TELECOMMUNICATIONS INFRASTRUCTURE: PROBLEMS AND SOLUTIONS

POLICY ISSUES IN NETWORK DEVELOPMENT FOR THE UK INFORMATION SERVICES COMMUNITY

Bill Tuck

Paper presented to the 16th UKSG Annual Conference, Southampton, March 1993

Summary

What kind of network infrastructure do we need in order to support information services into the 21st century? One way of answering this is to consider our present systems to see what the problems are. By addressing these issues, rather than the technology itself, we may begin to see how networks should evolve. There are two main issues that need to be resolved. One is a problem of incompatible technologies. The other is a problem of incompatible politics.

Incompatible technologies (such as the TCP/IP versus OSI problem) mean that applications running over one system cannot easily be made to operate over another. Incompatible politics (eg. the `academic' vs. `commercial' problem, as in JANET acceptable use policy) means that the information world has been divided into autonomous domains that are prevented from talking to each other, even when the technological means to do so are present. As a result network-based information services appear fragmented, with suppliers frequently unable to connect to customers. The consequences of this fragmentation are higher costs and poorer services. Only by resolving such issues can we build a network for supporting the next generation of information services.

Introduction

What kind of telecommunications infrastructure do we need in order to support the next generation of information services, such as electronic document delivery, multimedia databases and network publishing? This is the general question that I would like to try to answer, but in doing so, I want to use it as an opportunity to examine the limitations of our current communications facilities and also of our understanding of the problem itself.

Taking `document delivery sevices' as a suitable example, the following could be cited as stock answers to some basic questions on the future requirements for a telecommunications infrastructure:

- What? High speed networks (eg. SuperJANET)
- Why? The serials crisis
- How? Fast document delivery services
- When? Soon (SuperJANET has already begun)
- Where? Very limited number of delivery sites
- Who? Selected university and research organisations

There is no doubt that the new technology of high speed networks, as represented by the SuperJANET and NREN initiatives, will have a significant impact on services that might be offered [1,2]. Transmissions from image databases or video sources require very high bandwidths to be feasible. On the other hand, there is some danger that in our obsession with higher speed we may loose sight of the real purpose of the network. Perhaps the first question that should really be asked is: "Who are the suppliers and who are the customers for these new information services?" The true function of the network, after all, is simply to link these two together.

Dr W R Tuck, PhD, is an IT Research Consultant, 3 Thornhill Square, London, N1 1BQ

In the case of electronic document delivery, potential suppliers must include libraries, publishers, booksellers, database hosts and perhaps many other different kinds of organisation with an interest in marketing information. It is also likely to be the case that a new technology will generate interest from sources other than those traditionally concerned with information supply.

Customers will include researchers, students, libraries, government departments and industrial and commercial organisations. One obvious problem is that the range of potential suppliers and customers is so great that to connect them all is likely to require much more than a single sector network. JANET (or SuperJANET) on its own, for example, will not suffice as it excludes too many interests, on both the supply and the customer sides.

This is one of the fundamental problems in trying to design a new generation of data networks. They must be capable of connecting to many different domains. Technologies for the implementation of bridges, gateways, relays and translators will become increasingly important in a world where network connectivity is fundamental. The intention in this paper is that by focussing on the particular requirements of electronic document delivery services, a deeper understanding of some of the issues relating to connectivity and the telecommunications infrastructure may emerge.

Some basic facts and figures

The British Library Document Supply Centre (DSC) delivers something in the order of 10,000 journal articles per day [3]. At an average of ten pages each, this amounts to approximately 10 gigabytes of data per day. Of this output, perhaps around 40% will go to academic or research centres with a connection to the JANET network and so could, in principle, be carried electronically. This corresponds to 4 gigabytes per day.

The annual total of articles requested by the academic sector is, in fact, near the 800,000 mark, which indicates that the 4 gigabytes figure for daily throughput should be taken as a good indication of the average loading [4].

Recently published figures indicate that the JANET network itself carries a total load of around 10,000Gb of data per day [5]. Thus, even if all serials requests sent out from DSC were delivered by JANET, the total load increase would be less than 0.1%.

Of course, this entire flow originates from just one point, the Document Supply Centre at Boston Spa, creating the possibility of a bottleneck on the input side. A single megastream line can transfer 2M bits per second (or approximately 20Gb per day). On this basis one megastream connection should be able to handle the total output from DSC, even allowing for fluctuations in demand and other factors.

It has been argued that in order to do document delivery efficiently, we need network speeds of *10M* bps or more. This begins to look somewhat implausible in the light of the above calculations. What could be argued is that the increased bandwidth would shorten the transmission time, from around 4 mins per document on a 64K bps channel to perhaps a few seconds if 10M bps were available on the input channel. This might be used to shorten the response time — the time between request transmission and document delivery — from a few hours to a few minutes, or just seconds. On the other hand, by far the largest source of delay is the manual procedure of finding and scanning material; higher transmission speeds will do nothing to lessen these. Unless material is available already in electronic form, therefore, higher bandwidths on the network will have little effect on overall response times.

If speed is not such a problem, then what about costs? Will new technology bring significantly lower costs? If we again relate this to the question of document delivery services then some interesting facts emerge.

The charging policies for access to JANET are rapidly changing as the network grows. Increasing numbers of sites connected and increasing bandwidths available should lead to lower charges, as economies of scale come more effective. Therefore, the following figures are likely to be already invalid. They are nevertheless useful as the significance is in their relative rather than actual size, and this is likely to remain constant.

Figures quoted at the beginning of 1993 indicated that the cost of megastream access to JANET would be around £50K per year. If such a channel were used to transmit 800,000 article requests from DSC to academic sites, then the transmission cost per document would be around 6p (compared to 24p per document for normal postage).

On the receiving side, however, only a 64K bps (kilostream) link would be required to handle the 10,000 or so requests that a large research university would generate. The January 1993 charge for such a connection to JANET was around £20K per year, which translates into a transmission cost of £2 per document. There are clearly very significant economies of scale. If megastream connections can be justified then the cost per megabyte is very much better than for kilostream connections (which are, in turn, much more attractive than for simple 9.6K bps links, then quoted at £6K per year, for example). Printing might add a further 50p per document, at around 5p per page.

Moving up another notch to 10M bps connections, or higher, should give even more attractive economies of scale — if only the bandwidth (and costs) can be shared between different applications or services. This is perhaps one of the most significant factors in favour of high-speed networks — they are potentially much cheaper per megabyte of data transmitted.

On the other hand, transmission costs are likely to be a decreasing fraction of total costs, for the latter will be dominated by the fixed costs of equipment and labour. In the case of document delivery systems, equipment costs could include the hardware and software necessary for X.400 e-mail connections, as well as for scanners and printers. Between £2K and £5K per annum would be a likely cost. This in turn becomes relatively insignificant compared to the labour input required to operate a manual scanning facility (which could amount to perhaps £20K per year or more).

The labour and equipment cost for a site making 10,000 article requests per year could therefore be as high as £25K per year, or around £2.50 per document. This is comparable with present charges, and is unlikely to change unless automatic input becomes possible. Transmission costs are the most variable and could range from 12p to £2 or more, depending on the network connection available. Other costs, such as subscriptions, are likely to remain the same. At a total of over £5 per document this is not promising. For electronic document delivery to be viable, therefore, the basic need is for

- economies of scale
- automation of input

It is really only the first of these that is aided by high-speed networks. What SuperJANET will buy is not just higher speed circuits — as we have seen, these may not really be necessary — but economies of scale. The trick will be to ensure that there is a sufficient volume of traffic (of all kinds, not just document delivery) to justify the extra bandwidth.

Lastly, of course, it must be pointed out that in the case of document delivery, the issue of copyright control is likely to prove an even more fundamental constraint than bandwidth availability on any rapid development of this application in the near term. The arguments for economies of scale and for automated input nevertheless remain valid.

Real requirements

The issue of cost will, of course, be fundamental in determining the viability of any new network technology. As the figures above indicate, there is potential for reducing costs by harnessing economies of scale — faster networks are cheaper per megabyte transmitted. This, however, requires that the scale of service is feasible. In other words, we may need to increase the total volume of documents in order to justify the increased bandwidth. One hope is that improved services — such as shorter response times, easier access, etc. — will lead to higher demand.

Along with cost, many other factors will be important in determining the viability of network technologies:

- universal connectivity
- flexible bandwidth (eg. input may need 10Mbps, output only 64 Kbps)
- easy, standardised access
- stability, maintenance and security

Compared to the telephone or television networks, tele*data* networks generally rate poorly on these criteria. Take the issue of connectivity. Access to a telephone is not just a universal right, but has become in some sense even a social duty. The modern world, civilisation itself, has come to depend on it in a way that is fundamental. The same is beginning to be true of the fax machine, which has yet to make a serious impact on the social world but its penetration of the business world is almost complete. This is far from being the case with electronic mail. Access to teledata networks, on which e-mail depends, is still far from universal, even if penetration within certain cultures, such as academic research, is already significant. If the teledata networks are to play as important a role in the future as many suggest, then universal connectivity and ease of access are fundamental requirements.

Connectivity and access

'Universal connectivity' means that we must be able to connect suppliers to users, and not merely user groups to themselves. Less than half of DSC's market, for example, is with the academic sector. In this sense, our present networks tend to be parochial, and limited to particular communities, even when the technical means are there to interconnect them.

Debate and discussion within JANET, for example, has tended to see the academic community as self-contained and self-sufficient. While the concern was with basic services such as computing resources or e-mail, that was a not an unreasonable position. The concern now, however, is with access to information. From this point of view, JANET is neither autonomous nor self-sufficient. Despite recent moves to make the network more open, many significant information suppliers still lie outside, including publishers, database hosts, public libraries and commercial information sources. In fact, apart from the universities, very few other organisations have opted to join the network, even though it may now be permitted. The reasons for this need to be examined.

At the same time, not all members of the academic community have equal means of access. It is more difficult if you are a member of a history department rather than a computer science department and worst of all if you are a part-time student working largely from home or a remote workplace.

What is needed, therefore, is a more comprehensive model of client-customer relations. It is only on the basis of such a model that realistic plans for a national research and education network can be made. The present system is one that has been inherited from a previous generation, with different needs or objectives and is not necessarily the best way of organising future services.

Coping with technology change

What then are the important policy issues in planning for a new generation telecommunications infrastructure? At its most general level, there are just three things: *access, cost* and *standards*. The three are, of course, related. To be effective, access must be universal. For access to be universal, costs must be low. For costs to be low, standards must be effectively marketed.

In the case of SuperJANET, the group to which services are potentially to be offered is very broad — researchers, teachers, administrators and students within the higher education sector. The means of access will be equally varied, both in terms of network connection (from dial-up modem to high-speed LAN) and of terminal machine (DOS PC to high-performance workstation). In order to accommodate this range of requirements, policies must be flexible.

The issue of access is further complicated by historical factors. Initial demand for connection to the academic networks came from the science/ engineering sector with its penchant for high-performance terminals running Unix and local area networks based on the TCP/IP protocol suite. Emerging demand, however, is from the arts/laws sector, for which PCs on Novell LANs are the favoured arrangement. The new universities likewise are much more orientated towards PC networks, further shifting the balance in this direction.

It is worth reminding ourselves that the original concept of JANET (only some ten years ago) was as a wide-area network linking a number of university-based mainframe computers. At each site, access to these mainframes was traditionally via large numbers of dumb terminals, linked in over serial lines, concentrators, multiplexers, data switches, modems etc. Such a model would now seem hopelessly inadequate; in little more than ten years the basic infrastructure has changed dramatically. Nor is there any reason to believe that the next ten years will be any different in this respect.

Such a high rate of technology change makes long-term planning very difficult. It also creates the problem of 'legacy systems', where for reasons of cost it is not easy to replace equipment that may no longer be optimal. At the same time, new sites are able to adopt newer forms of technology. The result, once again, is to increase the range of technologies that may need to be supported.

While the objective of 'standards' is to assist such planning and to allow for coherent and orderly development, the procedures involved are often too slow to be able to cope with the rapid changes that are happening in the 'real' world. This is perhaps the fundamental problem: the mismatch between the ideal world within which official 'de jure' standards are negotiated and the real world of commercial enterprise, mass markets and 'de facto' standards.

The problem of standards: 'Official' standards, 'open' standards and the marketplace

The issues around which everything finally revolves are cost, functionality, and interconnection. The tendency, of course, is to play each of these off against the others. For example, a high level of interconnection might be bought at the expense of low functionality and high cost.

In the absence of a monopoly, cost tends to reflect the market. Since the advent of the PC costs for computing resources have generally depended on market size — the larger the market, the lower the cost for equivalent functionality. This is as true for X.25 or ethernet cards as for wordprocessing software and is reflected in the relatively low cost and wide availability of products (both hard and soft) for PCs, compared to those for Unix workstations. The same will generally apply also to networks and is the principal reason why ethernet is cheaper than token ring, for example.

In such a situation it is very difficult for a manufacturer to introduce a product that does not conform to the prevailing 'de facto' standards. Situations in which this is possible tend to be where the functionality advantages, such as with super computer, outweigh any need to conform to some generally accepted standard (of operating system, for example).

In the case of networks, the same principle will apply. De facto standards, both proprietary and open, are readily available. Competition within the marketplace has led to very significant cost reductions over the past few years, to the point where the market for products based on the TCP/IP open protocol suite, for example, generally exhibits commodity-based pricing — much of it being bundled in with the basic hardware (as on Sun) or as part of the operating system (Unix). The same is true of products based around the proprietary Novell IPX and Netware protocol suites. Price competition means, however, that systems that do not enjoy similar mass-market appeal are squeezed out. Such has been the fate of OSI. It is currently very much more expensive to implement an OSI network than either of these more popular alternatives. The paradox (or rather, irony) is that the stated objective of ISO, the international standards authority responsible for developing the OSI protocol suite, is to encourage cost reduction through free competition in an open market.

This begs the obvious questions: "Is OSI really necessary?" or "If OSI had not been invented, would we invent it now?". The honest answer is, "probably not". In fact, there is beginning to appear a strong element of PC (here meaning "political correctness") in our continued insistence on OSI at all costs. The logical or economic case for OSI must be recognised as weak, as witnessed by the fact that the Scandinavian countries, largely through the absence of any pressure from the central organisations of the EC, have virtually abandoned OSI in favour of TCP/IP for their academic and research networks.

The main result of such trends has been to confuse the market. It is not always clear to the end-user why certain choices of network configuration are more strongly endorsed by advisory or administrative organisations. This is particularly the case when cost, functionality and even connectivity may seem to be more advantageously obtainable from a non-OSI competitor.

Why we need OSI

Nevertheless, good arguments for OSI remain. The strongest case is perhaps for X.400 and related protocols (such as X.500, EDI and ILL) at the applications layer and above. X.400/MOTIS is the CCITT/ISO electronic mail standard. This has been widely implemented over a number of hardware platforms as an OSI application layer protocol. Although intended to operate across OSI protocol networks, such as X.25, it is also possible to run it over TCP/IP networks by employing intermediate layer ISO protocols, as in the ISODE suite.

Furthermore, major software companies such as Microsoft, Lotus and Novell, have begun to offer gateways from their LAN-based e-mail products (including cc:Mail and MHS) to wide area X.400 services. This should mean that X.400 will be very widely available as a common interchange standard, independently of whether or not it is adopted for local e-mail services.

As an example, this potential for universal availability is one reason why X.400 may prove an ideal carrier for document delivery services. Alternative transmission methods, such as file transfer protocols, are less likely to be universally available in standard form (whether FTAM or ftp). There is likewise less imperative to provide gateways for file transfer between LAN-based and wide area services.

Another attraction of X.400, in this example, is that document requests can be sent by the same e-mail channel. The requesting address and the forwarding address (i.e. the local library) are then automatically available for replies. In fact, the addressing facility within X.400 is perhaps the principal reason why it offers such good potential for the management of document supply services. By using an application programming interface (API) built on top of the protocol, coordination of quite complex functions becomes possible. Automatic printing and automatic notice or confirmation of document receipt (to requestor or supply centre) enable much more sophisticated monitoring and control of the total process. Request status, error messages, invoicing, service bulletins and even current awareness information could all be handled in the same way.

Document transmission is, in reality, only a small part of a document supply service. In the long run, it may be more important to seek to optimise the overall function, rather than to concentrate simply on increasing the transmission speed on the delivery phase. After all, the real bottlenecks in our current document supply services seldom lie with the postal system.

The potential for universal connectivity represented by X.400 means that all suppliers and users can be connected together, and not merely user groups to themselves. In this sense it should be viewed as a universal bridge between different networks, whether SNA, X.25 or TCP/IP based. Furthermore, because it is a store-and-forward technology it is easier to provide as a gateway function. End-to-end communications, such as file transfer protocols, on the other hand, are more difficult to implement across different networks.

The interconnection of different networks has two aspects. One is the technical problem of building appropriate relays, gateways and translators. For this function, OSI protocols have much to offer as common conversion standards. The other aspect concerns 'political' segregation and the problem of finding suitable funding mechanisms. It is the latter that is behind the problem of interconnecting JANET and Internet.

Flexible bandwidth, flexible connection

In addition to interconnection of different network domains, there is a strong need for access to 'bandwidth on demand'. In other words, it should be possible to request a certain level of network access, from 9.6K bps through to 10M bps, depending on the application, and to pay for only what is needed at the time. Present services, in contrast, are based on installing point-to-point connections at a certain fixed bandwidth and at a fixed price. Sudden or short term requirements for a higher bandwidth cannot be met by such systems. For example, it is reasonable to want to download a copy of an entire 1 hour video from a central database. The volume of data involved would be considerable, perhaps several gigabytes and existing networks are simply not designed to cope with sudden demands of this kind.

The problem is partly technical and partly inherited. Voice telephony has very different requirements from document (or video) delivery. A telephone conversation requires a constant 64K bps channel. Wide area networks are based on the same technology and rely upon multiple voice lines to provide the interconnection between local area networks. This is the reason why kilostream operates at 64K bps and megastream at 32x64K bps (=2M bps). The bursty traffic requirements of document delivery systems, where one might wish to transfer several gigabytes within a few seconds, but only infrequently, do not fit easily within this structure.

Newer technologies, such as SMDS (Switched Multimegabit Data Services) and ATM (Asynchronous Transfer Mode) offer at least a partial solution to this problem. It is for this reason that they have been heralded as opening up the way for a truly 'integrated services digital network', where television, telephone and teledata can all be run over the same infrastructure. Significantly, both NREN and SuperJANET have opted to employ ATM and SMDS as their basic technology.

While present plans are largely concerned with experimental networks, the real benefits may come with the possibility this affords for true mass-market services. Already telephone services over cable television networks have become commonplace. It is only a relatively small step, technologically speaking, from there to a totally integrated system. Digital TV (and particularly HDTV) is a possible instigator of such a move. The advantage of a mass-market teledata network, of course, lies in its potential for very significant cost reductions. Whether this will actually happen is another question. There are few examples at present of successful mass-market teledata services — France's Minitel/ Teletel being perhaps the most obvious.

Conclusions

The global telecommunications infrastructure is as fundamental to modern society as the transport networks were to a previous generation. Not just 'control information', orders, invoices and settlements, but the products themselves, newspapers, books, journals and videos, will soon begin to flow across these highways. Network connections will be as important to the development of local economies as road or rail access was in the 19th century.

How we design these networks will therefore exert a considerable influence on the patterns of economic and social development. To be denied access, by high costs or restrictive policies, is to be disenfranchised. There is a need for debate, therefore, on how such changes are to be initiated. What is to be the role of publicly-funded programmes (such as SuperJANET) and what might safely be left to the market? Just as the setting up of a network of public libraries was the 19th century's response to information disenfranchisement, the public 'virtual library' network may be needed today.

References

[1] JNT. *Network News (JANET Newsletter)*, No 38, December 1992. (ISSN 0945-0636), Joint Network Team, Rutherford Appleton Laboratories.

[2] US Office of Science and Technology Policy. *The National Research and Education Network Program*, A Report to Congress, December 1992, (available in directory pub/fnc via anonymous ftp at expres.cise.nsf.govC).

[3] Lang, B. and Russon, D. (1992) Unlocking the treasure house: information for enterprise in the British Library, *Information Technology and Public Policy*, 10(3), 171-173.

[4] British Library. *Document Supply Centre Annual Report for 1992*, Boston Spa, 1993.

[5] JNT. *Joint Network Team Report for 1991-1992*, Rutherford Appleton Laboratories, 1992.

[6] Cleveland, G. *Electronic Document Delivery* UDT Series on Data Communications Technologies and Standards for Libraries, Report no.2, Ottawa, IFLA. 1991.

[7] Heijne, M. *Survey of Existing Projects and Services in Document Delivery*. SURFnet bv, Utrecht, 1992.

[8] Tuck, W. New directions for document delivery: Quartet's experiments with Adonis, *Interlending and Document Supply*, 17(3), 94-100, 1991.