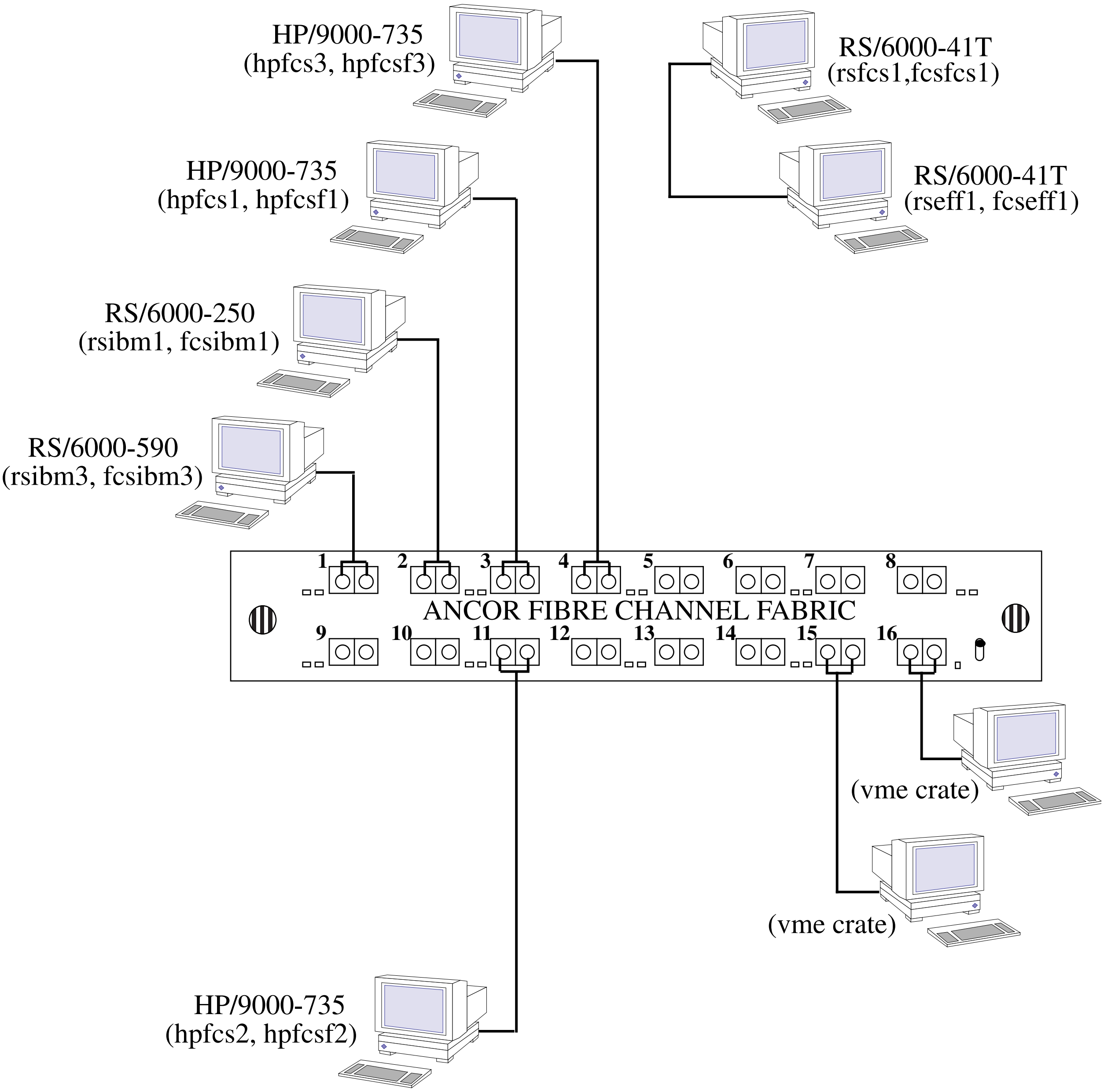
March 30 1995

FIBRE CHANNEL EQUIPMENT UNDER TEST

Fabrice Chantemargue, Susana Munoz, Gaetano Maron
(chante@sunrans.cern.ch, munoz@sunrans.cern.ch, maron@lnl.infn.it)

I. Hardware Configuration

The figure just below displays the hardware configuration that is presently under test.



Most of the Fibre Channel equipment is still implementing version 3.0 of Fibre channel, excepted the two interfaces that are located into the RS/6000-41T workstations, which are presently implementing version 4.2. That is the reason why these two interfaces are not connected to the Fibre

Channel Fabric, but rather back to back (there is indeed no backward compatibility).
We expect to move from version 3.0 to version 4.2 by end of march 1995.

Each VME crate is composed of a CPU board, a memory board, an Ancor VME 64 Fibre Channel interface and is running under the VxWorks Operating System. As no driver for the Fibre Channel interface was provided by Ancor, one has been evolved by Marco Pellato (INFN).

II. Workstations' features

Table 1 briefly summarizes the main characteristics of the workstations that are on use.

Table 1: Workstations' main features

System	CPU type	Clock (Mhz)	Cache (Ext+I/D)	SPEC (92Int)	SPEC (92FP)
RS/6000-250	MPC601	66	32	62.6	72.2
RS/6000-590	Power2	66.6	32/256	121.6	259.7
RS/6000-41T	MPC601	80	32	78.8	90.4
HP/9000-735	PA7100	99	256/256	109.1	167.9
VME crate	M60030	20			

SPECint92 is derived from the results of a set of integer benchmarks, and can be used to estimate a machine's single-tasking performance on integer code. SPECfp92 is derived from the results of a set of floating point benchmarks, and can be used to estimate a machine's single-tasking performance on floating point code.

III. Communication Benchmarks

All measurements, whose aim is to determine overheads, communication latencies and throughputs at the user level, have been achieved on the sender side. We defined overheads, communication latencies and throughputs as follows:

- Overhead: it is defined as the time from when an application sends a message to the time when it is ready to send another message. In practice, it corresponds to the measurement of the time taken for performing the write of a one byte message (the smallest message that can be written).

- Communication latency: it is defined as the time from when an application (on the sender side) sends a message to the time when another application (on the receiver side) receives this message. In practice, we measure a round trip communication latency for a one byte message and then divide it by two.

- Throughput (or transmission rate): it is defined as the amount of data that can be sent within one second.

The program used to test out communication performances, when using *TCP/IP* protocols, was *Netperf* (version 1.7.1). On all machines, the *rfc1323* mode was turned on by different means according to the workstations: the *no* command for IBM workstations, *adb* for HP workstations. Moreover the *PHNE_4854* patch (only for HP machines) had been installed. For all tests, the *TCP_no_delay* option on the sender side was set (*-D* option): this option actually provides the best performances regarding the throughput on ANCOR interfaces and is necessary for measuring relevant overheads and relevant communication latencies with all interfaces. With IBM interfaces, the best performance regarding the throughput is obtained when not using this option (the reason why, is still an open question).

For light-weight protocols, no program is available from the public domain; therefore we evolved and ran our own application.

A. With ANCOR FC interfaces

1. Using TCP/IP protocols

On all machines, overheads and communication latencies have been measured when a socket buffer size of 1 Kbytes was set.
For communications between IBM machines, the throughput was measured when setting the socket buffer size to 1 Mbytes. For communications between HP machines (and therefore between HP and IBM machines), it was measured with a socket buffer size of 349525 bytes: this number corresponds indeed to the maximum possible selection on the HP system, according to the present kernel configuration; this is of course a severe limitation in performance.

a) Between IBM workstations (back to back)

(1) Overhead

RS/6000-250 - RS/6000-590: **610** usec.
RS/6000-590 - RS/6000-250: **365** usec.

(2) Communication latency

RS/6000-250 - RS/6000-590: **1025** usec.
RS/6000-590 - RS/6000-250: **1025** usec

(3) Throughput

RS/6000-250 - RS/6000-590: **10.3** Mbytes/s.
RS/6000-590 - RS/6000-250: **12.1** Mbytes/s.

b) Between IBM workstations (with the Fabric inbetween)

(1) Overhead

RS/6000-250 - RS/6000-590: **610** usec.

RS/6000-590 - RS/6000-250: **365** usec.

(2) Communication latency

RS/6000-250 - RS/6000-590: **1104** usec.
RS/6000-590 - RS/6000-250: **1092** usec

(3) Throughput

RS/6000-250 - RS/6000-590: **10.3** Mbytes/s.
RS/6000-590 - RS/6000-250: **12.1** Mbytes/s.

(4) Fibre Channel Fabric Latency

Differentiating communication latencies measured with the Fabric and without the Fabric gives us an estimation of the Fabric Latency. We found experimentally two different values (according to the workstation on which the measurement was performed), which are close to each other:

79 usec (1104 - 1025) or **67** usec (1092 - 1025).

Ancor reported a Fabric latency of **64** microseconds.

c) Between HP workstations (with the Fabric inbetween)

(1) Overhead

HP/9000-735 - HP/9000-735: **890** usec.

(2) Communication latency

HP/9000-735 - HP/9000-735: usec.

(3) Throughput

HP/9000-735 - HP/9000-735: **10.4** Mbytes/s.

2. Using a light-weight protocol

a) Between IBM workstations (*Direct Channel*, with the Fabric inbetween)

(1) Overhead

RS/6000-250 - RS/6000-590: **640** usec.
RS/6000-590 - RS/6000-250: **376** usec.

(2) Throughput

RS/6000-250 - RS/6000-590: **11.9** Mbytes/s.
RS/6000-590 - RS/6000-250: **16.3** Mbytes/s.

b) Between HP workstations (*LLA*, with the Fabric inbetween)

(1) Overhead

HP/9000-735 - HP/9000-735: usec.

(2) Communication latency

HP/9000-735 - HP/9000-735: **216** usec.

(3) Throughput

HP/9000-735 - HP/9000-735: **12.5** Mbytes/s.

c) Between VME crates (*customized protocol*, with the Fabric inbetween)

Note that the data are moved in 32 bit wide.

(1) Overhead

VME crate - VME crate: **850** usec.

(2) Throughput

VME crate - VME crate: **6.9** Mbytes/s.

d) Between VME crate and HP workstation (*customized protocol* on VME side, *LLA* on HP side, with the Fabric inbetween)

Note that the data are moved in 32 bit wide.

(1) Overhead

VME crate - HP/9000-735: **800** usec.

(2) Throughput

VME crate - HP/9000-735: **5.95** Mbytes/s. (for 32750 byte message)

B. With IBM FC interfaces

1. Using *TCP/IP* protocols

a) Between IBM workstations (back to back)

(1) Overhead

RS/6000-41T - RS/6000-41T: **421** usec.

(2) Communication latency

RS/6000-41T - RS/6000-41T: **1162** usec.

(3) Throughput

RS/6000-41T - RS/6000-41T: **13.6** Mbytes/s.

2. Using a light-weight protocol (*Raw Socket*)

a) Between IBM workstations (back to back)

(1) Overhead

RS/6000-41T - RS/6000-41T:

(2) Throughput

RS/6000-41T - RS/6000-41T:

IV. Ethernet Medium

This paragraph has been added in this note just for pointing out that for transmitting small messages, the Ethernet medium was, for the time being, more suitable than the FC medium.
The measurements have been performed under the same conditions (using Netperf, 1 byte message, 1024 byte socket buffer).

1. Using *TCP/IP* protocols

a) Between IBM workstations

(1) Overhead

RS/6000-41T - RS/6000-41T: **230** usec.

(2) Communication latency

RS/6000-41T - RS/6000-41T: **715** usec.

V. Miscellaneous

There is a HP-IBM communication problem for the time being. The reason is that we need a new version of the Ancor firmware for HP interfaces. This new version may be ready by end of March 95.