The CERN FDDI Pilot Project

B.E. Carpenter, J.M. Joosten

(Presented at the Joint Networking Conference, Killarney, May 1990)
THE CERN FDDI PILOT PROJECT

Brian E. Carpenter
Johannes M. Joosten

CERN, European Laboratory for Particle Physics
CH-1211 Geneva 23
Phone + 41 22 767 - 6111
Fax + 41 22 767 - 7155
Email brian@priam.cern.ch, jmj@priam.cern.ch

LAN traffic on the CERN site, which spans many kilometres, has now as predicted reached 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, a substantial proportion of backbone traffic is already flowing through a pilot FDDI installation.

Keywords: FDDI

Brian E. Carpenter has been Group Leader of the Communications Systems group at CERN since 1985, following ten years' experience in software for process control systems at CERN, which was interrupted by three years teaching undergraduate computer science at Massey University in New Zealand. He holds a first degree in physics and a Ph.D in computer science, and is an M.I.E.E.

Johannes M. Joosten has led the Communications Hardware team in the CERN Communications Systems group since its formation. Prior to that he was in charge of the hardware design of CERNET and OMNET, networks implemented at CERN in the 1970's, and of various special-purpose processors for real-time physics data filtering. He holds an engineering degree in applied physics.

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 fourteen European countries (a fifteenth, Finland, is about to join) 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.
The CERN FDDI Pilot Project

The requirement for FDDI

Before talking about our FDDI project, it is useful to understand something about the CERN site and the general purpose network environment.

In fact, CERN has two main sites a few kilometres apart, as well as a number of small sites in the surrounding countryside which are access points for experiments buried deep underground. The largest machine is LEP, whose circular tunnel has a diameter of 8.5 km. The machine has four interaction regions (numbered PA2, 4, 6 and 8), where large experiments are taking place. A fibre infrastructure connects all the relevant areas together. Most of the long-distance fibres are reserved for G.703-compatible TDM channels.

The majority of the offices and laboratories (including the computer centre) are located on one of the main sites. In addition to older copper cables, there are several fibre optic trunks currently being used for native Ethernet and for TDM. The trunks consist of both 50 micron and mono-mode fibres. This infrastructure is now also used for FDDI. Typical distances here are 1 km.

The CERN standard LAN is Ethernet, currently consisting of about 35 individual Ethernets split into about 700 individual cable segments. The 35 Ethernets are linked using MAC-level bridges, until recently by connecting all of them (directly or indirectly) to a common backbone Ethernet known as the High Rate Segment (HRS). Our network management system automatically generates a real time map of the Ethernet bridge topology, giving information on traffic, up/down status, etc. Because of excessive detail, such a map cannot be given here, so a simplified topology is shown in Figure 1. The distances between Ethernets vary from zero up to 20 km; 8 links exceed 2 km. All distances less than 2 km can be easily dealt with by FDDI. For the longer distances (e.g. LEP experimental areas) solutions have to be found.

Over the last two years, busy-hour traffic on the HRS backbone has grown from around 4% to almost 30% of the nominal 10 Mbit/s of Ethernet (with half-minute peaks of 60%). The highest levels have been observed during data-taking by the LEP experiments. Over the same period, the number of Ethernet host connections has grown from about 500 to more than 2000, of which about 1500 are busy on any one day. These 2000 systems use at least 26 different Ethernet manufacturer codes between them. All of the quoted numbers, of course, are continuing to increase.

These data give a clear indication of our need for an FDDI backbone to replace the near-saturated HRS. 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 3090/600E). We project that the bandwidth required between these data sources and the workstations will exceed Ethernet capacity during 1990. Here again, FDDI is the evident solution.

Protocol aspects

The problem we have to solve is how we can 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 RFC 1103), and connectivity between Ethernet based hosts and hosts on FDDI (see Figure 2). The solution is seen in so-called translating MAC-level bridges between Ethernet and FDDI. Such bridges understand enough of Ethernet 2.0 and ISO LLC1 packet formats to translate them dynamically in certain cases. In particular, TCP/IP packets conforming to RFC 894 on Ethernet 2.0 must be translated to RFC 1103 format on FDDI, and vice versa. Translation is not required for LLC1 packets originating on Ethernet, and Ethernet 2.0 packets bridged via FDDI back to Ethernet 2.0 must be re-translated (see Figure 3). As a result of inconsistencies in
approach between various companies, we were forced to learn much more than we wanted to know about the fine details of these packet formats, and to intervene with several companies and with the Internet and IEEE standardisation committees. The use of encapsulating bridges would not require such gymnastics, but would not allow for full 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 physically 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.

We were in doubt for some time as to the best approach to the lack of final standardisation of Station Management (SMT). Finally, essentially everybody agreed to take SMT 5.1 as the de facto standard.

**FDDI activities**

In order to introduce multivendor products at CERN, a reference ring was first set up, based on AMD FASTcards interfaced to PC-ATs. This "interoperability laboratory" should of course always follow the latest definition of the standard, as far as practicable.

Following this, first two Apollo DN10000 workstations were connected and their safe coexistence with the AMD FASTcards was verified. Then TCP/IP was tested between the two Apollos (giving disc-to-disc file transfer rates around 1 Mbyte/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 recently, in the same way as the Apollos. However, it should be noted that we found that not all systems complied rigorously with the standard.

Apart from this, the Ethernet back-bone has been replaced by a second ring with concentrators and translating bridges on test from DEC. A powerful management system forms part of the tested equipment. This ring will initially function as a pure back-bone for the segments on the main site, where distances are less than 2 km. The ring has a total length of 10 km and is entirely based on 50 micron fibres. The certification of these fibres gave some concern since 5 dB had to be deducted from the 11 dB power budget, leaving a relatively small margin. When both rings give entire satisfaction, they will be merged.

The LEP pits are at distances precluding a normal FDDI connection. Although equipment will be available to span long distances over mono-mode fibres, there are reasons why we do not want to go in that direction: a single ring calling at all LEP experiments and the computer centre would be 61 km around (i.e. 300 µs round-trip). This delay is going to count and might affect the efficiency of certain protocols. Making such large rings is also considered as bad engineering practice. It is furthermore not our policy to use the long distance fibres for special applications: in principle TDM should be used.

There are various solutions in view. The idea would be to keep small rings, locally interconnected with MAC-level bridges, and bring the remote traffic to them via split bridges (or brouters) and TDM (see Figure 4). No such products exist yet, but work is already being done in some places to provide solutions in this spirit.
The CERN FDDI Pilot Project

Concluding remarks

- AMD, Apollo, IBM and Sun equipment inter-operate on a dual ring, using SMT 5.1 (with minor problems).
- TCP/IP according to RFC 1103 works well between Apollo, IBM, and Sun.
- An Ethernet backbone has been smoothly replaced by DEC FDDI concentrators and translating bridges.
- TCP/IP between an Ethernet host and an FDDI host works correctly via a translating bridge.

We have learnt that

- Interoperability does not necessarily mean compliance with the standard.
- A dual ring should be kept small, with only concentrators on it.
- Concentrators must be cheap, reliable, remotely managed.
- 50 micron fibres can be used, but the power budget is small and patch cables eat decibels.
- Fibre installation is not trivial or cheap. FDDI is not yet "plug and play," although the technology is moving.
- FDDI over RG58 coaxial cable to the workplace is probably necessary for FDDI to become economical except as a backbone.
FIGURE 2: THE FRAMING PROBLEM
TRANSLATION

TRANSLATING BRIDGE

IEEE 802 ↔ IEEE 802
ETHERNET ↔ SNAP OUI=0
SNAP OUI=0 ↔ SNAP OUI=0
(RFC1042) ↔ (RFC1103)
SNAP OUI=N ↔ SNAP OUI=N

ETHERNET PACKET FORMATS

ETHERNET: DST-SRC-TYPE-DATA-CRC

IEEE 802: DST-SRC-LENGTH-LLC-DATA-CRC

DSAP-SSAP-CONTROL

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

AA-AA-C3 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 Organizationnally Unique Identifier

FIGURE 3: TRANSLATION (iii)
FIGURE 4: OPTIONS AND ALTERNATIVES