[redacted]

CERN-COMPUTING & NETWORKS DIVISION
CN/91/6
July 1991

From Megabit to Gigabit:
Possible Transition Scenarios

François FLUCKIGER

(Published in Computer Networks and ISDN, Proceedings of the 2nd Joint European Networking Conference, Blois, May 1991)

ORGANISATION EUROPÉENNE POUR LA RECHERCHE NUCLÉAIRE
EUROPEAN ORGANIZATION FOR NUCLEAR RESEARCH

Laboratoire Européen pour la Physique des Particules
European Laboratory for Particle Physics

GENÈVE, SUISSE
GENEVA, SWITZERLAND

From Megabit to Gigabit: Possible Transition Scenarios (*)

Francois FLUCKIGER

CERN, CN Division, CH-1211 Geneva 23, Switzerland
tel: +41 22 767 49 89, tlx: 41900 CER CH, e.mail: fluckiger@vxcern.cern.ch

Abstract:

Over the last three years, segments of the European Academic and Research Community have introduced international megabit lines to support international collaborative research programmes, especially where intensive numerical analysis is required. How to satisfy the user requirements which include not only pure data services but also multi-media facilities and group or inter-personal communications based on voice and video? And how to pave the way for entering the gigabit era and to plan a smooth transition?

This paper aims to bind theory and practice, and is based on practical experience on megabit and hundred-megabit services and techniques gained within the European Particle Physics Community and by CERN. "Strategic technological components" for the transition are considered: Synchronous TDMs, CLNS Routers, Frame Relay Switches and ATM Switches, as well as their possible combination over definite periods of time called "phases". The paper concludes by suggesting two "scenarios" for transition, taking an engineering point of view, each scenario being composed of a succession of overlapping phases.

Keywords:

megabit, gigabit, CERN, TDM, delays, bit rates, IP Routers, Frame Relay, ATM, photonic switches.

(*) Published in Computer Networks and ISDN, Proceedings of the 2nd Joint European Networking Conference, Blois 1991

1. Brief summary of required services

To illustrate the requirements of international researchers working in a collaborative way, the example of the Particle Physics Community (also called HEP: High Energy Physics community) will be taken. The model of the community is as follows: accelerator or experimental centres, such as CERN, DESY or Gran Sasso in Europe, FNAL or SLAC in the USA offer experimental facilities. Physics institutes or physics departments at universities set up collaborations to realize an experiment at a given accelerator centre. Up to 80 such experiments may be carried out concurrently at big centres. Modern collaborations group 500 to 1000 people, spread over the world, over periods of 10 to 15 years, where they need facilities for distributed collaborative design and development, as well as distributed data analysis and simulations.

These required facilities fall into two categories: computer communications, and group or inter-personal communications. All research disciplines mainly organized on an international basis require today these two types of services. The requirements of HEP in computer communication are however acuter. For example, around LEP, the new giant accelerator at CERN, four major experiments have been set up. Each of them will, in 1991, produce 30000 tapes - i.e. 6 terabytes - which have to be processed at outside processing centres. To figure out what the 24 terabytes produced in one year by these four experiments represent, let us imagine a T1 (1.5 Mbps) line linking CERN and a remote processing centre. Assuming this T1 line is used permanently 24 hours a day / 7 days a week, at a sustained bit rate of one megabit per second, the transmission of these 24 terabytes would take no fewer than 7 years. Although this particular requirement may appear unique, all disciplines require now a network offering much higher bit rates and much shorter latencies.

Conventional group or inter-personal communications are also a must today: e-mail obviously, but also distributed document management, lists, news and bulletin boards facilities. However, non-data communication services are also vital for collaborative research: telephony, teleconferencing, videophony, videoconferencing; and, already required, multimedia station communications.

Such facilities are not available on an international basis. How to offer them today? How to evolve towards the provision, tomorrow, of a full range of multi-media services. The next sections address these two issues.

2. Where is A&R today with Megabit Networking?

2.1 Current Status

Considering the international infrastructure available today to European researchers, the answer to the above question is: "Not very far!". Although several countries have set up partial or general 2 Mbps A&R networks (e.g. Finland, Italy, Switzerland, UK), only 5 megabit international lines are in operation today, centred around CERN (CERN, International Treaty Organization located on the French-Swiss border, is considered here as a country): they connect to Lausanne and Geneva (CH), Bologna (IT), Saclay and Lyons (FR) and Cornell Uni. (US). Figure 1 shows all A&R international lines with a bandwidth at least equal to 256 Kbps.

A recent study [1] has provided figures for the total aggregate bandwidth available to each European country for communication with other countries, taking into account all A&R initiatives (such as IXI, EASInet, EARN, EUnet...) and disciplines (e.g. Space and Particle Physics research). Figure 2 visualizes the results: except the particular case of CERN and Switzerland, major countries dispose of about 2 Mbps aggregate bandwidth with the outside, spread over dozens of thin lines.

Hence, although the theme of this paper is transition from megabit to gigabit, the above suggests that:

The main concern of A&R international networking today is still transition from kilobit to megabit.

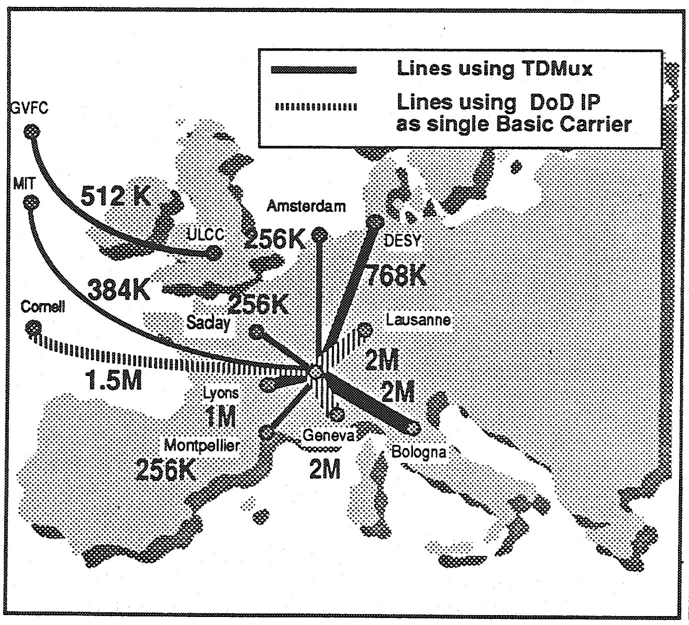


Fig 1: International A&R Leased Lines at speed ≥ 256Kbps

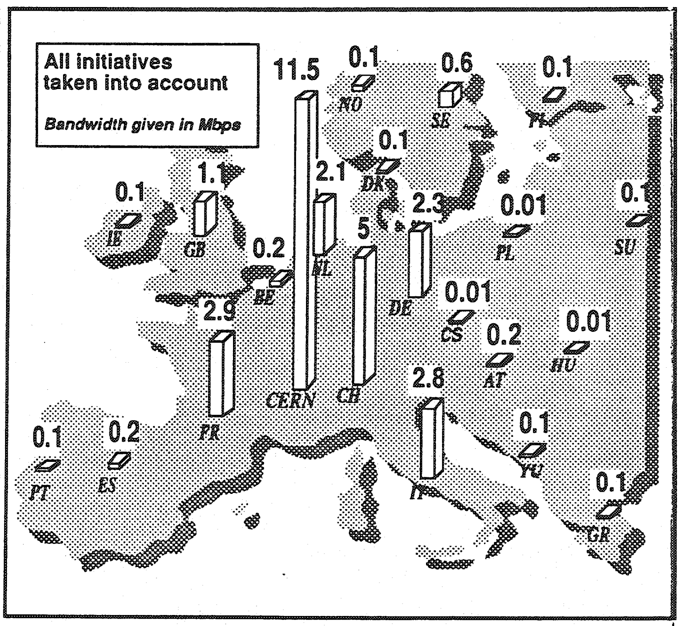


Fig 2: Total bandwidth per country for International A&R Networking

2.2 Current Technology

Figure 1 shows that only two bearer techniques are used over international lines with bandwidth at least equal to a quarter of megabit: Synchronous Time Division Multiplexing (TDM), dividing the bandwidth in independent sub-channels, and DoD Internet Protocol (IP).

The popularity of TDM technology is easy to understand: although TDM provides fix bandwidth allocation per service, the network managers benefit from all flexibility to adapt the bandwidth, as well as to set-up, when required, dedicated channels. Such channels may support specific applications, tests and trials of new technologies without disturbing the production, as well as audio and video. The technology is also well-proven, even somewhat boring, and affordable. TDM, as any multiplexing technique, consumes a few percent of the bandwidth and adds some delays, as explained later. The conclusion is that:

> *Synchronous TDM is the off-the-shelf technology of choice today for multi-service and multi-media networking.*

From an operational perspective, a side effect has been noted from several operational sites: TDMs offer powerful tools for physical line analysis, monitoring and trouble shooting.

The second bearer technology in use over these lines is DoD IP. Work is underway to support video and voice, but today international IP lines in Europe are devoted to data communication only. Whether used native or over TDMs, DoD IP is the main data communication protocol in use internationally in Europe. Figure 3 shows the distribution of the bandwidth staticly allocated to each bearer protocol, aggregated over all the leased lines connected to CERN (total bandwidth: 11 Mbps). Data were not available for the other international lines, but the distribution shown applies to 11 out of a total of 15 Mbps available today for international networking, all A&R initiatives being taken into account [1].

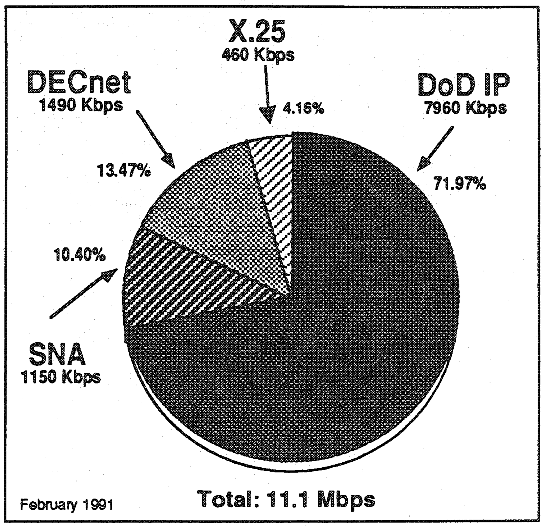


Fig 3: Static allocation of Bandwidth over the International Lines Ending at CERN

2.3 Practices on Delays

Greg Chesson said in his key-notes speech at the 2nd RARE Joint Conference *"the two main features of networks are bit-rates and latencies"*. This paper presents results on these two parameters based on CERN experience, and drawn from a production environment. Let us first discuss latencies.

Figure 4 shows results of round trip delays measured from CERN over DoD IP lines by means of "IP pings" between routers. Measures are made in normal working conditions over the production network. This calls for the following immediate comments:

1. The round trip delay to Bologna, 12 ms, demonstrates that (IP) packets can be delivered today in 6 ms over short continental distances with standard routers. With such performances, the propagation time represents 40% of the total delay.
2. The identical result, 20 ms, obtained to Amsterdam and Lyons (the latter geographically closer to CERN) indicates that the use of private TDMs may add delays, but more significantly, the suspected presence of numerous intermediate PTT equipment (probably TDMs) may completely alter the expected results.

The conclusions may be summarized as follows:

- 10 ms round trip delay over international lines is operationally achieved today
- Standard IP routers switch in less than **2 ms**
- Terminating TDMs may add **1 or 2 ms/TDM**
- Intermediate PTT TDMs may adds **dozens of ms**
- Finally: the propagation delay is no longer marginal:

> *Over continental distances, the technology dependant component of the delays will soon become minor.*

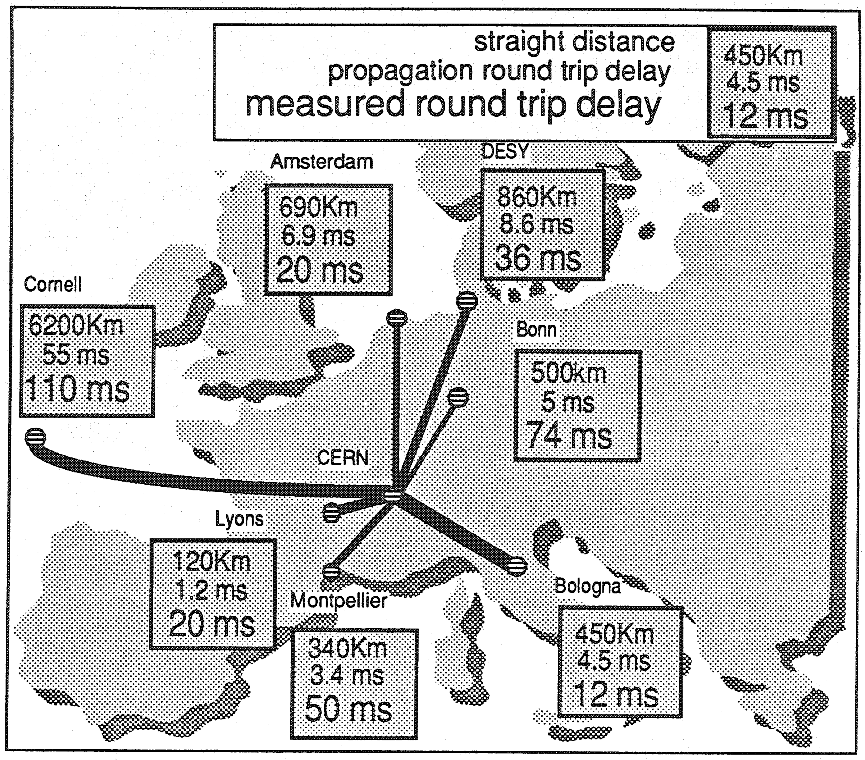


Fig 4: Measured "IP Ping" Round Trip Delay

The above discussed modest megabit bandwidths: let us tackle the 10/100 megabit range. Figure 5 gives round trip delays (still using "IP pings") measured on the one hand over the CERN production Ethernet (10 Mbps) and on the other hand over the CERN production FDDI (100 Mbps). In both cases, the Ethernet and the FDDI were acting as backbone LANs connecting other Ethernet segments through bridges. The result is that when the average round trip delay over the Ethernet is **1.5 ms**, that over the FDDI, in principle 10 times "faster", amounts to **1.8 ms**. For more information, see [2]. Such results are confirmed by measures conducted in Aachen [3], where the Ethernet round trip delay was found equal to the CERN one, but where the FDDI figure reached 11.2 ms (due to the use of a different type of bridges). Hence, in theory, FDDI is "faster" than Ethernet whilst in practice it is "slower". This simply shows that, as any qualifier, "fast" and "slow" need precise definition, and that:

High Bandwidth is different from Speed.

To conclude this discussion on latencies, let us simply observe the following: considering that over a 50 km LAN, the minimal round trip delay (propagation time) is 1/3 ms, considering that it amounts to 80 ms between Geneva and California, and considering that fast CPU may not afford to waste cycles waiting, one may argue that **conventional** Remote Procedure Call (RPC), as a tool for distributed processing, has a limited future (see also [4] in these proceedings).

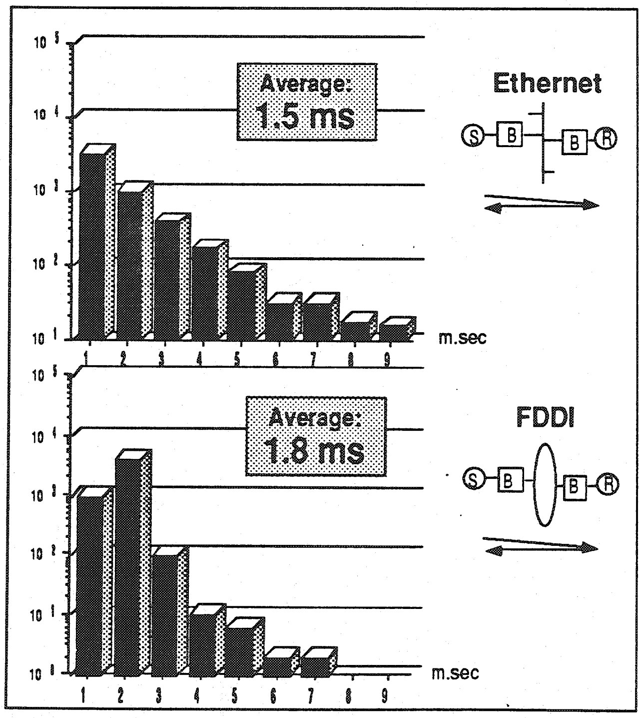


Fig 5: Return Trip Delay over Ethernet and FDDI measured at CERN

3. Transition Scenarios

3.1 Methodology

In order to reduce the number of options in discussing transition paths, the following methodology is adopted:

1. First, a small number of ***Strategic Technological Components*** is identified. They are mainly "technological objects" such as "an ATM switch" even though such objects are not necessarily implemented in this embedded form.
2. One or several Strategic Components may be used concurrently over a certain period of time called a ***Phase.***
3. Finally, a set of successive and overlaping phases is called a ***scenario.*** It constitutes a complete path from current megabit/mono-service situation to the gigabit/multi-service era.

The suggested phases and scenarios have been designed to apply to a wide area national or international A&R environment , and result from a planning and engineering approach. Other considerations, such as political constraints, although real, form another dimension of this complex problem and are not taken into account here.

3.2 Strategic Components for the Transition

Four components are retained as strategic for the transition. The paper will not discuss in detail their respective merits and features. More information can be found in [5] in these proceedings.

1. **Synchronous TDMs**
   as discussed above. They can operate at up to 155 Mbps.
2. **Connectionless Routers** (DoD IP or ISO IP)
3. **Frame Relay Switches**
   which can connect to packet routers as well as to X.25 switches.
4. **ATM Switches**
   This covers for simplification both switches and concentrators. They can connect to LANs via routers or bridges, and to MANs.

Many other objects or notions often refered to when discussing high bandwidth networking are purposely missing from this short list, because not relevant to the bit rates discussed, or not considered as strategic. This includes: **X.25** and Narrowband **ISDN** (because both narrowband), **DQDB** (because a MAN technology, which can be used but is not strategic for the transition), **SMDS** (because a PTO offer for MANs), **SDH** (because a transmission technology, which will be used but is not strategic) and **STM** (because CCITT has chosen ATM for B-ISDN).

3.3 Transition Phases

There are numerous possible combinations of the above components to form transition phases. The following restricts them to a reasonable number. The most notable phases are depicted by symbolic drawings (Figures 6.1, 6,2. 6.3) which are simplification of the reality, especially in the way in which the components have to be connected. The figures are based on planes, each representing a technological network.

1. **TDM Phase:** TDM + N-IP
   TDM is here the basic multi-service technology which supports various data services including the narrowband IP plane (N-IP: current technology, up to 2/8 Mbps), voice and video planes.

2. **B-IP Phase:** B-IP
   A new generation of broadband IP switches (B-IP, up to 34/45 Mbps), is used here. Whether they are DoD IP or ISO IP based is not strategic. The assumption is made that these protocols have been extended to offer various types of service and support the video and audio planes. Note that especially for the latter which requires delay equalization, these extensions seem to be a matter of research or at best experimentation.

3. **Frame Relay Phase:** FR + N-IP
   This is as Phase 1, but Frame Relay replaces TDMs. The commercial Frame Relay offer is currently limited to 2 Mbps. Voice support is for further study.

4. **N-ATM Phase:** TDM + N-IP + N-ATM
   Here, the flexibility of TDMs is used to provide in parallel to the production N-IP plane, a narrowband ATM plane (N-ATM). This N-ATM plane uses the 1st generation of ATM switches available by the end of the year, with bit rate possibly limited to 2 Mbps, and not all the classes of service being available. The interest of such a phase lies in the early creation of operational expertise with ATM technology, without disturbing the production.

5. **B-ATM Phase:** B-ATM + N-IP
   Here, the 2nd generation of ATM technology is available: 155 Mbps+, and all classes of services. Various sub-phases can be envisaged:

   - **LAN connected via Routers**: Here an IP plane is maintained.
   - **LAN connected via Bridges**: new concepts of LAN networking need then to have emerged, if IP packets are no longer exchanged over LANs. Far away.
   - **LAN connected via DQDB Mans**: if this technology gets improved and/or the carrier offer attracts customers, some LANs may connect in that way to the ATM plane. Will not form a universal access method.

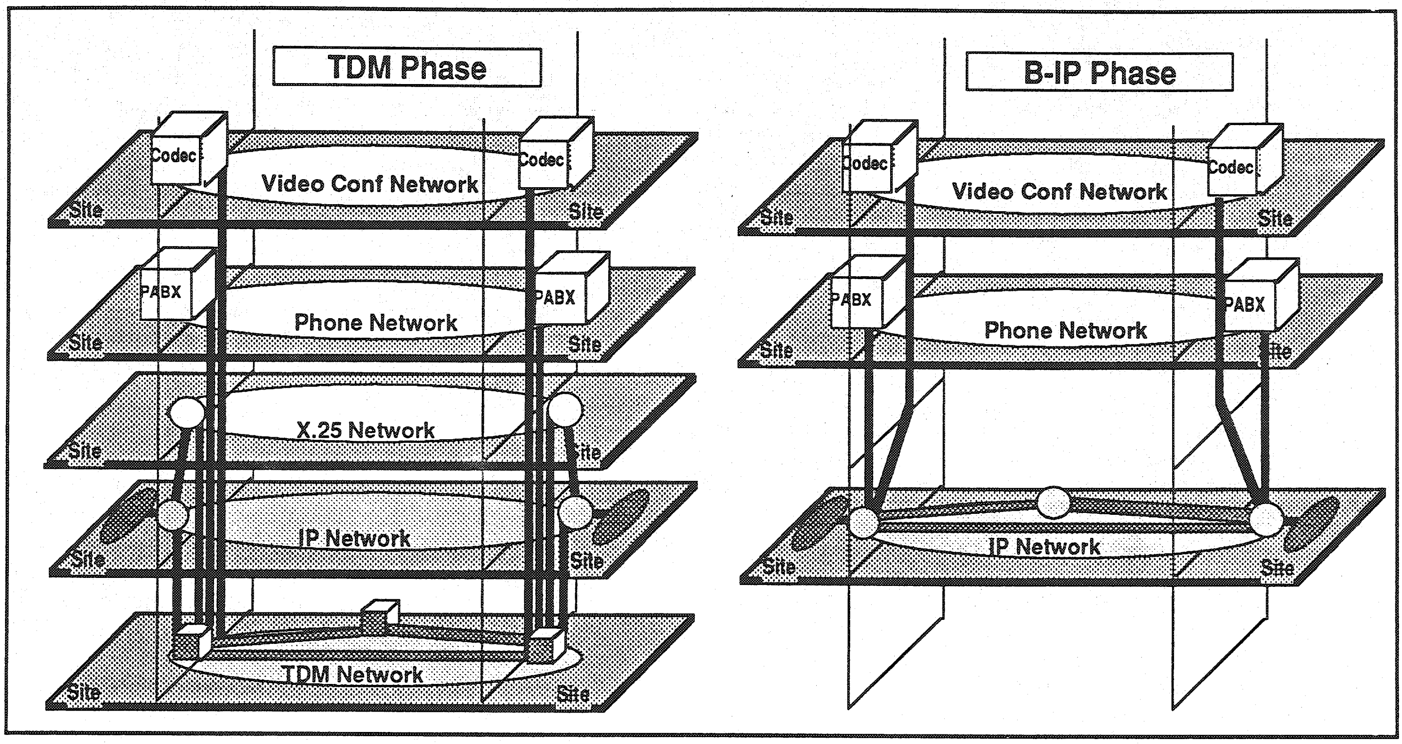


Fig 6.1: TDM and B-IP Phases

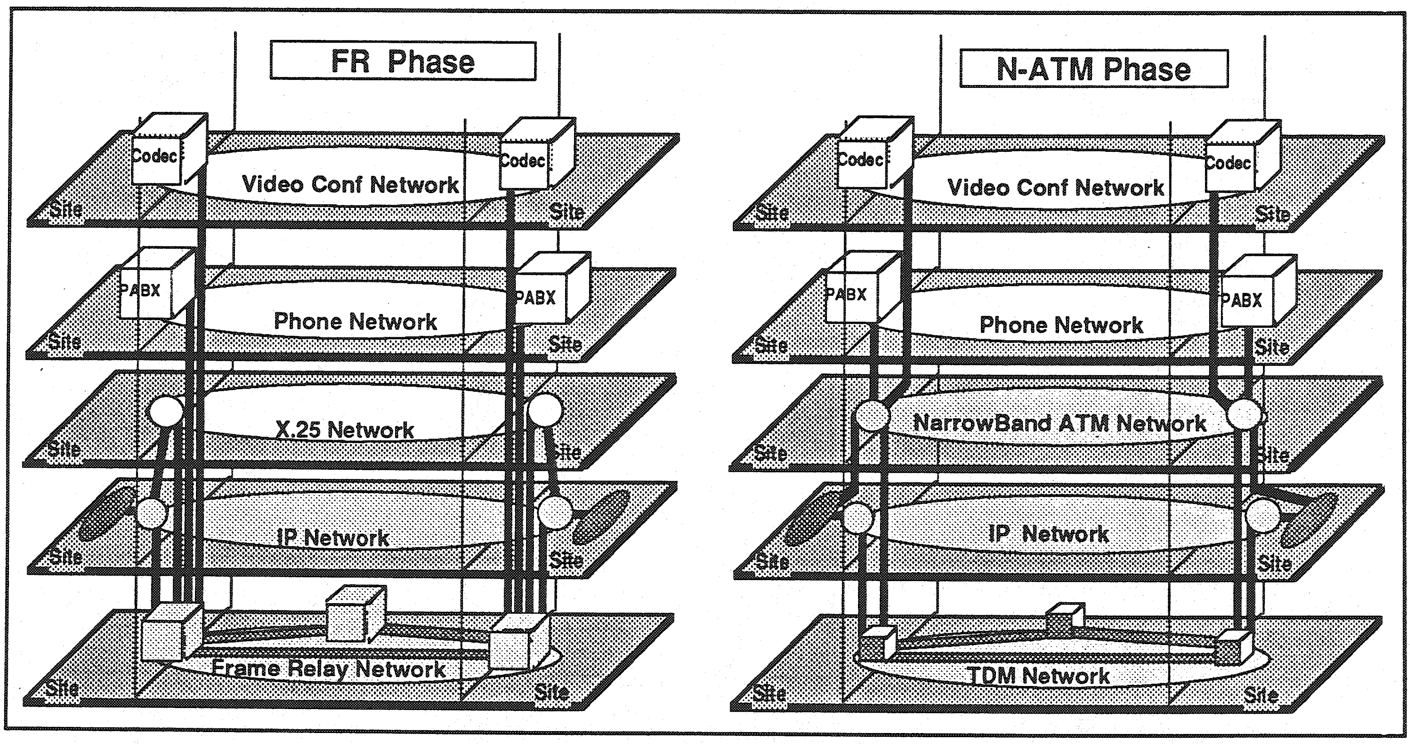


Fig 6.2: FR and N-ATM Phases

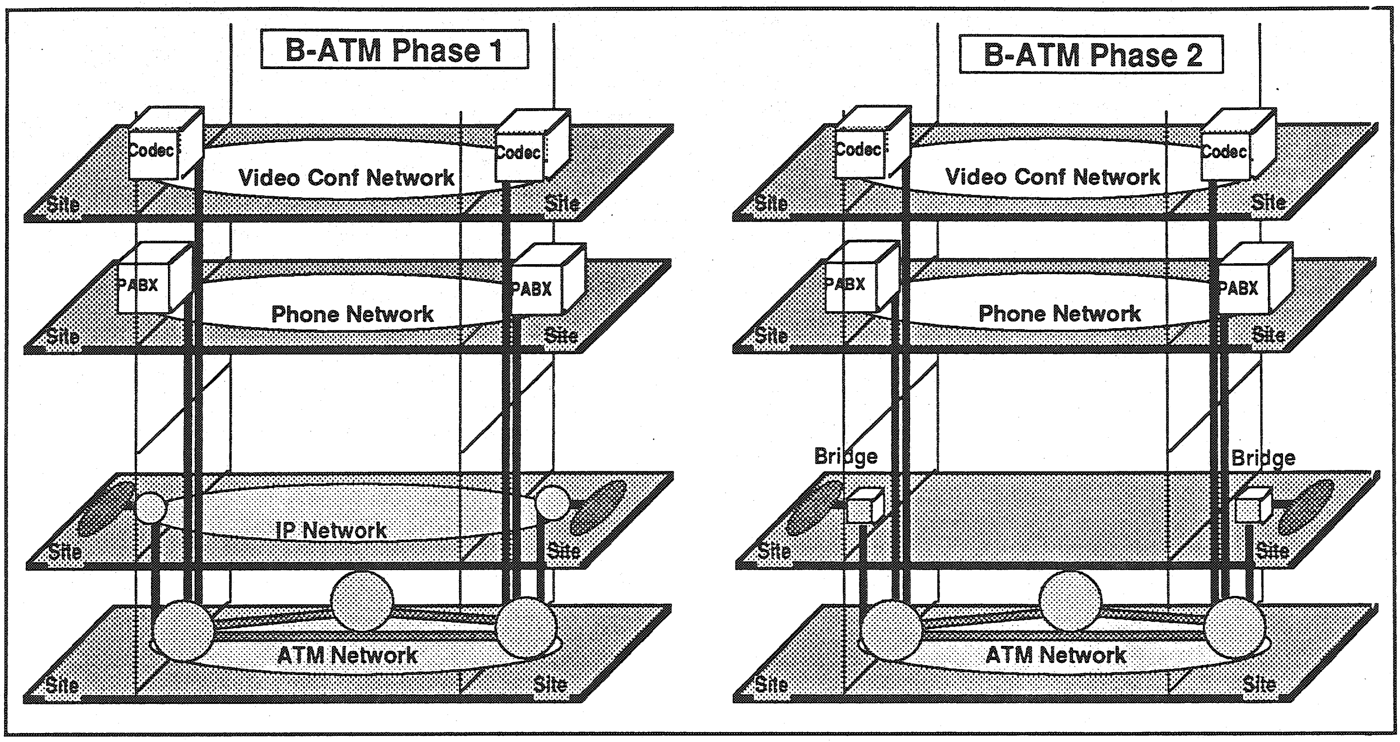


Fig 6.3: B-ATM Phases

As shown in the next section, ATM is considered today as the most promising technology. However, several fundamental uncertainties should not be neglected. Two of them are briefly described:

Multi-streaming and CONS

As shown on Figure 7, actual multi-streaming, whilst fully used over CLNS LANs, is currently widely under-exploited over existing Connection Oriented Network Service (CONS) WANs such as X.25. In theory, up to 4096 virtual channels are available, each capable of supporting one application connection. But in practice, all application connections are generally squeezed over a couple of circuits established between routers. Avoiding it is a research topic, and this is concerning when noting that ATM is CONS oriented: will ATM potential capabilities be fully exploited?

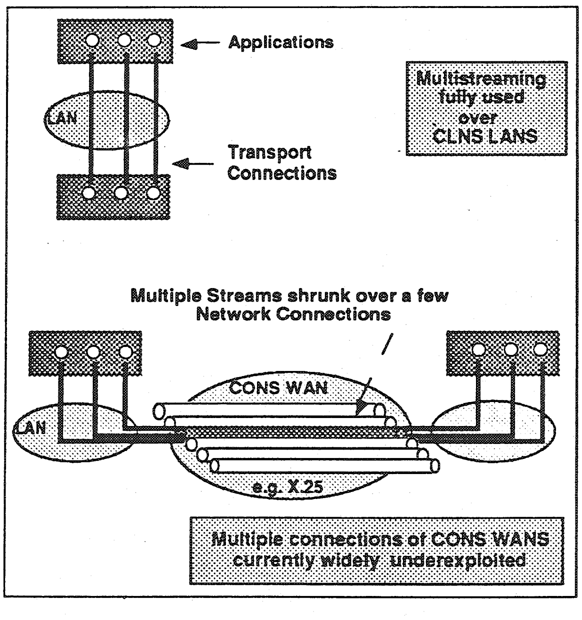


Fig 7: Multistreaming over CONS Network

Addressing in huge CONS Wide Area Networks

Assuming that the previous problem has been fixed: each application connection using its proper ATM virtual circuit, one has created a new problem, as depicted on Figure 8. A huge number of objects has now to be addressed through the ATM network, and this can only be reasonably achieved via automatic directory services (such as the Domain Name Servers used in conjunction with DoD IP). Except in the UK over JANET, this problem has never been solved satisfactorily so far with CONS network, and all users of X.25 networks have generally to manually manage address tables. This has been one of the main barrier to wide expansion of X.25 DTE connectivity, since one of the key principles of networking is:

No large scale network can work without Name Servers.

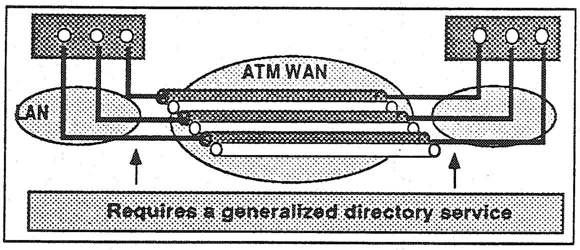


Fig 8: Full exploitation of ATM Connection Service

3.4 Practices on High Bit Rates

Let us discuss briefly achievable bit rates as measured at CERN.

Memory to Memory Bit Rates over FDDI
The graph given in Figure 9 is based on [6] where more detailed results are given. It shows that bit rates of up to 29 Mbps can are achieved using standard TCP/IP software. However, this consumes all the CPU of the systems:

A single pair of workstations can today load one third of an FDDI, ...
but this eats all their CPU!

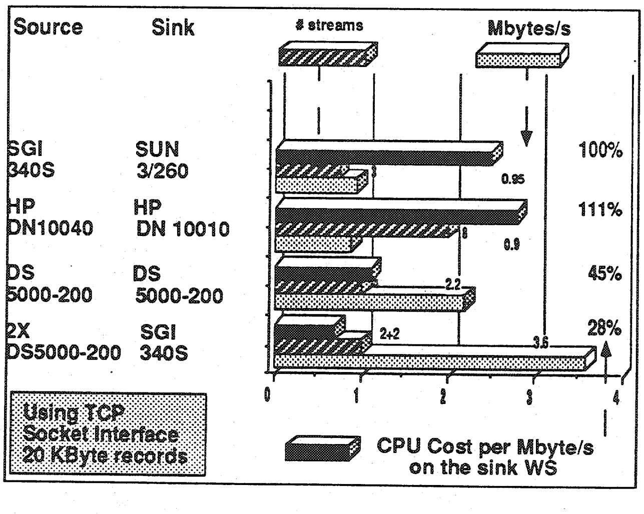


Fig 9: Memory to Memory Data Rates measured at CERN over Ultranet

Memory to Memory Bit Rates over Ultranet
Ultranet is a proprietary solution for very high speed local networking (800 Mbps at CERN) and is based on a multi-hub topology and out-board protocol implementation. It is used as a backplane in the SHIFT project at CERN (see [7]). Figure 10 shows that with very large blocks (2 Mbytes), it is possible to achieve 100 Mbps between two workstations (Silicon Graphics). The CPU cost on the sink system when getting the maximum bit rate ranges from 50% (SUN 3) to 10% (SGI 340S) and 1% (Cray X-MP/48).

A pair of workstations can today transmit at 100 Mbps, ... and only loading 10% of their CPU.

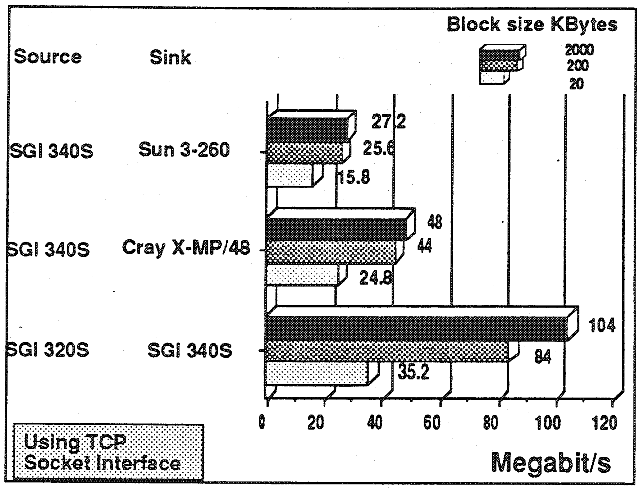


Fig 10: Memory to Memory Data Rates measured at CERN over Ultranet

This type of solution where each device benefits from a dedicated fibre to the hub indicates possible avenues for gigabit networking. The full exploitation of fibre technology will require optical terminations at the end user site, the famous "pieces of glass" called dark fibres, the user installing his own optical/electrical converters. But full benefits from dark fibres may only be obtained when an "optical continuity" is provided between end-users. This requires not only the availability of photonic switches (switching photons instead of electrons) but also that of photonic repeaters or amplifiers. P.E Green said in [8] when discussing photonic based technology: *"One can sum up the current state of third generation network technology by saying that a number of approaches are being addressed world wide that are of sufficient promise that there is no reason to doubt that 1000-node MANs (50 km diameter) at 1 sustained gigabit per second per node can be built within three to five years."*

Although there is no doubt that one or several photonic based phases will be part of actual transition scenarios by the end of this decade, the uncertainty on the timescale is judged too high by the author to include them in the suggested scenarios.

3.5 Suggested Scenarios for Transition

We conclude this paper by suggesting two scenarios for transition from megabit to nearly gigabit and from data services to multi-services (a full gigabit transition would have to include photonic based phases as discussed above).

Scenario 1: Broadband IP based Scenario

This scenario applies to existing networks with a strong IP base, and not willing to install TDMs. By the middle of 1992, they should be able to start a **"B-IP"** phase by installing fast IP switches (from 2 to 34 Mbps). However, the production support of video and voice may not be possible before 93/95. Beyond 100 Mbps, a **"B-ATM"** phase should start possibly by 1994. However the "B-IP" phase will have to be maintained in parallel at least until the end of the decade to support part of the production data services

Scenario 2: Narrowband ATM based Scenario

This scenario relies on the existence of a narrowband TDM phase (**"N-TDM"** from 2 to 8/16 Mbps) from now to about 1997. This phase will offer immediate multi-services on the one hand, and on the other hand, will allow overlap with an **"N-ATM"** phase using the first generation of switches. "N-ATM" will prepare the operational production of the **"B-ATM"** phase which could start by 1994. The **"N-IP"** phase will be maintained in parallel for part of the data, but it is not clear yet how more demanding data services (requiring bit rates of several hundred of megabits) will be best supported.

Two final comments have to be made:

1. No Frame Relay phase has been retained in the suggested scenarios. It is the opinion of the author that this technology (limited at 2 Mbps with current offer), although promising, is somewhat "too little, too late" for our large scale community. When planning phases, it should be kept in mind that in large environments, a typical delay to move from first installation of an off-the-shelf technology to fully operational production is about 2 years. No doubt however that Frame Relay will be justifiably used in certain segments of the community where its versatility will be appreciated.

2. The assumption has been also made that broadband IP (CLNS) technology will not evolve such as supporting lines beyond 100 Mbps. Hence, beyond this point, "ATM like" technology has to be used. In practice, it may happen that "hyper-CLNS" or even hybrid technologies emerge as direct competitors to ATM, thus directly affecting the latest part of the scenarios.

These scenarios target large A&R environments, but may also apply to enterprise environments. It is clear that a number of simplifying assumptions had to be made throughout this paper, and that many technological predictions are regularly contradicted by the reality. This is what makes the telecommunication planning and engineering difficult, ... but also fascinating!

Acknowledgements

I am in debt to many colleagues at CERN for fruitful discussions when preparing this paper: B.Carpenter, J.Joosten, O.Martin, as well as L.Robertson, B.Segal, A.Trannoy and more generally the SHIFT Project Team which carried out most of the reported measurements.

References

[1] European Engineering and Planning Group (EEPG), Final Report, RARE, May 1991

[2] J.Antonioz-Blanc, J.Joosten, J.Rochez, FDDI measurements, internal CERN report

[3] P.Davids, Th.Meuser, O.Spaniol, Measurements and experiences of a large FDDI installation, 3rd IFIP Conference on High Speed Networking, Berlin, March 1991

[4] B.Pehrson and S.Pink, Media Synchronization and High Speed Networking in MultiG, 2nd Joint European Networking Conference, Blois, May 1991.

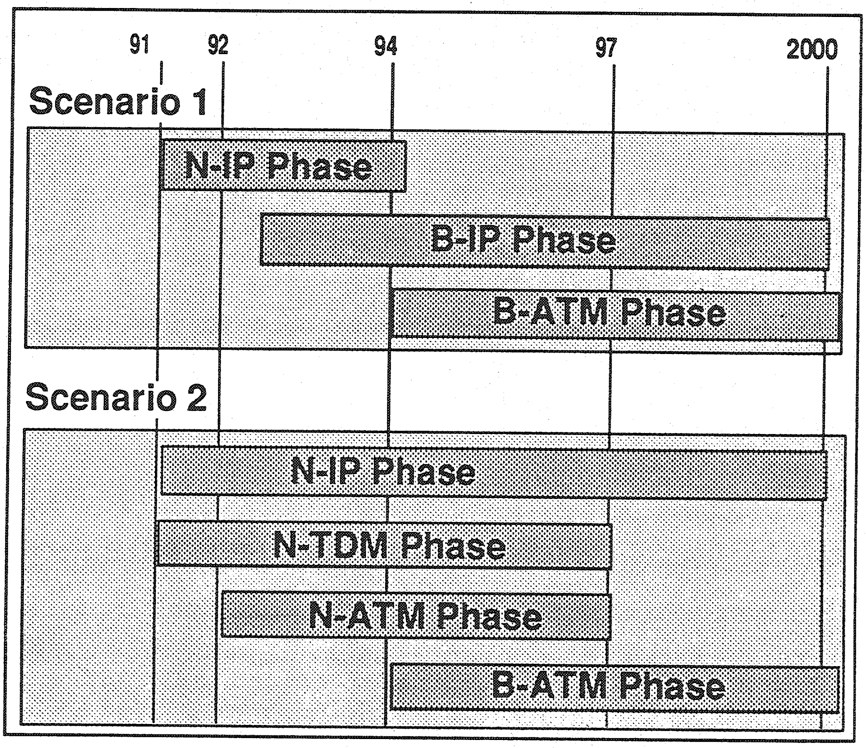
[5] J.Burren, High Speed Communications, a Tutorial on the Jargon and Technologies, 2nd Joint European Networking Conference, Blois, May 1991.

[6] L.Robertson, Experience with a prototype IBM 3172, TCP/IP and FDDI, SHARE Europe Conference, Lausanne, Spring 1991.

[7] J.P.Baud at al, SHIFT, The Scalable Heterogeneous Integrated Facility for HEP Computing, Computing in High Energy Physics Conference CHEP91, Tsukuba, March 1991.

[8] P.E.Green, Exploiting Photonic Technology for Gigabit Computer Networks, 3rd IFIP Conference on High Speed Networking, Berlin, March 1991

Author:
François Fluckiger graduated from the Ecole Supérieure d'Electricité and the Institut d'Administration des Entreprises. He has now more than 18 years of experience in design, development and large scale project management in networking. He was employed for 5 years by SESA, Paris, where he contributed to the TRANPAC X.25 network, before joining CERN in 1978. He has been in charge of CERN External Networking since 1983 and is currently Deputy Group Leader of the CERN Communication Systems Group. He was also member of the COSINE project team and Chairman of the RARE WG2. He currently chairs the European HEPnet Technical Committee.


91
92
94
97
2000
Scenario 1
N-IP Phase
B-IP Phase
B-ATM Phase
Scenario 2
N-IP Phase
N-TDM Phase
N-ATM Phase
B-ATM Phase