 dB
	MM 62.5 1000BaseLX (with MCP)	SC duplex	1.25 Gb/s	550 m/500 MHz-km	2.4 dB
	MM 50.0† 1000BaseSX	SC duplex	1.25 Gb/s	550 m/500 MHz-km†	3.6 dB†
	MM 50.0 1000BaseLX (with MCP)	SC duplex	1.25 Gb/s	550 m/500 MHz-km	2.4 dB
9729 WDM‡	SM	FC between 9729 pair; adapters as noted above	≦200 Mb/s	50 km‡ (except sysplex timer, 40 km)	15 dB‡
			≦1.06 Gb/s	20 km‡ 40 km RPQ only	12 dB‡

Note: The following RPQs are available for extending the distance, link budget, or otherwise exceeding the announced specifications given above for S/390 fiber optic links:

8P1785 – provides for ESCON channel extension (max. 4.5 km multimode, 24 km single mode).
8P1786 – provides for HiPerLinks channel extension (max. 20 km single mode).
8P1955 – provides for ETR and HiPerLinks channel extension via the 9729 (max. 40 km total distance).
8K1903 – provides for ETR channel extension via a modified pair of 9036 Model 3 repeaters (max. 26 km total distance).
8P1967 – limited-availability RPQ which provides multimode (50 MB/s) HiPerLinks, which were discontinued in May 1998.
8P1984 – provides for FICON link extension (max. 20 km single mode).

Note: This table lists maximum unrepeated distance and link budget for each type of channel; longer distances are possible using repeaters, switches, or channel extenders. SBCON is the non-IBM trademarked name of the ANSI industry standard. All industry-standard links (ESCON/SBCON, ATM, FDDI, Gigabit Ethernet) follow published industry standards. IBM has not yet announced when Gigabit Ethernet will be available on the S/390 platform, or whether it will support the full industry standard as listed below. Minimum bandwidth requirement to achieve these distances is listed for multimode (MM) fiber only; this specification does not apply to single-mode (SM) fiber. MCP denotes mode-conditioning patch cable, which is required to operate some links over MM fiber. Bit rates given below may not correspond to effective channel data rates in a given application due to protocol overheads and other factors. SC duplex connectors are keyed per the ANSI Fibre Channel Standard specifications. Deviations from these specifications, including longer distances, may be possible and are evaluated on an individual basis by submitting a request for price quotation (RPQ) to IBM. Various types of non-fiber optic interconnects are also supported on System/390, including token ring, Ethernet, Fast Ethernet, and parallel channel (bus-and-tag).

Each of these three applications has different requirements for data rates, distances, and fiber types. The I/O applications have relied exclusively on IBM Enterprise Systems Connection (ESCON*) data links for many years [1–7]. Although ESCON continues to be an important part of the S/390 architecture, the G5 systems have introduced a new type of I/O attachment known as fiber connection (FICON*). We describe the technology behind FICON links, including the strategy to migrate from predominantly multimode cables in the current ESCON environment to single-mode cables for future I/O requirements. Future applications for ESCON and FICON include data warehousing (many large companies already have terabyte and petabyte databases), audio/video and digital image storage (such as the IBM digital library project), and high-performance communications for workload balancing.

Networking applications on S/390 continue to support the open systems adapter (OSA) interface, which provides direct attachment to industry-standard LAN protocols. These include various types of copper-wire interfaces such as token ring, Ethernet, and Fast Ethernet, as well as optical fiber interfaces for fiber-distributed data interface (FDDI) and asynchronous transfer mode (ATM). These network interfaces are used to implement large Internet and intranet servers, and provide access to the server for groupware applications such as Lotus Domino** and Web browser access to legacy data. Using secure-transaction software from S/390, servers attached to the Internet also enable electronic commerce applications. An S/390 server can support thousands of Lotus Notes clients, making it among the largest servers available. Additionally, the S/390 servers can be logically partitioned into a maximum of 15 independent external zones, each of which is capable of running its own operating system environment. Thus, a G5 server might have one logical partition running OS/390* applications, another running Java**-based applications, and still another running UNIX** applications (S/390 is officially specified to be UNIX-compatible). The new Gigabit Ethernet (IEEE 802.3z) interface for G5 is an important addition to the networking protocols supported by S/390, as part of the OSA-Express offering. Gigabit Ethernet [8] is one of the few new industry standards to support short-wavelength laser links (850 nm) over multimode fiber, in addition to long-wavelength laser links (1300 nm) over both single-mode and multimode fiber. The standardized link configurations are given in **Table 1**; note that the bandwidth of multimode fiber is a key specification, in addition to the distance and link loss budget.

The G5 ten-way turbo model has become the first enterprise server whose performance exceeds 1000 MIPS (million instructions per second). Using fiber optic coupling links and the Parallel Sysplex architecture [9, 10], up to 32 of these processors can function as a single entity with a combined processing power of over 25 000 MIPS. This approach relies on two types of fiber optic links: the sysplex timer, which allows synchronous operation of multiple processors; and intersystem channel (ISC) links, which interconnect many processors in parallel through a coupling facility (these are also known as HiPerLinks or coupling links). When this architecture was first introduced, coupling links were originally offered in two versions: 50-MB/s data rates over multimode fiber, or 100-MB/s data rates over single-mode fiber. With the announcement of G5, support for multimode fiber has been withdrawn (except as a limited-availability RPQ). IBM has introduced a new feature which allows the installed multimode fiber coupling links to be reused with 100-MB/s adapter cards; we describe the technical details of this approach, known as an optical mode conditioner or mode-conditioning patch cable (MCP). The MCP also has applications for FICON and Gigabit Ethernet data links on multimode fiber.

We also describe several major fiber optic announcements made concurrently with the G5 which involve the support of other IBM divisions. Within large data centers, it is becoming increasingly difficult to manage large multipurpose fiber infrastructures; S/390 has worked closely with IBM Global Services to address this problem. We describe some of the innovations in IBM Fiber Transport Services (FTS), including the use of multifiber optical connectors and small-form-factor optical connectors in the new fiber quick connect structured cable solution. For applications such as disaster recovery which span multiple locations, it is now possible to extend a Parallel Sysplex over distances up to 40 km and still manage the complex from a single location. Known as a

*indicates channels which use short-wavelength (850-nm) optics; all link budgets and fiber bandwidths are measured at this wavelength. Unless noted, all other links are long-wavelength (1300 nm).

†indicates FDDI channels; the media access connector (MAC) was only supported on OSA I; actual line rate is 125 Mb/s, reduced to an effective data rate of 100 Mb/s by protocol overhead.

‡indicates 9729 Optical Wavelength Division Multiplexer (WDM), supported by the IBM Network Hardware Division (NHD), Raleigh, North Carolina. Supports adapter cards for MM ESCON, sysplex timer, SM coupling link, MM and SM ATM 155 and FDDI; also non-fiber adapters including token ring, Ethernet, Fast Ethernet. Link budget is measured at a wavelength of 1550 nm; the 9729 operates over 20 wavelengths spaced 1 nm apart near 1550 nm. Subtract 1 dB from link budget if dual-fiber switch option is installed. Total link distance, including adapter card attachments on either end of the 9729 pair, must be less than the maximum given. Maximum distance for sysplex timer is limited by timing considerations rather than link budget to shorter operating distances, and must NOT exceed 40 km (by RPQ approval only) due to potential data integrity exposure.

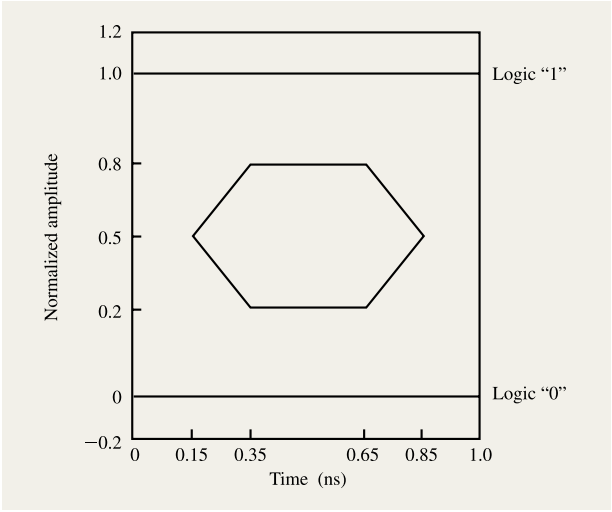


Figure 1

Gigabit Ethernet eye mask, also used for gigabit FICON links [8, 12].

Table 2 FICON output optical interface.

Parameter	*Minimum*	*Nominal*	*Maximum*	*Units*
Average power*	−8.5		−4	dBm
Center wavelength*	1280	1310	1355	nm
Spectral width (RMS)†			1.7	nm
20–80 rise time*‡			0.26	ns
20–80 fall time*‡			0.26	ns
Extinction ratio*§	11			dB
Relative intensity noise (RIN)			−120	dB/Hz

*Launched into single-mode fiber, based on any valid 8B/10B code pattern, measured using a four-meter single-mode duplex jumper cable, includes only power in the fundamental mode of the single-mode optical fiber.
†Spectral width may be increased on the basis of center wavelength and distance tradeoffs; link budget analysis is required for such a change.
‡Minimum-frequency-response bandwidth range of the optical waveform detector is 800 kHz to 3 GHz.
§Measurement can be made with a dc-coupled optical waveform detector of 800 MHz minimum bandwidth and whose gain flatness and linearity over the range of measured optical power provide an accurate measurement of the high and low optical power levels.

Geographically Dispersed Parallel Sysplex (GDPS), this continuous-availability solution relies on the fiber optic technology of the IBM 9729 Optical Wavelength Division Multiplexer, which provides the longest distance supported for a Parallel Sysplex architecture in the industry.

Table 3 FICON input optical interface.

Parameter	*Minimum*	*Maximum*	*Units*
Saturation level	−3	−22	dBm
Sensitivity	−22		dBm
Return loss	12		dB

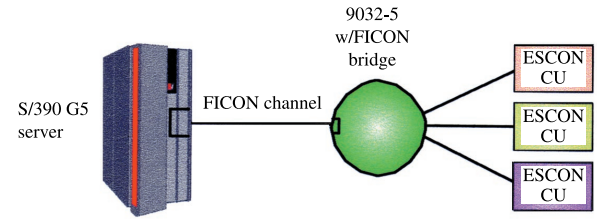
*Based on any valid 8B/10B code pattern measured at or extrapolated to 10^{-12}-bit error rate; must meet specification under worst-case conditions.
†Measured using a four-meter single-mode duplex jumper cable, includes only power in the fundamental mode of the single-mode optical fiber.

2. I/O applications: ESCON and FICON

In May 1998, IBM announced the availability of a new interface, FIber CONnection (FICON) for G5 machines and their successors. The FICON channels (Feature Code 2314 on the G5 Enterprise Servers) are based on the Fibre Channel Standard approved by the American National Standards Institute [11], with a new FC-4 (mapping layer) protocol optimized for enterprise applications. FICON was introduced to provide higher-speed I/O links and increase the overall I/O capacity of the G5 processors, to keep pace with improvements in processor speed and capacity. Since the higher-speed channels provide more efficient use of the channel capacity, S/390 is able to offer increased capability without exceeding the current 256-channel architecture. Initially, twelve FICON channels will be supported on a G5 server; in the future, IBM will increase both the number of FICON channels supported and the capacity of each FICON channel.

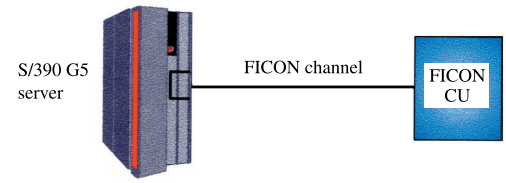
The physical layer specifications of the input and output interface of the FICON channel on single-mode fiber are given in **Tables 2** and **3**, while the input and output interfaces of the FICON channel on multimode fiber are given in **Tables 4**, **5**, and **6** [12]. All FICON channels comply with international laser eye-safety standards for inherently safe operation, without the need for special eye protection or user training. FICON links must also comply with the Gigabit Ethernet "transmit-data" eye mask, shown in **Figure 1**. The FICON specification has already been endorsed by other major datacom providers as an *ad hoc* industry standard, and IBM intends to pursue formal standardization of the interface in the near future. The link data rate is 1.0625 Gb/s (100 MB/s), consistent with the ANSI Fibre Channel Standard; future FICON channels will offer increased data rates. Although the ANSI Fibre Channel Standard does not support the use of long-wavelength lasers on multimode fiber, FICON allows the use of existing multimode cable infrastructures at reduced distances. Following the IEEE Gigabit Ethernet standard conventions, the use of multimode fiber will require an optical mode-conditioning jumper cable (see Section 3). Both single-mode and multimode FICON links use the same type of optical fiber as that specified today

FICON channel with FICON bridge:
- Largest channel capacity within a single image: 360 ESCON equivalent channels.
- Can reduce infrastructure costs for channels, cages/frames, cables, trunks, patch panels, power requirements.
- Can reduce connectivity complexity.
- Solves addressing limit constraints.
- Negligible data rate droop at long distances: vaulting, Geographically Dispersed Parallel Sysplex.



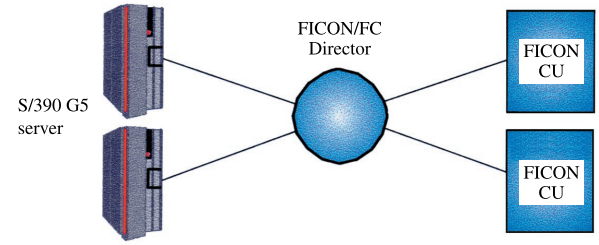
(a)

Native FICON direct attachment:
- High bandwidth to support larger-capacity control units and data-intensive applications.
- Fewer control unit interfaces, to reduce management complexity.
- Supports large and small data transfers concurrently: decision support and OLTP against same database.
- Negligible data rate droop: 20 km without repeaters; 100 km with repeaters.



(b)

Native FICON with switched connectivity:
- Full FICON dynamic switching.
- Fewer directors and director ports.
- Negligible data rate droop: up to 40 km with FICON/FC Director; up to 200 km with repeaters.



(c)

Figure 2

Examples of FICON network configurations: (a) FICON bridge; (b) direct attachment; (c) native attachment with switched connectivity.

Table 4 Comparison of ESCON and FICON data links

FICON	*ESCON*
New frame protocols	None
Protocol interlock reduction	None
Instantaneous data rate 100 MB/s bidirectional	Instantaneous data rate 17 MB/s unidirectional
Nonsynchronized command execution	Synchronous command execution—handshakes between channel and control unit
256 control unit images per link	16 control unit images per link
256 physical control units per link	120 physical control units per link
4K device addresses per control unit (bridge) 16K device addresses per control unit (native)	1K device addresses per control unit (total)
16 control unit images (bridge) 256 control unit images (native)	16 control unit images (total)
128K frame transfer buffer (data rate droop at 100 km)	1K frame transfer buffer (data rate droop at 10 km)
4K–5K I/Os per second (depending on block size)	400–500 I/Os per second (depending on block size)

Table 5 Test results for mode-conditioning patch (MCP) cables: Loss-dependent data.

Assembly loss 0.5 dB max. complying with EIA-455-171, method B	Pass mean = 0.236 sigma = 0.086 max = 0.406
Ship shock −40°C to +60°C, 10 cycles complying with EIA-455-71, 0.6 dB max.	Pass mean = 0.303 sigma = 0.100 max = 0.54
Random connection loss 0.7 dB max.	Pass mean = 0.228 sigma = 0.105 max = 0.468
Heat/humidity age 60°C, 95% RH, 336 hours complying with EIA-455-4B and 5B, 0.6 dB max. loss	Pass mean = 0.247 sigma = 0.061 max = 0.32
Off-axis pull 20 N at 1 meter, 45-degree angle 1 dB max. loss complying with EIA-455-6	Pass mean = 0.276 sigma = 0.055 max = 0.35
Mated axial pull 90 N at 1 meter, 0.6 dB max. loss complying with EIA-455-6	Pass mean = 0.09 sigma = 0.05 max = 0.20
Connector impact EIA-455-2 light duty, method B 0.7 dB max. loss	Pass mean = 0.31 sigma = 0.16 max = 0.67
Thermal cycle 500 cycles, 10–60°C, 0.6 dB max. loss	Pass mean = 0.249 sigma = 0.056 max = 0.361

for other S/390 applications. Either 50-μm- or 62.5-μm-core-diameter fiber may be used to construct a link, but cable infrastructures which mix 50-μm and 62.5-μm multimode fiber are not supported. FICON links use the same SC duplex optical connector as that currently used by other datacom links on the S/390 platform, including ATM, Parallel Sysplex coupling links, and single-mode ESCON.

IBM intends to provide a FICON Director and native FICON device attachments in addition to a FICON bridge feature which will allow encapsulation of ESCON data on a FICON link; these network configurations are shown schematically in **Figure 2**. A FICON bridge card (Feature No. 5260) installed in an ESCON Director 9032 Model 005 allows the attachment of a FICON channel to the Director. This card acts as a FICON-to-ESCON converter; up to eight ESCON channels at 50% utilization can be time-division multiplexed onto a single FICON link. The bridge card is used instead of one of the standard ESCON port cards. By replacing ESCON channels with the higher-performance FICON channels in this manner, the effective number of channels available on a G5 is increased from 256 to 340 (for heavy sequential workloads, the gain may be somewhat less). The FICON channels are also compatible with the ESCON Multiple Image Facility (EMIF), which allows multiple logical channels to be combined on a single physical channel. A comparison of the ESCON and FICON architectures and the enhancements introduced with the FICON channel is presented in Table 4.

While FICON supports 100 MB/s full duplex, for a total aggregate of 200 MB/s, the initial aggregate channel data rate will be a maximum of about 70 MB/s depending on data block transfer size; applications using large block sizes or transferring large numbers of sequential records will experience the most improvement from FICON. Because of protocol overhead, all types of I/O channels experience some performance droop, or degradation in the effective data rate per channel, at longer distances. This is partly due to the additional latency required for round-trip communications (about 10 microseconds per km), the number of acknowledgments required by the data transfer protocol, and the cache size at either end of the link. While there is no way to totally eliminate overhead effects, FICON protocols are designed to permit use of the channel by other operations during overhead periods. For small data transfer sizes, overhead will remain the dominant portion of service time (data transfer time is a small percentage of service time on ESCON links, and their bandwidth is not fully realized in many cases because of the data transfer sizes and protocol overheads). For larger data transfer sizes, FICON links are designed for high effective data rates. Using a combination of improved protocols which minimize overhead and larger cache size, the FICON links do not experience significant performance droop at distances less than 100 km; by contrast, ESCON channels begin to show measurable droop at distances of about 9 km. Configuration support for migration to FICON will be available; however, just as parallel copper-wire interfaces continue to coexist with the ESCON environment nearly ten years after its introduction, it is expected that ESCON and FICON will continue to coexist and complement each other for many years to come.

Table 6 Test results for mode-conditioning patch (MCP) cables: Coupled power data.

Test procedure	*50.0-μm fiber*	*62.5-μm fiber*
Ship shock −40°C to +60°C, 10 cycles EIA-455-71	Pass mean = 14.42 sigma = 0.108 max/min = 14.31/14.53	Pass mean = 32.98 sigma = 1.69 max/min = 32.62/37.21
Random connection	Pass mean = 14.28 sigma = 0.265 max/min = 14.02/14.55	Pass mean = 33.95 sigma = 1.49 max/min = 32.53/36.89
Heat age EIA-455-4B	Pass mean = 14.33 sigma = 0.003 max/min = 14.33/14.34	Pass mean = 33.76 sigma = 1.73 max/min = 32.26/37.14
Off-axis pull	Pass mean = 15.57 sigma = 0.115 max/min = 14.49/16.54	Pass mean = 34.15 sigma = 0.418 max/min = 31.06/35.64
Mated axial pull EIA-455-6	Pass mean = 16.21 sigma = 0.59 max/min = 14.92/17.31	Pass mean = 32.75 sigma = 0.423 max/min = 28.57/36.04
Thermal cycle 10–60°C, 500 cycles, 5 min. hold time	Pass mean = 14.21 sigma = 0.07 max/min = 14.11/14.31	Pass mean = 33.69 sigma = 1.82 max/min = 32.21/37.27

CPR measured according to EIA-455-14A, all values in dB:
Specified CPR for 62.5-μm fiber = 28 to 40 dB;
Specified CPR for 50.0-μm fiber = 12 to 20 dB.

3. Optical mode conditioners

Because of the bandwidth limitations of multimode optical fiber, future multigigabit fiber optic interconnects will be based on single-mode fiber cables. For this reason, most new fiber installations include at least some single-mode fiber in the cable infrastructure. However, many applications continue to use multimode fiber extensively; a recent survey of building-premise cable installers reported that most LAN infrastructures currently installed are composed of about 90% multimode fiber. As the fiber cable plant is upgraded to support higher data rates on single-mode fiber, we must also provide a migration path which continues to reuse the installed multimode cable plant for as long as possible. The need to migrate from multimode to single-mode fiber affects all of the major S/390 datacom applications:

- I/O applications currently using multimode fiber for ESCON will have to migrate the cable plant to single-mode fiber in order to take full advantage of the higher bandwidth of FICON links. Future FICON enhancements which extend this protocol to multigigabit data rates will also require single-mode fiber.
- Networking applications such as ATM have traditionally used different adapter cards to support multimode and single-mode fiber. The emerging Gigabit Ethernet standard (IEEE 802.3z) is the first industry standard to propose the use of both fiber types with the same adapter card.
- Parallel Sysplex links were originally offered as either 50-MB/s data rates over multimode fiber or 100-MB/s data rates over single-mode fiber. With the announcement of G5, support for multimode fiber has been withdrawn (except as a limited-availability RPQ). There is a need to support 100-MB/s adapter cards over installed multimode fiber to facilitate migration for customers who have been using the 50-MB/s option.

To address these concerns, IBM has worked with key fiber optic cable suppliers to develop a special fiber optic

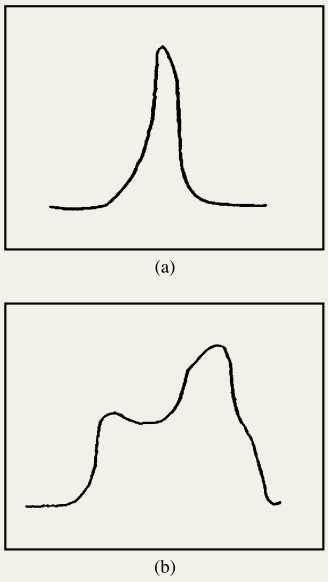


Figure 3

Differential mode delay example: (a) Optical impulse applied to multimode fiber cable; (b) impulse response of multimode fiber cable.

adapter cable known as a mode-conditioning patch cable (MCP). This cable contains both single-mode and multimode fibers, and it should be inserted on both ends of a link to interface between a single-mode adapter card and a multimode cable plant. This allows the maximum achievable distance for multimode fiber (550 meters, as shown in Table 1) and enables some applications to continue using the installed multimode cable plant. The MCP for Parallel Sysplex links is available as Feature Code 0107; additional feature codes for the other applications will be announced shortly. The remainder of this section describes the technical issues associated with this approach, and the results of MCP testing.

The bandwidth of an optical fiber is typically measured using an over-filled launch condition, which results in equal optical power being launched into all fiber modes [1]. This is also known as a mode-scrambled launch, and is approximately equivalent to the conditions achieved when using a Lambertian source such as an LED. By contrast, laser sources, which are more highly collimated, tend to produce an under-filled launch condition; this can result in either larger or smaller effective bandwidth relative to that of an over-filled launch, and is sensitive to small changes in the fiber's refractive index profile. As discovered in recent gigabit link tests [13], bandwidth measured using over-filled launch conditions is not always a good indication of link performance for laser applications over multimode fiber. As illustrated in **Figure 3**, when a fast-rise-time laser pulse is applied to multimode fiber, significant pulse broadening occurs because of the difference in propagation times of different modes within the fiber. This pulse broadening is known as differential mode delay (DMD); it is observed as an additional contribution to timing jitter (measured in ps/m) and can be large enough to render a gigabit link inoperable. DMD values are unique to the modal weighting of a source, the modal delay and mode group separation properties of the fiber, mode-specific attenuation in the fiber, and the launch conditions of the test. DMD is made worse by the excitation of relatively few modes with similar power levels in widely spaced mode groups and a high percentage of modal power concentrated in lower-order modes. The impact of DMD increases with link length. There is, unfortunately, not a simple relationship between the industry-specified over-fill-launch-measured bandwidths of the fiber and the effective bandwidth due to DMD.

The radial over-fill launch method was developed as a way to establish consistent and repeatable modal bandwidth measurement of a given fiber coupled with a given source [13]. A radial over-fill launch is obtained when a laser spot is projected onto the core of the multimode fiber, symmetric about the core center with the optic axes of the source and fiber-aligned; the laser spot must be larger than the fiber core, and the laser divergence angle must be less than the numerical aperture of the fiber. When these conditions are satisfied, the worst-case modal bandwidth of the link is taken to be the worse of the over-fill and radial over-fill launch-condition measurements (although for most applications, the radial over-fill launch will be the worst case). There is a good correlation between the radial over-fill launch bandwidth and the DMD-limited bandwidth of a fiber; thus, high-speed laser links implemented over multimode fiber will likely experience bandwidth values closer to those found with the radial over-fill launch method rather than the more commonly specified over-fill launch method.

To allow for laser transmitters to operate at gigabit rates over multimode fiber without being unduly limited by DMD, a special type of fiber optic jumper cable was developed to "condition" the laser launch and obtain an effective bandwidth closer to that measured by the over-fill launch method. The intent is to excite a large number of modes in the fiber, weighted in the mode groups that are highly excited by over-fill launch conditions, and to avoid exciting widely separated mode groups with similar power levels. This is accomplished by launching the laser light into a conventional single-mode fiber, then coupling into a multimode fiber which is off center relative to the single-mode core, as shown in **Figure 4**. There are two ways in which the offset launch can be introduced. One version requires manufacturing a splice between the

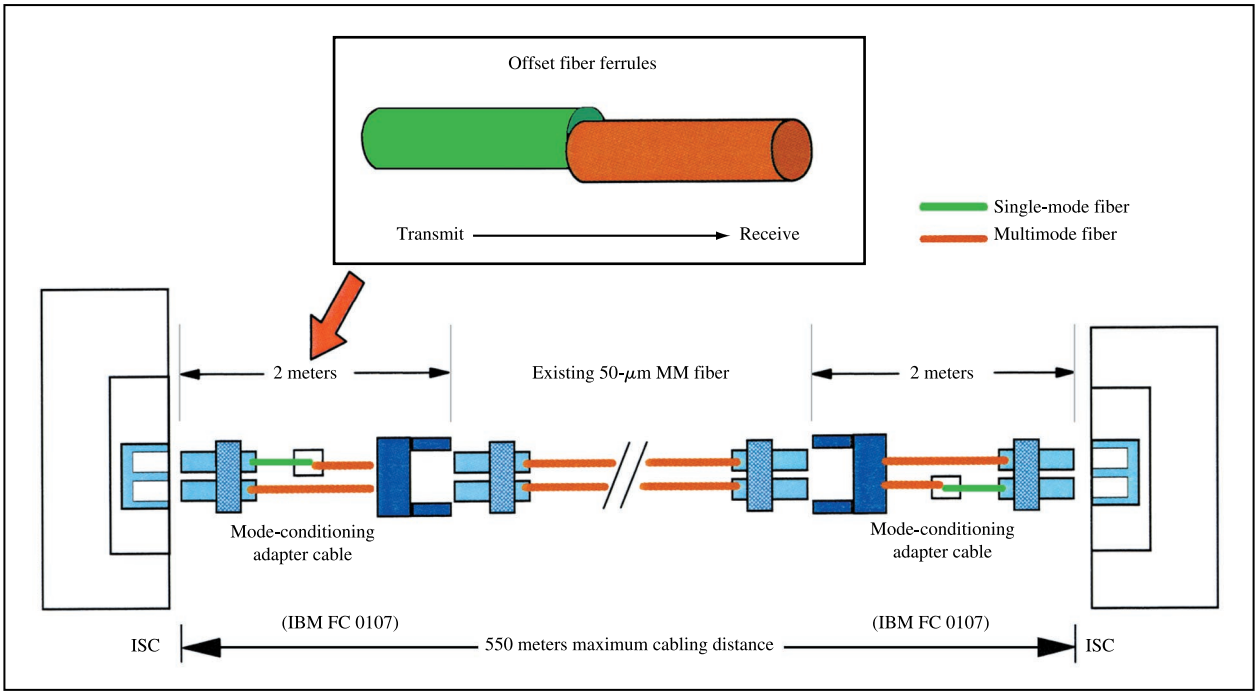


Figure 4

Off-center ferrule design for mode-conditioning patch cables and their application to gigabit HiPerLinks or Gigabit Ethernet.

single-mode and multimode fiber with a controlled amount of lateral offset between the fiber cores. A tolerance analysis of this approach revealed that some installations could experience unacceptable variability in the splice elements, resulting in poor alignment and ineffective mode conditioning. For this and other reasons, the preferred embodiment uses standard ceramic ferrule technology with an offset in the ferrule alignment. Different offsets are required for 50.0-μm and 62.5-μm multimode fiber cores. Evaluations conducted by the Gigabit Ethernet Task Force, Modal Bandwidth Investigation Group, have verified that single-mode-to-62.5-μm multimode MCPs with lateral offsets in the 17–23-μm range can achieve an effective modal bandwidth equivalent to that of the over-fill launch method across 99% of the installed multimode fiber infrastructure. Similar work has shown that single-mode-to-50-μm multimode offset launch cables with lateral offsets in the 10–16-μm range will achieve similar results.

The MCP is illustrated in **Figure 5**; its form factor is similar to that of a standard two-meter jumper cable, except that it contains both single-mode and multimode fibers and includes a small package for the offset ferrules near one end. During the manufacturing process, the offset ferrules are actively aligned and then sealed with a potting compound to provide thermal and mechanical

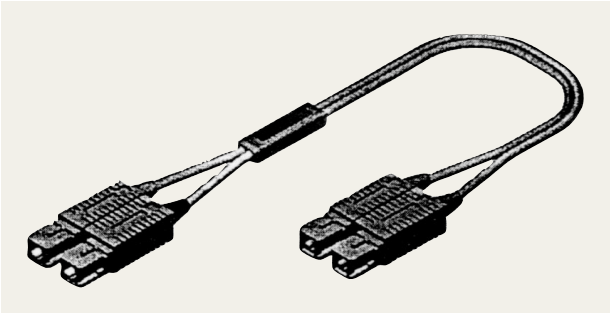


Figure 5

Mode-conditioning patch cable (MCP). The small box just behind one of the SC duplex connectors contains the offset ferrules; the transmit fiber from the gigabit laser is single-mode; all other fibers in the cable are multimode.

stability. A schematic of the active alignment apparatus is shown in **Figure 6**; a wide-field charge-coupled-device (CCD) camera is used to measure the two-dimensional spatial distribution of optical power at the output of the MCP. Typical results of this measurement are shown in **Figure 7**; plot (a) illustrates the optical power distribution

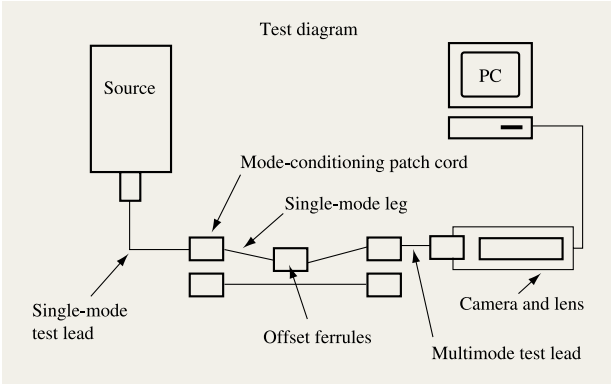


Figure 6

Schematic of alignment apparatus for MCP cable manufacturing.

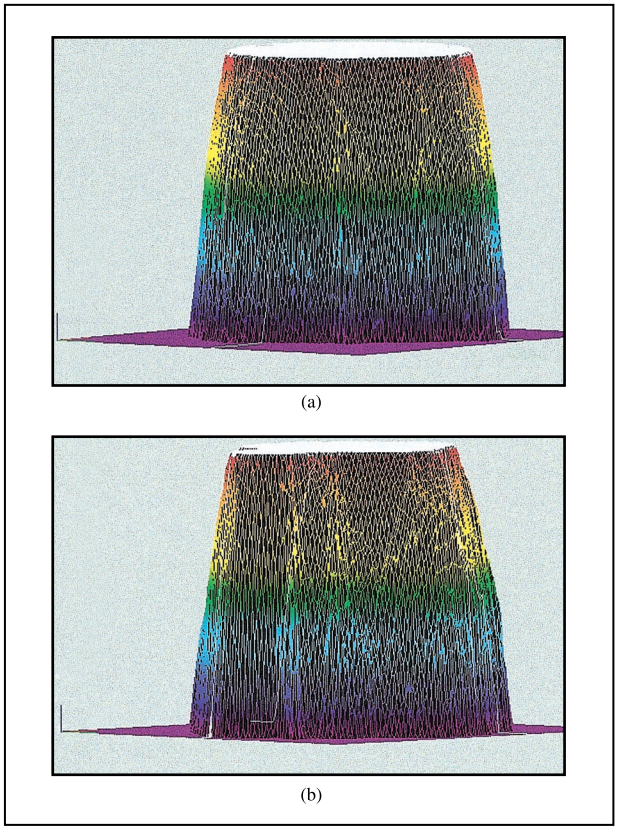


Figure 7

MCP optical power profiles: (a) Single-mode fiber; (b) multimode fiber.

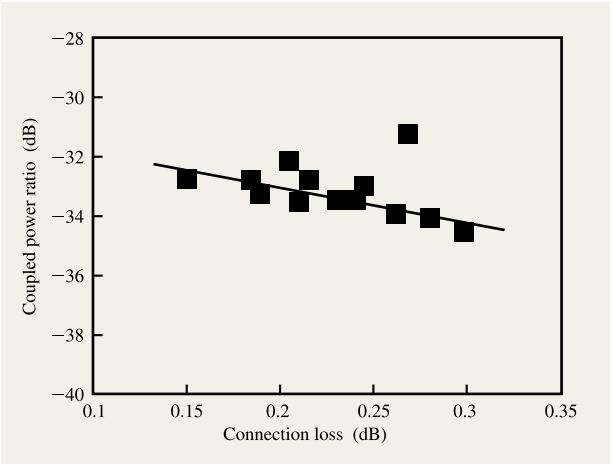


Figure 8

Coupled optical power vs. MCP connection loss.

for a long-wavelength laser source coupled directly into single-mode fiber; plot (b) shows distribution for the same laser launched into an MCP and then into a three-meter multimode jumper cable. It can be seen from these figures that the MCP-conditioned launch provides a uniform distribution of optical power among all the modes of the multimode fiber, and that the MCP-conditioned launch is virtually indistinguishable from the laser launch into single-mode fiber. Once the ferrules have been aligned and sealed into their optimal position, a simple assembly-loss type of measurement can be used to evaluate MCP performance, rather than measuring the complete two-dimensional coupled-power profile. As shown in **Figure 8**, there is a good correlation between the connection loss of the offset ferrules and the coupled-power ratio. The MCPs were tested under a variety of stressful conditions; experimental results are summarized in Table 5. In addition, several links were evaluated on the S/390 TeraPlex test bed attached to Parallel Sysplex links in a G5 sysplex; no performance problems were found during this testing.

4. IBM Fiber Transport Services

Given the many different types of fiber optic data links in a modern enterprise data center, the design of an optical cable infrastructure to accommodate both current and future needs has become increasingly complicated. IBM Site and Connectivity Services has been working with the S/390 development organization on the design of structured cabling systems to support multigigabit cable plants. In this section, we briefly describe several recent innovations in fiber optic cable and connector technology for the IBM structured cabling solution, known as Fiber Transport Services (FTS), which is available on G5

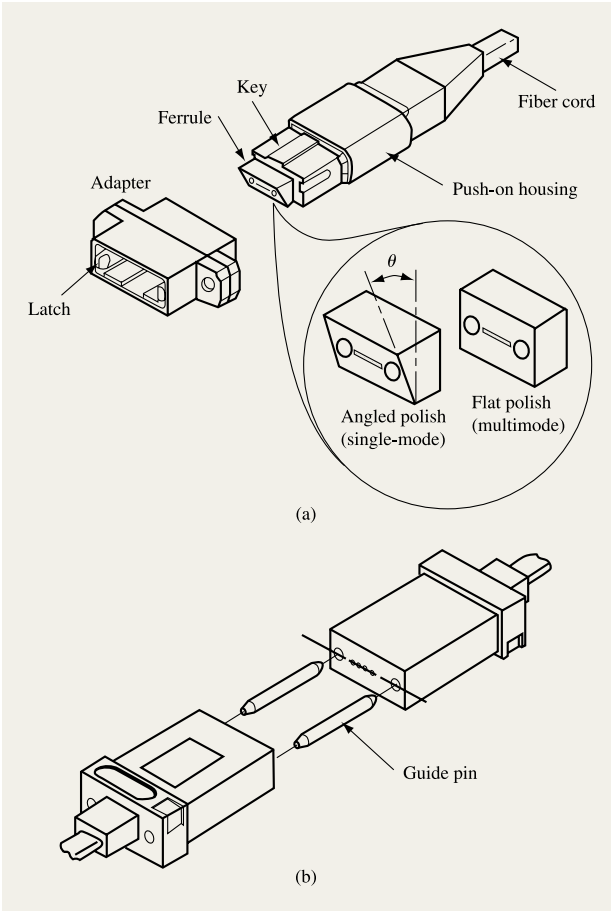


Figure 9

Multifiber termination push-on connector: (a) MTP connector and coupler for multimode and single-mode applications; (b) detail of multifiber ferrule showing metal alignment pins.

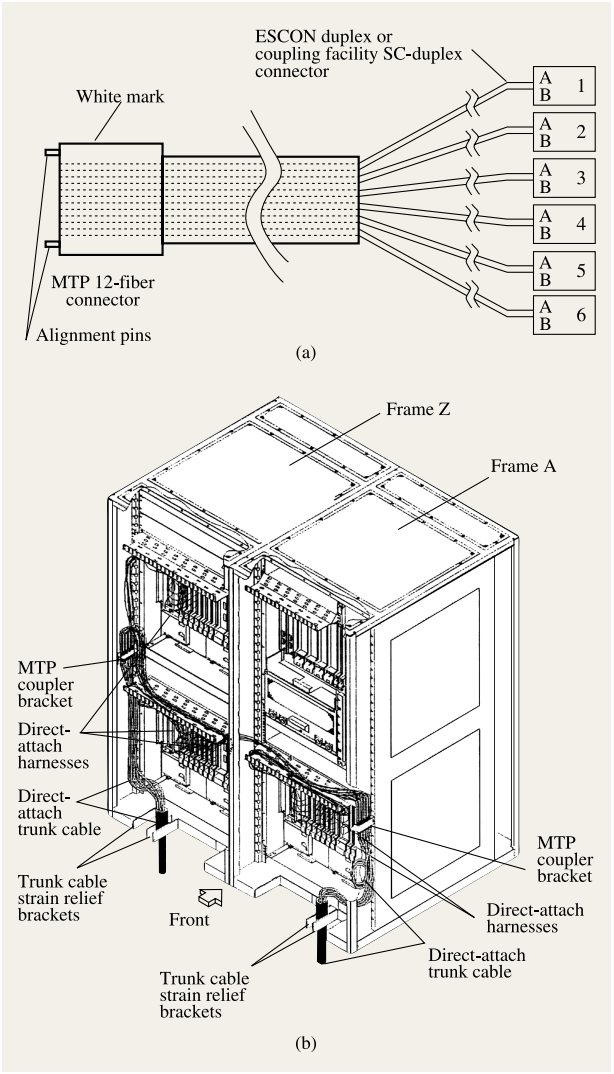


Figure 10

(a) FTS direct-attach cable harness using MTP connectors with fan-out to six duplex connectors; (b) FTS trunk cables and harness installed in a 9672 Model R3.

processors and will also be backward-compatible with all previous host processors, coupling facilities, and directors.

A central concept of FTS is the use of multifiber trunks, rather than collections of two-fiber jumper cables, to interconnect the various elements of a large data center [14–17]. FTS provides up to 144 fibers in a common trunk, which greatly simplifies cable management and reduces installation time. Cable congestion has become a significant problem in large data centers, with up to 256 ESCON channels on a large director or host processor. With the introduction of smaller, air-cooled CMOS-based processors and the extended distance provided by optical fiber attachments, it is increasingly common for data processing equipment to be rearranged and moved to different locations, sometimes on a daily basis. It can be time-consuming to reroute 256 individual jumper cables without making any connection errors or accidentally damaging the cables. To relieve this problem, this year FTS and S/390 have introduced the fiber quick connect system for multifiber trunks. The trunks are terminated with a special 12-fiber optical connector known as a multifiber termination push-on (MTP) connector [1] (**Figure 9**). Each MTP contains twelve fibers or six duplex channels in a connector smaller than most duplex connections in use today (barely 0.5 inches wide). In this way, a 72-fiber trunk cable can be terminated with six MTP connectors; relocating a 256-channel ESCON Director now requires replugging only 43 connections. Trunk cables terminated with multiple MTP connectors

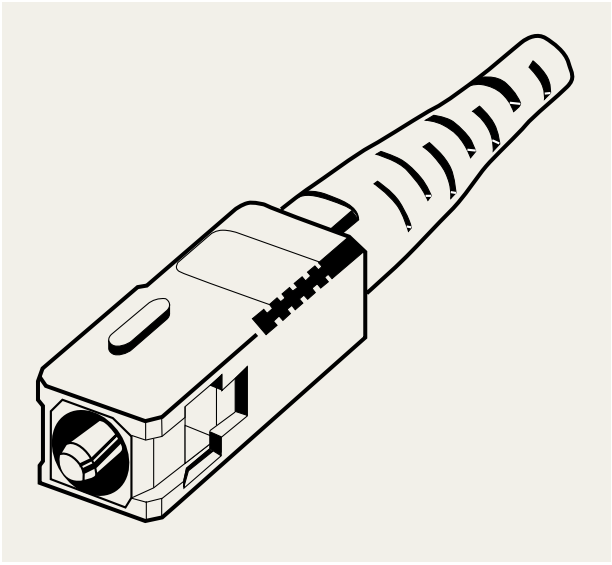


Figure 11

SC–DC fiber optic connector.

are available in four versions: 12-fiber/6-channel, 36-fiber/18-channel, 72-fiber/36-channel, and 144-fiber/72-channel. Optical alignment is facilitated by a pair of metal guide pins in the ferrule of a male MTP connector, which mate with corresponding holes in the female MTP connector. Under the covers of a director or G5 processor, the MTP connectors attach to an MTP coupler bracket (similar to a miniature patch panel); from there, a cable harness fans out each MTP into six duplex connectors which mate with the fiber optic transceivers (**Figure 10**). Since the qualification of the cable harness, under-the-covers patch panel, and trunk cable strain relief are all done in collaboration with the S/390 development organization, the FTS solution functions as an integral part of the applications and is the preferred cabling solution for the S/390 platform.

At the other end of the FTS trunk, individual fiber channels are fanned out at a patch panel or central patching location (CPL), where duplex fiber connectors are used to reconfigure individual channels to different destinations. These fan-outs are available for different fiber optic connector types, although ESCON and subscriber connection (SC) duplex are most common for multimode and SC duplex for single-mode. Fanning out the duplex fiber connections at a CPL also offers the advantage of being able to arrange the CPL connections in consecutive order of the channel identifiers on the host machine, greatly simplifying link reconfigurations.

As the size of S/390 processors was reduced and the number of channels increased, the size of the CPL soon became a limiting factor in many installations. To keep the CPL from occupying more floor space than the processors, a more dense optical connector technology was required for the fiber quick connect system. To meet this need, Global Services has adopted a new, small-form-factor fiber optic connector as the preferred interconnect

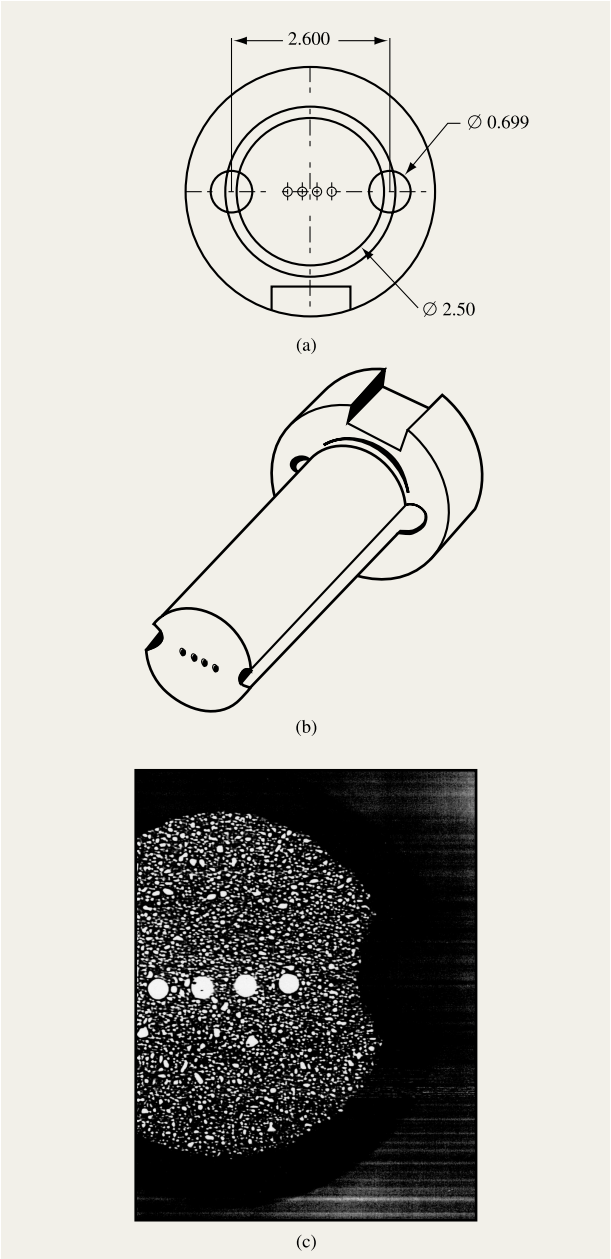


Figure 12

(a) Schematic of SC–DC ferrule design; (b) photograph of SC–DC ferrule; (c) optical microscope photograph of SC–DC end face showing four potential fiber locations and alignment groove.

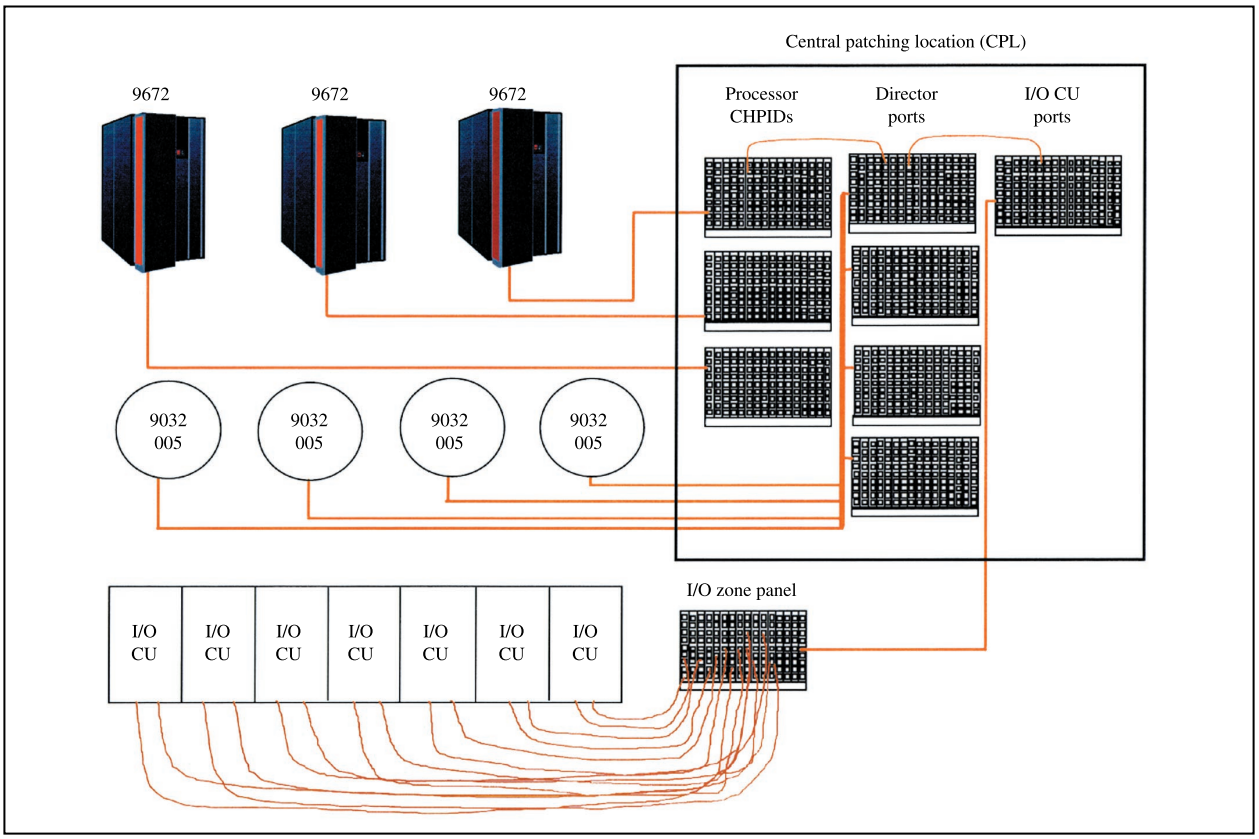


Figure 13

Typical FTS installation in an S/390 environment.

for multimode patch panels, the SC–DC** (abbreviation for SC dual contact, or two optical fibers in one SC ferrule and connector housing). Shown in **Figure 11**, the SC–DC [18, 19] is only about half the size of a conventional SC duplex connector, allowing for significantly more dense patch panel designs. Each SC–DC connector uses standard SC connector hardware containing a special multifiber ferrule with two fibers mounted on 75-μm fiber-to-fiber spacing or pitch. The outer body is compatible with industry standard SC housings and latching, with the exception of an offset key to prevent accidental misplugging of the SC–DC into a conventional SC duplex coupler. The SC housing was chosen for its proven design and easy push-pull insertion; additionally, it complies with the industry standard mezzanine height specifications (IEEE Standard 1301.4-1996). The SC–DC connector was originally developed by Siecor Corporation, and has been proposed as an EIA/TIA standard (FOCIS Proposal 11, EIA/TIA Accession No. 1460). Because the majority of structured cable systems use multimode fiber, the multimode SC–DC has been introduced first; in the future, single-mode SC–DC connectors may be offered as well.

A schematic of the SC–DC ferrule is shown in **Figure 12**, along with a photograph of the ferrule and an optical microscope photograph of the ferrule end face. The ferrule is thermoset-molded using a filled epoxy (the same material as that used on the MTP multifiber ferrule) which provides tight dimensional control. The outside diameter of the ferrule is 2.5 mm, with fiber holes on a 750-μm center-to-center pitch. The ferrule can potentially accommodate up to four fibers on a 250-μm center-to-center pitch for increased density in future applications. An alignment groove with a 350-μm radius is positioned along each side of the ferrule at 2.6-mm center-to-center spacing. These semicircular grooves provide alignment with guide ribs in the SC–DC coupler to facilitate alignment of the fibers; they also allow the ferrule to mate with other multifiber connectors containing guidepins at this standard spacing. The critical features of the ferrule are the alignment grooves and fiber holes. When two ferrules are mated inside a coupler, they are aligned

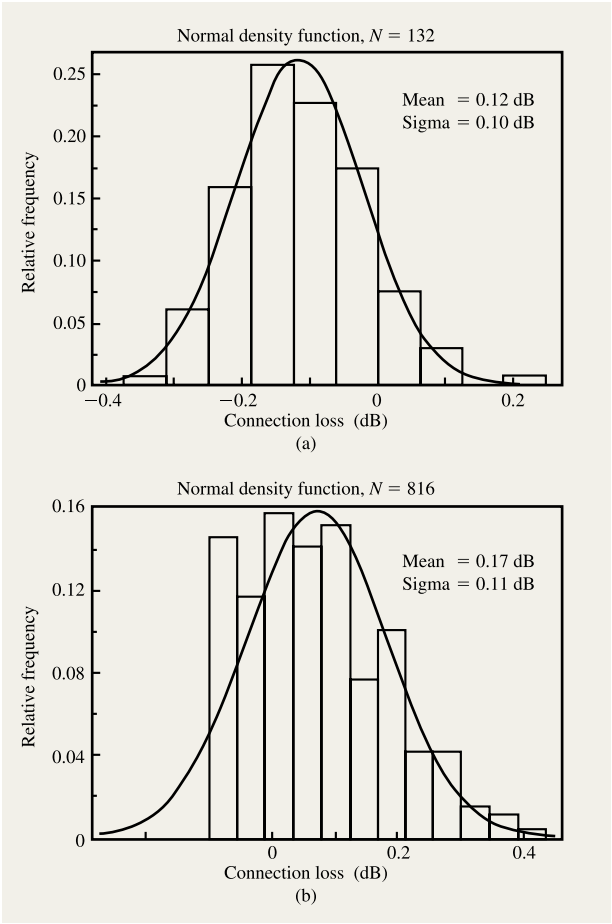


Figure 14

Statistical assembly loss measurements on a sample of 132 SC–DC connectors (a) before and (b) after environmental stress (thermal cycle −40° to +60°C, 10 cycles, complying with EIA-455-71).

Table 7 SC–DC fiber optic connector performance data.

Test procedure	*Results*
Connection loss 0.6 dB max.	Pass mean = 0.05 sigma = 0.03
Ship shock −10°C to +40°C, 10 cycles max., 0.6 dB change	Pass mean = 0.12 sigma = 0.10
Random connection 0.6 dB max.	Pass mean = 0.17 sigma = 0.11
500 × insertions 0.5 dB max. change	Pass mean = 0.03 sigma = 0.08
Axial pull 20 N at 1 m 0.6 dB max. change	Pass mean = 0.06 sigma = 0.11
Heat age/ humidity 60°C/95% 336 hours 0.6 dB max. change	Pass mean = 0.01 sigma = 0.06
Thermal cycle 500 cycles 10–60°C 0.6 dB max. change	Pass mean = 0.10 sigma = 0.10

axially and radially by ribs or pins in the coupler sleeve. The standard 2.5-mm ferrule diameter makes it possible to use existing termination and polishing equipment for both factory-terminated and field-terminated SC–DC connectors. The field-installable version of this connector, known as the SC–DC Unicam**, uses the pre-stubbed fiber concept [19], which eliminates epoxy and polishing on field-terminated connectors. A pre-polished Unicam fiber stub is attached in the field using a mechanical splice element filled with index matching gel, contained within the connector body. The Unicam can be field-installed in less than one minute.

Although its physical size is only half as large as that of conventional duplex connectors, the SC–DC has been designed and qualified to meet the same rigorous application requirements. To date, nearly 100 000 of these connectors have been installed in large data centers, including the IBM TeraPlex; a typical system configuration is shown in **Figure 13**. The connectors and cable assemblies have been subjected to an extensive qualification by IBM, including insertion loss, return loss, durability testing, and environmental stress. A summary of some experimental data on this connector is shown in **Table 7**; in addition, the connector meets or exceeds all standard Bellcore and EIA/TIA requirements for optical, mechanical, and environmental performance. Of particular interest are the cable assembly loss values for the SC–DC, which directly affect the link loss budget and total available distance. Assembly loss measurements were performed on 132 samples of SC–DC connectors, before and after environmental stress (ten cycles from −10°C to +40°C, complying with EIA-455-171). Statistical results are shown in **Figure 14**, along with a Gaussian curve fit; the connector performance exceeds all application requirements.

5. Wavelength multiplexing and GDPS

In 1994, IBM announced the Parallel Sysplex architecture for the S/390 platform. This architecture uses high-speed fiber optic data links to couple processors together in parallel [20], thereby increasing capacity and scalability. Processors are interconnected by a coupling facility, which provides data caching, locking, and queueing services; it may be implemented as a logical partition rather than a separate physical device. The gigabit links, known as intersystem channel (ISC), HiPerLinks, or coupling links, use long-wavelength (1300-nm) lasers and single-mode fiber to operate at distances up to 10 km with a 7-dB link budget, as defined in Table 1 (HiPerLinks were originally announced with a maximum distance of 3 km, which was increased to 10 km in May 1998). These links use the SC duplex connector. Future HiPerLinks will offer higher data rates on SM fiber. When HiPerLinks were originally announced, an optional interface at 531 Mb/s was offered using short-wavelength lasers on MM fiber. The 531-Mb/s HiPerLinks were discontinued in May 1998 for the G5 and its successors, consistent with the long-term strategy of using single mode for S/390. A feature is available to accommodate operation of 1-Gb/s HiPerLink adapters on multimode fiber, using a mode-conditioning jumper cable at restricted distances (see Section 3).

The physical layer design is similar to that recommended by the ANSI Fibre Channel Standard, operating at a data rate of 1.0625 Gb/s, except for the use of open fiber control (OFC) laser safety on long-wavelength (1300-nm) laser links. Open fiber control is a safety interlock implemented in the transceiver hardware; a pair of transceivers connected by a point-to-point link must perform a handshake sequence in order to initialize the link before data transmission occurs. Only after this handshake is complete will the lasers turn on at full optical power. If the link is opened for any reason (such as a broken fiber or an unplugged connector), the link detects this and automatically deactivates the lasers on both ends to prevent exposure to hazardous optical power levels. When the link is closed again, the hardware automatically detects this condition and reestablishes the link. The HiPerLinks use OFC timing corresponding to a 266-Mb/s link in the ANSI standard, which allows for longer distances at the higher data rate.

There are three possible configurations for a Parallel Sysplex. First, the entire sysplex may reside in a single physical location, within one data center. Second, the sysplex can be extended over multiple locations with remote fiber optic data links. Finally, a multisite sysplex in which all data is remote-copied from one location to another is known as a Geographically Dispersed Parallel Sysplex, or GDPS. GDPS provides the ability to manage remote-copy configurations, automates both planned and unplanned system reconfigurations, and provides rapid failure recovery from a single point of control. There are different configuration options for a GDPS. The single-site workload configuration is intended for those enterprises which have production workload in one location (site A) and discretionary workload (system test platforms, application development, etc.) in another location (site B). In the event of a system failure, an unplanned site failure, or a planned workload shift, the discretionary workload in site B is terminated to provide processing resources for the production work from site A (the resources are acquired from site B to prepare this environment, and the critical workload is restarted). The multiple-site workload configuration is intended for those enterprises which have production and discretionary workload at both site A and site B. In this case, discretionary workload from either site may be terminated to provide processing resources for the production workload from the other site in the event of a planned or unplanned system disruption or site failure.

Multisite Parallel Sysplex or GDPS configurations may require many links (ESCON, HiPerLinks, and sysplex timer) at extended distances; an efficient way to realize this is the use of wavelength-division multiplexing technology. Multiplexing wavelengths is a way to take advantage of the high bandwidth of fiber optic cables without requiring extremely high modulation rates at the transceiver. Traditionally, optical wavelength-division multiplexing (WDM) has been widely used in telecom applications, but has found limited usage in datacom applications. This may change, as a number of companies are now offering multiplexing alternatives for datacom networks which must make more efficient use of their existing bandwidth. This technology may even be the first step toward development of all-optical networks [21]. For parallel sysplex applications, the only currently available WDM channel extender which supports ETR and HiPerLinks in addition to ESCON channels is the IBM 9729 Optical Wavelength Division Multiplexer [22–24] (**Figure 15**). It allows the transmission of up to 20 independent data streams (ten full-duplex channels) over one single-mode fiber. An entry-level version of the product is also available which supports only eight independent data streams (four full-duplex channels). Using different adapter cards, a mixture of multimode links (ESCON, ETR, FDDI, Fast Ethernet, and ATM 155) and single-mode links (HiPerLinks) can be plugged into the device, which is protocol-independent. The data is remodulated using distributed-feedback laser diodes with wavelengths spaced 1 nm apart near the 1550-nm region. The optical signals are then combined, using a diffraction grating with embedded fiber "pigtails," and coupled into a single-mode fiber; another unit at the far end of the link demultiplexes the signals. If a single channel is disconnected from the 9729 input, this condition is detected and the corresponding 9729 output channel is also deactivated; both channels automatically

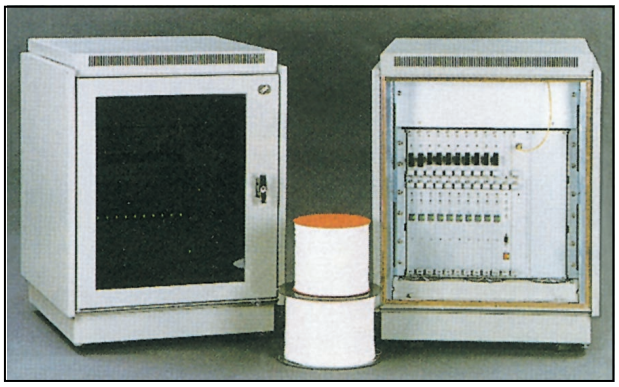


Figure 15

IBM 9729 Optical Wavelength Division Multiplexer.

Table 8 IBM 9729 optical power variation with temperature.

Temperature (°C)	*Wavelength, card A3*	*Wavelength, card B3*
23	1545.001	1544.002
10	1545.001	1543.997
40	1545.000	1543.990
10	1545.000	1543.994
50	1544.995	1543.989
0	1544.997	1543.996
23	1544.999	1543.998

Max. change for card A3 over 10–40°C is 0.001 nm.
Max. change for card B3 over 10–40°C is 0.004 nm.

resume operation when the link is restored (this function is required to prevent data integrity problems on the attached systems). The maximum unrepeated distance is 50 km for data in the 200-Mb/s range (such as ESCON or OC-3) with a 15-dB link budget, and 20 km for data in the 1-Gb/s range (such as HiPerLinks) with a 12-dB link budget (see Table 1). Note that despite the low data rate of ETR links, they are limited to shorter distances (26 km) because of timing considerations. Upon special request (RPQ 8P1955), it is possible to extend these distances to 40 km for both ETR and HiPerLinks, allowing the construction of multisite Parallel Sysplex or GDPS configurations. For protection against a broken optical link between the units, an optional dual-fiber switch card is available, consisting of an optical switch and a second fiber link. The units automatically detect whether the primary fiber link is broken, and switch operation to the secondary fiber until repairs can be made. The 9729 is managed through a serial data port, and can report its status or receive simple commands, such as switching to the second fiber link, from a personal computer or workstation running a software management package. This type of product is a cost-effective way to utilize leased fiber optic lines, which are not readily available everywhere and may be very high-cost. For example, the average cost of a leased fiber (sometimes known as dark fiber) is $150 per mile per month; for a 10-km configuration with 10 full-duplex channels, the 9729 saves more than $340 000 in leasing costs the first year alone.

For dense WDM applications such as the 9729, wavelength drift of the lasers due to temperature and other environmental variations must be strictly controlled. For this device, each WDM laser uses thermoelectric coolers to stabilize laser temperature and hence wavelength. Although it is possible to adjust the laser wavelength during operation, nominally the lasers remain quite stable. Thermal testing of the 9729 has been performed for two cases. As mentioned earlier, one 9729 unit is required at each end of a link; these are designated as A and B units, respectively. The first case involved a link with the A unit in a thermal cycle chamber while the B unit remained at room temperature, and the second involved both A and B units in separate temperature chambers. Thus, the effect of either one or both ends of a 9729 link operating at extreme temperature could be studied. Ambient operating temperature in both cases was cycled over the range 10°C to 40°C while pseudorandom data was running over all 9729 links; any drifts in the laser temperature and wavelength would induce bit errors on the link. During both tests, no data errors were observed during an eight-hour test at 200 Mb/s. The control voltage of the thermoelectric coolers was monitored during the testing to verify that the laser temperatures remained stable. Starting at room temperature (23°C), the units were cycled to 10°C, then back to 40°C and back to 10°C again. The units were then cycled outside their specified operating range to 50°C and back to 0°C before returning to room temperature. All temperatures were held for at least 30 minutes, the duration of increasing temperature transitions was 10 minutes maximum, and the duration of decreasing temperature transitions was 25 minutes maximum. Representative data of wavelength vs. temperature is shown in **Table 8**; it can be seen that the laser temperatures held stable during this measurement.

To illustrate the application of a 9729, consider the construction of a GDPS between two remote locations for disaster recovery, as shown in **Figure 16**. A Parallel Sysplex comprises four building blocks: the host processor (or parallel enterprise server), the coupling facility, the sysplex timer (ETR), and disk storage. Many different processors may be interconnected through the coupling facility, which allows them to communicate with one another and with data stored locally. The coupling facility provides data caching, locking, and queueing (message-passing) services. By adding more processors to the configuration, the

overall processing power of the sysplex (measured in millions of instructions per second, or MIPS) is increased. It is also possible to upgrade to more powerful processors by simply connecting them into the sysplex through the coupling facility. Special software allows the sysplex to break down large database applications into smaller ones, which can then be processed separately; the results are combined to arrive at the final query response. The coupling facility may be implemented either as a separate piece of hardware or as a logical partition of a larger system. The HiPerLinks are used to connect a processor with a coupling facility. Since the operation of a Parallel Sysplex depends on these links, it is highly recommended that redundant links and coupling facilities be used for continuous availability.

Since all of the processors must operate synchronously with one another, they all require a multimode fiber link to a common ETR reference clock. The ETR is a critical component of the Parallel Sysplex; the sysplex will continue to run with degraded performance if a processor fails, but failure of the ETR will disable the entire sysplex. For this reason, it is highly recommended that two ETRs be used, so that if one fails the other can maintain uninterrupted operation of the sysplex. For this to occur, the two ETRs must also be synchronized with each other; this is accomplished by connecting the two ETRs with two separate fiber links called control link oscillator (CLO) links. Physically, the CLO link is the same as an ETR link except that it carries timing information to keep the pair of ETRs synchronized. Note that because the two sysplex timers are synchronized with each other, it is possible that some processors in a sysplex can run from one ETR while others run from the second ETR. In other words, the two timers may both be in use simultaneously, running different processors in the sysplex, rather than one timer sitting idle as a backup in case the first timer fails.

Thus, in order to build a GDPS, we require at least one each of the processor, coupling facility, sysplex timer, and disk storage at both the primary and secondary locations, shown in Figure 16(a) as site A and site B. Recall that one processor may be logically partitioned into many different sysplex system images; the number of system images determines the required number of HiPerLinks. The sysplex system images at site A must have HiPerLinks to the coupling facilities at both site A and site B. Similarly, the sysplex system images at site B must have HiPerLinks to the coupling facilities at both sites A and B. In this way, failure of one coupling facility or one system image allows the rest of the sysplex to continue uninterrupted operation. A minimum of two links is recommended between each system image and coupling facility. Assuming that there are S sysplex system images running on P processors and C coupling facilities in the GDPS, spread equally between site A and site B, the total number of HiPerLinks required is

$$\text{No. of HiPerLinks} = S \times C \times 2. \quad (1)$$

In a GDPS, the total number of intersite HiPerLinks is

$$\text{No. of intersite HiPerLinks} = S \times C. \quad (2)$$

The sysplex timer (ETR, or 9037) at site A must have links to the processors at both sites A and B. Similarly, the 9037 at site B must have links to the processors at both sites A and B. There must also be two CLO links between the sysplex timers at sites A and B. This makes a minimum of four duplex intersite links, or eight optical fibers without multiplexing. For practical purposes, there should never be a single point of failure in the sysplex

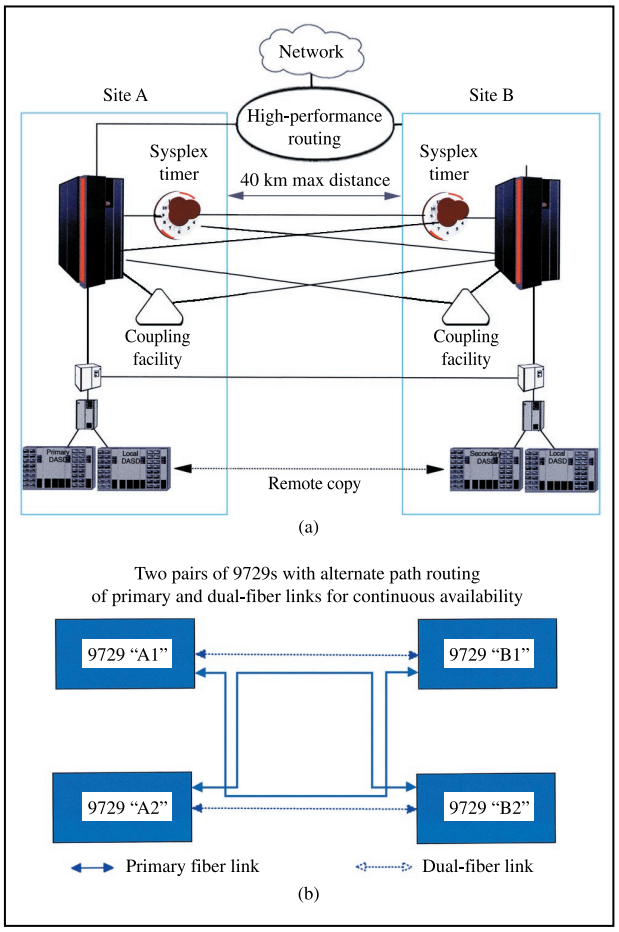


Figure 16

(a) Geographically Dispersed Parallel Sysplex (GDPS) configuration illustrating full redundancy on all intersite data links; the intersite data links would be routed through the 9729. (b) Recommended configuration for dual-fiber routing to provide continuous availability.

implementation; if all of the fibers are routed through the same physical path, there is a possibility that a disaster on this path would disrupt operations. For this reason, it is highly recommended that dual physical paths be used for all local and intersite fiber optic links, including HiPerLinks, ESCON, sysplex timer, and CLO links. If there are P processors spread evenly between site A and site B, the minimum number of sysplex timer links required is

$$\text{No. of ETR links} = (P \times 2) + 2 \text{ CLO links.} \quad (3)$$

In a GDPS, the number of intersite sysplex timer links is

$$\text{No. of intersite sysplex timer links} = P + 2 \text{ CLO links.} \quad (4)$$

These formulas are valid for CMOS-based hosts only; note that the number of sysplex timer links doubles for ES/9000* multiprocessor models.

In addition, there are other types of intersite links such as ESCON channels to allow data access at both locations. In a GDPS with a total of N storage subsystems (also known as direct-access storage devices, or DASD), it is recommended that there be at least four or more paths from each processor to each storage control unit (based on the use of ESCON directors at each site); thus, the number of intersite links is

$$\text{No. of intersite storage (ESCON) links} = N \times 4. \quad (5)$$

In addition, the sysplex requires direct connections between systems for cross-system coupling facility (XCF) communication. These connections may be provided by either ESCON channel-to-channel (CTC) links or HiPerLinks. If coupling links are used for XCF signaling, no additional HiPerLinks are required beyond those given by Equations (1) and (2). If CTC links are used for XCF signaling, at least two inbound and two outbound links between each system are required, in addition to the ESCON links for data storage discussed previously. The minimum number of CTC links is

$$\text{No. of CTC links} = S \times (S - 1). \quad (6)$$

For a GDPS with S_A sysplex systems at site A and S_B sysplex systems at site B, the minimum number of intersite channel-to-channel links is

$$\text{No. of intersite CTC links} = S_A \times S_B \times 2. \quad (7)$$

Since some processors also have direct local area network (LAN) connectivity via FDDI or ATM/SONET links, it may be desirable to run some additional intersite links for remote LAN operation as well.

As an example of the application of these equations, consider a GDPS consisting of two system images executing on the same processor and a coupling facility at site A, and the same configuration at site B. Each site also contains two primary and two secondary DASD subsystems. Sysplex connectivity for XCF signaling is provided by ESCON CTC links, and all GDPS recommendations for dual redundancy and continuous availability in the event of a single failure have been implemented. From Equations (1)–(7), the total number of intersite links required is

$$\text{No. of CTC links} = S_A \times S_B \times 2 = 2 \times 2 \times 2 = 8;$$

$$\text{No. of timer links} = P + 2 = 2 + 2 = 4;$$

$$\text{No. of HiPerLinks} = S \times C = 4 \times 2 = 8;$$

$$\text{No. of storage (DASD) links} = N \times 4 = 4 \times 4 = 16; \quad (8)$$

or a total of 36 intersite links. Since a pair of 9729s can support up to ten links, four pairs are required in this example.

It is preferable to use at least two pairs of 9729s at all times to accommodate alternate physical paths for the intersite fibers, which avoids a single point of failure. While it is possible to use the dual-fiber switch option on the 9729 to accommodate alternate fiber paths, this does not guarantee uninterrupted operation of the sysplex if an intersite link breaks. The 9729 cannot switch to the dual fiber fast enough to prevent the HiPerLinks from being interrupted by their open fiber control, which then takes up to ten seconds to reestablish the links. ESCON and ETR channels will also experience loss-of-light disruptions. Even when all of the links reestablish, the application will have been interrupted or disabled, and any jobs which had been running on the sysplex will have to be restarted or reinitiated, either manually or by the host's automatic recovery mechanisms, depending on the state of the job when the links were broken. The dual-fiber switch enables the system to recover when an intersite link is broken, but a pair of WDM channels (two pairs of 9729s using different fiber paths) is the only way to ensure uninterrupted operation of the sysplex when an intersite link is broken. A good design practice would be to use dual-fiber cards in each pair of 9729s and two diverse physical paths. The primary intersite link between the first pair of 9729s follows the same path as the secondary intersite link between the second pair of 9729s, as shown in Figure 16(b). If either path is disrupted, the attached systems continue uninterrupted operation over one pair of 9729s, and applications running over the other pair of 9729s can be restarted using the backup fiber.

While the performance of any large-scale computer system is highly application-dependent, we can infer some of the effects caused by extended distances. For the case of I/O requests to DASD on an ESCON link, assume that a typical storage read or write operation at the primary site takes 3 ms. The latency of an intersite fiber optic link is about 10 μs/km round trip; this must be multiplied by

the intersite distance and the number of acknowledgments required by the data-link protocol to determine the impact of intersite distance on performance. If we assume a conservative datacom protocol (such as ESCON) that requires six acknowledgments per operation, the additional delay at a distance of 40 km is (10 μs/km/round trip) $\times$ (40 km) $\times$ (6 round trips) = 2.4 ms. The time required for a DASD read operation from site B to DASD in site A is then 3 + 2.4 = 5.4 ms. Similarly, a data-mirroring application might require a write operation to the DASD in site A that would then be remote-copied to DASD in site B. This operation would take 3 ms for the local write, 2.4 ms for latency, and 3 ms for the remote write, or 8.4 ms total. If the data must first be requested from site B before this operation can begin, this adds another 2.4 ms for a total of 10.8 ms. In a similar fashion, performance of all ESCON and HiPerLinks degrades with distance; there is no general formula to predict this impact, so it must be individually evaluated for each software application and datacom protocol.

We have built the system configuration shown in Figure 16, with two pairs of 9729s to provide full redundancy, at the IBM TeraPlex Complex on the S/390 system test floor in Poughkeepsie, New York, in order to evaluate its performance. The test system used a bipolar S/390 mainframe as the site A processor, and a CMOS Generation 5 server as the site B processor. Each 9729 pair was equipped with the dual-fiber switch and configured with five HiPerLinks, two sysplex timer channels, and three ESCON channels. Spools of fiber were used to simulate the intersite links, with at least eight ST-type optical connections in a 20-km link to verify that there was no effect from connector return loss or modal noise (each laser transmitter on the 9729 is equipped with an optical isolator). Various applications were run on the sysplex to simulate typical commercial data link traffic under OS/390/MVS operating systems; the processors were logically partitioned into 15 processing zones, the maximum allowed for this machine type. To stress the sysplex, all fiber optic links between the processors, coupling facilities, sysplex timers, storage disks, and 9729s were set to their maximum specified distances and loss values as given in Table 1; the maximum distance between the 9729 pair was 40 km. System code on the processors was used to log bit errors on all links as well as any other conditions indicating either failure or degraded performance on any given link.

All fiber optic links operated error-free (extrapolated to 10^{-15} bit-error rate) over a 48-hour test at maximum distance and link loss. To test for saturation effects, the link between the 9729s was then reduced to three meters, and error-free operation was demonstrated for eight hours. Successive hour tests were performed at intersite distances of 4 km, 8 km, 12 km, 20 km, and 40 km (note

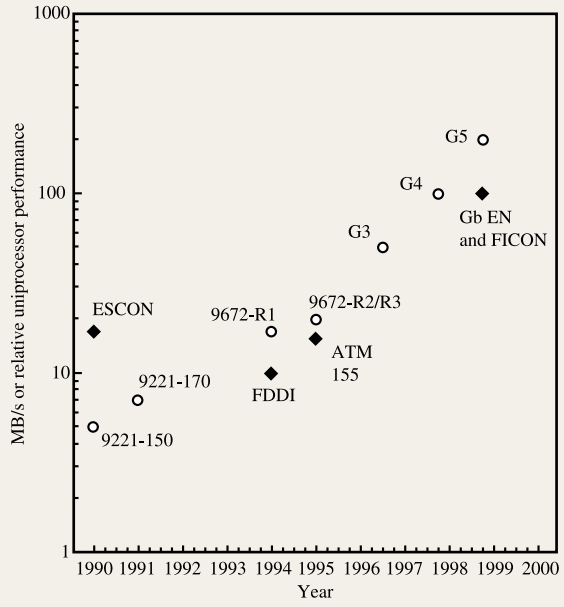


Figure 17

Growth of optical data link capacity in MB/s (black diamonds) and relative S/390 host processor capacity in MIPS (white circles) over time.

that GDPS installations beyond 20 km are supported by IBM only by special request). Thus, the links are loss-limited rather than dispersion-limited. It is extremely important to note that although the HiPerLink and sysplex timer link distances can be increased to 40 km, the sysplex timer links (in particular the timer-to-timer CLO link distance) *cannot* exceed 40 km. Extending the CLO link distance beyond 40 km could result in the two timers being out of synchronization with each other, thereby resulting in a potential data integrity exposure. The out-of-synchronization condition is not detectable during normal system operation. The 9729s were then populated with ESCON links only (no HiPerLinks or sysplex timer channels), and it was verified that distances up to 50 km with 15 dB loss between the 9729 pair ran error-free for eight hours. The system was then reconfigured again with three ESCON links, two sysplex timer links, and five HiPerLinks. It was verified that with redundant dual paths implemented on all fiber links, breaking any individual link in the sysplex configuration did not interfere with system operation. It was further shown that swapping channels on the 9729, or replacing cards on the 9729 while it was powered on, did not induce crosstalk or interference in adjacent channels.

6. Conclusions

Fiber optic communication links have come to play a central role in large computing applications; with the introduction of the G5 processors, fiber optics remains a key enabling technology for improved I/O performance, open systems networking, and the Parallel Sysplex architecture. The new FICON standard has provided higher-performance I/O to keep pace with the increased capacity of the G5 processors (in MIPS). The FICON bridge card uses an equivalent 8-to-1 aggregation of ESCON channels to provide increased effective channel capacity on the G5 and facilitate migration from ESCON to FICON environments. Single-mode optical adapter cards for FICON, Gigabit Ethernet, and Parallel Sysplex can now reuse installed multimode fiber with the optical mode-conditioning patch cords. New, smaller fiber optic connectors and multifiber connectors allow the IBM Fiber Transport Services to support more links in less space and simplify reconfiguration of patch panels and relocation of hardware in the data center. Finally, using wavelength-division multiplexing technology both to reduce the total number of fibers required for intersite links and to increase the distance of those links, it is possible to construct geographically dispersed Parallel Sysplexes at distances up to 40 km. This technology has enabled new applications, particularly for disaster recovery and data mirroring at extended distances.

The continued growth in MIPS of the S/390 platform requires the ongoing development of higher-performance fiber optic channels in order to achieve balanced system performance. **Figure 17** shows the relative growth in MIPS of CMOS-based uniprocessors over time, as well as the trend of increasing data rates on the attached channels. This figure illustrates why it was necessary to develop a new I/O channel type, FICON, to keep pace with the higher performance of the G5.

The FICON architecture introduced in 1998 also provides a foundation for higher-speed FICON channels on future large systems, allowing the I/O to continue its growth with processor performance. The need for higher-speed networking protocols such as Gigabit Ethernet is also apparent from this figure, although networking adapters are not driven primarily by the speeds of the servers as much as by their widespread acceptance on LAN clients, switches, and hubs. Therefore, we can expect that while networking applications will continue to grow in the future, their performance will tend to lag behind large-server performance. Coupling links for a Parallel Sysplex are not shown on this chart, since their performance also does not scale well with processor performance; in fact, the coupling link speed may have to be increased only after many generations of processors. In the future, as processor performance continues to increase, fiber optic interconnects will continue to increase in number and improve in performance to provide a total systems solution for the S/390 platform.

Acknowledgments

The authors gratefully acknowledge the contributions made to this paper by many talented individuals, including Walt Mostowy of IBM Global Services, Dayton, New Jersey; Jean Trewhella, Eric Hall, and Frank Janiello of the IBM Thomas J. Watson Research Center, Yorktown Heights, New York; Ernie Swanson and George Middlebrook of IBM Network Hardware Development, Raleigh, North Carolina; Victor Gregurick and Hamid Bagheri of IBM Global Procurement, East Fishkill, New York; Sam Thomas, former IBM co-op student from the Pennsylvania State University, University Park, Pennsylvania; John Fox of ComputerCrafts Inc., Hawthorne, New Jersey; Markus Giebel, Karl Wagner, and Todd Hudson of Siecor Corporation, Hickory, North Carolina; Charlie Hubert, Mario Borelli, Mark Maruzzi, Audrey Helffric, Ken Scea, Mike Humes, Ken Trowe, and the Connectivity Solutions development organization of the IBM S/390 Division, Poughkeepsie, New York.

References

1. *Handbook of Fiber Optic Data Communication*, C. DeCusatis, E. Maass, D. Clement, and R. Lasky, Eds. (IBM Puborder SR23-8194), Academic Press, Inc., New York, 1998.
2. Topical issue on IBM System/390: Architecture and Design, *IBM J. Res. Develop.* **36,** No. 4 (1992).
3. C. DeCusatis, A. Huffman, G. DeMario, and D. Stigliani, "ESCON Fiber Optic Interface for Mainframe Computing," *Opt. News* **2,** 33 (1991).
4. IBM Corporation, *ESCON I/O Interface Physical Layer Document*, Order No. SA23-0394, 1995; available through IBM branch offices.
5. *ANSI Single Byte Command Code Sets CONnection Architecture (SBCON)*, ANSI Standard No. X3T11/95-469, American National Standards Institute, Washington, DC, 1996.
6. IBM Corporation, *Introduction to Enterprise Systems Connection*, Order No. GA23-3833, 1991; available through IBM branch offices.
7. IBM Corporation, *Enterprise Systems Architecture/390 ESCON I/O Interface*, Order No. SA22-7202, 1991; available through IBM branch offices.
8. *Supplement to Carrier Sense Multiple Access with Collision Detection (CSMA/CD) Access Method and Physical Layer Specifications: Media Access Control (MAC) Parameters, Physical Layer, Repeater and Management Parameters for 1000 Mb/s Operation*, IEEE 802.3z (Gigabit Ethernet Standard), IEEE Standards Department, Piscataway, NJ, 1998.
9. IBM Corporation, *S/390: MVS/ESA and Parallel Sysplex Enterprise Computing*, Order No. SBOF-7352, 1993; available through IBM branch offices.

10. IBM Corporation, *Parallel Sysplex White Paper*, Order No. G326-3025, 1993; available through IBM branch offices. See also "Geographically Dispersed Parallel Sysplex: The S/390 Multi-Site Application Availability Solution" (full version and executive summary) at *http://www.ibm.com/pso/* or the IBM Redbook *S/390 I/O Connectivity*, Order No. SG24-5444.
11. *Fibre Channel Physical and Signaling Interface (FC-PH)*, Document No. X3.230, Rev. 3, American National Standards Institute, Washington, DC, 1994.
12. IBM Corporation, *S/390 Fiber Channel Connection (FICON) I/O Interface Physical Layer*, Order No. SA24-7172, 1998; available through IBM branch offices.
13. J. Fox, A. MacGregor, and S. Hogg, "Bandwidth Reduction in Gigabit Ethernet Transmission Over Multimode Fiber and Recovery Through Laser Transmitter Mode Conditioning," *Opt. Eng.* **37,** 3156–3160 (1998).
14. IBM Corporation, *Fiber Transport Services Physical and Configuration Planning*, Order No. GA22-7234, 1998; available through IBM branch offices.
15. IBM Corporation, *Planning for Fiber Optic Channel Links*, Order No. GA23-0367, 1993; available through IBM branch offices.
16. IBM Corporation, *Maintenance Information for Fiber Optic Channel Links*, Order No. SY27-2597, 1993; available through IBM branch offices.
17. C. DeCusatis, "Fiber Optic Data Communication: Overview and Future Directions," *Opt. Eng.* **37,** 3082–3099 (1998).
18. K. Wagner, D. Dean, and M. Giebel, "The SC–DC/SC–QC Fiber Optic Connector," *Opt. Eng.* **37,** 3129–3133 (1998).
19. M. DeJong, "Cleave and Crimp Fiber Optic Connector for Field Installation," *Technical Digest*, *Conference on Optical Fiber Communications*, 1990, paper THA1, p. 139.
20. IBM Corporation, *Coupling Facility Channel I/O Interface Physical Layer*, Order No. SA23-0395, 1994; available through IBM branch offices.
21. G. L. Bona, W. E. Denzel, B. J. Offrein, R. Germann, H. W. M. Salemink, and F. Horst, "Wavelength Division Multiplexed Add/Drop Ring Technology in Corporate Backbone Networks," *Opt. Eng.* **37,** 3218–3228 (1998).
22. IBM Corporation, *9729 Operators Manual*, Order No. GA27-4172, 1996; available through IBM branch offices.
23. "IBM Corporation 9729 Optical Wavelength Division Multiplexer," *Photonics Spectra* (special issue, 1996 Photonics Circle of Excellence Awards) **30,** No. 6 (June 1996).
24. C. DeCusatis, D. Petersen, E. Hall, and F. Janniello, "Geographically Distributed Parallel Sysplex Architecture Using Optical Wavelength Division Multiplexing," *Opt. Eng.* **37,** 3229–3236 (1998).

Received October 8, 1998; accepted for publication June 23, 1999

Casimer M. DeCusatis *IBM System/390 Division, 522 South Road, Poughkeepsie, New York 12601 (decusat@us.ibm.com).* Dr. DeCusatis is a Senior Engineer in the IBM System/390 Division, Poughkeepsie, New York, where he has participated in many data communications development teams for ESCON transceivers and cables, the InterSystem Channel used on six generations of Parallel Sysplex architectures, the OETC and Jitney parallel optical links, the 9729 Optical Wavelength Division Multiplexer (winner of the 1996 *Photonics Spectra Circle of Excellence Award*) and FDDI/ATM/Gigabit Ethernet interfaces for the Open Systems Adapter family. He received the M.S. and Ph.D. degrees from Rensselaer Polytechnic Institute in 1988 and 1990, respectively, and the B.S. degree *magna cum laude* in the Engineering Science Honors Program from Pennsylvania State University in 1986. He is a co-inventor on 16 patents and author of more than 50 technical papers as well as the book *Acousto-optics: Fundamentals and Applications* (Artech House, 1990). He has contributed chapters to several books and has served as co-editor of the *Handbook of Fiber Optic Data Communications* (Academic Press, 1998) and as editor of the *Handbook of Applied Photometry* (AIP Press and Springer-Verlag, 1997). Dr. DeCusatis is currently President of the Institute for Optical Data Communications. His other research interests include signal processing using wavelets, which has been the subject of recent invited talks in St. Petersburg, Russia, and Gdansk, Poland. Dr. DeCusatis is a member of the Optical Society of America, IEEE, SPIE, Sigma Xi Research Society, and ten academic honor societies including Tau Beta Pi and Eta Kappa Nu; he has also been profiled by various biographical publications including *Who's Who in Science and Engineering*.

Daniel J. Stigliani, Jr. *IBM System/390 Division, 522 South Road, Poughkeepsie, New York 12601 (danst@us.ibm.com).* Dr. Stigliani received the B.Engr. degree in general engineering from Stevens Institute of Technology and the M.S. and Ph.D. degrees in electrical engineering from the University of Illinois. In 1969, he joined the IBM Federal Systems Division in Owego, New York, working on a broad range of projects in optical signal processing and communications. In 1974, Dr. Stigliani transferred to the System/390 Division, where he was responsible for the development of optical fiber communications for data processing applications. He has received three division awards, an IBM Outstanding Technical Achievement Award, a publication award, and two IBM Invention Achievement Awards. He has published numerous technical papers and has co-authored two books in the field of optical communications and computers. Dr. Stigliani is currently a Senior Technical Staff Member in the S/390 System Design organization, responsible for future processor data communications infrastructure and fiber optic applications.

Walter L. Mostowy *IBM Global Services, 1551 South Washington Avenue, Piscataway, New Jersey 08854 (mostowy@us.ibm.com).*

Mark E. Lewis *IBM Global Services, 522 South Road, Poughkeepsie, New York 12601 (melewis@us.ibm.com).*

David B. Petersen *IBM System/390 Division, 800 North Frederick Avenue, Gaithersburg, Maryland 20879 (petersen@us.ibm.com).*

Noshir R. Dhondy *IBM System/390 Division, 522 South Road, Poughkeepsie, New York 12601 (dhondy@us.ibm.com).*
Mr. Dhondy is an Advisory Engineer in the S/390 Parallel Sysplex Development organization in Poughkeepsie, New York. He joined IBM Kingston in 1968 as an Associate Engineer in the Electromagnetic Compatibility Department. He has been involved with various aspects of hardware design and was the lead designer of the 9037 Sysplex Timer; he continues to work on external time reference-related development for the S/390 Parallel Sysplex. Mr. Dhondy has published three invention disclosures and has received an IBM Outstanding Technical Achievement Award and a divisional president's award. Mr. Dhondy received his B.Tech. degree in electrical engineering from the Indian Institute of Technology, Bombay, India, and his M.S. degree in electrical engineering from the University of Pittsburgh.