

CERN-CN 91-8  
C1 (9131)

CERN LIBRARIES, GENEVA



CM-P00065711

CERN-COMPUTING & NETWORKS DIVISION

CN/91/8

July 1991

# The CERN Multivendor FDDI Project

J. ANTHONIOZ-BLANC  
J.M. JOOSTEN  
J. ROCHEZ

(Presented at the International FDDI Congress, Düsseldorf, 24-26 September 1991)

# THE CERN MULTIVENDOR FDDI PROJECT.

J. Anthonioz-Blanc

J.M. Joosten

J. Rochez

CERN, European Laboratory for Particle Physics  
CH-1211 Geneva 23

LAN traffic on the CERN site, which spans many kilometres, had as predicted reached in 1989 a level close to the capacity of Ethernet. CERN chose FDDI as its future backbone network and as the solution for high-performance workstations several years ago, and since mid-1989 has been gaining experience with various FDDI products and prototypes. At the time of writing, an FDDI backbone for Ethernets is already operational for more than a year. The penetration of FDDI in the office has for various reasons not happened yet.

## 1. *INTRODUCTION*

CERN, the European Laboratory for Particle Physics, has its seat in Geneva, Switzerland. The site extends over the Franco-Swiss border. CERN studies the basic subnuclear particles and forces of matter. It operates a complex of accelerators: a 28 GeV proton synchrotron (PS) and a 450 GeV super proton synchrotron (SPS) which can also be operated as a proton/antiproton collider at up to 900 GeV. The biggest accelerator in the world, the 2x51 GeV electron/positron collider (LEP) was officially inaugurated in November 1989 and has already produced important physics results. LEP will subsequently be upgraded to 2x100 GeV.

The research programme at CERN involves the use of large particle detectors and extensive data processing facilities. CERN is funded by sixteen European countries and has a staff of about 3000. In addition, the laboratory is used by about 5000 scientists from universities and laboratories from all over the world, but mainly from member states for their research projects.

## 2. *THE CERN GENERAL PURPOSE NETWORK ENVIRONMENT*

In order to understand the context of our FDDI project, it is useful to know something about the CERN site and the general purpose network environment, which is based on Ethernets and MAC-level bridges. Although many routers and gateways to other networks (IP, Apollo Domain, DECnet, AppleTalk) are connected to it, they are not mentioned in this context.

CERN has two main sites where offices and laboratories are concentrated. These two sites are about 5 km apart. There are also a number of sites in the surrounding country side which are access points for experiments buried deep underground at interaction points of particle accelerators. The largest machine is LEP, a ring with a diameter of 8.5 km.

## The CERN multivendor FDDI project.

The majority of the offices and laboratories (including the computer centre) are located on one of the main sites. From the computer centre (the primary star point) fibre optic trunks are fanning out to the different locations. These trunks were initially used for Ethernet and TDM (G703), but are now also used for FDDI. They consist of 50 micron and single mode fibres and typical distances here are 1 km. A secondary star point connects all the other relevant areas together and thus to the primary star point. Here the longest single span is about 7 km. The longest distance to cover is about 15 km. Obviously, this is not the perfect topology for a dual FDDI ring.

Figure 1 represents a simplified picture of our network before the introduction of FDDI: all the Ethernet segments were interconnected via MAC-level bridges to an Ethernet backbone (HRS) in the computer centre. The map shows also some topological information. The numbers indicate distances in kilometers. Where the distances permit it we use local bridges, else split-bridges are used over 2 Mbit/s TDM links. All distances less than 2 km can be easily dealt with by the FDDI standard. For the longer distances we had to find cost effective solutions.

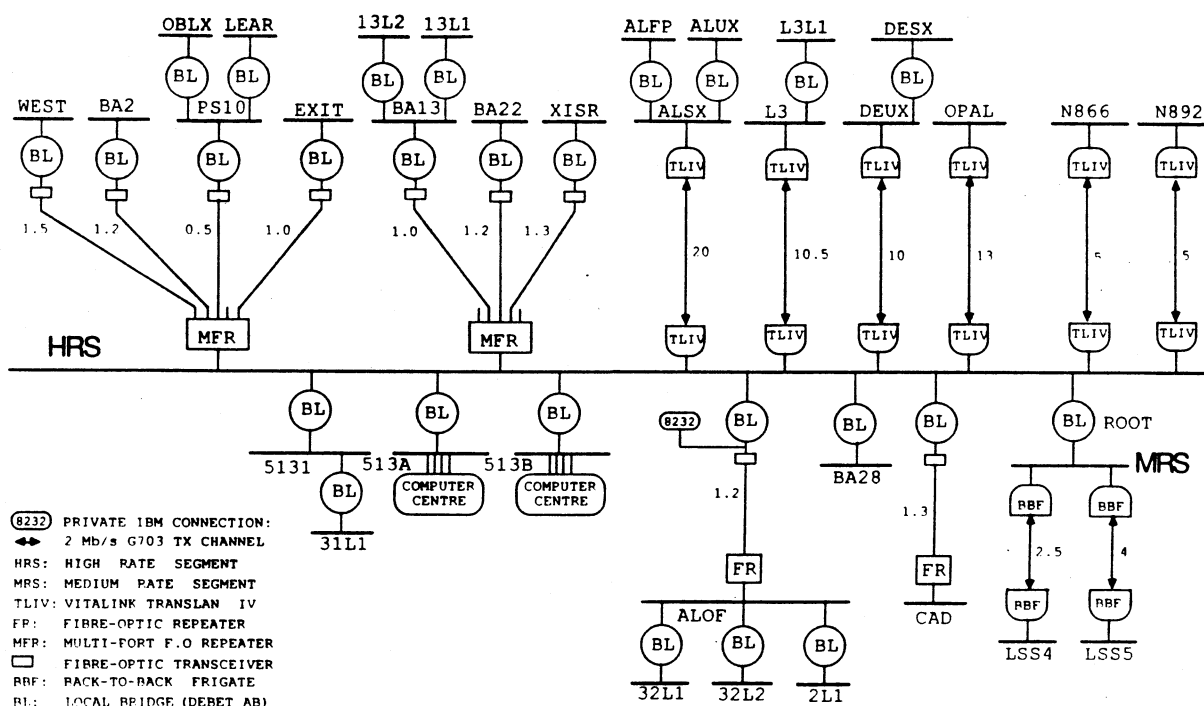


FIGURE 1: THE BRIDGED NETWORK

There are still home-made split-bridges in this network (Back-to-Back Frigates), from the time that commercial equipment was not yet available. These bridges have features that are still missing in respectable products (like unknown destination throttle).

What is not shown is that each segment is again subdivided with multiport repeaters (about 120 in total) and fan-outs. Routers and Gateways to other networks are also omitted from this picture.

### 3. *NETWORK EVOLUTION*

In the two years preceding the introduction of the FDDI backbone in April 1990, busy-hour traffic on the Ethernet backbone has grown from around 4% to 30% of the nominal 10 Mbit/s of Ethernet (with half-minute peaks of over 60%). The highest levels have been observed during data-taking by the LEP experiments. Over the same period, the number of Ethernet host connections had grown from about 500 to more than 2000, of which about 1500 were busy on any one day. Currently these numbers are about 3000 and 2300 respectively. These 3000 systems use some 35 different Ethernet manufacturer codes and 30 protocol types between them.

These data give a clear indication of our need for an FDDI backbone to replace the near-saturated Ethernet backbone. In addition, the interactive analysis of experimental physics data is increasingly carried out using scientific workstations with well in excess of 10 MIPS of computing power. The data source for these workstations is either a disc server for a cluster of stations, or the mass storage devices in the computer centre (attached to an IBM 3090/600J). Already, we see that the bandwidth required between these data sources and the workstations is exceeding the Ethernet capacity. Here again, FDDI is the evident solution.

### 4. *FDDI BRIDGING REQUIREMENTS AND RELATED PROBLEMS*

The problem we had to solve is how we could introduce an FDDI ring as a replacement of the Ethernet backbone, keeping multi-protocol transparency, while allowing traffic between hosts directly connected to FDDI (using TCP/IP according to RFC1188) and connectivity between Ethernet based hosts and hosts on FDDI (figure 2). The solution was seen in so-called translating MAC-level bridges between Ethernet and FDDI. Such bridges understand enough of Ethernet 2 and ISO LLC1 packet formats to translate them dynamically in certain cases (SNAP encapsulation). In particular, TCP/IP packets conforming to RFC 894 on Ethernet 2 must be translated to RFC 1188 format on FDDI, and vice versa. Translation is not required for LLC1 packets originating on Ethernet and Ethernet 2 packets bridged via FDDI back to Ethernet 2 must be re-translated (see figure 3).

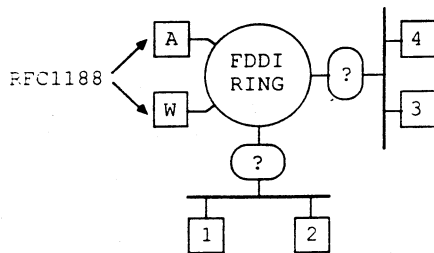
Using the (Organisationally Unique Identifier) OUI=0 to translate Ethernet packets, implied that that systems on Ethernet using TCP according to RFC 1042 would have to accept also the RFC 894 (Ethernet 2) format (the bridge translates these packets when going from FDDI to Ethernet). It was found out later that AppleTalk phase 2 did not adhere to this convention and was therefore not working over translating bridges. In fact AppleTalk phase 2 uses the zero OUI for its ARP, but refuses to accept the packet translated to Ethernet 2 format.

This problem has been solved by providing in every bridge a list of Type fields for which a packet with a SNAP, OUI=0, does not get translated to Ethernet 2 format when passing from FDDI to Ethernet. In order to ensure continued functioning of AppleTalk phase 1, its packets are passed over FDDI with a special SNAP encapsulation (OUI=F8). In fact, the idea to use lists of Type fields that are an exception to the rule, was proposed by CERN to the Internet Engineering Task Force and several major manufacturers in order to preserve the possibility to use RFC1042, but at the time it was considered not to be necessary.

The use of encapsulating bridges, although attractive for their simplicity, would not allow for a multivendor connectivity between hosts on FDDI and Ethernet.

The possibility to communicate transparently between an Ethernet based host and a host on FDDI posed an interesting problem with the representation of addresses in memory and in the TCP/IP Address Resolution Protocol (ARP). MAC addresses are physi-

## The CERN multivendor FDDI project.



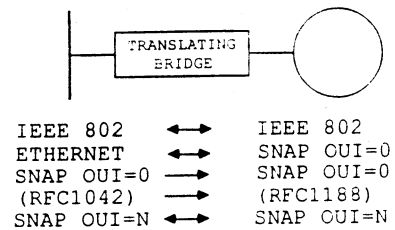
ETHERNET: -LITTLE ENDIAN

- ETHERNET 2 FRAMES (TYPE FIELD)
- IEEE802.3 FRAMES (LENGTH FIELD)
- MAX FRAME SIZE 1526 BYTES

FDDI:

- BIG ENDIAN
- IEEE FRAMES (NO LENGTH FIELD)
- MAX FRAME SIZE 9000 SYMBOLS

FIGURE 2: THE FRAMING PROBLEM



### ETHERNET PACKET FORMATS

ETHERNET: DST-SRC-TYPE-DATA-CRC

IEEE 802.3: DST-SRC-LENGTH-LLC-DATA-CRC

↓  
DSAP-SSAP-CONTROL

SNAP: DST-SRC-LENGTH-LLC-P.ID-DATA-CRC

↓ ↓  
AA-AA-03 OUI-TYPE

LLC Logical Link Control  
DSAP Destination Service Access Point  
SSAP Source Service Access Point  
SNAP Sub Network Address Protocol  
P.ID Protocol Identifier  
OUI Organizationally Unique Identifier

FIGURE 3: TRANSLATION

cally transmitted with different bit orderings on Ethernet and FDDI, and as a result tend to be stored differently in memory. It has been necessary to specify that a canonical bit ordering be used in the data field of ARP packets.

Another problem that was encountered later related to dropped packets due to the mismatch between the length in the data field and the actual length of the real packet. This is for instance possible with IEEE 802.3 packets and it was encountered with the Shiva (Kinetics) Fastpath, which padded an extra byte to make an even number. A similar problem occurred with IP packets on FDDI, where somewhere in the data there is a field giving the "Total Length" of the datagram. This field is used by a bridge when an FDDI IP packet has to be fragmented to fit into two or more Ethernet frames. After lengthy discussions it was decided that in both cases the Length value in the datagram would be taken as the reference and the actual packet shortened to fit it. If the actual packet is too short to do this, it would be dropped.

## 5. THE MULTIVENDOR FDDI TEST SCENARIO

In order to introduce multivendor products at CERN, a reference ring has been set up based on the AMD FASTcard interfaced to a PC-AT. This interoperability laboratory should of course always follow the latest definition of the standard. The Pdmo package running on these systems proved to be very useful. However, Ping and Telnet are unfortunately not available up till now.

Following this two HP/Apollo DN10000's were connected and coexistence with the AMD FASTcards was verified. Then TCP/IP was tested between the two Apollo's. Next two IBM 3090 prototype connections were introduced and TCP/IP between IBM and Apollo proved to be working straight away. Furthermore a SUN has been successfully connected. All these systems were running SMT 5.1. However, it should be noted that we

## The CERN multivendor FDDI project.

found that not all systems complied rigorously with the standard, but it was sufficient for operation under normal conditions.

Apart from this, the Ethernet back-bone has been partially replaced by another ring with two concentrators and 11 translating bridges on test from DEC. A powerful management system called ELMS (Extended Lan Management Software) formed part of the field test. This ring is operational since April 1990 and would initially function as a pure back-bone for the segments on the main site, having the right distances for FDDI. The ring had a total length of 11 km and was entirely based on 50 micron fibres. The idea was that when both rings gave full satisfaction, they would be merged. In practice we found that a gradual move from the test ring to the production ring was more realistic. In fact, before being connected to the production ring, systems will have to operate successfully for some time on a "pre-production ring". Currently only one IBM (prototype) main-frame and an HP/Apollo file server connection has been moved to the production ring, which runs a mixture of SMT 5.1 and 6.2.

The certification of the 50 micron fibres gave some concern since 5 dB was deducted from the 11 dB power budget, leaving a relatively small margin. However, recent tests revealed that by connecting 50 micron fibres directly to the optical transmitter and receiver, we were able to span 7 km and still have a comfortable margin. This can be explained by the superior bandwidth of 50 micron fibre (1-1.5 GHz) and the fact that a LED outputs a lot of power in the low order modes. The 5dB safety margin can be considered too conservative in this particular case. Furthermore the electro-optical components proved to be performing better than required by the standard.

Currently the production ring spans 60 km and is extremely reliable: there are essentially no unplanned ring initialisations. LEM values of 15 are common, even on the longer fibres. See figure 4 for some details.

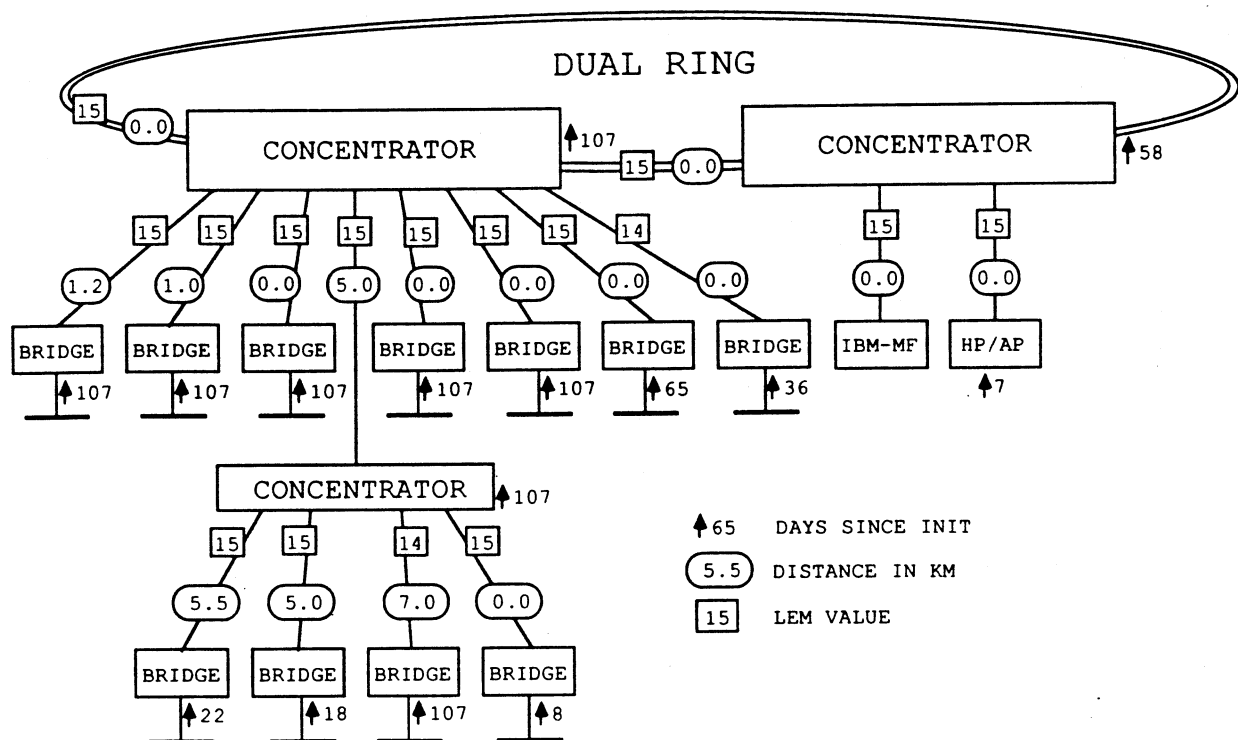


FIGURE 4: PRODUCTION RING

## 6. *BROADCAST STORMS*

The replacement of an Ethernet segment by an FDDI ring as a backbone had a side effect on our scheme of quenching broadcast storms. We made a device that recognises a storm on Ethernet and starts to collide on the source address of an offending system. One such a device on the Ethernet backbone is sufficient, if one accepts that the bridge and the segment with the source of the storm get congested. On FDDI this cannot be done, hence the necessity to protect every FDDI bridge with a quencher. It would be of interest if such a device would be part of a bridge.

Of course, a broadcast storm originating on FDDI will still be a big problem. This is one of the reasons why we think that every host on FDDI should be connected via a remotely manageable concentrator capable of disconnecting a misbehaving system. It would even be better if a concentrator could find out in real time about such a problem and take immediate action, while informing the network management station as well.

The use of SNMP (Simple Network Management Protocol) to manage concentrators can unfortunately lead to a situation where control over the network is lost during a broadcast storm. The reason is that SNMP is using IP and hence an SNMP managed concentrator will have to look at every broadcast on the network. The problem is that during a broadcast storm, the concentrator CPU can get overloaded such that it does not any more execute the command from the management station to disable a specific port (if it still has a buffer to receive it at all). It is very important that the concentrator hardware and software gets conceived in such a way that the predicted scenario is not materialising.

## 7. *FURTHER FDDI TEST ACTIVITIES*

In 4Q90 two DEC DS5000 workstations have been successfully introduced in our test ring. Figure 5 represents a typical configuration of this ring, and shows the test and the analysis tools we were using. From components readily available on the market, some home-made hardware and a fast logic state analyser, we made a line level monitor with trigger capabilities (Linelyser). With this device we were able to pin down early problems with the address representation and the Hardware Type bit in ARP packets.

Furthermore, a feature of a bridge allowed us to forward local FDDI packets onto an Ethernet with a classic Ethernet monitor. It proved to be invaluable in tracing problems and understanding the results of throughput measurements, protocol behaviour etc.

The test ring was also connected via a bridge to the CERN network, providing connectivity and injecting the CERN background multicast and broadcast traffic.

Last, but not least, yet another bridge was used to inject into the ring test traffic generated on Ethernet with our existing Ethernet test tools.

Currently there are 15 different FDDI devices from 9 different manufacturers that have been connected to our test rings. Wherever applicable performance measurements were done with a programme that could send and receive large blocks of data using TCP. Several parallel streams could be run concurrently. The results of these measurements were more or less in agreement with a well known "law" that states that a system can do about one Mbit/s I/O per MIP.

Below we list some of our findings in this multi-vendor environment.

## The CERN multivendor FDDI project.

- There is quite some confusion about the big-endian vs canonical representation of MAC-addresses, OUI's etc.
- In order to be able to communicate via bridges with Ethernet based systems the "Hardware Type" value in an ARP packet had to be set to "1" (the value for Ethernet) and not to "6" (specified for an 802 network). This caused initially some misunderstandings.
- Adding padding bytes to 802.3 and FDDI packets caused problems with bridges that dropped such packets.
- The use of OUI=0 by AppleTalk phase 2, forced the bridge manufacturers to provide exception Type lists.
- SIF operations produced faulty values for counters. There is also confusion about the interpretation of the SMT specification (e.g. the "supported version" field).
- SIF configuration information was also a source of inconsistencies
- A system with a dual MAC caused under certain WRAP conditions a MAC to disappear from the ring resulting in IP inaccessibility (the ARP tables were not updated). To correct for this problem both MAC's should be showing up in series on the ring.
- Memory to memory TCP transmissions between systems of two different manufacturers, showed an up till now not understood performance asymmetry: In one direction we obtained 2 Mbyte/s while in the other direction we only obtained 90 Kbyte/s. It could be due too bad negotiation of window sizes, loss of packets causing time-outs or anything else.
- An active ringmap done by a system showed that a certain other system did not reply to the NIF broadcast request with FC .NE. NSA, if the 'A' bit was already set by a preceding system.
- The performance of a dual access station dropped to very low values when the PHY A cable was disconnected or when connected as a slave to the master port of a concentrator.
- We found a system having an erratic behaviour which caused bursts of ring initialisations and beacons, while its status display did not show any suspicious behaviour. The problem on FDDI is that there is no easy mechanism to know who started an ring initialisation.
- Systems can be under power, while not in the ring because the interface has not been started up. This is one of the good reasons to have all systems on a ring connected as slaves to concentrators.

## 8. *THE CAMPUS-WIDE BACKBONE RING*

The LEP pits are at distances precluding a "standard" FDDI connection. As described earlier we managed to connect FDDI reliably over these distances of 5/7 km, but it is considered too risky for a permanent solution. Fortunately, single mode solutions are becoming available and CERN has been field testing successfully DEC single mode equipment in concentrators over a "bad" (many patches) fibre spanning 10 km.



## The CERN multivendor FDDI project.

Currently our production ring spans some 60 km, but we consider making such large rings a bad engineering practice, especially if many systems are going to be connected to it. Partitioning into several smaller rings is required.

It is furthermore not our policy to use the long distance fibres for special applications: in principle TDM should be used. Split bridges interconnected with full duplex 34 Mbit/s TDM links are probably fast enough, save fibre and allow for proper partitioning of rings in geographically distant areas. However, it is understood that the dedicated use of fibres on our site is also cost effective because of the relative ease of installation and the limited distances involved (5-15 km).

### 9. *INTRA-BUILDING FDDI*

In contrast to the successful introduction of FDDI as a backbone for the CERN Ethernets, FDDI has not yet any impact on the level of the individual user. There are several reasons for this:

- The need for higher speed is not generally there yet. The majority of the users only need lightly loaded Ethernets.
- FDDI interfaces are not always available for the already installed systems.
- Getting more than Ethernet performance is still a problem.
- The cost of FDDI interfaces is still too high for widespread utilisation.
- There is no infrastructure yet and installing it will be expensive.
- We have still some doubts about reliable multi-vendor rings.
- Partitioning of rings with bridges is still a problem.
- Good multivendor FDDI management systems are not readily available yet.

The effort by different vendors to make FDDI work over shielded twisted pair will result in cheap equipment that will profit from already installed cable plants. However, there is no twisted pair standard yet, although the major players in this field claim that they will interoperate and that their equipment will conform to the standard (they make the standard...). This and also the choice between 50 micron and 62.5 micron fibre with the possibility to have "low cost" fibre-optic components risks to give confusion and incompatibility. Furthermore, if a cable plant has to be installed, one should think about future higher speed networks, basically meaning a choice for fibre. Considering these point and the fact that CERN has essentially no IBM Type 1 cable plants, our current policy is to use 62.5 fibres for our infrastructure.

We believe also that the last word has not yet been said about the office outlets for fibres. To consistently equip buildings is currently expensive when using the standard MIC connector. There are interesting low cost alternatives, having the drawback of not being part of the FDDI standard.

Nevertheless, there will be shortly several FDDI rings on the site to solve problems at certain hot spots. These are mainly clusters of workstations and file servers used by the physicists for data analysis.

## 10. *CONCENTRATORS*

Concentrators are a fundamental part of the Token Ring concept. Hence the issue of concentrators is not unique to FDDI. The merits of concentrators have been well documented and our current policy is that only concentrators will be connected to the dual ring. These concentrators will reside in protected places, shielded from unauthorised intervention. End users will be slaves on concentrators that are connected in a tree structure to these protected root concentrators. With this approach we hope to reap the following benefits:

- Transparency in the optical flux budget: no bypass switches, which we consider a hazard, especially in combination with patch panels.
- Reliability: the chance of partitioning a ring by systems that are powered off or that have not started up their FDDI interfaces is reduced to zero. Bypass switches only reduce the chance of partitioning, while complicating the task of the network implementor.
- Easy extension by just adding another concentrator in a tree.
- Manageability: a misbehaving system can be eliminated from the ring with a simple command from the management system. This is the only way to stop a broadcast storm on FDDI, apart from physically interfering with the offending system

Concentrators must of course be very reliable, remotely manageable and cheap.

## 11. *CONCLUSIONS*

FDDI as a backbone for a large multi-vendor, multi-protocol Ethernet based network has been in operation for more than one year, with excellent results. General acceptance of FDDI at the user level has not materialised yet, although single-vendor rings are being introduced for some specific user groups.

Multivendor FDDI will be shortly part of the scene, thanks to the effort of the interoperability laboratories, the test agreements between different vendors and the work of several institutes like CERN, providing sophisticated test environments. However, to be a success, an integrated network management system for such an environment is required.

Working and affordable FDDI interfaces are not sufficient, one needs also an infrastructure. Shielded Twisted Pair (STP) FDDI provides an economic solution, but has also drawbacks. In principle we will equip our buildings with 62.5 micron fibres