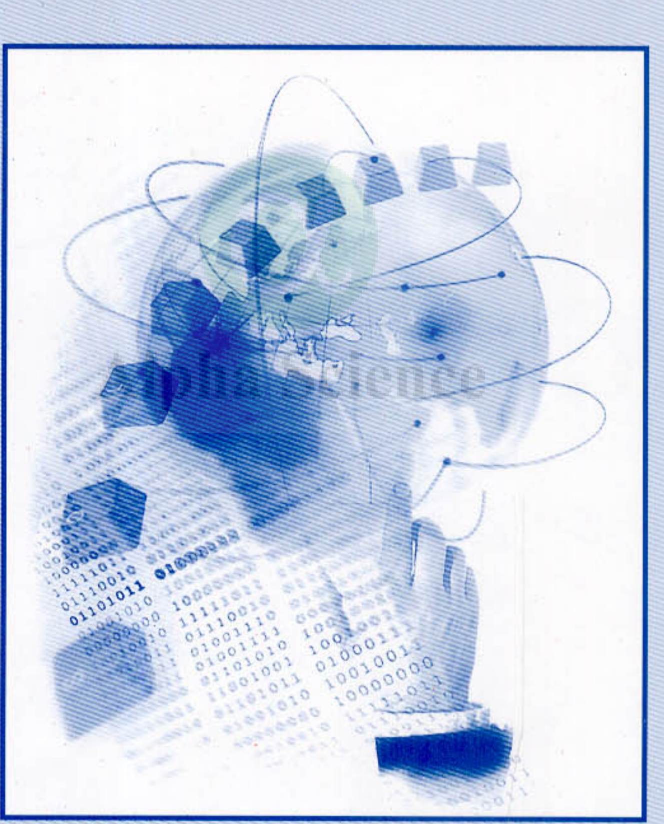
EMERGING COMMUNICATION TECHNOLOGIES AND THE SOCIETY



INSA

N. BALAKRISHNAN

Narosa

Emerging Communication Technologies and the Society

Emerging Communication Technologies and the Society

Edited by

N. Balakrishnan, FNA

Indian National Science Academy
New Delhi

Narosa Publishing House
New Delhi Chennai Mumbai Calcutta
London

Guest Editor
Prof. N. Balakrishnan
Chairman, Supercomputer Education & Research Centre
Indian Institute of Science, Balgalore 560 012, India

NAROSA PUBLISHING HOUSE

6 Community Centre, Panchsheel Park, New Delhi 110 017
22 Daryaganj, Prakash Deep, Delhi Medical Association Road, New Delhi 110 002
35–36 Greams Road, Thousand Lights, Chennai 600 006
306 Shiv Centre, D.B.C. Sector 17, K.U. Bazar P.O., Navi Mumbai 400 705
2F–2G Shivam Chambers, 53 Syed Amir Ali Avenue, Calcutta 700 019
3 Henrietta Street, Covent Garden, London WC2E 8LU, UK

ISBN 81-7319-341-X
E-ISBN 978-81-8487-283-5

Published jointly by Indian National Science Academy,
Bahadur Shah Zafar Marg, New Delhi-110002 and Narosa Publishing House,
6 Community Centre, Panchsheel Park, New Delhi-110017. Printed at
Replika Press Pvt Ltd, Delhi 110 040 (India).

Foreword

It is now believed that the Electronic revolution and Computer revolution that followed it have had their creditable innings and the future will be shaped predominantly by the Communication revolution. This revolution has also made the technology of computers and information processing become closer to the man and the society.

The Indian National Science Academy, which is a premier national body, has the very best minds actively concerned about the developments in all areas of Mathematical, Physical, Biological, Agricultural, Medical and Engineering Sciences, amongst its fellowship. It has been actively directing its efforts to bring the fruits of research and developments in Science and Technology to the society. This is in tune with the objectives of the Academy set forth by the founders which clearly enunciated that the Academy should promote Science and use it for the benefit of the society. Hence, it is most appropriate that the Indian National Science Academy chose to conduct a Seminar on "Emerging Communication Technologies and the Society" during 15–16 March, 1999 at its premises.

The world has seen many technological revolutions that have impacted the society a great deal. But the communication revolution has been the one that is characterised by high-speed change, has touched every member of the society and ubiquitous. The rapid pace at which the communication technologies emerge, the need to manage and control the data traffic on communication networks in an optimal way, the newer opportunities that arise for applications of social relevance such as Digital Library, Distance Education and Information Kiosks and the need to make communication technologies cost effective enough to reach every citizen of this vast nation are some of the issues that need a greater and deeper look at this time. I am happy to see that a galaxy of eminent speakers have addressed all these issues in a very coherent way and the Proceedings of the above seminar have been brought together in this volume. I am sure that the insights provided in the Proceedings will assist everyone concerned about taking the benefits of the advances in communication technology to the common man. The Academy is grateful to all the authors and to Professor N Balakrishnan for editing this volume. The Academy is also grateful to Professor M.G.K. Menon and Dr. S. Varadarajan for their continued support and encouragement.

New Delhi
November 24, 1999

G. Mehta
President, INSA

Preface

Communication is one of the key ingredients for the advancement of any civilization. In recent times, the communication technology had shown a discontinuous growth of high order and is a result of the convergence and confluence of many mini and major revolutions. The first is the Electronic revolution that resulted in the emergence of microprocessors that are becoming smaller, faster and cheaper. The processor performance has been taken for granted. The paradigm shift from the dinosaurs of the mainframes to the mammals of microprocessors has taken place. There has also been a shift from MFLOPS and MIPS that defined the processor performance to functionality - the ability to use the processor for many applications beyond mere number crunching. The word 'computer' has become a misnomer and it has now become a information utility.

The world of computers that was monomedic - a world where only numbers and alphabets were *lingo franca,* soon turned multimedic. The multimedic abilities of the computers made it possible to touch every human being. This has removed the elitism associated with the monomedic computers of yester years and paved the way for a true information Society.

When one was astonishing at the fast pace of change in Computer performance, the advances in Communication technology are even more astounding and leave many breathless. With the two rapidly-changing technologies like those of Computers and Communication merge, the result is an impact on the society, the likes of which the world has never witnessed. The birth of internet World Wide Web, the near Giga PC's and multimedia Technologies that threaten to break the language barrier have all contributed to the fact that Information will be available to anyone, anywhere and any time. This, in brief, has sounded death knell of time, distance, cast or creed. This is the key ingredient of borderless society or a Global Village.

Understanding the impact of such revolutions on the society and, in particular, on India has been a serious pursuit of many an intellectual mind. This realization of the potential of the Emerging Communication Technologies and their impact on Society has formed the motivation for the Seminar conducted by Indian National Science Academy during March 15–16, 1999 as a part of its mandate to take the fruits of science to the mankind. For many years to come, the corporate and national backbones will have the optical fiber as a transport medium. The lightwave communication that promises seemingly limitless and exponentially increasing enhancement in reach and richness has been the focus of the paper by V.S. Arunachalam. The future of the fiber has been elegantly brought out by Kumar Sivarajan. For rural and urban India, the access to technologies will range from the Plain Old Telephone System (POTS), cable modems and the Satellite Communication. The Indian leadership in Satellite Communication and cost-effective telephony have formed the central theme of the papers by K. Kasturirangan, S. Rangarajan and A. Jhunjhunwala, while the use of Cable TV to disseminate Information has been addressed by Mohan Tambe. In any networked application of social relevance, the traffic monitoring, control and management and adaptability of the network to carry differentiated class of data ranging from data and voice to bandwidth hungry multimedia applications. Anurag Kumar, S.V. Raghavan and S. Swaminathan have looked at these issues. The major impact of the communication would be seen only when applications of social relevance are also designed. For a country like India, besides social information systems, the Universal Digital Library and Net-based Distance Education are two of the key applications that will propel the nation in the next millennium to be a scientific superpower. Raj Reddy, Michael Shamos and Mukul K. Sinha have eloquently addressed these. The overview of the Seminar theme has been given by the visionary M.G. K. Menon who has seen the revolutions grow from their infancy to adulthood and, in fact, been responsible for making IT happen in India. In brief, the Seminar has addressed all the relevant issues that too by experts of high repute from academic institutions in India and abroad as well as by those from the industry.

We are happy to present the edited proceedings of the Seminar. We thank the speakers who not only delivered high quality and perceptive lectures, but also ensured that the manuscripts were ready in time for the publication. A special word of thanks to Dr. S. Varadarajan whose untiring efforts made it possible to have such a galaxy of speakers. Prof. G. Mehta with his stimulating leadership and Prof. M.G.K. Menon with his encouragement made this seminar a memorable one. Thanks are also due to Mr. A.K. Tagore of INSA for his meticulous work on interfacing with the publishers.

N. BALAKRISHNAN
Convener & Guest Editor

INSA Editor's Note

Communication revolution has drawn closer the two vital technological initiatives namely the computer Technology the Information Technology. This revolution has a tremendous impact on the society breaking barriers of distance, time and individuality. In fact, it has brought in communication exchange/information dissemination at an incredible speed using gadgets at micro miniaturized level with higher and higher efficiency.

The present volume entitled "Emerging Communication Technologies and the Society" is very timely which is an outcome of the proceedings of the Seminar conducted by the Academy on the above subject during March 15–16, 1999, under the Convenrship of Professor N. Balakrishnan. This book, we hope, would be found as one of the most informative documents of topical interest by the scientific community.

We are grateful to the authors for contributing articles based on their presentations in the seminar which made it possible to frame this book. We are indebted to Professor N. Balakrishnan for his untiring efforts to conduct the Seminar and editing this volume. We are also thankful to Dr. S. Varadarajan for his able guidance to bring in the eminent speakers and their articles for this volume. The remarks by Professor M.G.K. Menon during his inaugural address in the seminar were inspiring and stimulating. We are truly grateful to him for his overview of the overall scenario. The editorial assistance rendered by Shri A.K. Tagore and Shri M. Ranganathan is thankfully acknowledged.

New Delhi
November 22, 1999

S.K. Malik
Editor of Publications, INSA

Remarks
Professor M.G.K. Menon, FRS, FNA

There have been truly spectacular developments in the field of Communication Technology over the past three decades—with digital technologies and micro-eletronics, lasers and optical fibre systems, large scale introduction of wireless, growth of mobile/cellular systems, developments in space communications and much else. I envisage that similar development will continue into the foreseeable future, at least for another quarter of a century. It represents a rich field for scientific work, it is also a very important field for applications of relevance to society. The Seminar, organised by the Indian National Science Academy (INSA) during March 15–17, 1999 on the theme "Emerging Communication Technologies and the Society", is therefore relevant, topical and most opportune.

The title of this Seminar has two important components: the first relates to Emerging Communication Technology, and the second the manner in which these developments relate to society. I would like, in this inaugural address, to give some prespectives concerning both these aspects.

Information Technology (IT) is a very commonly used phrase. For most, it represents computers and workstations. Few go deeper into the physics, chemistry, material sciences & engineering, and particularly the micro-electronics, that lies at the heart of the systems that one actually uses; nor of the mathematical underpinning of the software and programming; this is all taken for granted. There is little appreciation of the fact that Information Technology, in its present form, is based on connectivity. Computers and workstations that store and work on information have to be interconnected. Internet and the use of the World-Wide-Web (WWW) characterises the connectivity and the access that we enjoy today. The information superhighway, based on telecommunication technology, is an essential component of IT. In addition, the broadcasting media, namely radio and television, and all information contained in textual form, have also merged into the overall interconnected IT system. It is this seamless convergence, leading to a connectivity of all devices that deal with information, that has led to the spectacular onward march of IT today.

Information is, of course, the very basis for our existence and of civilisational development. The entire biological system is based on information which is contained in, transmitted, and made use of through genetic material; and the greatest of the efforts of modern biology relates to the manner in which one is trying to unload this data base—an effort that is needed before one can make use of it. Through human history we have seen successively higher capabilities in information storage, display and transmittal e.g., rocks and stones, the use of papyrus and palm leaves, beautifully handicrafted and illuminated manuscripts, and finally the advent of printing which enabled information to be truly widespread throughout society; the latter enabled society to move to a more equitable level in terms of access to knowledge. But, in the past half a century, we have moved ahead in phenomenal fashion through wholly science-based and technology-driven capabilities. This is leading us to completely new dimensions in the way we can deal with information e.g. storage, handling, accessing, collation, correlation and computing, and new forms of networking, of connectivity and accessing. It is the scale and speed involved in all of these which is giving rise to the emerging information society, which represents a new phase of human history. Many of us have had the privilege of witnessing these discoveries as they took place, and interacting with the discoverers.

One tends to forget that these developments have taken place only in the very recent past. To highlight this, I would like to draw attention to a Survey, called "Science of the Times", produced by the "New York Times". It states: "The New York Times, it is generally recognized, publishes the best scientific reportage in the country. Here is a broad-based and engrossing survey, drawn from the pages of the Times, that is invaluable for anyone interested in keeping abreast of the most pressing scientific issues and discoveries of today—and tomorrow." In the issue of the late 1970s, there is not one reference to Information Technology! I also recall that the year that I set up the Electronics Commission in India, (in January 1971), was the year when e-mail, the microprocessor, the floppy disk and Microsoft all came into existence. Today each one of these has permeated into ordinary households. This is the speed with which information technology, from a base of esoteric developments in different disciplines of science has now become the equivalent of an environment in which human society has to function in the future.

A few broad remarks concerning the development of Information Technology, globally as well as nationally, would be appropriate at this point. From a figure of US $ 5 billion in 1960 to US$ 690 billion in 1997, IT industry is expected to cross US $ 1000 billion in revenues annually from this year. It would be the first industry in the world to do this. Globally, IT is growing at eight and a half per cent per year, faster than any other industry. As IT permeates various sectors of human life, and business and industry, the role of software and IT-enabled services is becoming increasingly more and more important. Today the latter constitutes about 60 per cent of the revenues of the total IT industry.

In India also, IT is in a phase of tremendous growth, far more than globally, because of the very low base from which we have started. The compounded annual growth rate of IT industry as a whole has been around 40%. In the case of software, the growth has been particularly high, around 55%. Software exports, which were around 0.3 billion US dollars in 1993–94, have crossed US dollars 2.65 billion in 1998–99.

IT has caught the public imagination in India. A drive through any metropolitan centre in India will demonstrate this. Apart from what the Central Government is doing to promote IT, the individual State Governments are vying with each other to make IT a major component of their developmental activities.

In view of the enormous potential for the growth of IT, both hardware and software, as also of all applications, and its potential to transform a wide range of human activities, and modernise the economy, the Govt. of India set up in May 1998, a National Task Force on Information Technology and Software Development. This Task Force submitted its first report at the end of June 1998; this covered various aspects relating to the promotion of “Software”. The Government of India accepted all these recommendations and notified these in the Gazette of India towards the end of July 1998. The Task Force has submitted two further reports: the second on “Hardware” and the third dealing with the “Human Resource Development”, ”Citizens-IT Interface”, “Research, Design and Development”, “Content Creation and Content Industry”, “Microelectronics”, Financing the IT Sector” and “Mission Mode Creation of Fibre-optic Infrastructure” it also had in it certain broad policy recommendations. The latter two reports have received approval in discussion in Government, but have yet to be finally approved before being notified as official policy.

It will be noticed that the Telecommunications sector was not dealt with by the National Task Force. Indeed, telecommunications was covered by a National Telecommunication Policy announced in 1994, which visualised India emerging as a major manufacturing base for telecommunication equipment. It also visualised the opening up of telecommunication services to the private sector.

Whilst there has been development in telecommunications over the last few years, with a significant increase in the wide band optical fiber systems which now extend to 75,000 route kms, the introduction of cellular phones, significant capabilities through our space programs, and large scale induction of electronic switching systems, particularly based on C-DoT technology in the rural areas, (apart from imported systems/ technologies), the progress in the telecommunication sector is far from satisfactory. This is certainly so considering the strides that IT is making globally, and the telecom sector, with it; as also in terms of India’s needs in the telecom sector if IT is to make headway here, as also play its role in fulfilling national goals and aspirations.

Infrastructure in India in all sectors has been slow in growing for a variety of reasons, which I shall not go into here. This has been the case also with telecommunications.

Recognising that the telecommunication infrastructure is a key element for rapid economic development in the country, and particularly for the growth of IT industry, the Central Government decided to look at the telecommunications sector separately. This has resulted in the New Telecommunication Policy (NTP) 1999. It is not possible to cover the details of NTP 1999 in this brief survey. However, it does represent a considerable advance from the past; and when implemented properly, can lead to significant growth.

An aspect that needs to be emphasized is the attitude framework which relates to the telecommunication sector. Firstly, telecommunication has been largely regarded as a means for voice communication e.g. synonymous with the telephone, and more recently for text. Its role as an essential component of IT, providing the information superhighway along which data that can flow easily at low cost, is not adequately appreciated. The second aspect relates to the fact that hitherto, telecommunication has been in the public sector, being entirely handled by the Department of Telecommunications. Certainly, for a long time to come, this Department will have an important and primary role in providing a major part of the telecommunications infrastructure and handle many different aspects that relate to ensuring order in the system. However, telecommunications cannot be a monopoly of the Department of Telecommunications. There is need for significant involvement of the private sector, not only in terms of financing, but also in the management and entrepreneurship that is essential to ensure dynamism and flexibility in the many diverse ways in which the information superhighway ultimately manifests itself. Internet service provision, cellular phones, paging, many aspects of basic services, creation and maintenance of networks, voice over data, the role of cable TV networks and set top boxes, are some examples of areas that need to be rapidly grown in the country with private sector participation.

At this point I would like to make a few general remarks on the telecommunications scenario in India today. India has around 18 million telephones, with a population of close to 1000 million people. This represents 1.5 telephones for 100 persons. About 25% of these telephones are in four metropolitan cities. The reason for this situation is that telecommunications has been wholly public sector responsibility which does not respond rapidly and flexibly to needs, markets and new opportunities opened up through science and technology. At the present cost per line, the investment needed for a large scale expansion of basic telephone services through copper and optical fibre cables are very high. Even in urban areas, telephones cost more than Rs. 30,000 per line to install. This becomes much greater for rural areas, since the population is spread out and lower in density. At present, the costs for national and international long distance calls are kept very high. This is because of the phenomena of cross subsidization. The assumption is that the common man or poor individual should be able to make local telephone calls at low costs; the total revenues required have then to be made up from STD and ISD calls.

What is required is a paradigm shift. First to consider telecommunications not as serving the needs of voice alone, but significantly as one required for data transmission. It will then be necessary to reduce STD and ISD rates so that telephone lines can be used extensively, at low cost, for data transfers, Internet access should be easy, fast and at low cost. There is need for a higher, and more realistic, rate for local voice calls which take up a great deal of time. However, these calls are seldom made by the truly poor or disadvantaged; a higher local call rate will have little impact on the poor.

It is equally important to ensure that costs for installing and operating telephones are greatly reduced. One should reduce the per line cost to between 10 to 15000 rupees. It is here that technological innovation, particularly of an indigenous nature, applicable to the situation encountered here, is relevant. Such innovation can enusre that telecommunication is not only an urban luxury, but available nationally in urban as well as rural areas.

Once low cost lines, wide band capabilities, low cost internet access and large scale data transfer capabilities become a reality, the traffic situation will change completely from the present. There will be certainly an interim period from the view point of financial outlays and returns. But very quickly, taking note of the manner in which telecommunications is becoming the heart of the IT revolution in the form of the information superhighway, one will find that the situation is of a very high volume, low margin, but a very profitable enterprise. Even more, it will open up the rural areas of the country so that IT based entrepreneurship can develop all over. Internet use, teleinformatic kiosks, e-commerce and much else will bloom. This is the vision that we should work towards.

There is also need for clear appreciation of the fact that a great deal of research, design and development is possible within the country which can provide products and services of relevance to local needs, can be superior for the specific purposes that one has in mind in the country, and cost effective, compared to what can be provided by foreign vendors. The latter must undoubtedly be involved in a spirit of cooperation and collaboration, particularly taking note of technology development and global experience, but not on the basis of being wholly dependent on them.

To illustrate this, I would like to point out some major successful indigenous research efforts in various sectors of telecommunications: on electronic exchanges at the Telecommunications Research Centre of the Department of Telecommunications; the development of the Automatic Electronic Switch for the Army Radio Engineered Network at the Tata Institute of Fundamental Research Mumbai, and the Electronics and Radar Development Establishment at Bangalore that was the basis on which the Centre for development of Telematics progressed so rapidly and successfully in its development of electronic exchanges; work in the Indian Space Research Organization (which has very significant research capabilities in telecommunications) which has provided the country with satellite based systems, terms of transponders, earth stations and the like; capabilities for the design and fabrication of large antenna systems that was developed from late 1960's; and many more.

From papers to be presented at this Seminar you will see the extent to which a range of institutions, which are not direct Government organisations, but educational institutions as well as those in the private sector, are now starting to contribute significantly to research, design and development in the telecommunications sector. It is important to support and promote research capabilities in this field over a wide spectrum of both areas and institutions.

There are many areas relating to IT, of which telecommunications is a part, that I have not been able to touch. These include the manner in which telecommunications has moved in the direction of digital systems, mobile systems and highly personal systems. Mobile/cellular telephony represents one of the great areas of growth over the past decade; from 16 m mobile users in 1991, it has grown to 330 m this year, globally. Along with these developments, many new issues are coming to the fore, relating to secrecy, security, privacy, cyberlaws that will have to be enacted, and the manner in which all of us will have to learn to live, in this wholly new environment that will be upon us in the next century. I have not also been able to deal with the opportunities that exist in the area of education, in the creation of content that will ensure that we use IT to conserve the best of our culture and be able to access it widely, and ensure the development of our indigenous pluralistic ethos. These are all areas of unique opportunity for science and the applications of science to meet human needs and advance civilization.

Contents

1. Lightwave Communications: The Endless Frontier

V.S. Arunachalam

Department of Materials Science & Engineering,
Engineering and Public Policy and Robotics Institute
Carnegie Mellon University, Pittsburgh PA 15213-3890, USA

Human society is built on communications and thrives on their usage. There is a continuing effort to improve its reach and richness to provide us humans with abundant information that gets transformed, when relevant, into knowledge. In this paper we discuss the recent breathtaking changes in communications technology and suggest that, by its very range and richness, it offers opportunities for growth that were not dreamt even a decade ago. The technologies fuelling the Information Revolution are still nascent and are robust for further growth in performance. A few scientific opportunities available in further enhancing the performance are discussed and their consequences in meeting the societal and individual objectives of people are described in this paper.

1. Introduction

Philosophers from both the East and the West consider *Sadhana* or inquiry as a unique and exclusive characteristic of humans. Mathematician Descartes attributed the very existence to thinking (*I think, therefore, I am*). If inquiry is the essence of human individuality, communication is central to human society. The impulse to communicate to others what one sees, feels and thinks is so strong and basic that we have evolved ingenious ways to effect that urge. Human voice is given meaning through language that modulates the sounds and transforms them into speech. Languages and scripts were invented to provide channels for the more permanent formats of communication. The sensory organs that we are endowed with receive, pre-process and transmit signals irrespective of the type of sensation (aural or visual, for instance) to the brain for perception. Perception is the result of comprehensive model building in the brain based on stored information called memory and the received signals. Information is built as a consequence of such deliberate activities and, when exchanged among humans, forms the content of communications. The quality depends very much on the received signals and the interpretive capability of the human brain. A waterfall may appear merely as a majestic sight to an observer and encourage her to say so, while in another it may inspire a song or a painting. It is therefore not unusual to find the processed information that forms the contents to be different for different people though the input signals could have been the same. Our society is built on a rich repertoire of such communications.

Apart from the stored memory, model building in the brain depends very much on the received signals, their quality and bandwidth. We are able to comprehend and appreciate an event such as climbing a mountain peak better when voice comments are supplemented with visual images. And, as individuals we are confined to only one location at a time, though keen to know of happenings of the past and present in other locales. To satisfy this longing, it is necessary that we have access to process the information from outside at a richness level as if we were actually there at the time when the event was unfolding. No wonder, a continual passion of our society through all ages is to build a robust and efficient communications structure with a long reach, excellent signal quality to fill the needs of as many human sensors as possible, and all these at a speed that is almost instantaneous.

2. Lightwave Communications

In this paper, we present a summary of the recent spectacular growth of communications, their reach at global levels, the speed and the richness (bandwidth) provided by the various channels of transmission.

Figure 1 shows the reach and richness of various communications systems from the past century onwards. The richness and speed of some of the new lightwave communications systems that may become available in the first decade of the coming century promise to be so impressive that the bandwidth of an optical fiber system would approach 1,000 Terabits/sec (10^{15} bits/sec). The burgeoning capability from this capacity can be imagined from the presently

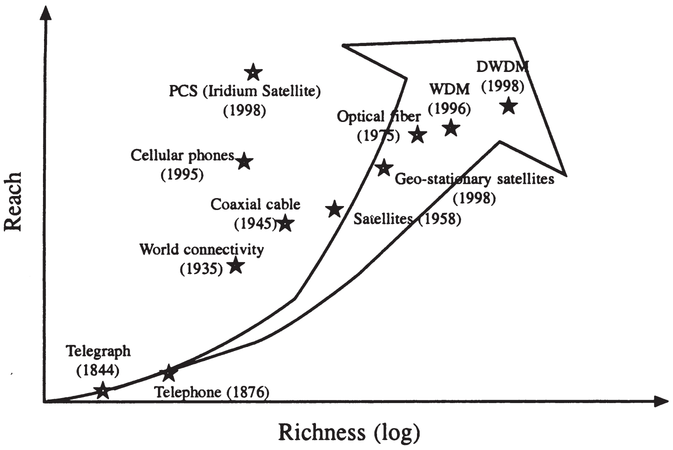


Fig. 1. The reach and richness of communications technologies

available capacity of about 40 Gigabits/sec that supplies about 500,000 voice channels or 800 high quality television channels.

The two elements that make this capacity possible are the increase in the processing speed of microprocessors and the capacity of lightwave systems. The first is enabled by the increase in the number of transistors per chip, and the second, by the availability of minimum loss fibers with broad transparent bandwidth and *in-situ* fiber amplification. Figure 2 shows the growth of the processing speed of microprocessors. The validity of Moore's Law that predicts the doubling of processing power in every 18 months has been so impressive that many anticipate that this law would continue to be applicable even after reaching the silicon processing limit of 0.1 μm. At first sight, such processing speeds may appear esoteric and unnecessary for communications systems, but when we consider the bandwidth requirements of one single high-definition multi-media presentation integrating seamlessly a few locations and with other traffic, the speeds appear essential. Data traffic has also increased so spectacularly over voice that in a few years it would overtake voice traffic that is only growing at an annual rate of over 10%. Compared to this, the data traffic is growing at the rate of 80% annually. The Internet is fast becoming ubiquitous consuming over 3,000 terabytes every month.

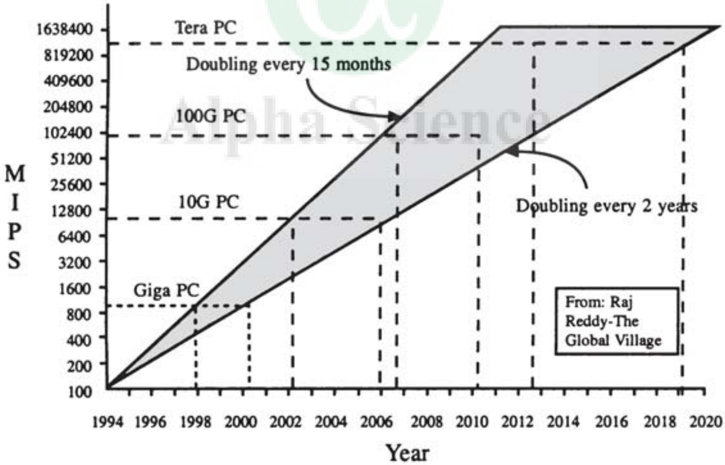


Fig. 2. Microprocessor processing speed as a function of time

The growth of optical communications is more spectacular. The idea of using light energy as the communications medium is not new. Alexander Graham Bell, over a century back, considered the possibility but abandoned it because of the absence of powerful light sources and non-attenuating channels for their transmission. Two major scientific and technological breakthroughs, viz., lasers and optical glass fibers in the 1960s and 70s have changed the status of lightwave communications radically. Semiconductor lasers with very high output, in excess of 100 mw, laser arrays and distributed feedback systems have eliminated the problem of lack of powerful light sources. Optical glass fiber that was highly dissipative (over 100 dB/km) when proposed originally by Kao and Hocham, has now become a far more transparent medium, almost matching its theoretical limit. The attenuation loss at 1550 nm has come below 0.2 dB/km.

Figure 3 shows the attenuation of silica fibers as a function of wavelength. There are three well-defined plateaus (850, 1300 and 1550 nm). As the attenuation is the lowest at 1550 nm, and the bandwidth larger, it is now increasingly used for long-haul communications. There are three major sources of loss in a silica glass fiber. Some are intrinsic to the material such as absorption due to atomic and molecular vibrations and scattering. Impurity atoms and defects also contribute to this loss. The second source is due to dispersion that is caused by the different wavelengths of the light pulse travelling at different speeds. This broadens the pulse to an extent that after some distance the signal is lost. This type of dispersion loss is known as intermodal dispersion and is compensated by a graded cladding encapsulating the fiber core with materials of different refractive index. In such fibers, except a single mode, all other modes are lost making it into a single mode fiber without such dispersion losses. Even in single mode fibers, there is dispersion loss due to chromatic aberration. This is a material property and varies widely as a function of wavelength. For silica fibers, the chromatic aberration is almost zero around 1300 nm, but becomes fairly large (18–20 ps/km-nm) at 1,550 nm. The third component of loss is due to fabrication and laying of optical fiber cables, mechanical bending, splicing, stressing the cables and

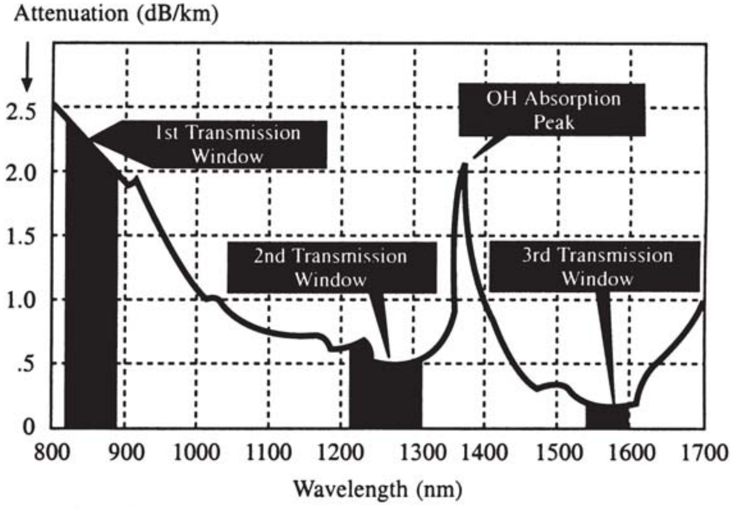


Fig. 3. Optical fiber attenuation as a function of wavelength

aging. These are becoming smaller thanks to advanced fabrication and encapsulation technologies and innovative slicing and fusing options.

3. Issues on Reach and Richness

In spite of these impressive strides, lightwave communications would still not have met the requirements of global reach and veritable richness but for some of the new innovations of the past few years. For extending the reach, it is necessary to lengthen the path between regenerators where the optical signals are converted to electrical ones, amplified and then reconverted for onward transmission. Minimum attenuation is attained by a careful choice of core and cladding chemical compositions and processing controls. The OH^- absorption that separates the 1,300 and 1,500 nm is now being minimized and some laboratory results show absorption due to hydroxyl ions to be totally eliminated. In spite of these, the repeater distances are around 60 km and further extensions would have to depend on other chemical compositions and innovative network designs.

There have been studies to calculate the theoretical limits of attenuation of various glass forming compositions. These show that a reduction to below 0.01 dB/km is indeed theoretically possible. This means that future fibers could have spans as long as a few thousand kilometers without regeneration. But it is unlikely to be realized in the immediate future because of the difficulties of manufacturing very high purity fibers that have far more stringent purity requirements than the silica fibers. These fibers based on heavy atom halides would also shift the band of operation to longer wavelengths for which the other components of the optical communications systems such as the laser generators and detector systems are presently not available.

Dispersion losses are minimal at 1,300-nm band, but are significant and large at 1,500-nm. It is possible to shift the zero dispersion to longer wavelengths with appropriate doping of the silica core and cladding. Such zero dispersion shifted fibers are now commercially available. However, these suffer from major disadvantage in that non-linear effects such as four wave mixing and self-phase modulation become prominent when a number of wavelengths near the 1,550-nm passband are used. A better option would be to shift the dispersion so as to minimize it in the passband, but not eliminate it altogether. The dispersion so accrued could be eliminated by combining the fiber with dispersion compensating fibers that show large, but negative, dispersion values. Such a fiber combination appears attractive for reducing the dispersion losses. There is a disadvantage in using such compensating fibers in that they show very large attenuation. A judicious assessment of the attenuation loss and dispersion gain has to be made in designing the fiber network system.

A major breakthrough in increasing the reach and richness of the optical fiber network has come from the development of erbium doped solid state fiber optical amplifier. The erbium atoms in silica fiber are excited to higher energy levels with appropriate laser pumping at wavelengths transparent to silica. When weak signals enter the erbium-doped fiber, the erbium atoms transfer their energy to the signal by laser action. The *in-situ* amplification, without first converting them to electrical signals, can be large (around 30 dB), independent of wavelength in the erbium pass band and are also virtually error-free. The emergence of the optical amplifier has made wavelength division multiplexing (WDM) where a number of wavelengths separated by 50 GHz are used as carriers, a commercial reality. Figure 4 shows the passband at 1,550 nm and the fiber amplification band.

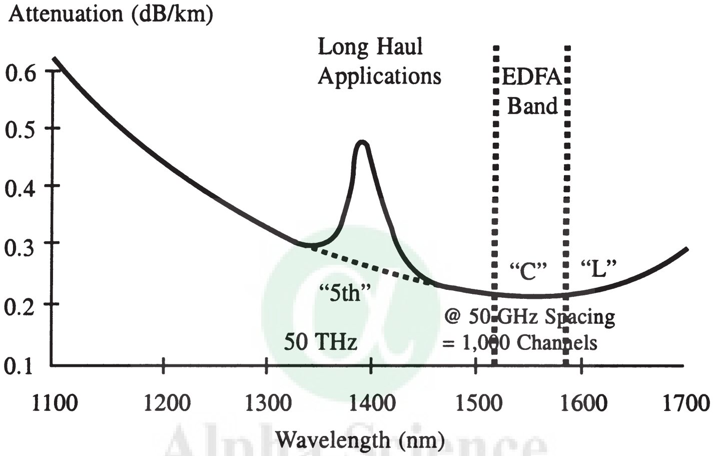


Fig. 4. Wave division multiplexing in optical fiber

Further increases in WDM to still larger number of wavelengths sharing the input laser power would depend on the availability of adequate bandwidth and amplification with a better signal-to-noise ratio. These are now seen as commercial realities when the fiber amplifiers are cascaded with distributed Raman amplifiers. Raman amplifiers are constructed by pumping with laser diodes to rare earth-doped silica fibers with appropriate germanium and fluorine additions. By backward pumping with these amplifiers it is possible to extend the bandwidth to almost 90 nm. Even though Raman amplification may not be large as fiber amplifiers, its distributed nature (the entire fiber acts as an amplifying medium), versatility to shift the gains spectrum and improving the signal-to-noise ratio makes it an attractive asset to increasing the bandwidth. It is then possible to operate over 1,000 wavelengths and providing one thousand terabits per second! What is more interesting is the speed with which laboratory successes are making to the market place. Figure 5 shows the narrowing of this gap in optical fiber communications. These results suggest that the innovations in lightwave communications are not incremental as in many others, but are radical and path breaking.

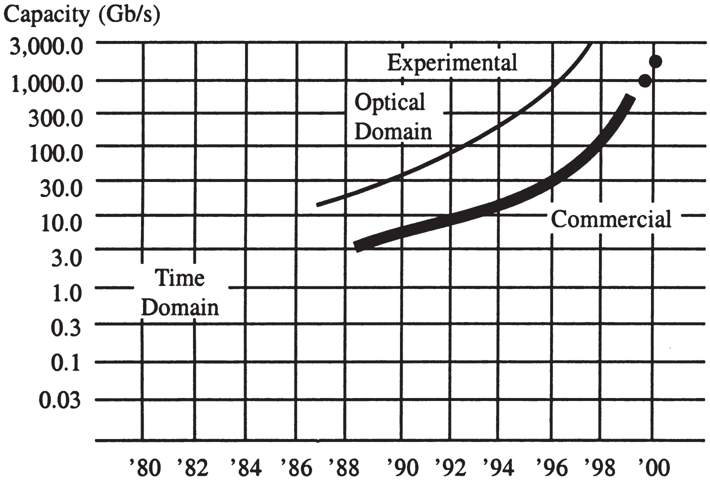


Fig. 5. Fiber capacity as a function of time (Both laboratory results from Lucent Technologies and commercial realization are shown)

We began this analysis wanting to scrutinize the available communications technologies that would provide all the bandwidth that human sensory organs would need and also overcome the limitations imposed by distance and time. Optical communications appear to provide the answer to these needs. The challenge for the coming years will be on enabling this technology to reach every consumer without losing its attractions and thereby solving the 'Last Mile' problem.

2. Future of Fibre-Optic Networking

Kumar N. Sivarajan

ECE Department, Indian Institute of Science, C.V. Raman Avenue,
Bangalore 560 012

The first generation of optical networks used optical fibre for its remarkable properties such as extremely low loss and high bandwidth, as a replacement for copper links. Examples of such networks are SONET and FDDI. Second - generation optical networks are based on incorporating optical switching and routing in addition to fibre optic transmission links. While these networks started as research curiosities in the late' 80s and early' 90s, commercial wavelength-division-multiplexing products incorporating wavelength switching/ routing have begun to appear in 1999. However wavelength-division multiplexed networking is still very much an emerging area where many questions remain to be answered, or even to be appropriately formulated. Some of these questions are discussed in this paper.

Researchers are also working on optical networks based on very short optical pulses and optical time-division multiplexing. The future of such networks is uncertain and no commercial products are in sight. Nevertheless, the area offers great potential and aspects of optical time-division multiplexed networks are also discussed in this paper.

Key Words: wavelength-division multiplexing, time-division multiplexing, optical network, wavelength routing, photonic packet-switching.

1. Introduction

The invention of optical fibre marks a revolution in the history of telecommunications. Compared to copper, the medium which it has mostly replaced today for communication links spanning more than a few hundred meters, it is capable of supporting transmission with greater fidelity, over larger distances, and at higher transmission rates.

Optical fibre as a transmission medium is almost immune to electromagnetic interference and other forms of impairment that signals traversing copper links have to suffer. As a consequence, the signals sent over optical fibre arrive virtually unchanged at the end of the communication link. For the transport of digital information, this translates into an extremely low *bit error rate* or the fraction of bits in error. A typical bit error rate is 10^{-12} and some systems are designed to achieve even lower error rates of 10^{-15}.

A remarkable property of optical fibre is its extremely *low loss*. Thus if a signal with a certain energy is injected into the fibre at one end of a link, this signal retains a substantial fraction of its initial energy even after traversing a considerable length of fibre. A typical value for the loss is 0.25 dB/km which means that a signal retains over 94% of its energy even after traversing a distance of 1 km along the fibre.

We will be concerned exclusively with the transport of digital information (1's and 0's) and this information is transmitted over optical fibre using pulses of light. A 1-bit is represented by the presence of an optical pulse, and a 0-bit by its absence. The duration of these pulses can be extremely short which in turn means that a great many pulses can be transmitted in a second. This translates into a higher transmission rate (bits/second) and this property is described by saying that the fibre has a *large bandwidth*. A typical transmission rate today is 2.5×10^9 bits/second or 2.5 Gb/s, and 10 Gb/s is also in use. Further increases in this rate are limited by the transmission and reception abilities of the sources and sinks of information at the ends of the links—the inherent capability of optical fibre is much higher.

In the mid-1990's, we saw the deployment of *wavelength-division multiplexing* (*WDM*). In a WDM system, multiple signals using non-overlapping frequency or wavelength ranges are transmitted simultaneously over optical fibre. WDM is a solution to two different problems. In places where the availability of fibre is limited, it enables higher transmission rates over a single fibre. Thus the transmission rate over a single fibre can increase beyond the limitation of a few Gb/s imposed by electrical sources and sinks. Secondly, it enables amplifier cost savings for long haul links: a single amplifier can boost all the signals in a fibre. If these signals were transmitted over saparate fibres, each would have required its own amplifier.

After WDM systems were in place, people began to realize that a significant fraction of the cost of the system was in the equipment using the WDM links. For example, the synchronous optical network (SONET) in North America, and its European and Indian counterpart, the synchronous digital hierarchy (SDH), is the transmission standard for telecommunication networks based on optical fibre links. When the optical fibre links employ WDM, it was observed that cost savings in SONET switching equipment could be obtained if some switching and routing capability was introduced at the optical or WDM layer. This led to the second generation of optical networks—the first being those employing optical links only such as SONET and SDH—whose discussion forms the subject to this paper.

2. Multiplexing

Before we go on to discuss optical networks, in this section, we review the concepts involved in multiplexing.

Time division multiplexing (TDM) and WDM are illustrated in Figure 1. In Figure 1(b), N streams of data on wavelengths $\lambda_1, \ldots, \lambda_N$ arriving on different optical fibres are multiplexed on to a single fibre. Contrast this with Figure 1(a) where the *N* streams are at the same wavelength but transmitted at different instants of time. Note that in this case, the pulse width in the multiplexed stream is no more than 1/*N*th of the pulse width in the incoming stream. Since we will be concerned with TDM systems where the input and output streams consist of optical pulses, we will refer to them as optical TDM (OTDM) systems.

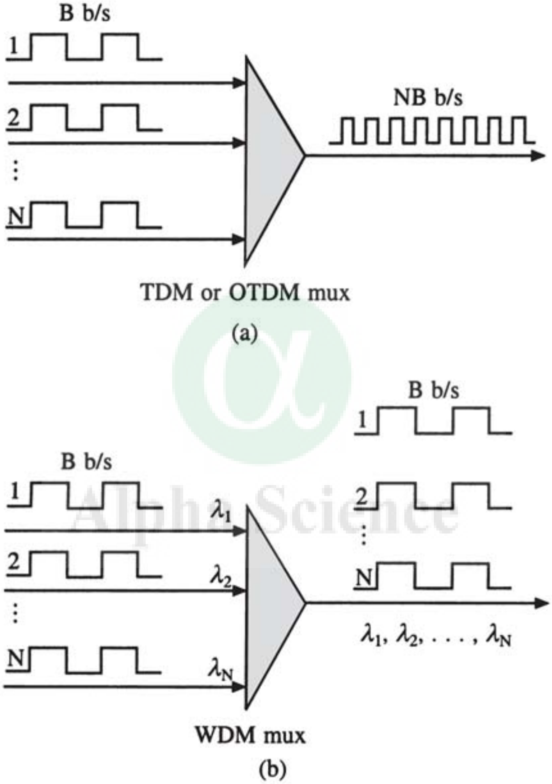


Fig. 1. (a) Optical TDM. When *N* input streams are multiplexed on to an output, the pulses from the various input streams are transmitted at different instants of time. (b) WDM. The pulses from different input streams are transmitted simultaneously but on different wavelengths.

The two fundamentally different forms of TDM—fixed multiplexing and statistical multiplexing—are illustrated in Fig. 2 for the case of two incoming streams. In the case of fixed multiplexing shown in Fig. 2(a), a periodic sequence of pulse positions or slots, is reserved for each incoming stream. The figure shows slots, 1, 2, 5, 6, 9, 10, ... reserved for one stream and the remaining slots for the other. Note that if one stream does not use its slots, for example if the corresponding source does not have anything to send, they still cannot be used by the other stream. In the case of statistical multiplexing, no specific slots are reserved for the incoming streams. Rather, the multiplexer transmits data from an input link to the output link as and when data become available. If only one link has data to transmit, it will (usually) be able to transmit them immediately. If multiple input links have data to transmit simultaneously, they are scheduled by the multiplexer in some order. This kind of multiplexing is particularly efficient when data on the input links arrive in "bursts", that is, there are no data for a while and then for a short period of time there is a relatively large amount of data. In the case of statistical multiplexing, a group of bits or pulses from each input called a *packet* is transmitted together. Hence statistical multiplexing, is also called a *packet*-multiplexing.

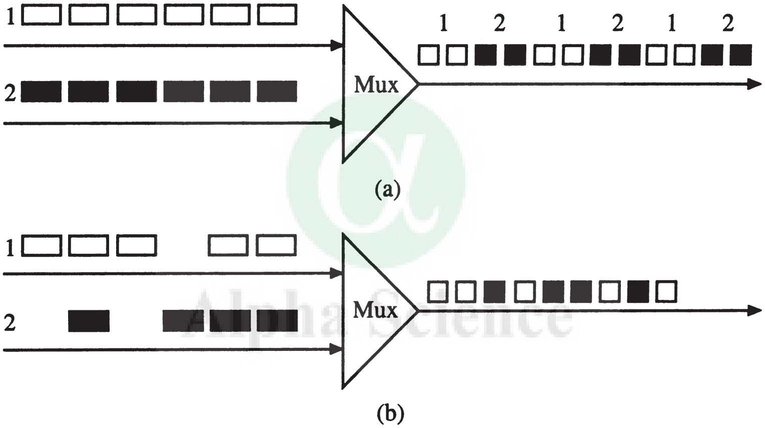


Fig. 2. (a) Fixed multiplexing. The positions of the pulses from a given input in the output stream are fixed. (b) Statistical multiplexing. The pulse positions for each input in the output stream are not fixed; the pulses from the inputs are transmitted as and when they become available.

In the case of fixed multiplexing, the data belonging to a particular incoming stream can simply be identified by their position in the outgoing stream. This is not possible in the case of statistical multiplexing and some identifiers must be carried among the data packets. These identifiers are called packet headers or simply *headers*.

3. Services

Any network can be viewed as offering one or more of three kinds of service: a circuit-switched service, a virtual circuit service and a datagram service. The most familiar example of a network offering a circuit-switched service is the telephone network. In fact, the telephone network offers the user a bandwidth of approximately 4 kHz which is designed to carry speech but can carry anything else the user wants to transmit in the same band such as fax or data. The network delivers the signals in this band from the sender to the receiver with a certain quality, or signal-to-noise ratio (SNR). This SNR is designed such that speech signals can be conveyed with acceptable quality. However fax and data can also be transported over the same network using signals in the same band provided their transmission system is designed to work with the SNR offered by the telephone network. The network itself does not distinguish between voice, fax and data. This brings out another property of the telephone network: it is *transparent* to the user signals.

Virtual circuit and datagram services, are only applicable to purely digital networks, and are usually referred to as *packet-switching*. The asynchronous transfer mode (ATM) network is an example of a virtual circuit network. Like in the case of a circuit-switched network, there is the notion of a connection between the sender and the receiver. However, unlike in the case of circuit-switched networks, transmission resources such as link bandwidth are not reserved in a dedicated fashion for each connection. Rather, the network tries to take advantage of the fact that not all users will have something to send all the time and reserve less network resources than would be required if resources were dedicated to each user. But all data belonging to a connection traverse the same set of links and arrive in sequence at the receiver.

In the case of a datagram service, each slice of data—a set of bits called a pa*cket* or a *datagram*—is independently routed by the network from the source to the destination. Thus each packet must contain all the information required by the network, in particular, at least the full address of the destination. The most familiar, and today, all-pervading, example of a datagram network is the Internet protocol (IP) network.

Second generation optical networks, which we define as networks where the switching and routing of data is accomplished optically, can offer any of the three services, as illustrated in Fig. 3. Today, the only commercially

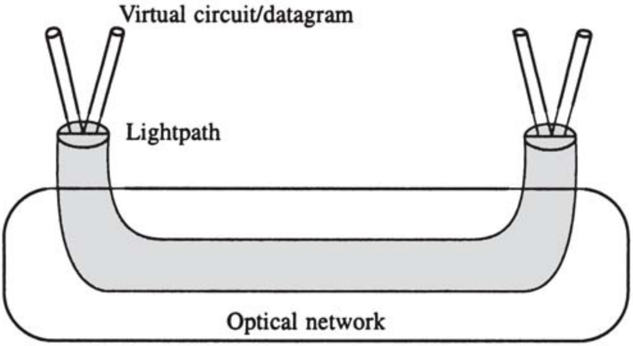


Fig. 3. Services offered by an optical network: circuit-switched, virtual circuit-switched and datagram.

deployed optical networks are those which offer a circuit-switched service. The circuit in this case is termed a *lightpath*. While such networks, in principle, can be based on OTDM (and fixed multiplexing), as well as on WDM, it is the WDM based networks that have been developed and deployed. OTDM networks that offer virtual circuit or datagram services usually employ packet multiplexing.

In WDM networks offering circuit-switched services, the routing or switching of signals within the network is based on the wavelength used by them; hence these networks are also termed *wavelength routing networks*. Like the telephone network, these networks can be designed to be *transparent*, which in this context means that the network doesn't care about the format of the data sent on each wavelength. However, transparency is usually limited, for example, to digital data only.

4. Wavelength Routing Networks

Figure 4 shows a wavelength routing network with five nodes labelled *A-E* and six WDM links. Each link represents a pair of fibres carrying WDM signals in opposite directions. Each lightpath must be assigned a wavelength on each of the links on which it is routed. Moreover, two lightpaths sharing a link cannot be assigned the same wavelength on that link. If W wavelengths are used on each fibre and a node has *M* incoming and outgoing fibres, each node has upto $M \times W$ incoming signals which must be switched or connected to the $M \times W$ possible outgoing signals. Thus each node functions as a *WDM crossconnect* with MW inputs and MW outputs. For simplicity of explanation, we can assume an incoming fibre, in addition to those shown in Fig. 4, is used by lightpaths originating at that node/crossconnect, and similarly, an outgoing fibre is used by lightpaths terminating at that node. With this assumption, for example, the crossconnect at node B has four incoming and four outgoing fibres.

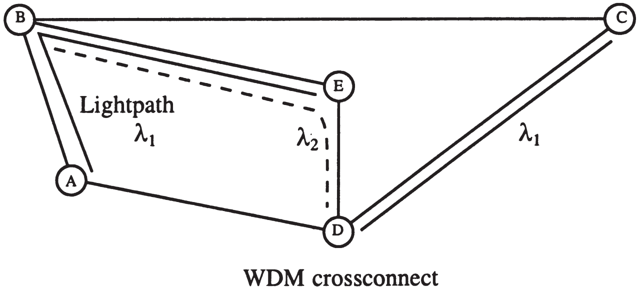


Fig. 4. A WDM wavelength routing network. Lightpaths are circuit-switched connections between pairs of nodes. The WDM crossconnects route the wavelengths from their inputs to their outputs.

A possible internal structure for a WDM crossconnect is shown in Fig. 5. In this crossconnect, the wavelength of each incoming signal is preserved—the crossconnect only controls the output to which the signal is routed. If this type of crossconnect structure is used at each node in the network, the wavelength assigned to a lightpath must be the same on all the links of its route since the crossconnects do not have the ability to change wavelengths. This constraint is called the *wavelength continuity constraint.*

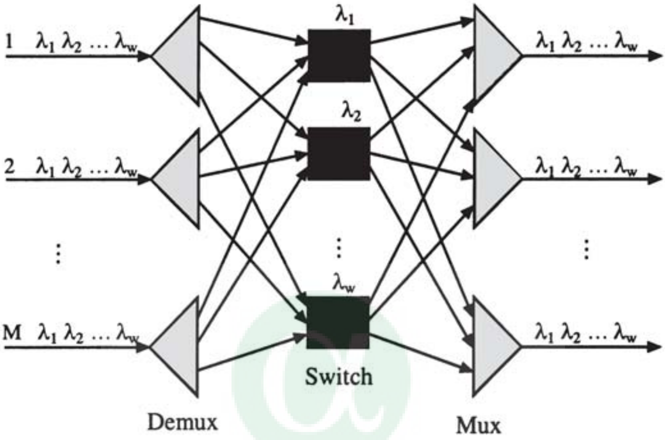

1 $\lambda_1 \lambda_2 \ldots \lambda_w$
2 $\lambda_1 \lambda_2 \ldots \lambda_w$
M $\lambda_1 \lambda_2 \ldots \lambda_w$
λ_1
λ_2
λ_w
Switch
Demux
Mux
$\lambda_1 \lambda_2 \ldots \lambda_w$
$\lambda_1 \lambda_2 \ldots \lambda_w$
$\lambda_1 \lambda_2 \ldots \lambda_w$

Fig. 5. WDM crossconnect with no wavelength changing capability. The wavelength of each lightpath passing through the crossconnect is unchanged.

Crossconnects that are capable of changing the wavelength of a lightpath can be built—they are either expensive or rely on technology which is not yet mature for commercial use. An example of such a crossconnect is shown in Fig. 6. This crossconnect is capable of achieving any interconnection between the MW inputs and the *MW* outputs, provided no two inputs are to be connected to the same output. As can be readily seen, we not only need to use wavelength converting devices but the required switch size is much larger. Large optical switches are very hard to build with today's technology mainly because the losses and crosstalk become intolerably high when the number of ports exceeds a few tens in number.

4.1 Engineering wavelength-routed networks

Considerable research efforts have been directed at determining the impact of the wavelength continuity constraint on the performance of the network. See, for example, Ramaswami and Sivarajan (1995), Barry and Humblet (1996), Birman (1996), Kovacevic and Acampora (1996). These papers attempt to quantify the effect of this constraint in terms of the increase in the lightpath blocking probability due to it, or equivalently, in terms of the number of additional wavelengths required on each fibre to mantain the same level of

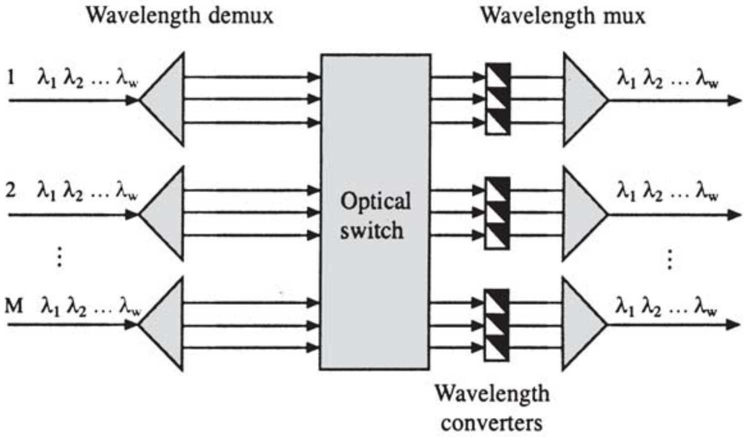

Wavelength demux
Wavelength mux
1 $\lambda_1 \lambda_2 \ldots \lambda_w$
2 $\lambda_1 \lambda_2 \ldots \lambda_w$
M $\lambda_1 \lambda_2 \ldots \lambda_w$
Optical switch
$\lambda_1 \lambda_2 \ldots \lambda_w$
$\lambda_1 \lambda_2 \ldots \lambda_w$
$\lambda_1 \lambda_2 \ldots \lambda_w$
Wavelength converters

Fig. 6. WDM crossconnect with wavelength changing capability. Any interconnection of the *MW* inputs to the *MW* outputs can be achieved.

lightpath blocking probability as would be obtained without this constraint. Thus the wavelength continuity constraint has the effect of increasing the number of wavelength on each fibre and thus the cost of the system. One can get rid of the wavelength continuity constraint by developing crossconnects that are capable of wavelength translation. However, as we have seen, such crossconnects are likely to be more expensive making systems based on them costlier. Thus the cost trade off that must be evaluated by system engineers is between the cost of using crossconnects with wavelength conversion capabilities and using more wavelength on each fibre.

The number of wavelengths required depends on the algorithms used by the network for routing lightpaths and assigning wavelengths to them. The design of routing and wavelength assignment (RWA) algorithms is thus an important area of research. Since optimal algorithms are difficult to design, a number of papers have been written on heuristic RWA algorithms; see, for example Chlamtac *et al.* (1992), Stern *et al.* (1993), Ramaswami and Sivarajan (1995), Sato *et al.* (1995), Birman (1996), and Wauters and December (1996).

The performance of RWA algorithms depends on the nature of the requests for lightpaths. Two classes of models have been considered in the literature on the performance analysis of RWA algorithms. The first is the same model that is used for telephony—the lightpath request arrival process is Poisson and the lightpath's holding time (duration before it is disconnected) has an exponential distribution. Such a model may not be appropriate for lightpath requests which are very high bandwidth connections (Gb/s or more) and also very long-lived. The other model that is used is a deterministic one and usually only considers the wavelength assignment problem. In this model, the routes of the lightpaths are assumed to be known and the number of lightpaths passing through a single fibre link is assumed to be limited to say L. If the wavelength continuity constraint is not present, *L* wavelengths on each fibre are sufficient to support all sets of lightpaths conforming to this model (no more than *L* through a link). With the wavelength continuity constraint, the number of wavelengths required is, in general, more than *L*. The determination of bounds on the number of wavelengths required in various cases has been discussed by Aggarwal *et al.* (1994), Gerstel *et al.* (1994), Gerstel and Kutten (1994), and Ramaswami and Sasaki (1997). However this model has the disadvantage that only a very small fraction of lightpath configurations, among those that satisfy the condition that no more than *L* use any link, may require a number wavelengths close to the maximum. In practice, these few configurations could probably be neglected and the number of wavelengths provided on each fibre, reduced.

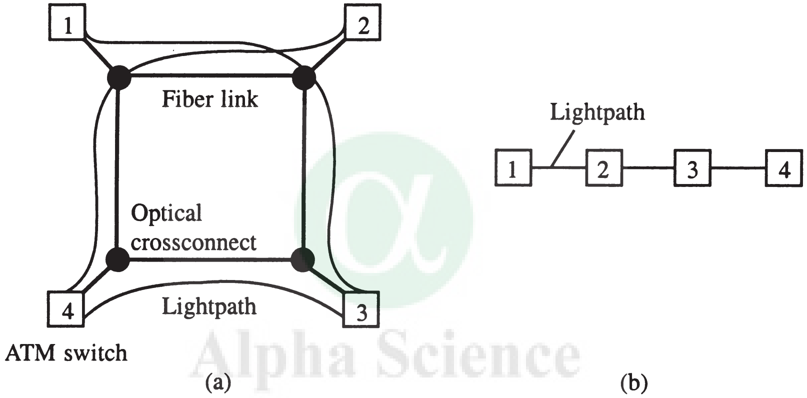


Fig. 7 (a) Physical or fibre topology of the network. (b) Logical or lightpath topology.

The development of appropriate traffic models for lightpath requests is another area (along with the development of RWA algorithms) that should receive the attention of researchers.

4.2 Logical topologies

Wavelength routing optical networks provide a circuit-switched service with some degree of transparency. These networks are beginning to be used today as reconfigurable infrastructures for more conventional networks such as SONET, IP or ATM networks. This concept is illustrated in Fig. 7 for an ATM network but is equally applicable to SONET and IP. In Fig. 7(a), the physical network consisting of optical fibre links and WDM crossconnects is shown. ATM switches set up lightpaths among themselves using this physical optical network. The ATM switches treat each lightpath in exactly the same fashion they would treat a fibre link interconnecting them. Thus the network topology seen by the ATM switches is as shown in Fig. 7(b) and is called the *logical topology* of the network.

If the physical, or optical layer, topology is owned by a telecommunications service provider, and different organisations set up their ATM or IP networks by leasing lightpaths from the service provider, we will have many logical networks embedded on the same physical network. This is the same scenario that exists today with respect to leased telephone lines but with two important differences. The first is that the logical topologies are reconfigurable since the lightpaths can be set up and taken down quite rapidly. Hence the logical topology can be changed in response to traffic demands, or faults. The second is that the lightpaths can be made much more transparent that today's leased lines (which must operate, at least, at a fixed bit-rate) and hence can more easily support a heterogeneous mix of networks.

The design of logical topologies is a subject of continued research; see, for example, Chlamtac *et al.* (1993), Ganz and Wang (1994), Jagannath *et al.* (1995), Mukherjee *et al.* (1996), Ramaswami and Sivarajan (1996), and Krishnaswamy and Sivarajan (1998).

5. Photonic Packet Switching Networks

Wavelength routing networks provide a circuit-switched service but many applications require a packet-switched service such as a virtual circuit or a datagram service. Today, packet-switching is provided only by electronic networks such as IP and ATM which can use the lightpaths provided a wavelength routing optical network in lieu of physical links. Many research efforts, going back to over a decade, have been directed at the problem of attaining a fully optical, or photonic, packet-switching network. We discuss such networks in the remainder of this paper.

5.1 Broadcast WDM networks

The earliest efforts at optical networking were directed at emulating the electronic local area networks such as the Ethernet. These networks are characterized by a broadcast medium (a bus in the case of the Ethernet). While optical bus networks have also been studied, a more convenient broadcast medium in the case of an optical networks is a *star coupler* shown as one of the components in Fig. 8. Consider a star coupler with n inputs and n outputs. The coupler *broadcasts* every input to every output simply by dividing the energy from each input among the outputs, and is usually a passive device. If the signals on each input are at different wavelengths as is the case in the WDM network illustrated in Fig. 8, each output of the star coupler contains the signals at all the wavelengths. If each output is connected to a receiver at a node, these nodes can *select* the desired signal by tuning their receivers to the appropriate wavelength. Hence the term *broadcast and select* is used to describe these networks.

The key networking problem in broadcast networks is that of coordinating the use of the broadcast medium by the users. Without coordination, two users could transmit on the same wavelength simultaneously and their signals would collide. Even if different users were assigned different transmit wavelengths so that collisions are avoided two users could contend for the same receiver, that is, transmit data (packets) for the same user simultaneously. The receiver can only receive one of the signals and the other(s) would be lost. To prevent such collisions and contentions, or to recover from them if they occur, a medium-access (MAC) protocol or algorithm is executed by the nodes.

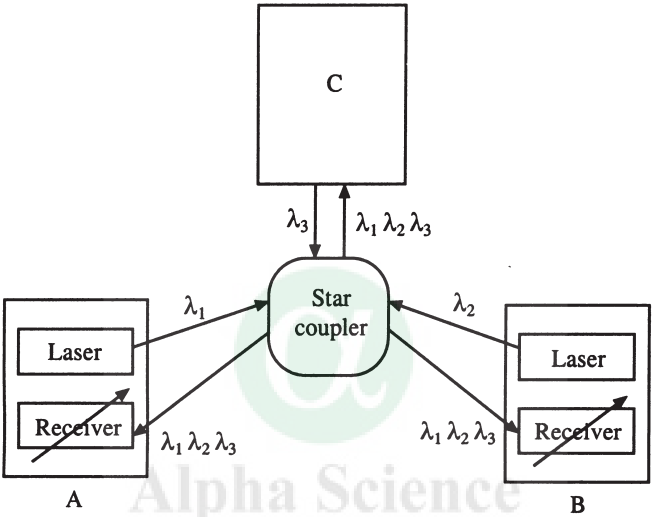

C
λ_3
$\lambda_1\,\lambda_2\,\lambda_3$
Star coupler
λ_1
λ_2
Laser
Receiver
$\lambda_1\,\lambda_2\,\lambda_3$
Laser
Receiver
$\lambda_1\,\lambda_2\,\lambda_3$
A
B

Fig. 8. A WDM broadcast and select network. This example shows each node trasmitting on a different wavelength. The star coupler passively splits the energy from each input among the outputs. Each node can select the signal it wants by tuning its receiver.

Many MAC protocots have been designed for broadcast WDM networks. The key difference between such protocols and those developed for electronic networks is that there are multiple channels or wavelengths that can be used simultaneously. Thus it requires more complicated (or ingenious!) protocols to make efficient use of the available bandwidth. Some papers on this subjects are those by Habbab *et al.* (1987), Chen *et al.* (1990), Mehravari (1990), Chen and Yum (1991), Chipalkatti *et al.* (1992), Humblet *et al.* (1992), Jia and Mukherjee (1992), Li *et al.* (1993), Pieris and Sasaki (1994), Weller and Hajek (1994). Hajek and Weller (1995), and Rouskas and Sivaraman (1997).

5. 2 Broadcast OTDM networks

There have been efforts to demonstrate broadcast OTDM networks along the lines of broadcast WDM networks. However, primarily because of the need to work with very short pulses to get meaningful results, these efforts have been less fully developed than WDM efforts.

Let us consider OTDM networks based on fixed multiplexing (see Figure 2 (a)), one pulse at a time. A block diagram of an optical multiplexer to accomplish this is shown in Fig. 9. A mode-locked laser is a device capable of generating very short optical pulses. Since such a laser is expensive, the output from this laser can be distributed to all the nodes in the system. This is feasible only when the geographical span of the network is limited (in a local-area environment). Each node has to delay the pulse stream from the mode-locked laser appropriately so that when they are subsequently combined by the star coupler, the streams do not overlap. These delays can be achieved by passing the pulse stream through a waveguide of appropriately different lengths at each node. The delayed pulse streams are then modulated according to the data at each node, that is, the pulse is transmitted if a 1-bit is to be sent, and the pulse is suppressed if a 0-bit is to be sent. At bit rates up to a few Gb/s per node, such modulators are available today. Such a modulator is essentially a 1-input, 1-output, on-off switch. The pulse stream is input to the switch and the data determine whether the switch is on or off. A *framing pulse* is needed to identify the pulse positions in the output stream obtained by combining the individual streams from each node. If there are *n* nodes in the network, a framing pulse is added after every *n* pulses such that the pulse immediately following the framing pulse is from node 1, the next pulse is from node 2, and so on.

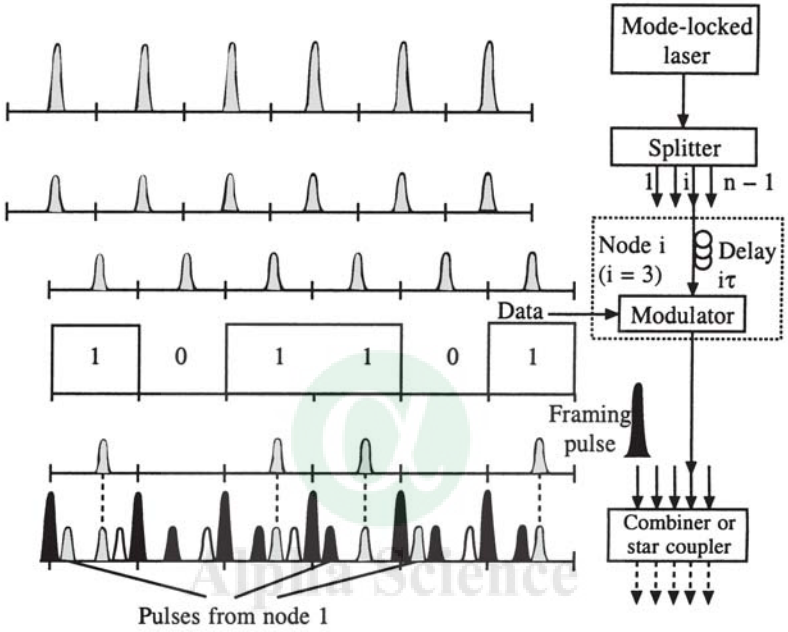


Fig. 9. Optical bit-multiplexing. One bit from each input is transmitted at a time on the output (after Midwinter (1993)).

The corresponding demultiplexer is shown in Fig. 10. In a broadcast network, due to the use of the star coupler, each node gets its individual copy of the multiplexed stream. The framing pulses are extracted by some mechanism. For example, the framing pulses could be transmitted at a higher power level and some thresholding device used at each receiving node to extract the framing pulses alone. This requires the node to make another copy of the received pulse stream using a splitter or a 2-output star coupler, as shown. The final, and most crucial operation, is to recover the desired bit stream from the multiplexed stream. If the bit stream from node *i* is to be recovered, the framing pulses are delayed by *i* pulse durations so that the extracted stream of framing pulses overlaps in time with the data pulse stream to be extracted. Now a device is needed which will output a pulse if and only if a framing pulse and a data pulse are present simultaneously. Such a device is called a *logical AND gate*. If a data pulse is present but a framing pulse is not, this data pulse is from an input stream that is not to be extracted, and the AND gate should not output a pulse. If a framing pulse is present but the data pulse is not, the transmitted bit is a 0, and again the AND gate should not output a pulse.

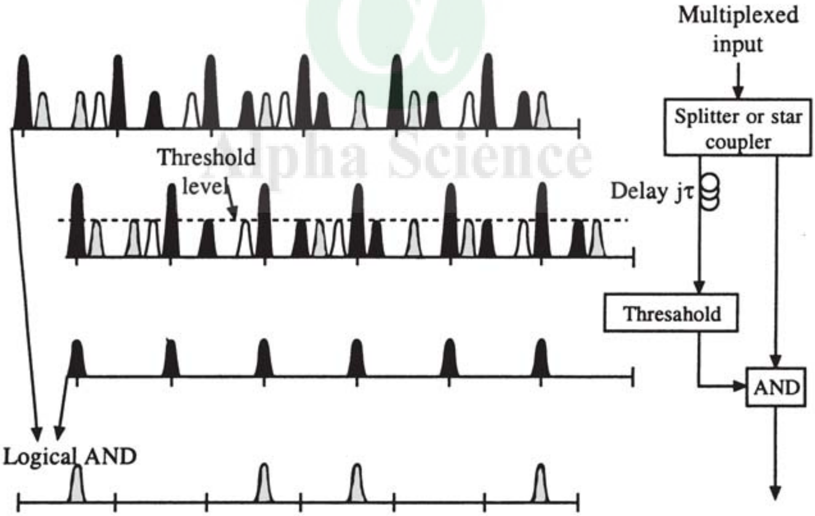


Fig. 10. Optical bit-demultiplexing. A bit-multiplexed input data stream is extracted using the framing pulses and an AND gate (after Midwinter (1993)).

5.2.1 AND gates

Different physical phenomenon have been used to demonstrate the AND gate function required to demultiplex bit streams. We briefly consider two mechanisms for building AND gates. For details, please consult the references below.

The first mechanism we consider is that of a loop mirror and is illustrated in Fig. 11. The loop mirror consists of a piece of optical fibre connected to a 3-dB directional coupler. A 3-dB directional coupler is a 2-input, 2-output device which splits each input equally to each output—a 2 × 2 star coupler. (In fact, a star-coupler can be made by suitably interconnecting a number of such directional couplers.) Assume a pulse is present on one of the inputs of the directionl coupler, say arm A. This pulse is split into two equal pulses which traverse the fibre loop in opposite directions. One output pulse also undergoes a phase shift of 90° with respect to the other. After traversing the same length of fibre, but in opposite directions, the two pulses reach the directional coupler again. The directional coupler is a reciprocal device (it works the same way if its inputs and outputs are interchanged) so that each of these pulses is again split equally to arms A and B. However due to the additional phase shift of 90° introduced now, the two split pulses interfere constructively on arm A and destructively on arm B. Thus the pulse that was input at arm A appears to be reflected, and hence the name *loop mirror* for this configuration.

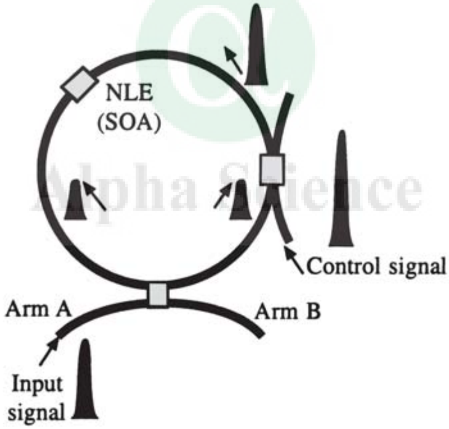


Fig. 11. An optical AND gate using a nonlinear optical loop mirror. The framing pulses serve as the control signal. The data pulses serve as the input signal. A pulse appears on arm B only if both the input signal and control signal are present.

Now if some nonlinear element is introduced in the loop such that one of the pulses undergoes a different phase shift relative to the other, the interference on arm B will no longer be fully destructive and an output pulse will appear. To build an AND gate, we would like to use a non-linearity which itself can be controlled through the presence or absence of a pulse. A semiconductor optical amplifier (SOA) is a device that can be used for this purpose. If a pulse of sufficiently high energy is passed through an SOA, the SOA exhibits nonlinear behaviour so that another pulse immediately following it through the amplifier undergoes a phase shift different from that undergone by a pulse preceding it. The framing pulses can be made to serve as the high energy pulses that drive the SOA to nonlinearity. This requires that the timing of the framing pulses be adjusted such that, say, the clockwise input or data pulse passes through the SOA before it, and the anti-clockwise pulse afterwards.

For further discussions of nonlinear optical loop mirrors and their applications to demultiplexing see, Doran and Wood (1988), Blow *et al.* (1990), Fermann *et al.* (1990), Eiselt (1992), Sokoloff *et al.* (1993), and Kane *et al.* (1994).

Another physical phenomenon that can be used to construct an AND gate is *soliton trapping*. Solitons are pulses with special shapes that have remarkable properties. Normally, pulses transmitted through optical fibre undergo spreading due to a phenomenon called chromatic dispersion—different spectral components of the pulse travel at slightly different speeds in optical fibre. (In many current WDM system, it is chromatic dispersion that sets the limit on the achievable bit-rate and distance combinations.) However, solitons are exceptions and can propagate in optical fibre for very long distances without spreading. Another property of solitons is the following. If two soliton pulses with orthogonal polarizations traverse a length of birefringent fibre (a fibre in which the refractive index for the two polarizations are different), they undergo a wavelength shift in opposite directions. This property is called *soliton trapping*. This phenomenon can be used to construct an AND gate—see Chbat *et al.* (1992)—if one of the polarizations is used for the framing or control pulses, the other for the input or data pulses, and a filter is placed such that it passes a pulse if it undergoes a wavelength shift but not otherwise. This is illustrated in Fig. 12.

While we have discussed only optical bit-multiplexing and demultiplexing, which is a special case of fixed multiplexing, structures to perform packet-multiplexing have also been developed. In these structures too, AND gates are required for demultiplexing. While packet-multiplexing can be used in a broadcast scenario, we will next discuss such networks over a mesh topology, that is, over a topology consisting of a set of nodes interconnected in an arbitrary fashion using fibre links.

Broadcast optical networks, though the earliest to be studied by researchers and prototyped in laboratories, have not seen commercial light, to date. The main reason appears to be a lack of applications requiring the ultra-high per-node bit rates that these networks can provide. Broadcast topologies are suitable deployment only over a relatively small geographical region—say, a campus. However, only the aggregate bit-rate in a campus approaches the per-node bit-rates offered by such networks. And interconnecting campuses requires mesh networks, not broadcast networks. Broadcast networks may find applications in the near future for niche applications, for example, in a film or television studio. If the per-node bit-rate requirements increase beyond the rates which electronic or first-generation optical networks (those based on optical links only) can provide, we may see a renewed interest in these networks.

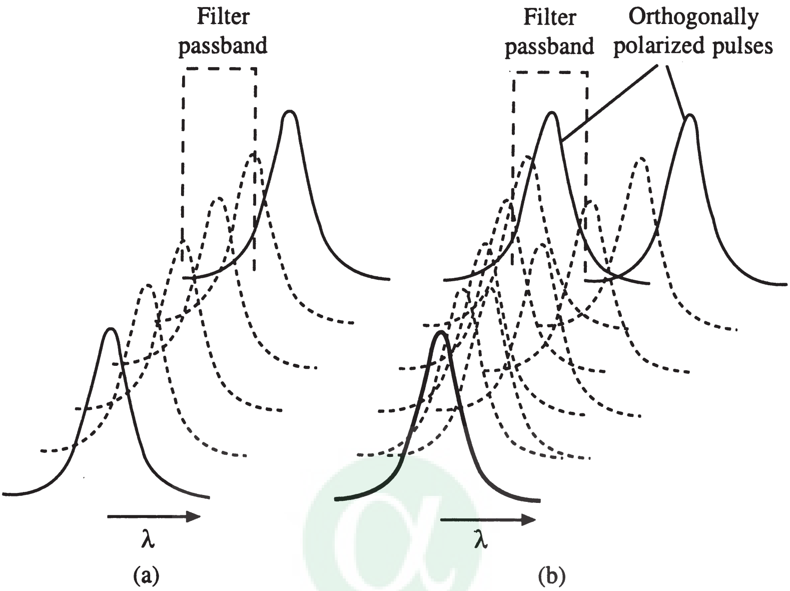


Fig. 12. An optical AND gate using soliton trapping. If both orthogonally polarized pulses are present, they undergo wavelength shifts in opposite directions due to the birefringence of the fibre. If only one pulse is present, it undergoes no wavelength shift.

5.3. OTDM mesh networks

Consider the optical packet-switched mesh network shown in Fig. 13. Each link carries packets bound for different destinations in the network. Each packet carries a header that contains the address of the destination node to which the packet is to be routed. Each node uses the information in the header to route the packets from their input links to the appropriate output links. If the routing is consistent at all the nodes (for example, all the nodes are routing the packet on the same path towards the destination) the packet ultimately gets to its destination. This is how the internet works today. Except that an OTDM network, when it is realized, will work at much much higher link speeds—at Tb/s. (A Tb/s is 10^{12} b/s.) Such networks are likely to represent the long-future of fibre-optic networks. But a number of technologies have to be developed before such networks can be realized. We discuss a few of them now.

5.3.1. Header recognition

In order to be able to route the packets, each node must be able to read the header bits or pulses in the packets, since these carry the destination information.

At Tb/s speeds, this represents a significant challenge. One technique that has been tried in laboratories is to transmit the header at a much lower speed than the data. A structure for address recognition based on a nonlinear optical loop mirror is discussed by Glesk *et al.* (1994).

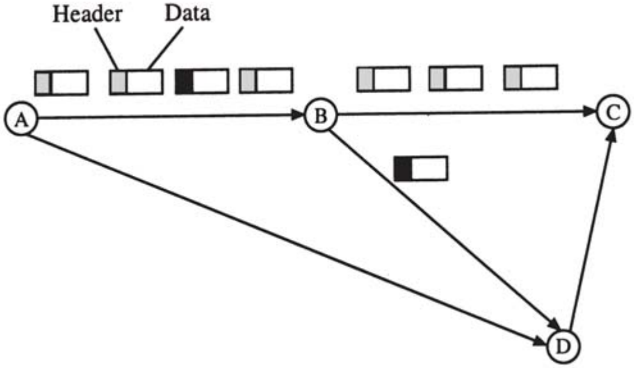


Fig. 13. An optical packet-switched mesh network. Each node routes packets towards its destination whose address is carried in a header.

5.3.2. *Buffering*

Assume two packets arrive simultaneously at a node, to be routed over the same output link. Since only one packet can be sent over the output link at a time, the other packet must be stored, or buffered, temporarily. Buffering of optical pulses is difficult to realize. This difficulty is a major impediment to the development of optical packet-switching networks. Attempts to develop optical buffering are underway; see Hall (1997) for a summary. Today, buffering is realized by using delay lines (fibres of waveguides) of appropriate lengths. However, these do not represent true storage devices. Many attempts have been made to build networks using such redimentary storage, or no storage at all. For example, packets can be misrouted (sent over the wrong link) if they cannot be stored. This technique is called *deflection* or hot-potato routing; see Borgonovo (1995) for an overview. Deflection routing consumes more network resources since packets use longer paths to their destinations. More importantly, deflection routing can be unstable: the packets could be deflected forever. The stability of such routing techniques requires further study.

6. Conclusions

Fibre-optic networking is an emerging area. While fibre-optic links have been deployed for many areas, we are seeing the first deployments of fibre-optic networks, in the form of wavelength-routing networks capable of providing a lightpath service, in 1999. The deployment of such networks is motivated by the cost savings that can be achieved in higher layer switching equipment (SONET/ATM/IP) by the use of wavelength switching in the optical layer.

Since applications use IP and the physical topology is increasingly based on WDM links, running IP over lightpaths provided by the optical WDM network, has become a hot topic of study today.

In the longer term, it is conceivable that optical networks will provide packet-switched services, in addition to circuit-switched services, thus moving into the domain of IP. This would require the development of OTDM packet switched mesh networks. This requires developments in both optical processing for header recognition and optical storage for packet buffering. Both areas are being studied by researchers but much remains to be done and we will probably be well into the next millennium before we see an OTDM IP network.

Acknowledgments

All the figures are from the book by Ramaswami and Sivarajan (1998). (Figures are copyright 1998 Morgan Kaufmann Publishers. All rights reserved.) Many of the issues discussed in this paper are explored in greater depth in this book.

References

1. Aggarwal A. *et al.*, (1994), Efficient routing and scheduling algorithms for optical networks, *Proc. 5th Annual ACMSIAM Symp. Discrete Algorithms*, pp. 412–423.
2. Barry R.A. and Humblet P.A., (1996), Models of blocking probability in all-optical networks with and without wavelength changers *IEEE JSAC/JLT Spl. Issue Optical Networks*, **14(5)**, pp. 858–867.
3. Birman A., (1996), Computing approximate blocking probabilities for a class of optical networks, *IEEE JSAC/JLT Spl. Issue Optical Networks* **14(5)**, pp. 852–857.
4. Blow K.J., Doran N.J. and Nelson B.P., (1990), Demonstration of the nonlinear fibre loop mirror as an ultrafast all-optical demultiplexer, *Electron. Lett.*, **26(14)**, pp. 962–964.
5. Borgonovo F., (1995), Deflection routing; *Routing in Communication Networks*, ed. M. Steenstrup (Englewood Cliffs: Prentice Hall).
6. Chen. M.S., Dono N.R. and Ramaswami R., (1990), A media-access protocol for packet-switched wavelength-division metropolitan area networks, *IEEE J. Sel. Areas Commun.*, **8(6)**, pp 1048–1057.
7. Chen. M. and Yum T.S., (1991), A conflict-free protocol for optical WDMA networks, *Proc. IEEE Globecom*, pp. 1276–1281.
8. Chipalkatti R., Zhang Z. and Acampora A.S., (1992), High-speed communication protocols for optical star networks using WDM, *Proc. IEEE Infocom.*
9. Chbat M.W. *et al.*, (1992), Ultrafast soliton-trapping AND gate; *IEEE/OSA J. Lightwave Technol*, **10(12)**, pp. 2011–2016.
10. Chlamtac I., Ganz A. and Karmi G., (1992), Lightpath communications: An approach to high Bandwidth optical WAN's, *IEEE Trans. Commun*, **40(7)**, pp. 1171–1182.
11. Chlamtac I., Ganz A. and Karmi G., (1993), Lightens: Topologies for high speed opticals networks; *IEEE/OSA J. Lightwave Technol.*, **11(5/6)**, pp. 951–961.

12. Doran N.J. and Wood D., (1988), Nonlinear-optical loop mirror, *Optics Lett.,* **13(1)**, pp. 56–58.
13. Eiselt M., (1992), Optical loop mirror with semiconductor laser amplifier, *Electron. Lett.,* **28(16)**, pp. 1505–1506.
14. Fermann M.E., Haberl F., Hofer M. and Hochreiter H., (1990), Nonlinear amplifying loop mirror, *Optics Lett.,* **15(13)**, pp. 752–754.
15. Ganz A. and Wang X., (1994), Efficient algorithm for virtual topology design in multihop lightwave networks, *IEEE/ACM Tran. Networking,* **2(3)**, pp. 217–225.
16. Gerstel O. and Kutten S., (1997), Dynamic wavelength allocation in all-optical ring networks, *Proc. of IEEE Int. Conf. Commun.*
17. Gerstel O., Sasaki G.H., Kutten S. and Ramaswami R., (1997), Worst-case dynamic wavelength allocation in optical networks, Technical Report RC 20717, IBM Research Division.
18. Glesk I., Sokoloff J.P. and Prucnal P.R., (1994), All-optical address recognition and self-routing in a 250 Gb/s packet-switched network, *Electron. Lett.,* **30(16)**, pp. 1322–1323.
19. Habbab I.M.I., Kavehrad M. and Sundberg C-E. W., (1987), Protocols for very high speed optical fiber local area networks using a passive star topology, *IEEE/OSA J. Lightwave Technol.,* **LT-5(12)**, pp. 1782–1794.
20. Hajek B. and Weller T., (1995), Scheduling nonuniform traffic in a packet switching system with large propagation delay, *IEEE Tran. Info. Th.,* **41(2)**, pp. 358–365.
21. Hall K.L., (1997), All-optical buffers for high-speed slotted TDM networks, *IEEE/LEOS Summer Topical Mtg. Adv. Semicond. Lasers and Appl.,* p. 15.
22. Humblet P.A., Ramaswami R. and Sivarajan K.N., (1992), An efficient communication protocol for high-speed packet switched multichannel networks, *IEEE J. Sel. Areas Commun.,* **11(4)**, pp. 568–578.
23. Jagannath S.V., Bala K. and Mihail M., (1995), Hirarchical design of WDM optical networks for ATM transport, *Proc. IEEE Globecom,* pp. 2188–2194.
24. Jia F. and Mukherjee B., (1992), The receiver collision avoidance (RCA) protocol for a single-hop WDM lightwave network, *Proc. IEEE Int. Conf. Commun.* pp. 6–10.
25. Kane M.G., Glesk I., Sokoloff J.P. and prucnal P.R., (1994), Asymmetric loop mirror: Analysis of an all-optical switch, *Appl. Optics,* **33(29)**, pp. 6833–6842.
26. Kovacevic M. and Acmpora A.S., (1996), On the benefits of wavelength translation in all optical clear-channel networks, *IEEE JSA/JLT Spl. Issue Optical Networks* **14(6)**, pp. 868–880.
27. Krishnaswamy R.M. and Sivarajan K.N., (1998), Design of logical topologies: A linear formation for wavelength routed optical networks; with no wavelength changers, *Proc. IEEE Infocom.*
28. Li C.S., Chen. M.S. and Tong F.F., (1993), POPSMAC: A medium access control strategy for high speed WDM multiaccess networks, *IEEE/OSA J. Lightwave Technol.,* **11(5/6)**, pp. 1066–1077.
29. Mehravari N., (1990), Performance and protocol improvements for very high speed optical fiber local area networks using a passive star topology, *IEEE/OSA J. Lightwave Technol.,* **LT-8**, pp. 520–530.
30. Midwinter J.E., (1993), *Photonics in Switching, Volume II: Systems* chap. 6 (San Diego, Academic Press).
31. Mukherjee B. *et al.,* (1996), Some principles for designing a wide-area optical network, *IEEE/ACM Tran. Networking,* **4(5)**, pp. 684–696.

32. Pieris G.R. and Sasaki G.H., (1994), Scheduling transmissions in WDM broadcast and select networks, *IEEE/ACM Tran. Networking,* **2(2)**, pp. 105–110.
33. Ramaswami R. and Sasaki G.H., (1997), Multiwavelength optical networks with limited wavelength conversion, *Proc. IEEE Info. com.,* pp 490–499.
34. Ramaswami R. and Sivarajan K.N., (1995), Routing and wavelength assignment in all-optical networks, *IEEE/ACM Tran. Networking,* **3(5)**, pp. 489–500.
35. Ramaswami R. and Sivarajan K.N., (1996), Design of logical topologies for wavelength routed optical networks, *IEEE JSAC/JLT Spl. Issue Optical Network,* **14(5)**, pp. 840–851.
36. Ramaswami R. and Sivarajan K.N., (1998), *Optical Networks: A practical perspective* (San Francisco, Morgan Kaufmann Publishers).
37. Rouskas G. and Sivaraman V., (1997), Packet scheduling in broadcast WDM networks with arbitrary transceiver tuning latencies, *IEEE/ACM Tran. Networking,* **5(3)**, pp. 359–370.
38. Sato K.-I., Okamoto S. and watanabe A., (1995), Photonic transport networks based on optical paths, *Int. J. Commun. Systems* (*UK*), **8(6)**, pp. 377–389.
39. Sokoloff J.P., Prucnal P.R., Glesk I. and Kane M., (1993), A terahertz optical asymmetric demultiplexer (TOAD), *IEEE Photon. Technol. Lett.,* **5(7)**, pp. 787–790.
40. Stern T.E., Bala K., Jiang S. and Sharony J., (1993), Linear lightwave networks: Performance issues, *IEEE/OSA J. Lightwave Techn.*, **11(5/6)**, pp. 937–950.
41. Wauters N. and Demeester P., (1996), Design of the optical path layer in multiwavelength cross-connected networks, *IEEE JSAC/JLT Spl. Issue Optical Networks,* **14(6)**, pp. 881–892.
42. Weller T. and Hajek B., (1994), Scheduling non-uniform traffic in a packet switching system with small propagation delay, *Proc. IEEE Info. com.*

3. Indian Telecom and Internet Tangle: What is the Way Out?

Ashok Jhunjhunwala[1], Bhaskar Ramamurthi[1] and Timothy A. Gonsalves[2]

[1]Department of Electrical Engineering

[2]Department of Computer Science & Engineering
Indian Institute of Technology Madras, Chennai 600 036

India aims to increase its telephone and Internet network manifold in the next decade. However, it is not viable to expand the Telecom network in India to such an extent at the prevalent per-line investment. Recent advancements in telecom technologies have been used in the West to enhance services and features while keeping the per-line investment to be nearly constant. In India, the advancements have to be instead geared towards reducing costs. The Telecommunications and Computer network Group at IITM has focussed its efforts in developing several new technologies keeping this in mind. These technologies enable affordable telecom and Internet access in both urban as well as rural areas promising rapid expansion. The new approach however requires openness to new ways of doing things and new organisational initiatives.

Introduction

India's success over the last one and half decades in providing high quality software professionals to the world has prompted us to take a relook at ourselves. The question that is being asked is whether we can sustain this success and make it grow—can IT in India become a force to reckon with in the world? The IT Task Force has recently set a yearly revenue target of Rs. 200,000 crores from this industry. Is it realisable? Can we have a few million people working in this industry in the next five to seven years? Can this industry be used as a spring-board for India to get into the forefront of modern technology?

These and many other similar questions are indeed being raised. But will these goals remain a mere dream or become a reality? This paper looks at one of the many, and probably one of the more important tasks that needs to be carried out for achieving such a goal—the task of providing an affordable telecom and Internet network all over India. Without wide-spread access to the telecom and Internet networks, one cannot even dream of expanding our IT industry.

The Reality

The reality today is grim. We have less than two telephones per hundred population, (as opposed to one for every two persons in the developed world). The Internet, which is the crucial component in today's IT revolution is accessible today only through a telephone. The number of Internet connections in India has barely crossed 400,000. There is a yearning for more telephones and Internet connections. But will this yearning be satisfied?

Not many of us know that today it takes approximately Rs. 35,000 to install a new telephone connection in India (this is per-line cost of the complete telecom network). Assuming 15% as finance charge on such an investment and even a low figure of 15% for operation, maintenance and obsolescence cost, and assuming no other charges such as license fees, connectivity costs etc., it would require 30% of Rs. 35,000 per year, or approximately Rs. 10,500 per year of revenue from each telephone, just to break even. How many in India can afford to pay such a large telephone bill? Is it more than 2–3% of our population? How then can we carry out an IT revolution?

Internet Tangle

Let us look at the way Internet access is provided today. One has to buy a modem and connect a computer to a telephone line. Then one dials an Internet Service Provider's telephone number and gets connected to an ISP router, which connects one to the Internet. This seemingly simple technique has a number of pitfalls.

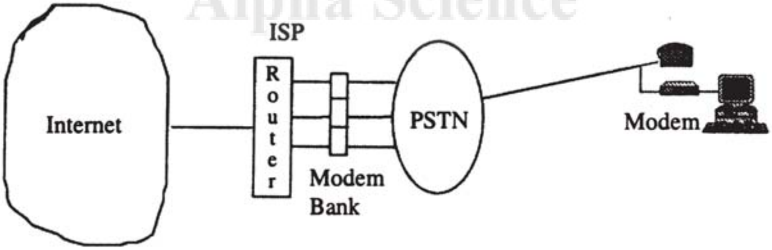


Fig. 1. Internet Access using Today's Telephone Network

(i) Internet sessions last a long time, usually an hour or more. On the other hand, telephone calls last barely a few minutes. The telephone network is designed assuming that a telephone is used on an average approximately only 10% of the time during busy hours. The existing telephone network can not handle a much higher traffic level. As more and more people have long Internet sessions, the telephone network will get totally congested and collapse.

(ii) Today a telephone call costs about Rs. 1.30 for three minutes. This implies that when we use the Internet for an hour, we pay a telephone charge of Rs. 26 per hour. This is the Internet access charge. The charges paid to the ISP are extra and may vary from Rs. 10 to Rs. 20 per hour. But even when ISP charges are reduced, it does not make too much of an impact to the total cost as the base charge of Rs. 26 per hour for access remains. Further this charge is not very visible. It comes seperately, as a part of the telephone bill. It is often not mentioned while discussing Internet service charges.

(iii) The data rate that we get through a modem depends on the telephone lines. Sometimes we may get 33.6 kbps; but very often the rate is 28.8 kbps, 14.4 kbps or even 9.6 kbps. When we dial an ISP from outside a city, the data rate could go as low as 2.4 kbps. All this while we are paying the telephone charges of Rs. 26 per hour in addition to the ISP charges.

(iv) An ISP does not easily get sufficient number of telephone lines for its subscribers to call in. Short of lines, an ISP can handle only a limited number of subscribers at a time. A subscriber finds it very difficult to get connected; the ISP numbers are often busy. The problem is exaggerated when a modem drops the call due to poor line quality, and one has to start all over again.

How do we expect to provide Internet connectivity to "millions" in such a situation? The high investment required to install a telephone line and the problems of providing Internet service on such a telephone network make our task very difficult. What is the solution?

Such complex problems do not have easy answers. The TeNeT (Telecommunication and Computer Network) Group of IITM has been groping hard for some answers to these questions over the past few years. We present in this paper our thoughts on the approach that needs to be taken towards solving such problems. We have pursued this approach over the years and this paper will briefly describe some of the systems and technologies developed under the leadership of the TeNeT Group, which could contribute significantly towards the solution. However, it is appropriate to point out that neither the approach suggested nor the technologies developed are sacrosanct. Many such efforts and multiple such answers are needed. We hope that the efforts of the TeNeT Group will add to other efforts in the country to enable installation of several hundred million telephone and Internet connections in India.

The Approach

The reason that the cost of installing a telephone line in India is close to Rs. 35,000 is because this cost is around $1,000 in the West and we have by and large imported the technology as well as the approach to build the telecom network. In the West innovations were used to bring down the cost per line of the telephone network. However, once the cost came close to $1,000 per line, with the economics and affordability in the developed economies being what it is, telephones became available to most people. It made little economic sense to bring down the cost further at this stage. The R&D efforts were therefore directed to increase value for money, and provide more features and services at the same cost. The emphasis, rightly enough, shifted to providing mobility, higher bandwidth and other features.

The requirements in India are different. Telephones at Rs. 35,000 are affordable only to a few. Our approach therefore has to be different. We need to use technological advances to reduce cost further and make telecom widely affordable in our country. We need to bring down the cost to say, Rs. 10,000 per line. At such a cost, the revenue required per line would be barely Rs. 250 per month, which would be affordable to 100–150 million of our people.

Now, such a drastic reduction of per-line cost while also providing affordable Internet access is difficult and will take time. But with rapid enchancements taking place in telecom technologies, it appears a worthwhile task for the R&D community in India. R&D tasks are by their nature not easy and Indian Telecom professionals should welcome such challenges.

A key to solving the Internet tangle lies in the differences between voice and Internet calls.

First, a little more detailed examination of the telephone and Internet traffic reveals that the average Internet traffic is very unlikely to exceed the voice traffic. During a voice conversation, the traffic is pretty much at full rate at all the time. Therefore a 64 kbps voice call for about 6 minutes in an hour (0.1 E) is expected to generate a total traffic (each way) of $64 \times 60 \times 6$ kbits or approximately 2.9 Mbytes of traffic. The telephone network is designed to handle such traffic per line.

Now, let us look at Internet traffic. Assume the line is used throughout the busy hour. This is the worst case—in reality the data Erlang is about 0.15–0.3 E. But Internet traffic is not constant or continuous. Instead it is bursty. A burst of data is transmitted as a packet at the peak rate of 64 kb/s and then the line is quiet till a response is obtained. Even after the response is obtained, the user takes time examining the received data on screen before transmitting fresh packets. No data is normally received during this time. Rarely is the actual data transmitted, in a long session, more than 10% of the peak rate. Thus taking 10% burstiness, the total data transmitted (each way) on an Internet connection would rarely be more than $64 \times 60 \times 60 \times 0.1$ kbits or 2.9 Mbytes.

This is indeed a revelation. Internet traffic is not more than voice traffic even in the worst situation. Therefore for Internet access at upto 64 kbps, existing telephone network has the traffic carrying capacity. The problem is that the telephone network is circuit-switched. Voice communication from one line occupies a PSTN circuit for only 10% of the time even during the busy hours. An Internet session, on the other hand, could occupy a PSTN circuit nearly 100% of time in a busy hour, whether data is actually being transmitted or not. This could result in excessive congestion and even a collapse of the network.

The solution will be to find a way of not occupying the telecom network resources 100% of time for Internet access. If a way is found to statistically multiplex Internet traffic from several on-line users and then carry the traffic through the circuit switched network, the ubiquitous telephone network would be capable of handling long Internet sessions. The key is that such multiplexing be carried out in the network as close to the users as possible. Such an approach has been used in the systems and technologies developed by the TeNeT group in recent years.

The Technologies

In recent years, developments in the area of fibre optics and microwave radio technology, have reduced the cost of the backbone telecom network to about Rs. 1,000–1,500 per line. Similarly with the access network getting separated from the exchange, the cost of the main exchange today is around Rs.1,200 per line. The key contributor to the telecom network cost today is the access network, which is sometimes as high as Rs. 22,000 in urban areas. The cost in rural areas may be several times higher.

The TeNeT Group and its associates has developed a Fibre Access Network (optiMA), Wireless in Local Loop system (corDECT) and a Direct Internet Access System (DIAS), which aim to significantly reduce access cost and at the same time enable large scale usage of Internet. Using these technologies it is possible to set up a total network today at an average cost of Rs. 16,000 per line. These systems flexibly enable rapid expansion of the telecom and Internet network both in urban and rural areas. Let us first take a brief look at these system technologies.

optiMA Fibre Access Network

This fibre-to-the-curb (or street-corner) system provides one of the most cost-effective means of deploying telephones especially in dense urban areas. The system consists of a Remote Terminal (RT) deployed at street-corners as shown in Fig. 2. The single cabinet with a built in power plant and battery, serves about 500 subscribers and could be located at a PCO or a street-corner kiosk. The subscribers are connected to the RT on copper (either POTS or ISDN) or on wireless (the last 500–800 m). The RTs on the street-corner are connected to a Central Office Mux (COMUX), located at the main exchange premises using a fibre-optic ring as shown in Fig. 3. The ring network enables the system to withstand any single failure of fibre link. The COMUX is connected to the main exchange on standard E1 lines using V5.2 protocol and to the radio exchange (DIU) to serve wireless subscribers. As shown in Fig. 3, a Remote Access Switch with Modems (RASM) connected to the main

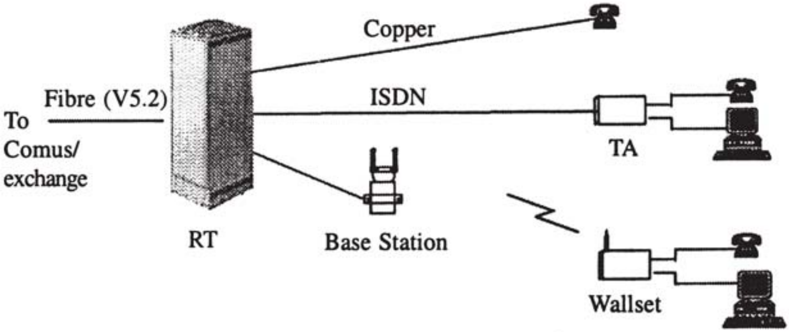


Fig. 2. Remote Terminal at the curb serving subscribers on copper or wireless

exchange on E1 would ensure that Internet traffic from different users are statistically multiplexed before entering the backbone telecom network. The RASM also enables a guaranteed 56 kbps Internet connectivity as the analog portion of the loop is now reduced to 500–800 m. The cost of the POTS access solution using optiMA is astoundingly low, approximately Rs. 7,500 per line. The wireless connection would cost around Rs. 13,000 and the use of multiwallset providing connections to four subscribers in a building would bring down the cost to around Rs. 7,000 per line.

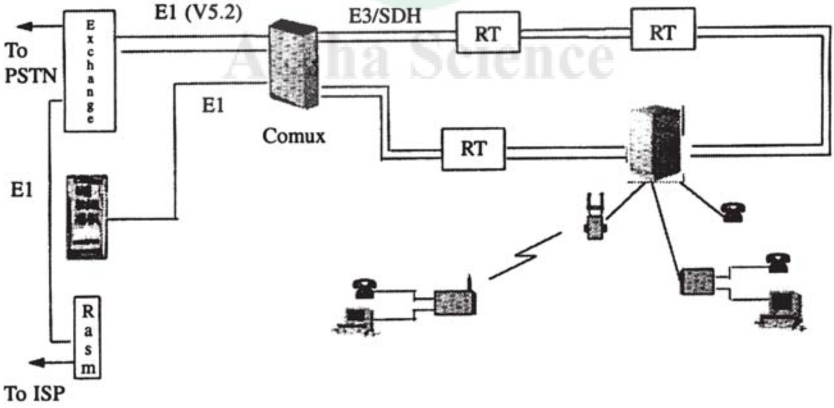


Fig. 3 Several RTs connected on ring network to a central office multiplexers and then to an exchange.

corDECT WLL

corDECT Wireless in Local Loop system, based on the DECT standard, has an interesting architecture, especially for its fixed part. The fixed part consists of a DECT Interface Unit (DIU) acting as a 1000-line wireless switching unit providing a V5.2 interface towards the main exchange, and weather-proof Compact Base Stations (CBSs). These are connected to the DIU on either three pairs of copper wire carrying signal as well as power, or fibre/radio using E1 links through a base station distributor (BSD). The subscriber terminal is a wallset (WS), with either a built-in antenna or a rooftop antenna providing a line-of-sight link to a CBS. The WS has an interface for a standard telephone (or fax machine, modem or payphone) and an additional RS232 interface for a computer, enabling Internet connection at 35 or 70 kbps. No modem is required for Internet access, since all links between the WS and DIU are digital, as shown in Fig. 4.

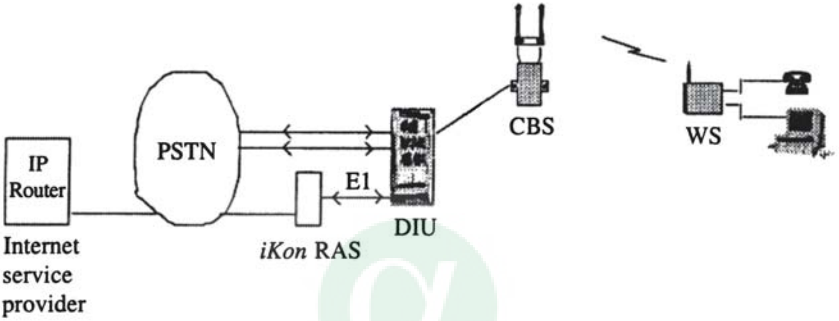


Fig. 4. corDECT Wireless in Local Loop

Efficient transmission of packet-switched Internet data on a circuit-switched network is achieved by combining a RAS with the corDECT system. The connection is digital all the way from subscriber to ISP. The Internet call from a WS to a RAS *does not enter the exchange at all,* but terminates in the access network itself. Only the concentrated IP traffic from the RAS to the ISP traverses through the exchange and PSTN.

All subsystems are built primarily using digital signal processors (DSPs), with the DIU having nearly 100 DSPs. This soft solution, while cutting down development time and affording design flexibility, also ensures that the cost of the fixed part is no more than 15% of the total per-line cost in a fully loaded corDECT system. This allows deployment flexibility for both dense urban and sparse rural areas.

corDECT Deployment Scenarios

A new operator who wishes to initially deploy 5,000 lines in a mid-sized town or city in the very first year would use the deployment scenario shown in Fig. 5. All the DIUs are collocated with the main exchange and connected to it using the V5.2/E1 interface. Each DIU is connected to a BSD located on a rooftop at a suitable part of the town using a point-to-point 8 Mb/s microwave link. At the BSD site, a cluster of about 12–15 CBSs (each serving 50–70 subscribers at 0.1 Erlangs each), along with the microwave equipment, are mounted on a 15 m rooftop tower to serve an area of 2–3 km.

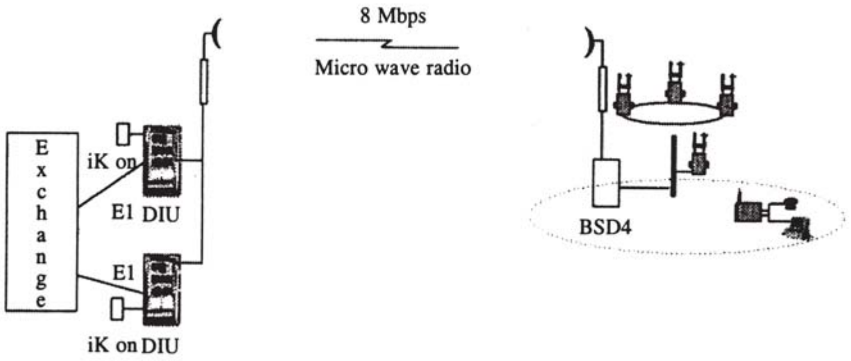


Fig. 5. Rapid corDECT deployment in urban areas

This deployment uses no cables and can be made operational in two to three months at a total deployed cost of Rs. 15,000 per subscriber.

Later, the operator could increase the number of lines by using an optical fibre grid to connect BSDs to the DIUs. A CBS cluster now serves 1,000 subscribers within a 700 m to 1 km radius. Here, many subscriber installations may not need line-of-sight links to the CBS. Once again, the total deployed cost of the access solution is under Rs. 15,000 per subscriber, including the cost of optical fibre cable and cable-laying.

The corDECT system also offers an excellent deployment opportunity for a small town and its surrounding rural areas at a similar cost. The mode of deployment is similar to that in Fig. 5, except that the DIU itself is at the tower base and there is no BSD. To serve about 1,000 subscribers, an operator needs a tower (about 35 m high) in the town centre. The microwave link connects the DIU to the nearest trunk exchange. The base stations now serve subscribers within a radius of 10 km using wallsets with rooftop antennas providing line-of-sight links.

Deployment in sparse rural areas is possible using the corDECT relay base station (RBS). A two-hop DECT link is used to provide connection to the subscriber. One link is from the WS to the RBS, which is mounted on a tower typically 25 m in height. The other DECT link is from the RBS to CBS, which is also mounted on a tall tower (say 40 m). Both the RBS and CBS use high-gain directional antennas, making a 25 km link possible. The 5 km maximum link distance due to the guard-time limitation of DECT is overcome by the use of auto-ranging and timing adjustment. This technique is used in the RBS to support a 25 km link, and to enhance the CBS range to 10 km. This provides a subscriber density as low as 0.5 subscriber/km^2 at a total cost of Rs. 18,000 per line.

Direct Internet Access System

The Direct Internet Access System (DIAS) allows service providers to provide high bandwidth Internet access to residential and corporate subscribers, in addition to voice services, *without any changes to the existing cabling infrastructure.* In contrast to current residential PSTN (Public Switched Telephone Network) and ISDN (Integrated Switched Digital Network) dial-up access, the DIAS provides an *Always On Internet Access* that is permanently available at the customer's premises. Using DSL techniques, seamless voice and data connectivity is provided to the customer over the same pair of copper wires.

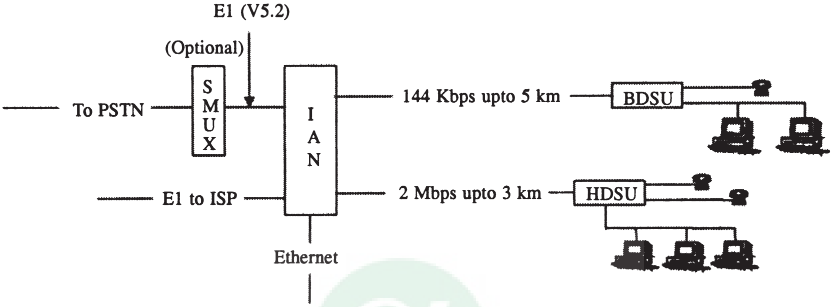


Fig. 6. Direct Internet Access System

Implementation of this system is done using the existing cable plant. All that is required is the installation of the IAN (Integrated Access Node) at the exchange and a DSU (Digital Subscriber Unit) at the customer premises.

The DIAS has a DSU that combines voice and data packets on a single twisted-pair wire at the subscriber's premises. At the service provider's premises, an IAN separates the voice and data traffic from a number of subscribers and routes them independently to the PSTN and the Internet respectively. The IAN is connected to the PSTN via the E1 V5.2 port, and to the Internet either through 2 E1 data ports or through the Ethernet port. Alternatively, the PSTN connectivity can be achieved through the POTS ports of the exchange with the addition of an optional subscriber multiplexer module, that converts a single E1 line into multiple POTS lines.

The DIAS system provides 2 types of voice and data services:

- The BDSU (Basic Digital Subscriber Unit) is designed for the SOHO (Small Office Home Office) and residential Internet user. It provides a permanent Internet connection at a maximum data rate of 144 Kbps, which drops to 80 Kbps when the telephone is in use and transparently goes back to 144 Kbps when the telephone is not in use.
- The HDSU (High bitrate Digital Subscriber Unit), which is designed for corporate subscribers, can provide voice connectivity for upto 8 telephones and a permanent data bandwidth of upto 2 Mbps, which drops by 64 Kbps for each telephone call.

Deployment in Urban and Rural Areas

The technologies and systems described above, used along with standard PDH and SDH fibre-optic and microwave links and widely-available main exchanges and routers, enable flexible planning and rapid commissioning of a telecom and Internet network in both urban and rural areas. The examples described here will illustrate this.

But before we proceed, let us describe what we call an *Access Centre (AC)*. The Access Centre is a self-contained unit which could provide voice and Internet connectivity to subscribers using either plain old copper, wireless or using DSL on copper. It combines corDECT, optiMA and DIAS access systems today and has the capacity of including other access systems tomorrow. Fig. 7 shows one such Access Centre. It is a single cabinet with a built-in power plant and battery back-up. corDECT BSDs are mounted in the cabinet to drive base stations and thereby provide wireless connectivity to subscribers. OptiMA RTs or, Versatile Remote Unit (VRU) provide POTS connectivity within a radius of a kilometre. Similarly DIAS IANs provide voice as well as Internet connectivity using BDSUs or HDSUs. Some Access Centres would also contain a DIU to which the BSDs in other ACs are parented. The Access Centre has been designed to be housed in a small 100 sq.ft. room and could easily be used by a franchise operator to provide service in a neighbourhood.

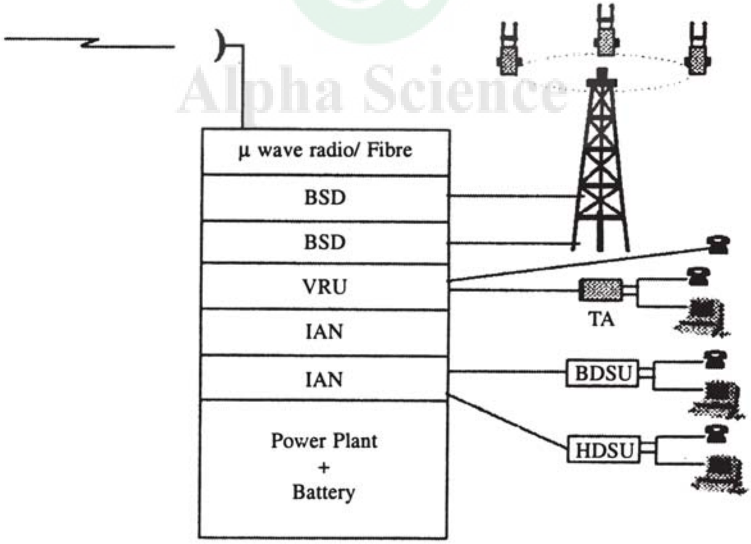


Fig. 7. Access Centre combining corDECT, optiMA and DIAS

A telecom and Internet network would contain a Switching Centre connected to several Access Centres using either PDH/SDH microwave or fibre links connected as shown in Fig. 8 or as a ring. The Switching Centre would

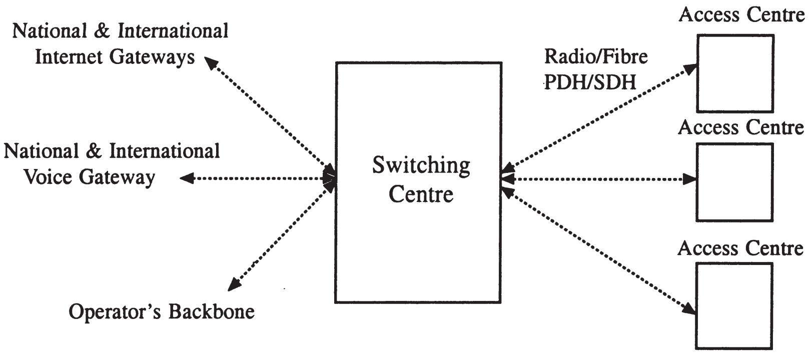


Fig. 8. A Switching centre serving multiple Access centre in fibre optic or successive digital units

consist of a main exchange, Internet routers, some Remote Access Switches and a Network Management system to manage the network. The Switching Centre would be connected to a national and International Internet network, national and international voice network as well as to the other switching centers of the operator located in other cities/districts.

The network described above is versatile, flexible and expandable. It provides different kinds of services including voice telephony, dial-up Internet connection and permanent Internet connections. The network can be deployed in large cities, in small towns and in sparse rural areas using the corDECT Relay Base Station.

Fig. 9 shows the typical deployment in a large urban centre. Access Centre are located at about 20 places, each serving roughly an area of 1.6 × 1.6 km.

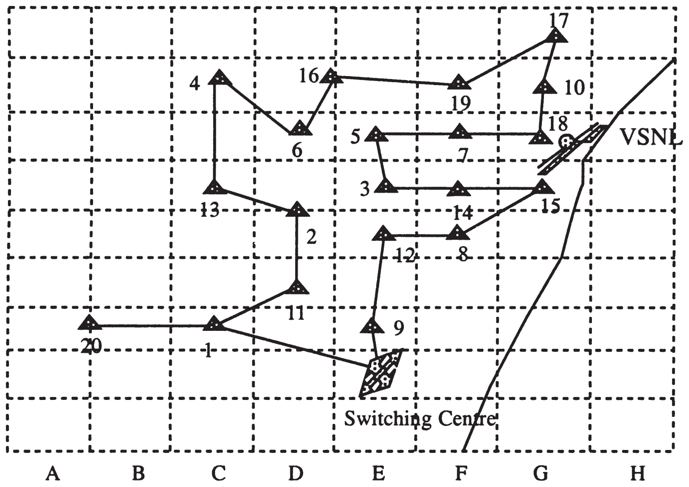


Fig. 9. Access Centres serving 1.6 km × 1.6 km grid and connected to a Switching Centre

About 1,000 subscribers are provided connectivity from each Access Centre. To start service quickly, a few ACs can be connected to the Switching Centre using point-to-point 8 Mbps radio links. In the mean-time, the work on a fibre-optic SDH ring network would commence to connect all the AC to the Switching Centre. It is possible to provide 20,000 subscribers connectivity in this manner in about a year at a cost of Rs. 20,000 per line. With larger deployment, the cost would decrease. It is important to note that this network not only provides voice but also Internet connectivity *without loading the telephone network*. This is because most of the Internet traffic is separated at the Access Centres and carried from thereon on a packet-switched network.

Figure 10 shows a plan of a similar network designed for a rural area. Thanjavur is one of the better-off districts of Tamil Nadu with rich agricultural land. The proposed network consists of two Switching Centres located at Thanjavur and Kumbakonam cities. Several Access Centres are located in small towns all over the district. Each AC serves not only the town, but also the surrounding rural areas. By using Relay Base Stations, a subscriber can be provided both voice as well as Internet connection as far as 25 kms from the town. The ACs are connected to the Switching Centres using point-to-point microwave radio link. Once again, about 15,000 subscribers can be served all over the district at a cost around Rs. 20,000 per line. The network can be set up in about a year.

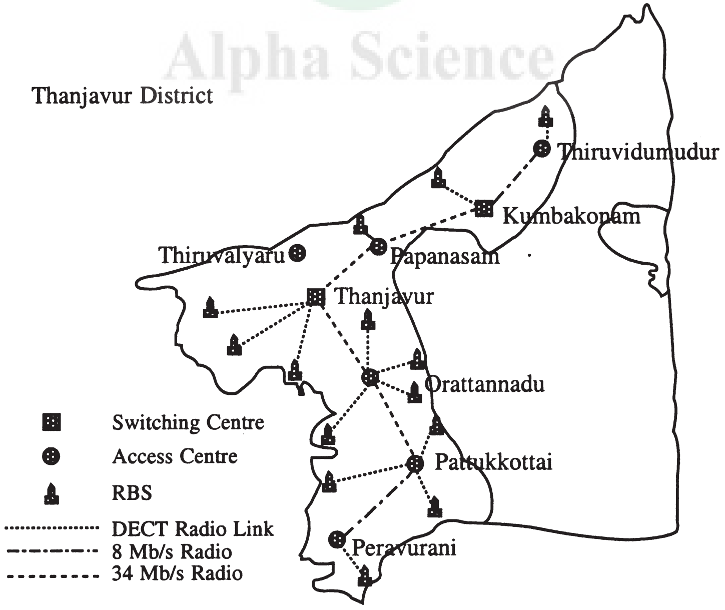


Fig. 10. Telecome network plan for Thanjavur District in Tamilnadu, India

Conclusion

For voice and Internet connectivity to become widely affordable and available, countries like India need a telecom network that costs much less than what prevails today. Such cost reductions are possible if one innovatively utilises the continuing technological developments in this area. Indian scientists and technologists have the capability of taking up such a challenge. Besides, if one is able to reduce the cost considerably, the market size in developing countries is huge. In fact, it is much larger than the current market in the developed world. The Government, the planning bodies, industry and R&D organisations have to get together to take-up such a challenge. The task will not only bring out the very best from within us, it will also be exciting and sufficiently rewarding. Further, success in this endeavour could propel India into the front-ranking countries of the world.

References

1. Jhunjhunwala, A., Ramamurthi, B. and Gonsalves, T.A., (1998), “The Role of Technology in Telecom Expansion in India, *IEEE Commun. Mag.*, **36**(11), pp. 88–94.
2. Jhunjhunwala, A., (1998), Can Telecom and IT be for the Disadvantage? *Rural Development,* **17**(2), pp. 321–327.
3. Morgan, S., (1998), The Internet and the Local Telephone Network: Conflicts and opportunities, *IEEE Commun. Mag.*, **36**(1), pp. 42–48.

4. Satellite Communications and its Impact on Society

K. Kasturirangan and S. Rangarajan
Antariksh Bhavan, New B.E.L. Road, Bangalore 560 094

Communication satellites have emerged as an important media for meeting the telecommunications, broadcasting and mobile communication requirements of the modern Society. Their applications extend from telephony, entertainment T.V. to development communication and disaster warning. Evolution of Satellite Communications and different elements of Satellite Communication are discussed in this paper. Applications of Satellite Communications in distance education, health care and disaster management are discussed. The Indian National Satellite System (INSAT) is described. The impact of Satellite Communication on the Indian Society as well as the global community is also discussed in the paper. Finally, the growth in Satellite Communication is discussed with emphasis on future directions.

1. Introduction

Communication requirements of modern societies are summarized as "instantaneous connectivity with anybody, anywhere and at any time" Communication via satellites helps to meet the above requirements. In addition to meeting the basic telecommunication requirements of regional and global societies, satellite communication systems play a major role in broadcasting. Satellites form an ideal medium to provide video, audio and data broadcast over large regions. Their applications include entertainment TV, distance education, news and document dessimination as well as the emerging internet applications. Dramatic changes are taking place not only in the technology and services, but also in policies, regulations and financing. Market forces dominate the direction and speed of change. Satellite communication is all set to enter a new vista viz, providing the 'last mile' connectivity to homes or offices. There are about one thousand communication satellites in the Geo-Stationary Orbit (GSO) providing telecommunication and broadcasting across the globe. Many Low Earth Orbiting (LEO) satellites are also either in the planning or initial operational stages to provide Global Mobile Satellite Communications. Many advances have taken place in the area of digital satellite communications which have enabled reduction of ground terminal sizes and an increase in number of TV channels. Developments in satellite communication has ushered in the global communication era and the widest possible reach for TV broadcast. Apart from entertainment, these technologies have a major role to play in developing countries like India in the area of distance education/training, telemedicine, emergency communications and disaster warning. Indian Space Research Organisation (ISRO) through the Indian National Satellite System (INSAT) is contributing significantly towards these developments.

2. Evolution of Satellite Communication

Arthur Clarke first pointed out that an artificial satellite at a distance of about 36000 km from earth would remain relatively stationary with respect to a rotating earth so as to provide a radio relay to more than one third of the globe. The successful relay of TV signals across the Atlantic in 1962 using Telestar, the first active repeater communication satellite, ushered the satellite communication era. By 1964, Syncom III, an experimental satellite, had been placed in the geostationary orbit. In the same year, INTELSAT, an international consortium having the objective of setting up a global satellite network for fixed telecommunications, came into being. By July 1965, the 'Early Bird' satellite (INTELSAT-1) was in operation, relaying telephone calls between Europe and North America. The remarkable developments in space communication in just three decades since then, has brought us to the threshold of achieving the capability of establishing human connectivity any- where in the world—on land, air or sea. The inherent advantage of satellite communication systems, which can cover wide areas from their vantage point in space and establish connectivity even with distant and inaccessible areas, makes them ideal for point to multi-point, and for multi-point to point applications. The superior quality and reliability of satellite links in combination with their high percentage of availability, have made them attractive even for point to point links. Satellite networks not only offer a high degree of flexibility to meet changing needs through reconfiguration, but also have the distinct advantage over other media, because of their ability to aggregate small requirements spread across vast territories to provide cost-effective specialised services.

The evolutionary nature of satellite communication is reflected in their capacity increase from just 240 voice ch, which on an average, can carry over 20,000 voice channels in 1965, to the present day satellites circuits, in addition to several TV channels. The shift from voice to data and video via satellite is given in Fig. 1. This remarkable growth is well reflected in the growth of INTELSAT, which has gone through seven generations in just three decades, and is now operating more than 900 earth stations located in over 180 countries. The emergence of many domestic satellite systems, including those of the developing countries like India in the last 15 years, has made satellite communication revolution a global phenomenon. Satellite systems have also become an important medium for the distribution of television programme material to terrestrial broadcasting stations, cable systems and more recently, direct to home TV services. During the eighties, a substantial market has also developed for satellite network operating with small, low-cost earth stations, or Very Small Aperture Terminals (VSATs) located in the users' premises.

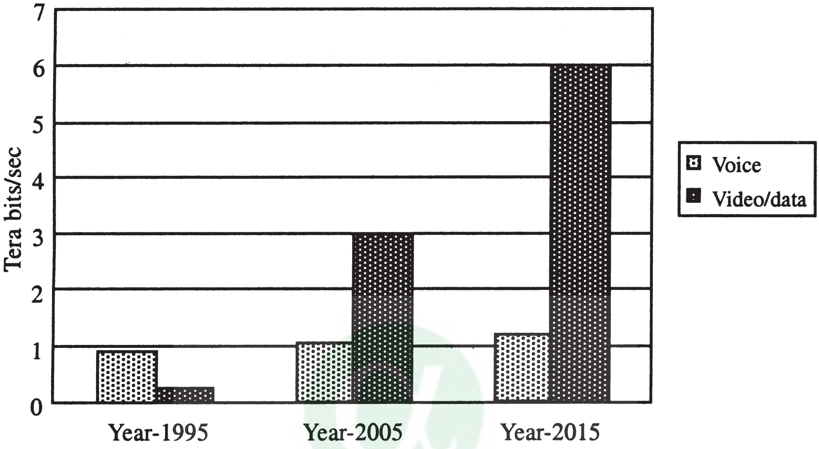


Fig. 1. Shift from voice to data & video

For a large number of developing countries, where the existing investment in ground infrastructure is low, it is only the satellite communication which can enable them to leap-frog into the new era without going through the conventional step-by-step technological process. In many large developing countries like India, China, Indonesia and Arab States, satellite technology has already proved to be most cost-effective for providing nationwide or even provincewide communication facilities. The immense potential of TV broadcast for imparting education in health, hygiene, better agricultural practices and family planning, particularly to remote rural areas, is now within the reach of all developing countries for transforming the life style of their entire rural population in a short time.

On the Mobile Satellite Services (MSS) front, after early experiments with mobile earth stations as part of NASA's Applications Technology Satellite (ATS) programme and MARISAT, the International Maritime Satellite Organisation (INMARSAT), an inter-governmental body, was established in 1979 to provide global maritime mobile satellite services. INMARSAT has grown from the initial 27 member countries to more than 75 member countries, and has now expanded its mandate to provide international aeronautical as well as land mobile satellite services (AMSS and LMSS). Apart from INMARSAT, a number of national systems such as Optus (Australia), AMSC (USA) MSAT (Canada), Solidaridad (Mexico) and INSAT (India) are also expanding to meet their domestic requirements even in the mobile communication area. Over the past few years, the concept of a constellation of low earth orbiting satellites to provide personal communication services through hand held phones has been gathering momentum, with IRIDIUM already operational and several other systems having been proposed for operationalisation in the next 2–3 years. Details of proposed global satellite personal communication systems are indicated in Table 1.

Table 1. New global satellite PCS systems

Parameter	Iridium	Globalstar	ICO-Global
No. of active satellite	66 + 6 spare	48	10 + 2 spare
No. of satellites per orbit plane	11	8	5
No. of orbit planes	6	6	2
Orbit attitude (km)	750	1,414	10,355
Orbit inclination	86.5	52	45
Number of spot beams/satellite	48	16	163
Reported cost ($billion)	4.7	2.5	4.6

Communication satellites are also being advantageously used in allied areas such as navigation, search and rescue and disaster management. Satellite-aided search and rescue service has become an integral part of the worldwide search and rescue system, because of its capability to locate the distress site accurately, thereby drastically reducing the time required for initiating search and rescue efforts and increasing the survival chances of people in distress. The low earth orbiting COSPAS-SARSAT system and the geostationary systems such as INSAT and GOES, are now operationally providing this valuable humanitarian service.

3. Elements of Satellite Communication

In order to derive maximum benefit from satellite communication systems, which have proved their attractiveness for providing several satellite services, International Telecommunications Union (ITU) as the regulatory body has divided these services into three broad categories viz:

(i) **Fixed Satellite Service (FSS)** which provides radio-communication service via one or more satellites between fixed points on earth. FSS covers services like telephony, telex, data and fax transmission, audio/video program distribution. The various constituents of the FSS systems are the space segment, satellite control centre, ground segment consisting of earth stations and network control and back-haul links. There is a wide choice for frequency bands of operation, transmission techniques and network parameters.

(ii) **Broadcast Satellite Service (BSS)** for direct reception of satellite broadcast by the general public either for individual or for community reception. BSS is a point to multi-point service for transmission of TV and audio signals via satellite to be received by small inexpensive receive-only terminals. BSS system essentially consists of the satellite and the control centre, TV uplink stations and back-haul links, and TV Receive-only terminals.

(iii) **Mobile Satellite Service (MSS)** which facilitates communication between mobile earth stations and fixed users or between different mobile earth stations via satellites. MSS services are catagorized into Land Mobile Satellite Service (LMSS), Maritime Mobile Satellite Service (MMSS) and Aeronautical Mobile Satellite Service (AMSS).

Projected growth of investment in various satellite systems have been grouped under various service categories in Table 2.

Table 2. Profile of satellite systems by service category

Service	Service Description	Year-1996	Projected 2005
FSS	conventional	$14 billion	$29.5 billion
Broadcast	DTH, DBS	$3	$17
Multimedia, broadband	Internet access, multicast...	—	$13
Mobile	Maritime, aero, global & regional	$2.5	$12.5
Other	Store/forward/paging	$0.2	$2.5
Total services		$19.7	$74.5

4. Applications of Satellite Communication

When satellite communications were introduced in the 1960s, it was considered an appropriate solution for international communications. However, the phenomenal developments in satellite communications over the past three decades have led to the establishment of large number of regional and domestic systems, including those of several developing countries, providing a host of telecommunications and broadcasting services. Satellite communications is the key technology that could bring the developing countries to participate in the build-up of the Global Information Infrastructure (GII). Satellite communications systems, featured by large footprints from country to continent size, avoid the need for terrestrial infrastructure, and shorten the time for establishing basic and advanced communications in rural and remote areas.

Satellite system have important unique features that include: (i) mobility—mobile users cannot be connected to the fibre network directly, (ii) flexibility—once a terrestrial infrastructure is built, it is extremely expensive to restructure it, and (iii) rural and remote connections—it is still not cost-effective to deploy high-capacity fibre networks in areas with low-density traffic and difficult topography, and (iv) broadcast (point-to-multipoint) capability—reaching millions simultaneously in a cost-effective way. Thus, satellites and wireless technologies will be important in the future implementation of the GII and for the National Information Infrastructure.

New satellite communications technology could intervene, in particular, in rural, low density traffic areas. Rural customers now cost between 10 and 30 times as much to serve with telephone wires as urban customers. Developing countries have only 1 to 3 per cent of the telephones that industrialized countries have, and only 10 per cent of the television sets. Satellite communication can dramatically improve the telecommunication and broadcast scenario of rural societies.

Although nearly 100 percent of the broadcasting industry employs satellite technology, broadcasting only comprises approximately 30 per cent of the communications 'satellites' overall traffic. At the end of 1996, there were over 1500 transponders (36 MHz equivalent) worldwide for television relay. By 2000, the number of television broadcasting transponders could grow to 3350 transponders. Worldwide, direct to-home (DTH) broadcasting is expected to meet the needs of more than 800 million TV households and more than two billion people are now without access to television.

Radio is the most ubiquitous communications medium in the world. There are over 2 billion radio sets in the world and over 100 million sets are sold every year. Satellite based Radio Networking (RN) is used for distributing national radio channels through local radio stations for better quality audio. Satellite based sound broadcast to small portable radio receivers is planned from current year on a global scale.

Very Small Aperture Terminals (VSATs) are proven technology for wide-area networking. They are designed for asymmetric traffic, are capable of simultaneous broadcasting to multiple users and are compatible with the internet's TCP/IP protocol. Generally, VSAT system, design aims to minimize the satellite power and bandwidth requirements while meeting the network performance requirements. This has resulted in different VSAT architectures that are optimized for data, voice or video applications. Typical Ku band VSAT terminals are 1.0 to 1.2 metres in diameter. These systems provide data access at 64 kbps up to 13.5 Mbps for business customers. Currently the total number of terminals worldwide is around 250,000. VSAT rural telephony products use existing satellites and apparently are most cost-effective than alternatives such as planned LEO/MEO and geostationary satellite mobile phone systems. There is also a big market for VSAT—based paging services. VSATs will allow communities to connect to a hub station or directly to the international networks, giving instant access to both basic telephony and more advanced applications—ranging from simple internet to full multimedia applications such as telemedicine and distance education.

New proposed services via satellite include voice, data, video, imaging, video teleconferencing, interactive video, TV broadcast, multimedia, global internet, Messaging, and trunking. A wide range of applications is planned through these services, including distance learning, corporate training, collaborative work groups, telecommuting, telemedicine, wireless backbone interconnection (i.e. wireless Local Area Network-Wide Area Network), video distribution, direct-to-home video, and satellite news gathering, as well as the distribution of software, music, scientific data, and global financial and weather information. Satellite-based systems are also indispensable for emergency communications services. Satellites also provide paging and messaging services. The largest space-based data network has sold over 175,000 terminals. This service is operated by Omnitracs, Euteltracs and related systems around the world. This system has been in operation since 1987, but a number of other space-based systems including the "Little LEOs" are also entering this service.

In the early 1990s it became apparent that digital technology could be efficiently and cost-effectively used for the delivery of television and audio signals. Of particular interest was the possibility of delivering many more channels over the same infrastructure (cable TV, satellite transponders, terrestrial spectrum) by using digital compression rather than existing analogue transmission. The arrival of digital radio offers exciting possibilities for the combination of radio and images, or links to internet sites, marketing compact disks or tickets etc.

Recent technological developments have enabled of a new type of satellite communications systems which is small, relatively cheap to manufacture and launch, and capable of orbiting much closer to the Earth. These new kinds of systems are generally known as Global Mobile Personal Communications by Satellite (GMPCS), an acronym which in fact encompasses a range of systems, some of them based on existing geostationary satellite technologies. They do have enormous potential to change the way in which the world communicates. At the moment, three major systems (Iridium, Globalstar and ICO) are in the advanced stage of implementation: all of these are expected to become fully operational by 2000.

5. Satellite Based Educational Systems

Satellite based audio-visual delivery systems have immense potential in transforming entire societies. This is particularly true for isolated remote rural area populations because of the immense reach of satellite media. Hence distance education and development communication have become important application areas of satellite communication. It was realised that satellite TV is a tool that could bring direct benefit to millions of people who are in need of the basic necessity of civilised living by providing them access to information and education. The fundamental question was how best broadcasting of sound or television could be effectively utilised for imparting formal as well as non-formal education in a cost-effective manner. The technology one adopts for carrying out a given task is obviously dependent upon the magnitude of the task itself and the base with which one starts. What is possible in a highly developed society with a high level of literacy rate may not be the proper solution applicable to nations having a massive illiterate population. It is equally true that fundamental requisites for proper education such as student-teacher interaction, learning by experimentation, sustaining literacy by reading, and even continuous training of teachers and industrial workers to overcome technological obsolescence, cannot be ignored in any society. The complexity of the educational issue is further compounded by the need to take into account geographical and linguistic specifics as well as religious and cultural preferences, to ensure wide acceptability of the technological solution by the large illiterate population in a country like ours.

While it was indeed a very creditable move on the part of space scientists to suggest the use of space technology for solving the problem of education in the developing countries, the process of practical implementation of educational projects all over the world has unfortunately not received the attention it deserves. By and large, the conventional revenue-earning applications of space technology such as telecommunication applications, TV broadcasting for entertainment, national and international programme exchange and military applications have taken precedence over educational needs. Even though satellite-based education has been kept alive by limited experiments and modest operational systems, the size and scope of these systems have yet to make a truly significant worldwide impact. While initial attempts were primarily restricted to mass education in the rural and underdeveloped areas, with the experience gained over the years, some attempts have been made to extend satellite education to cover both training and formal education. The initial system planning was based on the concept of direct satellite TV broadcast transmission of educational programmes dealing with health, hygiene, family planning, better agricultural practices and national integration to remote areas having little or no terrestrial infrastructure. Inexpensive community receivers provided by the government and located in each village or a hamlet, where individual homes could not afford the cost of such sets, proved to be the only alternative to serve the poor rural population. In technical parlance, this was called *technology inversion*, which made the spacecraft powerful, technologically sophisticated and expensive, keeping the cost of direct reception ground segment, which runs into millions of installations, simple and inexpensive.

6. Health Care and Telemedicine

Developments in satellite communication can provide a very wide spectrum of medical services and assistance to remote area population-ranging from dissemination of basic health care information, to rural medicos on paediatrics, internal medicine, orthopaedics, ophthalmology, gynaecology, nutrition and physical therapy, in assisting medical doctors even in well-equipped hospitals in performing difficult surgical operations through interaction with specialists elsewhere and providing timely assistance to the disaster-affected people. The use of ATS-6 satellite for the dissemination of basic health care information to the rural population in India and Alaska, the use of COSPAS-SARSAT search and rescue system which has already saved over 2000 lives since 1980, assistance to the Mexican earthquake victims in 1985, and the use of locale-specific disaster warning system in India using INSAT for saving thousands of lives and livestock during cyclone or flood disasters are but a few examples of the use of disaster-warning capability. The establishment of Satellite for Health and Rural Education (SHARE) project by INTELSAT in the 1980s, medically connecting many developing nations in Africa and over 30 nations of Latin America with sophisticated facilities available in well-equipped hospitals elsewhere, is the beginning of a global semi-operational system. Since then Satel Life, a non profit international organisation, has attempted to provide medical services using low earth orbiting satellites which can receive, store and transmit information to a large number of terminals located in East Africa and elsewhere.

The continued and rapid development of satellite communications with very high powered transponders, has resulted in Integrated Satellite Digital Network and remote area communication facilities with small, low-cost antennas on the ground. The increasing deployment of VSATs all over the world, has made it possible to transmit not only digital data, voice and messages, but also video images to even remote areas in a most cost-effective manner. To the array of already existing domestic, regional and international satellite network, mobile communication system using low and intermediate earth orbiting satellites is now being added, which in the course of next five years, is bound to change the global communication scenario by providing connectivity between persons anywhere in the world, through the use of low-cost portable ground terminals. The time is thus ripe for planning and realization of global telemedicine network, which can take advantage of the vast resources available to humanity as a whole without reference to a geographical location.

7. Disaster Management

Space systems, which derive their basic advantage from the altitude of their operation, have unambiguously demonstrated their capability in providing vital information and services towards all the three major aspects of disaster management, namely—prevention, preparedness and mitigation. The unique capabilities of remote sensing satellites to provide comprehensive, synoptic and multitemporal coverage of large areas in real time and at frequent intervals, have become valuable for continuous monitoring of atmospheric as well as surface parameters related to natural disasters. The vast capabilities of communication satellites, are now available for timely dissemination of information on early warning and real time coordination of relief operations.

The advent of Very Small Aperture Terminals (VSATs), Ultra Small Aperture Terminals (USATs) and phased array antennas have enhanced the capability further, by offering low-cost viable technological solutions towards management and mitigation of disasters. Though several countries, including a few developing nations, have taken advances of the advantages in space technology to address specific problems related to disaster management, most of the developing countries do not yet have the necessary infrastructure and trained manpower to absorb and utilise the benefits of space technology.

Disaster warning is the basic prerequisite for ensuring disaster preparedness, and in some cases, to aid in disaster prevention. Disaster Preparedness relates to actions designed to minimise loss of life and damage to property, and to facilitate timely and effective rescue, relief, rehabilitation measures in case of a disaster. Clearly, the most important application of satellites, is in detecting and delivering early warnings of impending disasters, and disseminating information on hazard awareness to the people in the areas which fall under the threat of potential danger. Communication satellites are particularly well-suited to deliver locale-specific disaster warnings to select groups of people, even in remote and inaccessible areas.

Satellite communication (satcom) capabilities-Fixed, Mobile and Specialised-are vital in a large number of disaster management situations, especially in data collection, distress alerting, position location and coordinating actual relief operations in the field. Countries that operate domestic satellite communication systems, either through their own or through leased space-segment capabilities, have no problems in maintaining and energising satellite-based communication links to disaster stricken areas using transportable or air-liftable small satcom terminals. From suitcase size L-band satcom terminals for exceptional land applications to, shoe-box size 'message only' STD-C terminals, INMARSAT has developed a whole range of attractive application possibilities that await harnessing of their full potential. As a part of future Global Maritime Distress and Safety System (GMDSS), INMARSAT is already providing ship distress alert; search and rescue coordination; marine safety information involving transmission and reception of navigation and meteorological warnings; and general radio communication, to shore-based communication networks.

COSPAS-SARSAT Search and Rescue satellite system for distress alert detection and position location, represents yet another laudable multinational effort that has successfully completed its demonstration phase and has now become operational. Its utility for a variety of land, sea and air distress situations has been amply demonstrated. Incorporation of 406 MHz Emergency Position Indicator Radio Beacons (EPIRBs), capable of working with COSPAS-SARSAT, is now a requirement in International Maritime Organisation (IMO)'s future Global Maritime Distress and Safety System (GMDSS). However, from the point of view of a much more widespread use and greater global acceptance, there still remain a few aspects that need to be addressed, specially for ensuring:

(i) continued availability of COSPAS-SARSAT space-segment capabilities on an assured basis to the entire global community,
(ii) greater international first-level participation, and
(iii) augmentation of space-segment capabilities to improve system availability.

India is a signatory of the COSPAS-SARSAT committee. Indian Space Research Organisation has set up two earth stations at Banglore and Lucknow for receiving distress alerts relayed through COSPAS-SARSAT satellites and alerting Search and Rescues agencies like Coast guards and Air Traffic Controls in the country. Besides, India played a key role in the Development and Evaluation phase of the geostationary component to the Search and Rescue systems, which provides distinct timing advantage over LEO satellites.

8. Indian National Satellite System (INSAT)

In the 1970s, the nation strongly felt the need to build up a viable space infrastructure for several national developmental applications. Communication as a vital link to progress was foreseen. The conduct of Satellite Instructional Television Experiment (SITE) during 1975–76 using the US Application Technology Satellites (ATS-6), which is considered as the largest sociological experiment anywhere in the world, demonstrated the potential of satellite technology as an effective communication medium. The Satellite Telecommunication Experiment Project (STEP) conducted during 1977–79 using Franco-German "Symphony" satellite provided the necessary inputs for the country towards the establishment of operational space system. Thus the seeds for the Indian National Satellite INSAT were sown and since then, the country has made considerable progress in the establishment and operation of INSAT System.

INSAT System is a joint venture of the Department of Space, Department of Telecommunications, Ministry of Information & Broadcasting and Department of Science & Technology. The system was commissioned in 1983 with the launch of INSAT-1B. At present, the INSAT System consists of the following satellites:

(i)	INSAT-1D	74°E
(ii)	INSAT-2A	74°E
(iii)	INSAT-2B	93.5°E
(iv)	INSAT-2C	93.5°E
(v)	INSAT-2DT	55°E
(vi)	INSAT-2E	83°E

INSAT-1D and INSAT-2A, which have already completed their designed mission life, are now in an inclined orbit. At present, there are about 70 transponders operating in the INSAT System which have been allocated for

National Television Channels, Regional Television Channels, jhabua Development Communication Project (JDCP) and Training and Development Communication Channel (TDCC) besides for a variety of telecommunication services including about 6000 Very Small Aperture Terminals (VSATs).

The application of INSAT system for telecommunications has helped in establishing cost-effective satellite links between places separated by long distances, achieving higher order of reliability with negligible sensitivity to terrain and terrestrial disasters, providing comparatively better flexibility in routing of circuits and diversity of media. The system has also helped in providing interim services at short notices. Today, INSAT provides about 5,500 two-way speech circuits over 166 routes. There are about 280 earth stations including 20 mobile terminals linked through the INSAT Network. More than 6,000 Very Small Aperture Terminals (VSAT), 800 NICNET terminals and more than 580 Remote Area Business Messaging Network (RABMN) terminals are operating besides other services like Remote Area Communication, High speed VSATs, etc. Mobile Satellite Service in S-band on an experimental basis has also been introduced.

The 1970s saw the introduction of television broadcasting in india. However, the significant progress of INSAT system and their successful commissioning and operation accelerated the goal of near total coverage of the entire country under television in a short time. Today, there are more than 800 television transmitter stations linked via the INSAT. Two National and twenty Regional television channels have been introduced. More than a lakh Direct Reception television sets have been installed. Besides the no. of channels there have been improvements in the geographical coverage too. INSAT-2E provides a foot print extending from Western Europe to Australia.

Educational television has been a high priority area since the INSAT system was established. Educational television programmes are now telecast by Doordarshan which include curriculum-based programmes telecast from Delhi, Mumbai and Chennai, enrichment programmes for school children produced in various languages by several States and Institutes of Educational Technology, General enrichment programmes for University students, and syllabus based programmes for students of the Indira Gandhi National Open University.

Based on several experiments carried out during the beginning of 1990 s to use INSAT satellites for interactive satellite based communication system for development, training and continuing education, a Training and Developmental Communication Channel (TDCC) using INSAT was established on an operational basis in 1995. TDCC provides a unique one-way video and two-way audio system of interactive education where the teaching-end includes a simple studio and up-link terminal for transmitting live or pre-recorded lectures. The participants at the class-rooms located nation-wide receive the lectures through simple dish antennas and have facility to interact with the lecturers using telephone lines. The TDCC system is now being used extensively by several agencies and industries like the Indira Gandhi National Open University, National Diary Development Board, State Bank of India and several State Governments (Gujarat, Karnataka and Madhya pradesh) for distance education in rural development, women and child development, Panchayati Raj and industrial training.

Jhabua Development Communication Project (JDCP) is a pilot project launched in November 1996 to demonstrate the effectiveness of communications channel support to the developmental activities in a backward rural district. Under the project, interactive training is being provided to the field officials and rural population in the predominantly tribal district of Jhabua. The programmes include watershed management, health, education and Panchayati Raj. Under JDCP, 162 television-receive terminals have been deployed in as many villages with talk-back facility in 12 block headquarters. The network is currently being expanded to more than 1000 villages spread over three districts of Madhya Pradesh. Social Research is an important component of this project. Formative research which helps in appropriate software program production, process research which helps in monitoring the actual transmissions and summative research which helps in assessing the impact of the programs are all part of the JDCP. The results of the two year project are very encouraging and point to significant improvement in the quality of life of this primarily tribal population in Jhabua.

INSAT system is being effectively used to network the All India Radio stations which has resulted in a perceptible improvement in the quality of the audio signals transmitted through the AIR ground network. This is so because, earlier the AIR stations used to receive transmissions through high power transmitters in the HF and UHF band which are prone to large variations due to atmosphere while at present, they use S-band transponders of INSAT.

The INSAT system provides, round the clock, half-hourly synoptic images of weather systems including severe weather conditions like cyclones, sea surface and cloud top temperatures. The system also helps in the collection and transmission of meteorological, hydrological and oceanographic data from unattended remote platforms and issue timely warnings of impending disasters like cyclones, floods and storms. The Very High Resolution Radiometer (VHRR) instrument—which have been a part of the INSAT system right from inception—are regularly used for weather forecasting and dissemination of advanced warning on any impending cyclone over the East and West Coast of India. INSAT-2E meteorological payload incorporates a Water Vapour channel besides multi-band CCD camera for finer resolution of the imagery. The INSAT VHRR instrument helps in tracking the motion of the cyclonic storm. This information is transmitted from Chennai, Calcutta and Mumbai to the INSAT satellites through the S-band transponders. These signals are received by 250 Cyclone Warning Dissemination System (CWDS) receivers deployed in 10 coastal States of India. The INSAT DRT transponders are used for data collection from remote unattended platforms. The India Meteorological Department has deployed 100 Data Collection Platforms (DCPs) all over the country. These DCPs collect the meteorological data such as rainfall, wind speed, wind direction, temperature, humidity, etc., from the various locations and they are transmitted to the Meteorological Data Utilisation Centre at New Delhi.

INSAT is a vital element in the international COSPAS/SARSAT Satellite Aided Search and Rescue system that provides distress alerts and location information to the rescue authorities for maritime, aviation and land users who are in distress. While the COSPAS/SARSAT system uses low earth satellites in polar orbit causing delays in distress detection, particularly in the equatorial regions, INSAT satellites which are placed in geostationary orbit enable detection of distress events in real time.

INSAT system is also used by the National Physical Laboratory for disseminating Standard Time signals to several time receivers located in different places of the country.

9. Growth in Satellite Communications

There has been tremendous growth in satellite communications over the past four decades. We have come a long way from the Telestar relay of TV signals across the Atlantic. Today there are more than 1000 communication satellites in orbit. Another 500 satellites have been notified through International Telecommunications Union (ITU). Starting with about 200 voice channels in 1965 current communication satellites can easily support over 20,000 voice circuits. It is estimated that there will be more than 4000 transponders in C and Ku bands by the year 2000 for TV transmission. The number of TV channels will increase from 2400 in 1996 to about 15000 in the year 2000. In the Asia-Pacific region alone about 400 C band and 500 Ku band transponders will be added in the next two years. This is equivalent to about 4500 digital TV channels.

The evolution from geostationary to Low Earth Orbit (LEO) satellites has resulted in a number of proposed global satellite systems that can be grouped into three distinct types: “Little LEOs”, “Big LEOs” and “Broadband LEOs” which can be best distinguished by reference to their terrestrial counterparts. Little LEOs are the equivalent of paging systems, big LEOs provide the equivalent of cellular telephone services, and broadband LEOs are similar to fibre optic networks. A principal distinction between the two main types of system is that “Little LEO” satellites will offer a range of low-speed text and data services, while “Big LEO” satellites will offer users global voice, fax and, possibly, broadband services.

With the enormous growth in personal communications, Internet, data transfer, direct-to-home (DTH) television and even basic voice communications, the demands being placed on the radio spectrum and the GSO are leading to ever more creative ways of exploiting them, and to the opening up of higher frequencies, including the Ka-band which until now has not been much used. The growing congestion in the traditional frequency bands used by satellite systems for provision of global international services coupled with the new communications service offerings and the rapid developments in technology led to new systems using the 30/20 GHz bands being planned. Ka-band satellites represent the satellite industry's contribution to the convergence of telecommunications, computing and broadcasting technologies. They will lead to a fundamental restructuring of the world's communications satellite industry and lead to the development of global satellite operators with integrated L-band, C-band, Ku-band and Ka-band systems, using geostationary orbits and Low and Middle Earth Orbits. The Ka-band available to satellite operators involves a massive 2.5–3.5 GHz of spectrum—that is to say some 4 to 7 times that available to some C-band satellite operators. Thus already some 60 proposed Ka-band and V-band projects have emerged world wide alongside an intensified research and development effort to refine Ka-band satellite and associated technologies. The term Ka-band satellite is now generally recognised as a shorthand term for a new generation of communications satellites that will use on-board processing and switching to provide full two way services to and from small Earth stations comparable in size to today's satellite television dish. Such Ka-band satellite systems have also been described in other terms such as "multimedia satellites", "Asynchronous Transfer Mode (ATM) satellites", "broadband switched satellites", and "broadband interactive satellites".

The Ka-band satellites offer fundamentally different services from conventional communications satellites. Each Ka-band satellite will carry what is, in effect, a form of telephone "switchboard-in-the-sky". This will allow the satellite to operate like a telephone network—offering point-to-point circuits to both business and individual users at a cost far lower than available via satellite today. However, unlike a normal telephone network, such satellites will be able to offer all end users a wide variety of services from simple narrowband through to broadband applications. They offer the prospect of a truly modern, ubiquitous and global alternative to the world telephone infrastructure to underserved areas, but at a fraction of the cost of upgrading the latter to provide a comparable variety of services.

10. Current Trends and Future Direction

Less than five years ago, the satellite telecommunications industry's primary offering was fixed voice and video-broadcast services. Limited mobile services were available, but only at substantial cost. Today, satellite systems are rapidly diversifying their services and are targeting customers directly. The ability to instantly develop infrastructures and quickly provide services to wide geographic and sparsely populated areas have made satellite systems attractive and very cost-effective vehicles for consumers and investors alike. Seizing on this, the telecommunications, media, and entertainment communities have embraced satellite technologies and made them a vital and integral part of their overall infrastructures. The emerging applications of satellite networks are wide-ranging. A study in this regard indicates applications getting manifold in various countries. (Table 3).

The satellite telecommunications industry is evolving rapidly. Virtually every week, low- and medium-earth-orbit satellites are being placed in orbit to offer an array of mobile services that will revolutionize the way we communicate. The sector is growing robustly, and as more satellites are launched and technologies become more powerful and cheaper, revenue derived from satellites are anticipated to increase dramatically.

Earlier in the decade, the telecom sector was focused on GEO satellites operating in the L, C and Ku-bands. Its services were regional and dominated by international organizations and large telecommunications providers. Today and for the foreseeable future, however, the satellite industry will be shaped by private companies operating LEO voice, LEO/GEO interactive broadband, and GEO broadcast systems offering data, voice, and video capabilities. By harnessing new technologies that allow satellites to operate in the Ku, Ka and V-bands, networks, of LEO satellites, inconjunction with geostationary satellites, are to form the core of a global, integrated system enabling world-wide access to a broad spectrum of communications, ranging from voice and data to internet and video transmissions.

The growth of these activities around the world, combined with expanding bandwidth requirements, is expected to almost triple satellite revenues over the next five years. Already the finance community has committed more than $20 billion to satellite ventures, and analysts forecast a need for an additional $50–$75 billion in the next five to seven years.

11. Satellite Communications for Development

Satellite has become a prime means for long distance communications, television broadcasting, private data networks, maritime communications and disaster relief networks. Despite the initial financial costs, the implementation of a satellite system can be an efficient and economical solution for providing communications to wide-spread areas because distance has no effect on the cost of providing a satellite service. It is far easier and more cost-effective to build satellite ground stations and to use satellite links than to establish the infrastructure for ground-based microwave systems, particularly over large geographical distances and in difficult terrain. In addition, the implimentation of a satellite system offers enormous social and economic benefits. Communications satellites can provide tele-education and tele-medicine services by transmitting educational programmes and medical information directly to remotely located rural villages. In terms of national sovereignty, satellites can

Table 3. Applications of Emerging Satellite Networks

Application	Country			
	Europe	Japan	U.S.	Other
I. a. Internet access	Development	Development	Emerging	China: Emerging India: Emerging Israel: Emerging
b. Multicasting	Concept	Concept	Development	No data
c. Global Telephony	Emerging	Emerging	Emerging	Emerging in the Rest of the world
II. a. Telemedicine	Development	Development	Development	Russia: Development
b. Teleeducation	Development	Development	Emerging	Korea: Development
c. Data broadcasting	No data	Development	Development	No data
d. Digital broadcasting	Growth	Growth	Growth	Korea: Growth
e. Government	No data	Emerging	No data	China: Emerging India: Emerging
f. Telecontrol	No data	Emerging	No data	No data
g. Teleconferencing	Development	Development	Emerging	India: Emerging
h. Electronic commerce	Emerging	Emerging	Emerging	Emerging in the Rest of the world
i. High data-rate transfer	Research	Research	Research	No data
j. Disaster recovery	Research	Research	Research	No data

be used to transmit vital information as well as promote social integration by providing a means for the exchange of information between people in remote and urban areas. Satellites have become indispensable elements for the expansion of integrated digital networks, video programme delivery, and land and maritime mobile communications.

Communications are no longer only a medium of inter-personal expression but have become the medium for gaining access to information throughout the world. There is a strong correlation between telephone density and development. Today's communications markets clearly demonstrate the importance of establishing and maintaining a communications infrastructure as a key to the societal and economic development of a country. Currently, practically every country in the world benefits from a variety of communications services through participation in international, regional or domestic satellite communications systems. Indeed, satellite communications technology is now recognized as a critical tool for social and economic development as advances in technology continue to lower the costs of its utilizations.

Communications satellites are used for a wide variety of purposes including rural and wireless communication, news and data dissemination, emergency communication, navigation, disaster warning, television and radio programme distribution, search and rescue, tele-medicine and remote education. Satellite communication has created an enormous amount of opportunities that can enhance economic development. Moreover, these benefits have the potential to be accessed by all sections of society and further their sustainable development.

Communications satellites optimize the reality of two phenomena, the "global village" and the "information age ". Satellites can bring the power of information to virtually everyone on Earth and enable individuals to share experiences instantaneously. Modern telecommunications and electronic information systems are an indispensable tool in the continuing quest to meet basic human needs. Information technology, in general, is a great social leveller which can help erase social barriers and overcome economic inequalities. Modern telecommunications are as critical and fundamental to the sustainable development of developing countries as any other basic necessity.

In many developing countries where terrestrial systems are underdeveloped or even non-existent, satellite communications services, which have witnessed rapid advances in the last three decades, are of particular importance. The most significant contribution of satellites is to bring basic communications to the people. Accordingly, the developing countries could skip the costly 20th century wired infrastructure and proceed directly to the 21st century global information infrastructure. This approach to information flow will stimulate the economy and the national growth of each country. Satellite communications have been a major factor in facilitating the establishment and continuing expansion of the internet and world-wide web it has created.

As noted above, the satellite market combined with the deregulation of the world's telecommunications and the instant infrastructure offered by the satellites, is creating an unprecedented growth for satellite telecommunication services. The deployment of geostationary GEO satellite systems and constellations of Low Earth-Orbit (LEO) satellite systems promise to bring low-cost access to even the most remote areas of the world. Such access by the world's population to telephony, high speed data, Internet, distribution of video signals for cable and television programmers and other multimedia services will make the "global village" a reality. Society at large will greatly benefit in the remote access to a large variety of services in the fields of medicine (tele-surgery, tele-diagnosis) and education (tele-teaching).

5. Internet Packet Transport: Traffic Control and Network Engineering

Anurag Kumar

Department of Electrical Communication Engineering,
Indian Institute of Science, Bangalore 560 012

The Internet can be viewed as the transportation infrastructure for the deployment of Information Technology. The flow of traffic in any transportation network needs to be controlled to avoid congestion, to provide special service to flows that need it, and to protect against overloads. In this paper, we consider the control of store-and-forward (also called *elastic*) traffic in the Internet. We first motivate the need for feedback based congestion controls for elastic traffic. These controls provide for fair sharing of the network bandwidth, and hence determine how bandwidth is shared between elastic flows. We discuss two approaches for implementing such controls: in one, the network provides explicit feedbacks to the traffic sources, and in the other the sources adjust their rates based on implicit feedbacks. Then we discuss models that could help in sizing the Internet for guaranteeing minimum throughputs to elastic flows. These models lead to engineering guidelines and suggest the need for bandwidth management and overload controls. Finally, we present results from an experience in developing and deploying such controls in a campus network.

Key Words: congestion controls, bandwidth sharing, traffic engineering, overload in the Internet

1. Introduction

The Internet is widely regarded as the most powerful emerging technology for transporting information, and is now an integral part of the Information Technology revolution. Internet transport is versatile enough to be able to operate over a variety of physical transmission media, ranging in speeds from a few kilobits per second to more than a billion bits per second. In the Internet, all information (whether it is data, or digitised voice or video) is transported as packets of bits, that are moved from link to link by packet switches that are called routers.

The digital communication links and the routers provide a basic transport mechanism for moving data around. Figure 1 shows such a transport mechanism being used to carry traffic from various sources. If these sources are permitted to use the basic transport system directly, there will be chaos. Hence traffic controls are always implemented in packet networks. These controls help to prevent congestion, provide service differentiation and quality of service, and take action in case of traffic overload. In Fig. 1 these controls are shown conceptually as lying between the traffic sources/sinks and the basic packet transport. The controls are actually implemented as software in the network routers, and in the communicating end-systems. Distributed procedures serve to carry out the various control functions.

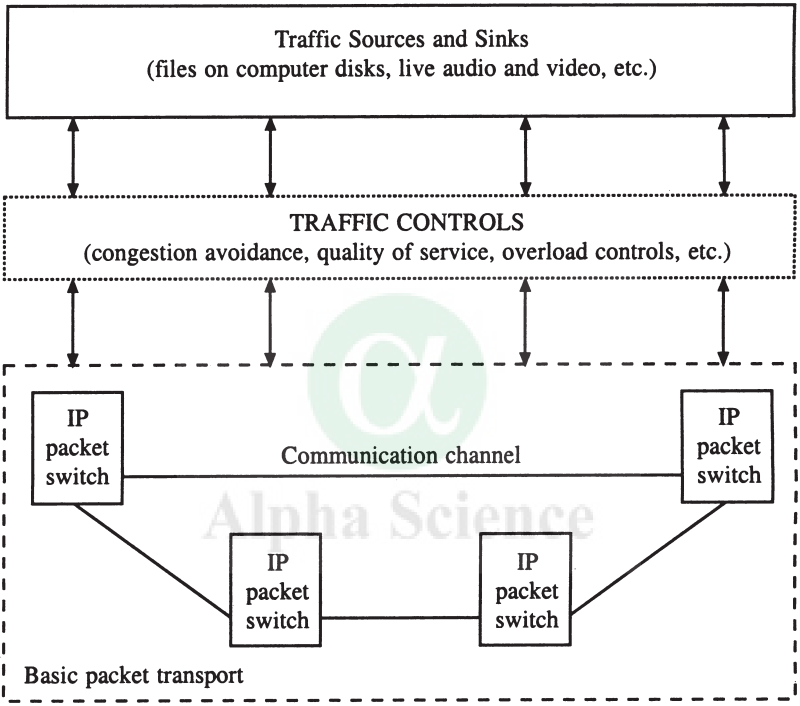

Traffic Sources and Sinks
(files on computer disks, live audio and video, etc.)
TRAFFIC CONTROLS
(congestion avoidance, quality of service, overload controls, etc.)
IP packet switch
Communication channel
IP packet switch
IP packet switch
IP packet switch
Basic packet transport

Fig. 1. Traffic controls over a basic packet transport

At the present time, the overwhelmingly large percentage of use of the Internet is for the transport of, so-called, *elastic* sources of information. Examples of such sources are data files, image and audio files residing on computer disks. Transport of such stored information is distinguished by the fact that there is no natural transfer rate that such a source of information requires. A file on a computer disk, if transported reliably over the network, can be transported at *any* positive rate (bytes per second), and will end up safely on another computer disk. It is in this sense that such transfers are elastic. This is in contrast with *stream* traffic, for example, interactive voice, or live video. Such traffic sources have *intrisic temporal patterns,* and hence cannot be controlled, i.e., cannot be asked to slow down or speed up[1]. This paper is entirely about elastic traffic.

It is clear that elastic sources are eminently suitable for transport over networks in which the available bandwidth is time varying. The Internet transports elastic flows at time varying rates for two reasons: (1) When several elastic connections share a link in the network, the link bandwidth must be *fairly shared* between these connections; hence, as the number of active connections changes, the fair rate for each one of them also changes with time. (2) The Internet is more and more being pressed into the transport of guaranteed bandwidth connections (e.g., Internet telephony) that need to be carried at a higher priority than elastic data; as the number of such connections varies with time, the bandwidth available to the elastic sessions also varies.

Since the bandwidth available to an elastic connection varies with time, there has to be a *feedback control* mechanism in the network by which the sources can vary their sending rates in response to the time varying available bandwidth. In this paper, we begin by discussing two approaches for the implementation of such feedback control in packet networks in general. In ATM (Aysnchronous Transfer Mode) networks, an explicit feedback based approach is used. In the Internet, the ubiquitous TCP protocol uses implicit feedback for such control.

As mentioned above, bandwidth sharing for elastic flows is intimately related to some notion of fairness. Given such notions of bandwidth sharing, a model for elastic transfer request arrivals, and a model for the amount of data each such transfer requires, we have a model that can be used to engineer the Internet. The notion of *network engineering* here is much the same as the one that is well known in the context of telephone networks. In telephone networks, given a model for the network, we are interested in the probability that a call is blocked, and the network is engineered (by trunk sizing) so that the call blocking probability is less than, say, 1%. In the same way, using an engineering model for the Internet, we can expect to engineer the Internet to provide a minimum transfer rate for elastic flows.

We will discuss results from the analysis of such a model for a simple network, such as that would be present at a small Internet service provider (ISP). The results immediately imply the need for bandwidth management controls if the load exceeds the engineered levels, and overload controls if the offered load exceeds the network capacity. Finally, we will present some results from actually implementing such controls at the Internet access link to an academic campus in India.

[1] It is interesting to note here that, in some cases, the coders at such sources can produce high or low bit rate outputs (with varying degree of quality) in response to the bandwidth availability in the network. However, even then these sources cannot be considered to be arbitrarily elastic.

2. Feedback Controls for Elastic Traffic

In Fig. 2, we show a simple scenario in which several hosts need to download files (say, using the World Wide Web (WWW) protocols; formally called *http*) from one of the web servers. Typically, the hosts would be on a high speed local area network (LAN), and would be connected to the web site by a wide area network (WAN). The WAN is shown here simply as a single link, depicted as a *bit pipe.* Let the bit rate of this link be R bits per sec (bps) (typically links have the same rate in both directions; but we are concerned here in the direction from the server to the *client* hosts).

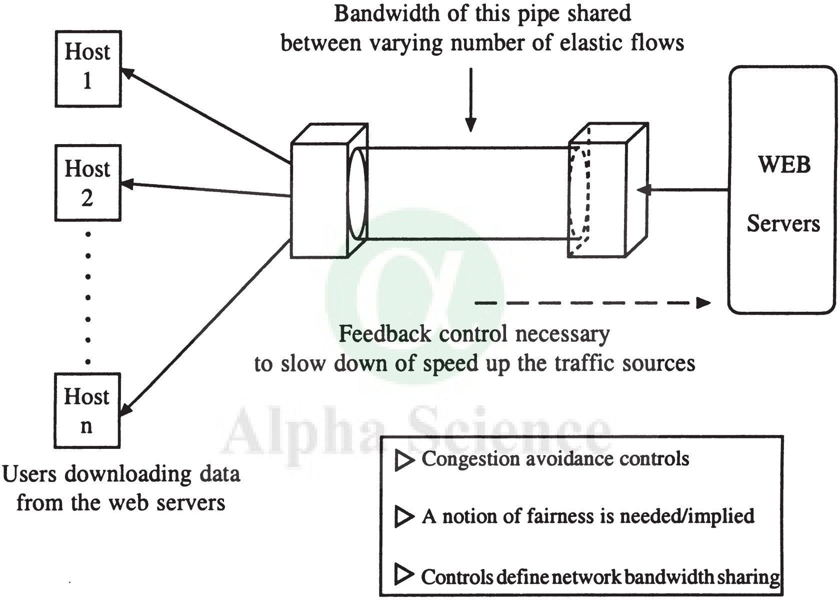

Bandwidth of this pipe shared
between varying number of elastic flows
Host 1
Host 2
Host n
WEB Servers
Feedback control necessary
to slow down of speed up the traffic sources
Users downloading data
from the web servers
▷ Congestion avoidance controls
▷ A notion of fairness is needed/implied
▷ Controls define network bandwidth sharing

Fig. 2. A simple illustration of the need for feedback congestion controls for elastic traffic.

Suppose Host 1 begins to download a file. Since the host is on a high speed LAN, the bottleneck for this transfer will be the WAN link. The host can then expect to obtain a file transfer rate of R bps. Now, suppose that before Host 1's transfer finishes, Host 2 requests a download (from the same or different web server). Suppose that nothing is done, then even if Host 2 gets a small positive transfer rate ε, the link will need to carry $R + \varepsilon$. This is not possible, and hence there will be data loss at the router between the web server and the link. Hence, clearly, when Host 2 starts its transfer, the rate given to the transfer to Host 1 should reduce. Furthermore, a natural requirement is that the rates should be fairly allocated to the two hosts. Thus when n hosts are each downloading data, the average rate that each should get should be $\frac{R}{n}$.

In general the servers would not know of each other's transfer requests (in fact, the servers would be geographically separated, unlike the simple example here). Hence, in order for the transfer rates to adapt, some kind of feedback needs to be provided to the traffic sources. Such feedback should result in the sources adjusting their rates to share the bandwidth fairly. We notice that bandwidth sharing between elastic flows is intimately linked to these feedback based congestion controls, and the notion of fairness they implement. This should be seen in contrast to the telephone network, where bandwidth sharing between various sessions simply entails the end-to-end allocation of a fixed amount of bandwidth to each session (e.g., 64 Kbps (Kilo bits per second) to each voice call).

3. Fair Allocation of Bandwidth: Concepts and Techniques

In the example of Fig. 2, owing to the simple topology, it was very clear what fairness meant; fairness meant equal allocation. Consider, however, the scenario in Fig. 3. Three sessions, 1, 2, and 3, share the bandwidth in a two link network. The link speeds are 1 Mbps (Mega bits per second) and 2 Mbps. Session 3 transits over both the links, Session 2 over the 2 Mbps link, and Session 1 over the 1 Mbps link.

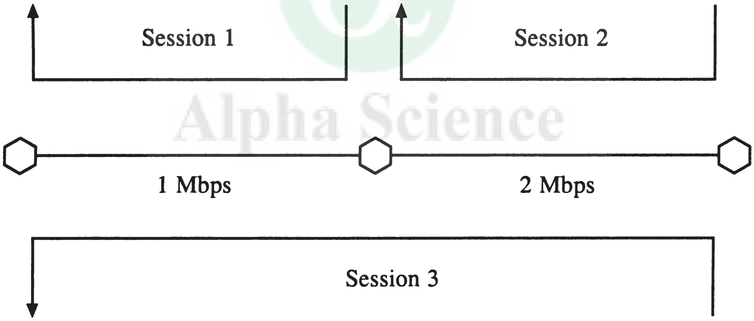


Fig. 3. An example demonstrating max-min fair bandwidth sharing in a network

If equal allocation was the objective, then we can obtain an allocation as follows. Starting with all flows at zero rate, increase the rates of all the flows equally, until some link saturates. In this example, this procedure will result in each session getting a rate of 0.5 Mbps, and the 1 Mbps link will saturate. It is easy to see that what we have is the *max-min* rate; i.e., among all the feasible rate vectors (for the three sessions), the minimum rate in any rate vector is no more than the max-min rate.

But clearly, it is inefficient to stop at this allocation. There is still spare bandwidth on the 2 Mbps link. We can allocate it to Session 2, thus obtaining the feasible allocation: Session 1 = 0.5 Mbps = Session 3, and Session 2 = 1.5 Mbps.

The rates that result from the above procedure are called *max-min fair.* For details, see the textbook treatment in Bertsekas and Gallager (1992). A recursive procedure for obtaining the max-min fair rate vector in a given network and session topology is evident:

Step 0. Start with all the sessions and the entire network; call this the Network.
Step 1. Increase the rates of all the sessions in the Network until some link saturates. Allocate the resulting rate to all the sessions that pass through the saturated links.
Step 2. Form a new network by removing all the saturated links and all the bottlenecked sessions, and by reducing the capacity of each unsaturated link by the amount of flow allocated to the bottlenecked sessions passing through that link. Replace Network with this network, and repeat Step 1.

The algorithm is guaranteed to stop in at most as many steps as there are links in the network, since at each step at least one link saturates or there are no more unbottlenecked sessions left. We have described, however, a centralised algorithm; i.e., the session and network topology information has to be available at some central place that computes the fair rates. This is clearly not practical in real networks considering their large extent, high link speeds, and propagation delays. A distributed algorithm is essential. In the next section we describe two approaches for implementing such distributed feedback controls.

Recently, other notions of fair bandwidth sharing have been introduced and studied. One of these, that arises from an economic formulation (maximising user utility subject to flow feasibility) is called *proportional fairness;* see Kelly *et al.* (1998). In some network topologies, proportional fairness is equivalent to max-min fairness.

3.1 Network Algorithms: Explicit feedback

Fig. 4 shows the same network and session topology as in Fig. 3. Notice that we have shown that the source of each session is being provided with an

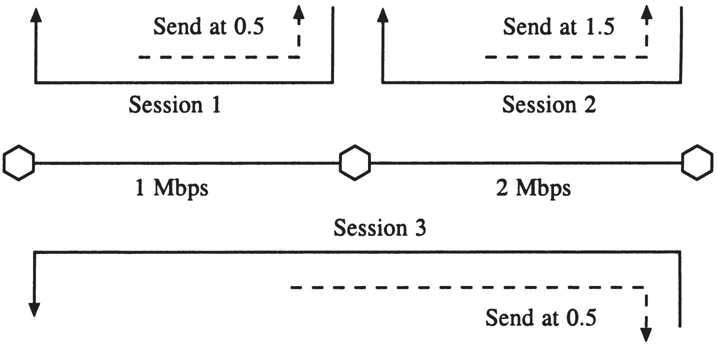


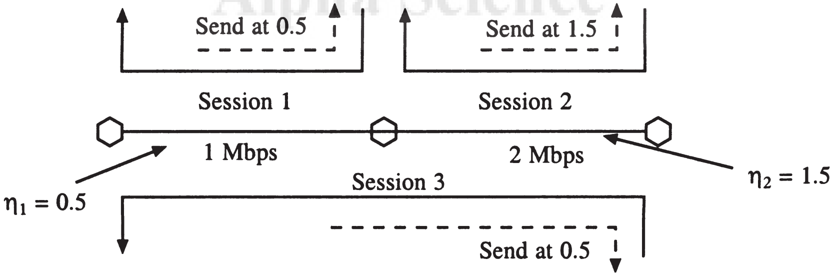
Fig. 4. Fair rate allocation by explicit feedback from the network.

ex*plicit* feedback of the fair rate of that session. Evidently, in this approach the network explicitly participates in the computation of the fair rate. Also this computation should be done in a distributed fashion, by the mutual cooperation of software running in the switches and the end systems. We describe the conceptual basis for one technique for carrying out such a distributed computation.

Fig. 5 shows associated with each link l a number η_l. The values shown are $\eta_1 = 0.5$ Mbps, and $\eta_2 = 1.5$ Mbps. The specific values depend on the problem instance. There is an η for each direction of flow on each link; since in the example all the flows are from "right" to "left", the η's here correspond to this direction of each link. Notice that the fair rate r_s for session s is given by

$$r_s = \min_{\substack{(j \in \text{links} \\ \text{used by } s)}} \eta_j$$

Let us now suppose that, in the network, somehow these link parameters have been computed. Each source can then determine its fair rate in a distributed way as follows. Each source sends a control packet in which one field is set to a large value. As the packet passes through each switch in its path, that switch replaces the value in this field with the minimum of the local η and the existing value in the field. When the packet reaches the destination node, it is looped back. Notice that when the packet reaches the source again, it carries the minimum of the η values in the path of the source, and hence the fair rate for the source.



Associated with link l is a number η_1

Fig. 5. Each link has an associated link parameter η_l, from which the max-min fair rates for each session can be computed

It remains to produce distributed algorithms that compute the η values for the links. This is an important research problem. A survey is provided in Bonomi and Fendick (1995); an important seminal algorithm is presented in Charny (1994); an algorithm that has seen a lot of real implementation and experimentation is reported in Jain *et al.* (1995). Our work, which addresses the important issues of the effects of propagation delays, and time varying available link capacities (owing to high priority stream traffic) is documented in Abraham and Kumar (1997a), Abraham and Kumar (1998a), Abraham and Kumar (1998b), Abraham (1998). Our approach yields algorithms that operate at each switch and use only local information; i.e., no information needs to be explicitly exchanged between the various switches in order to compute the η values. Each switch only needs to measure the aggregate elastic flow through each exit link, and estimate the available capacity (for elastic flows) at each such link. A simple asynchronous algorithm working independently at each switch then succeeds in iteratively calculating the link η values. Note that the coupling between the algorithms running at the switches is by the way the η values are used to obtain the max-min fair rates (see above), and by the sources actually sending at these rates. Our approach has the advantage of not requiring per flow computations at each link, or per flow updates on the control packets that pass through a link.

3.2 Network Algorithms: Implicit feedback

In Fig. 6 we show another approach to controlling the sending rates of elastic sessions. The network does not explicitly participate in the feedback control.

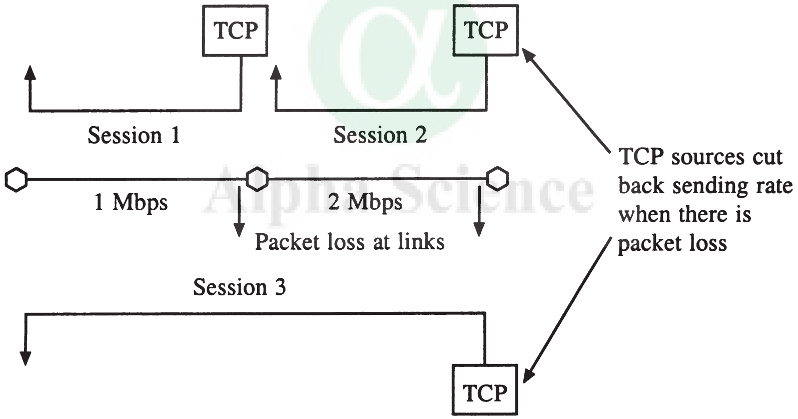


Fig. 6. Fair rate allocation by implicit feedback, caused by acknowledgements and packet losses.

The boxes marked "TCP" in Fig. 6 represent software running in each source device that autonomously adjusts the rate at which the source sends, in response to *implicit* feedbacks. Normally when a packet succeeds in getting through the network, the destination sends an acknowledgement back to the transmitter. The TCP software interprets an acknowledgement of successful data as an indication of bandwidth availability in the network, and slightly increases its sending rate. If the aggregate data rate at any link is more than the link's transmission rate, then data loss occurs. This results in the lack of acknowledgment for the lost data, which causes the TCP software to drastically reduce its sending rate.

The term TCP used above is the short form for Transmission Control Protocol, a portmanteau protocol in the Internet, that takes care of the following functions:

- Recovery of lost data, and resequencing: The basic packet transport in the Internet can lose or reorder packets. TCP recovers and sequences the packets before delivering them to an application that uses TCP.
- Sender-Receiver flow control: TCP allows a slow receiver (e.g., a mechanical printer with a small memory) to slow down a fast transmitter to avoid data loss.
- Congestion control: This is the topic that we have been discussing at length in this paper.

TCP carries out its functions by means of a window based transmission protocol. All data is sequence numbered, and at any time there is a sequence number (the left window edge) upto which the transmitter knows that all data it sent has been received. There is also a window upto which the transmitter can send new data beyond the left window edge. The flow control and congestion control functions are carried out by making this window time varying. The rate of data transmission is reduced by reducing the window size. Since new data is generated in response to acknowledgements, roughly, the effective data transmission rate is given by the window size (in bytes) divided by the round trip delay in the connection. Thus window adaptation indirectly results in rate adaptation. For details on the TCP protocol see Stevens (1994).

The design approach behind TCP is in keeping with the general Internet philosophy of a simple network connecting intelligent terminals. TCP expects little from the network, and is itself a very complex protocol that runs in the end-systems. Owing to its simplicity and its lack of assumptions, TCP basically hunts around for the optimal sending rate, and keeps overshooting (resulting in data loss) or undershooting (resulting in link starvation). In a sense, it implements a kind of "bang-bang" control. Since TCP does not know what rate to send at, it builds up its window slowly; if there is loss, the rate is drastically cut and then rebuilt slowly. Since the rate build up is driven by acknowledgements, the rate can only build up over round trip times. These factors contribute to TCP being conservative, and inefficient, and *unfair* to sessions that have large round trip delays. The latter effect occurs since connections with longer round trip delays need larger windows, and take longer to rebuild these large windows when there is data loss. For various experiments and analyses that demonstrate these performance aspects of TCP, see Kumar (1998a), Lakshman and Madhow (1997), Padhye et al. (1998).

4. An Engineering Model

A packet data network is a traffic handling system, and in this context "engineering the system" means to design and size the system to handle a certain amount of load (offered traffic) so as to meet some quality of service (QoS) requirement for what is being transported. For example, in the Internet context, assuming that only store and forward applications (e-mail, file transfers, and WWW) are carried, the offered load may be expressed as the rate of arrival of transfer requests between all pairs of sources and sinks, and the distribution of the volume of the transfer at each request. For example, λ_{ij} may denote the rate at which files are requested to be transferred from node j to node i. The volume of each transfer may be represented by the random variable X with distribution function $F(\cdot)$. This then specifies the network offered load. The QoS may be specified simply as: the average throughput of a transfer should be at least $\tau_{\min}$ Bps (Bytes per second). Given an Internet topology, and routing within the network, the problem then is to determine if the offered load can be carried with the required QoS.

A very similar problem arises in the telephony context. Consider Fig. 7 where we show the scenario of a small basic phone service provider's network. The provider sets up a small telephone exchange, and leases some telephone channels to the public phone network (in India the DOT (Dept. of Telecom) network). Some telephone channels then connect the provider's customers to this small exchange. The customers in turn distribute the service within their campuses using PBXs. The problem now is to determine the number of channels connecting the provider's exchange to the phone network and to its customers. Phone calls can be between the phone network and the subscribers, or between the campuses of the subscribers. A mathematical model is now needed to proceed with this network engineering problem. The following are the elements of the model:

- Call arrival model: The arrival rates of calls between the various pairs of points in the network are needed. Typically a Poisson arrival model is used for each of these arrival processes.

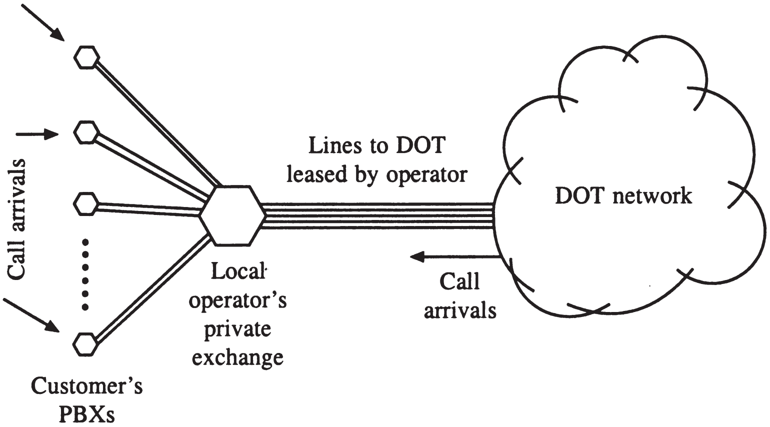


Fig. 7. A network model for a private access provider for telephone service.

- Bandwidth sharing model: Each call between a pair of points is allocated one channel in each link along the route between that pair of points.
- Call holding time model: Each call holds its allocated channels for a certain random amount of time and then terminates. Typically, it is assumed that the call holding times are exponentially distributed; this assumption is, however, not even mathematically necessary for solving the fixed routing model.

The QoS objective is to keep the probability of call blocking below some small number, say 1%. Given a network topology, the number of channels on each link, and the routing plan, the above is a well formulated problem whose solution has been known for decades. Hence, by and large, the above problem is considered to be a well studied and tractable one (see, for example, Bertsekas and Gallager (1992) and Ross (1995).

Let us now turn to the problem of engineering a small Internet Service provider (ISP). A typical small ISP is depicted in Fig. 8. The ISP leases a high speed link (say 2 Mbps, or 34 Mbps) to a higher level (bulk) ISP, who presumably has a very high speed backbone. Then our small ISP offers service to its customers via local leased lines or dial up. Requests for transfer of files arrive from the local customers, who want to download files from the Internet, and from the users on the wide area Internet, who want to download data from the sites hosted by the ISP's customers. A certain minimum average throughput is required for such transfers. We now need to analyse the performance of the ISP's network. The problem is similar to the one seen above for the phone network, with some important differences. The elements of the engineering model now are:

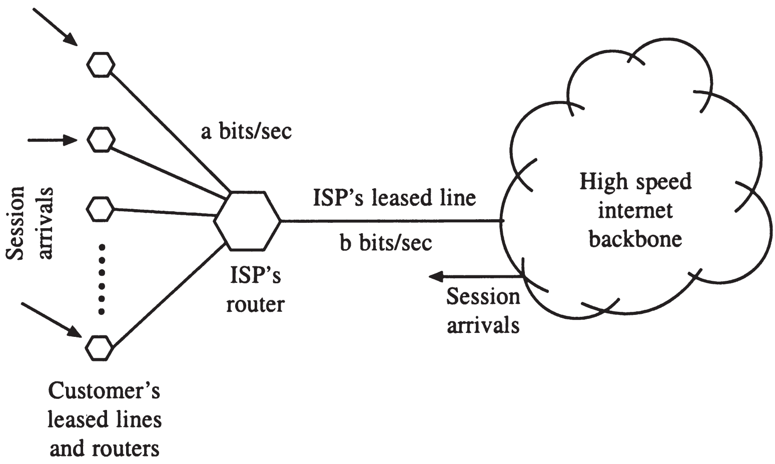


Fig. 8. A model for a small internet service provider (ISP).

- Session arrival model: E-mails and file transfers (via ftp) result in arrivals of requests to transfer individual files. The way in which the WWW is used is, however, somewhat different. Users start a WWW session, during the course of which they may down-load several short files, looking at them and thinking in between, and then occasionally download large files. To obtain a first cut model, we can simplify things and model the process of arrivals of all individual file transfer requests (including those within a WWW session). The model for arrival epochs of transfer requests between each source and sink can be taken to be Poisson; for more on this approach, see [Paxson and Floyd '95].
- Bandwidth sharing model: Assuming fixed routing, and an ideal protocol that enforces perfect max-min fair bandwidth sharing, we can assume that given any configuration of sessions at an instant, the rate that each session gets is the max-min fair rate at that instant. When there are large propagation delays, a large number of sessions with small amounts of data to transfer, and the protocols use slow-start, this model could be quite inaccurate.
- Session volume model: Each transfer request is modeled as having a certain file size, which is sampled from a distribution, independently of any other transfer. Note that this model does not translate to a session holding time model, as the rate at which the file is transfered is randomly varying.

The QoS objective may be to ensure that the average transfer rate over any route is more than, say, $\tau_{\min}$ Bps. This model, which has been the subject of some recent research activity, is known to be hard to analyse. In De Veciana *et al.* (1999), the authors have proved that (for exponentially distributed session volumes) the process of the number of active sessions is stable if the offered load on each link is less than the link rate. In Kherani (1999) some approximations have been developed for analysing session throughputs in small example networks, and the effect of propagation delays and slow start have been studied by simulations.

For the simple network shown in Fig. 8, however, calculations can be easily done if the number of customers is allowed to grow to infinity. It is as if each new session is started by a customer on an access link that is not already carrying a session. Let us assume that all transfers are from the Internet to the ISP's subcribers (this is usually the direction in which the ISP's access link will be congested). We denote the ISP's Internet access link speed by b bps, and the customers' access link speed to the ISP by a bps.

Fig. 9 shows two example curves obtained from the analysis of the model. We plot the average throughput per active session (normalised to a), defined by ($N(t)$ is the number of active sessions at time t, and $I_{\{N(t)>0\}} = 1$ if $N(t) > 0$, and $= 0$ otherwise)

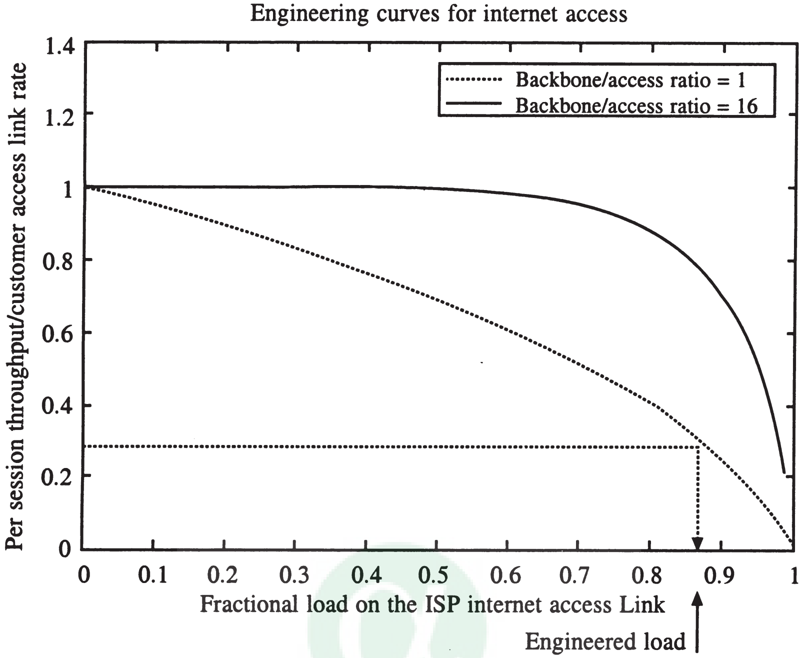


Fig. 9. Session throughput vs. access link load; curves obtained from the simple engineering model. The bent line with an arrow shows an example of maximum engineered load for given minimum session throughput.

$$\lim_{t \to \infty} \frac{1}{\rho t} \int_0^t \frac{I_{\{N(u)>0\}}}{N(u)} du,$$

versus the normalised load ρ on the backbone link. Note that the normalised load (ρ) on the backbone link is simply

$$\rho = \frac{1}{b} \times \text{session arrival rate (per second)} \times \text{mean session volume (bits)}$$

The ratio b/a is seen to be important; shown are curves for $b/a = 1$ and $b/a = 16$; $b/a = 1$ applies to all cases in which the access link cannot be the bottleneck, as would be the case when all the users are on a high speed LAN attached to the ISP router. The formula for $b/a = 1$ is obtained from the well known processor sharing model see Gelenbe and Mitrani (1980), and is given by

$$\frac{1-\rho}{\rho} \ln \frac{1}{1-\rho}$$

Notice that for $b/a > 1$ the curve is flat upto a larger value of ρ. Hence if an ISP has a large Internet access bandwidth (say 2 Mbps) and its customers access it at low speeds (say 64 Kbps), then the ISP can operate its access link at high occupancies and yet offer its customers a high average transfer rate.

Note, however, that for large *b*/*a* an operating point close to 1 is more sensitive to small variations in the load.

We make the following observations from Fig. 9. The average throughput per active session decreases as ρ increases, dropping to 0 when $\rho \to 1$. For $\rho \geq 1$ the system is unstable, in the sense that the number of active sessions will grow to infinity. In practice, of course, the number of users is finite; if more transfer rate is requested than the bottleneck rate *b*, the users will simply get backlogged, and will not be able to generate any more requests. It follows that if a certain minimum transfer rate is required, then the offered load cannot exceed some maximum value (see Fig. 9). Further, if the offered load is allowed to approach or exceed 1, the users will experience very poor throughput, and certain overload behaviour may manifest itself: users abandoning partially transferred files, TCP connections timing out, etc.; see Massoulie and Roberts (1998). These arguments motivate the need for some form of traffic control at the TCP connection level and at the aggregate flow level.

5. Edge Based Connection Control: An Experience

Fig. 10 show the Internet connectivity at the Indian Institute of Science campus. An access link connects the campus, via a router, to a major ISP (in the figure, VSNL is shown). Users access this bandwidth either via the campus LAN, or via serial links (dial-up or leased). Notice that the network is conceptually very similar to the one shown in Fig. 8. There are several 100s of active users on the campus, and unless adequately sized, the Internet access link could become a bottleneck.

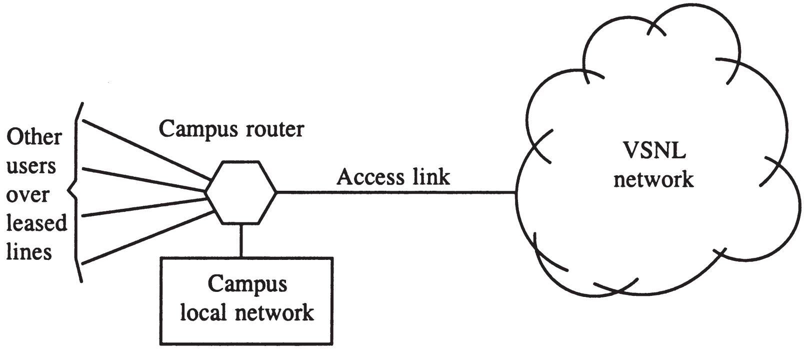


Fig. 10. Local network and wide area Internet access in a campus.

The access link into the IISc campus was (at the time of writing) just 128 Kbps. Fig. 11 shows the 24 hour occupancy of the access link (on February 17, 1999), in the direction into the campus. Observe that, except for an hour or two between 5am and 10am, and a period in the evening (meal times in the campus messes!), the link occupancy is close to 100% (the fact that it does not show 100% is probably because our measurement is not counting some low level link header bytes that are attached to every packet sent on the link). Clearly, the link experiences sustained overload during much of the day. Thus one can expect that, for much of the day, the detrimental effects of operating near $\rho = 1$, as pointed out at the end of Section 4, should be evident. This is indeed the case, as we will show from actual measurements presented below.

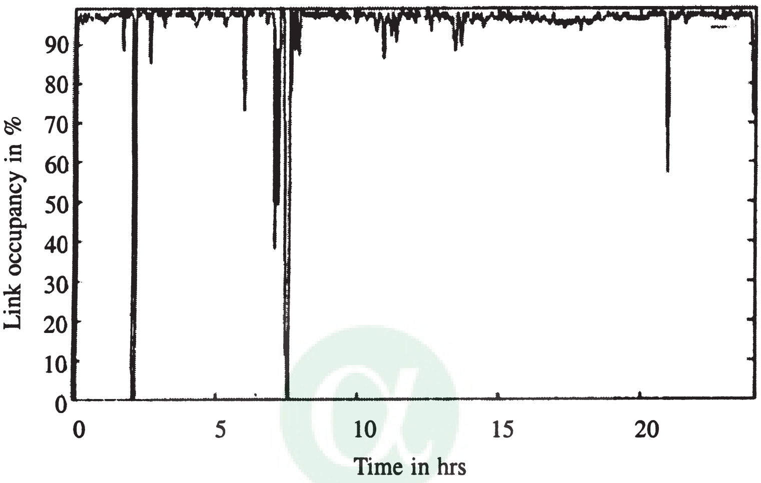


Fig. 11. Internet access link occupancy without overload controls.

In order to control the traffic, we adopted a nonintrusive approach shown in Fig. 12. We have developed a software package Kumar *et al.* (1998b) that monitors the access link occupancy, the amount of access bandwidth used by various services and IP address groups (as configured by us), and takes control actions if certain levels of utilization are exceeded. This software runs on a computer labelled "Overload Controller" in the figure. Notice from Fig. 12 that the traffic between the Internet and the campus users does not pass through the controller device. Instead, the controller sits on a broadcast network (simply an Ethernet hub), and can listen to all the packets flowing between the users and the Internet. The controller sends poll packets to the routers at either end of the access link; the responses to these packets carry traffic statistics collected at these routers. By picking up packets off the broadcast local network, the controller can look at individual packets (passing between the users and the access link) and classify them according to whatever categories we set (e.g., one category can be; "e-mail packets arriving from outside the campus for a user who is on the campus LAN"). One function of the software is to record and display the statistics collected in this way.

We also set policies based on access link utilisation. For example, we can set the maximum utilisation of the access link to be 90%; further, we can say that users on the campus LAN will get a minimun of 60 Kbps (in the inbound

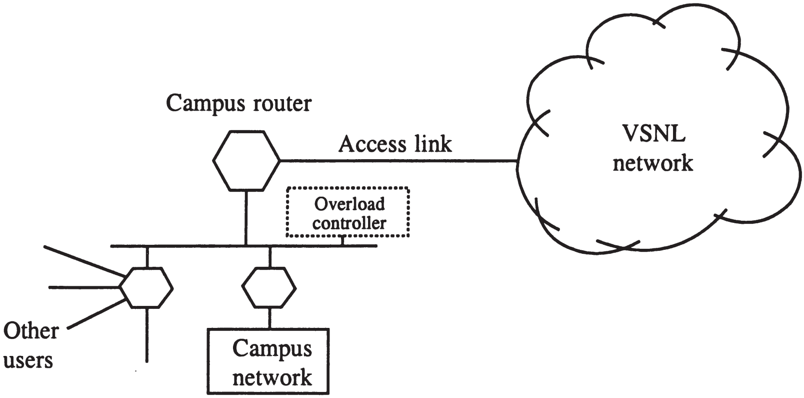


Fig. 12. A nonintrusive overload control device.

direction, out of the 128 Kbps) during the day. Out of this, 25 Kbps will be guaranteed for e-mail, and the rest can be used by WWW and file transfers. When the actual traffic flow violates these allocations then the connections are controlled. New TCP connections are disallowed, and existing connections are made to slow down their transfers.

In Fig. 13 we show a result of implementing our controls on WWW throughputs. The Figure shows the cumulative distribution of web throughputs with and without controls. The data was obtained from a web proxy[2] through which all web transfers take place. Notice that without control more than 95% of the web transfers get a throughput of less than 0.2 KBps (note that 128 Kbps = 16 KBps), whereas with control the throughputs range upto 2 KBps. Without control the average throughput is about 75 Bps, whereas with control the average is about 450 Bps. Further observations from the web

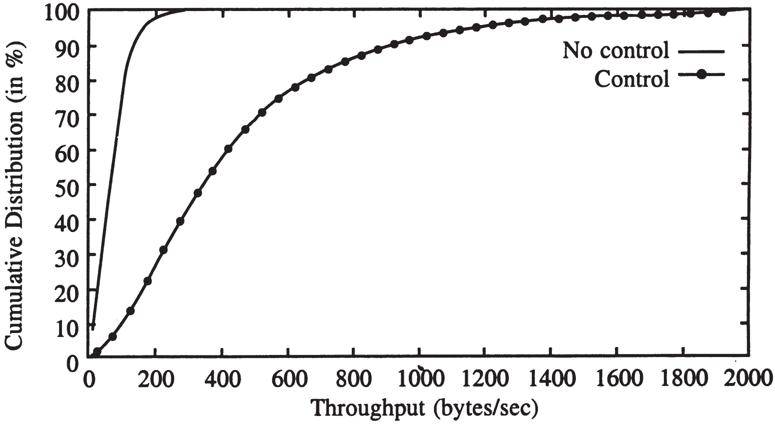


Fig. 13. Cumulative frequency distribution plot of web transfer throughputs with and without overload controls.

proxy show that, if there are no controls, then although the link carries more traffic, much of the additional traffic it carries does not contribute to "good throughput". This is mainly because there are a lot of partial transfers, either because of users abandoning their transfer requests or the connections timing out.

Another important observation that we made is that, without control, users on the campus LAN are able to use more of the access link bandwidth than users that are attached via the low speed leased or dial-up lines (see Fig. 10). With our bandwidth management controls, we are able to allocate any desired portion of the access link bandwidth to any subset of users.

In Fig. 14 we show the occupancy of the link when the controls are implemented. The occupancy now varies between more reasonable values.

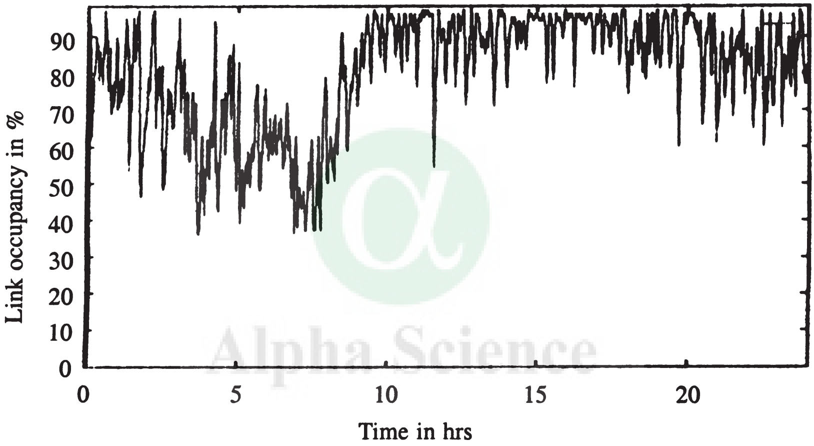


Fig. 14. Internet access link occupancy with overload controls in place.

As mentioned above, connection control is exercised by connection blocking and by slowing down ongoing connections. With the current level of offered load, and the 128 Kbps link, during peak hours on weekdays, we need to block upto 60% of the connections. There is thus a tradeoff. Without overload control, there is no connection blocking, and users get the satisfaction of always "getting through", but then suffer very low throughputs or stalled connections. Also the overloaded link carries a lot of useless data. On the other hand, if connection controls are exercised, depending on the level of overload, some percentage of connections get blocked, but those that get

[2] A *web proxy* is a machine that obtains data from a web server on behalf of a requesting client. An important advantage of such an approach is that the data can be copied into a *cache* in the proxy, just in case anyone else requests for the same data. Proxies are also an important point where access controls can be exercised.

through experience good throughputs. Further, the link is efficiently utilised. Obviously, no network should need to operate for long with a high percentage of blocking. If an operator makes bandwidth allocations, and then finds that the blocking probabilities begin to approach, say, 1% then it should be taken as an indication that bandwidth is insufficient, and either bandwidth reallocation or increase in the basic access bandwidth is needed. Our observations, however, clearly demonstrate the performance of elastic transfers in the Internet in a situation of severe overload, and the success of our approach in mitigating the situation.

6. Conclusion

In this paper we have considered the Internet as primarily transporting elastic traffic. We have traced through a complete understanding of the issues involved in controlling elastic traffic in the Internet, and engineering the network for such traffic. Starting from congestion control of elastic traffic, and its implications for bandwidth sharing, we have described network engineering models that incorporate such bandwidth sharing and that demonstrate the possibility of network overload. We have shown by actual experience the effects of network overload, and the results of our attempts to mitigate network overload in our campus.

Unlike the phone network, very little work seems to have been done on large scale modelling and control problems in the Internet. The framework presented in this paper should help to put all the associated problems and issues in context.

References

1. Abraham Santosh P. and Kumar Anurag, (1997), Max-Min Far Rate Control of ABR Connections with Nonzero MCRs, *Proc. IEEE Globecom*, pp. 498–502.
2. Abraham Santosh P. and Kumar Anurag, (1998), A Stochastic Approximation Approach for Max-Min Fair Adaptive Rate Control of ABR Sessions with MCRs, *IEEE Infocom '98,* San Francisco.
3. Abraham Santosh P. and Kumar Anurag, (1998), A Simulation Study of an Adaptive Distributed Algorithm for Max-Min Fair Rate Control of ABR Sessions, *Proc. CCBR '98*, Ottawa, Canada.
4. Abraham Santosh P., (1998), Asynchronous Distributed Rate Control Algorithms for Best-Effort Sessions in Integrated Services Networks with Minimum Rate Guarantees, *Ph.D. Thesis,* ECE Department, Indian Institute of Science, Bangalore, India.
5. Bertsekas D. and Gallager R.G., (1992), *Data Networks,* Second Edition, Prentice Hall of India, New Delhi.
6. Bonomi F., Fendick K.W., (1995), The Rate-Based Flow Control Framework for the Available Bit Rate ATM Service" *IEEE Network*, pp. 25–39.

7. Charny Anna, (1994), An algorithm for rate alocation in a packet-switching, network with feedback, *Master's thesis,* MIT, Cambridge.
8. De Veciana G., Lin T-J., Konstantopoulos T., (1999), Stability and Performance Analysis of Networks Supporting Services with Rate Control -Could the Internet be Unstable? *Prof. IEEE Infocom 1999,* New York.
9. Gelenbe E. and Mitrani I., (1980), *Analysis and Synthesis of Computer Systems,* Academic Press, London.
10. Jain R., Kalyanraman S., Viswanathan R. and Goyal R., (1995), A Sample Switch Algorithm", *ATM Forum/95–0178.*
11. Kelly F.P., Maulloo A.K., Tan D.K, (1998), Rate Control for Communication Networks: Shadow Prices, Proportional Fairness and Stability, *Jour. Operat. Res. Soc.,* **49**.
12. Kherani A.A., (1999), Rate Based Congestion Control for Ephemeral Best-Effort Sessions in Packet Networks, *Master of Engineering Thesis*, ECE Department Indian Institute of Science, Bangalore.
13. Kumar Anurag, (1998), Comparative performance analysis of versions of TCP in local network with a lossy, link *IEEE/ACM Transactions on Networking,* **6(4)**.
14. Kumar Anurag, Hegde Malati, Anand S.V.R., Garge Gopi, Bindu B.N., (1998), NETMASTER: A Nonintrusive Approach for Bandwidth Management and Service Control at a WAN Access Link to a Campus Network, *Technical Report*, Number ERNET-IISC-1998.2, ERNET Project, Indian Institute of Science, Bangalore.
15. Lakshman T.V. and Madhow Upamanyu, (1997), The performance of TCP/IP for networkswith high bandwidth delay products and random loss. *IEEE/ ACM Transactions on Networking,* **5(3)**, 336–350.
16. Padhye J., Firoiu V., Towsley D. and Kurose J., (1998), Modelling TCP throughput: A simple model and its empirical validation, *Proc. ACM SIGCOMM '98.*
17. Paxson V. and Floyd Sally, (1995), Wide Area Traffic: The Failure of Poisson Modelling, *IEEE/ACM Transactions on Networking,* **3**(3), pp. 226–244.
18. Massoulie L. and Roberts J., (1999), Arguments in Favour of Admission Control for TCP Flows, *Proc. Internat. Teletraffic Cong.* **16,** Edinburgh.
19. Ross K.W., (1995), *Multiservice Loss Models for Broadband Telecommunication Networks,* Springer-Verlag, London.
20. Stevens W.R., (1994), *TCP/IP Illustrated, Volume 1,* Addison-Wesley, Reading, Mass.

6. The Universal Digital Library: Intelligent Agents and Information on Demand

Raj Reddy and Michael Shamos

Carnegie Mellon University, 5325 wean Hall, Pittsburgh, Pa 15217, USA

This article presents concepts and issues related to creating the Universal Digital Library containing all the creative works of the human race. Such a system raises a number of technical, economic and legal issues which need to be solved before the vision can be realized. This article further presents the current status and future directions in digital libraries.

Library of the Future

The library of the future will be digital and have the following features:

- contain all recorded knowledge online (billions of items)
- distributed, maintained globally
- accessible by:
  - — any person
  - — in any language
  - — any time
  - — anywhere on earth
  - — via the Internet
- searchable, browsable and navigable by humans and machines, and
- act as the information resource for the 21st Century

This notion of the universal digital library is radical; in that it raises many technical, social and legal problems yet to be solved. However, the vision that "each home will be able to access electronic libraries and electronic museums around the world via networks, allowing users to readily search and obtain worldwide information on books, music and art based on their own particular interests" will have a profound impact on the way we live, work, and access information.

Digital vs. Traditional Libraries

The shift from traditional libraries to the digital is not merely a technological evolution, but requires a change in the paradigm by which people access and interact with information.

A traditional library is characterized by the following:

- emphasis on storage and preservation of physical items, particularly books and periodicals
- cataloging at a high level rather than one of detail, e.g., author and subject indexes as opposed to fulltext
- browsing based on physical proximity of related materials, e.g., books on sociology are near one another on the shelves
- passivity; information is physically assembled in one place; users must travel to the library to learn what is there and make use of it

By contrast, a digital library differs from the above in the following ways:

- emphasis on access to digitized materials wherever they may be located, with digitization eliminating the need to own or store a physical item
- cataloging down to individual words or glyphs
- browsing based on hyperlinks, keyword, or any defined measure of relatedness; materials on the same subject do not need to be near one another in any physical sense
- broadcast technology; users need not visit a digital library except electronically; for them the library exists at any place they can access it, e.g., home, school, office, or in a car

Everything Can Be Stored

The total number of different books produced since printing began does not exceed one billion. (The number of books now published annually is less than one million.) If an average book occupies 500 pages at 2,000 characters per page, then even without compression it can be stored comfortably in one megabyte. Therefore, one billion megabytes are sufficient to store all books. This is 10^{15} bytes, or one petabyte. At commercial prices of $20 per gigabyte, this amount of disk storage capacity could be purchased for $20 million. So it is certainly feasible to consider storing all books digitally.

Very Large Databases

A database of a billion objects, each of which occupies one megabyte, is large but not inconceivable. Once one is comfortable with sizes of this kind, it is feasible to imagine a thousand such databases, or to envision them all as portions of the same global collection. This amount of storage is sufficient to house not only all books, but all of the following:

- photographs
- legislative material, court decisions
- museum objects
- recorded music
- theatrical performances, including opera and ballet

- speeches
- movies and videotape

Distributed Holdings

When information is digitized and accessible over a network, it makes little sense to speak of its "location," although it is technically resident on at least one storage device somewhere, and that device is connected to at least one computer. If the information is available at multiple mirror sites, it is even less meaningful to speak of it being in a "place." While traditional libraries measure their size by number of books, periodicals and other items held, the relevant statistics for a digital library is the size of the corpus its users may access. This means that digital libraries will want to expand their "holdings" by sharing digital links with other libraries. Unfortunately, there seems to be very little sharing of this sort taking place at present.

How can we understand the unwillingness of libraries to share content? The question goes back to the old measure of the size of a traditional library—the number of books it holds. When a library expends funds to assemble digitized works, it loses a portion of its prestige (or thinks it does) by allowing other libraries to copy or access its data. Ultimately, however, *all* material should be accessible from *every* library.

Gore's Digital Earth

In 1998, Gore Vice President of USA stated that "A new wave of technological innovation is allowing us to capture, store, process and display an unprecedented amount of information about our planet and a wide variety of environmental and cultural phenomena. ... I believe we need a "Digital Earth." A multi-resolution, three-dimensional representation of the planet, into which we can embed vast quantities of geo-referenced data". He then called on scientists to create a digital map of the earth at a resolution of one meter. Such a project will require technical innovation beyond that required even for a digital library containing every book ever written. The area of the earth in square meters is about 5×10^{14}. Storing two megabytes of data per square meter (which would include terrain data, imaging, environmental and other pertinent information) will require 10^{18} bytes, an amount roughly equal to the amount of digital storage currently present on earth.

Digitization

Ultimately, everything that people are interested in accessing will have to be digitized. The reason is that digitial searching will become so easy, inexpensive, fast and ubiquitous that users will not tolerate, or will not access, traditional materials. Capture requires a concerted, shared, worldwide effort. The cost of digitizing is not trivial, so it makes little sense for any work to be digitized more than once. Yet without any registry of digitized works, many books are digitized multiple times, while others are ignored.

Converting text, images and objects to digital form requires much more than digital photography or even high-resolution scanning and requires, instead, the following:

- initial input, either scanning or keyboarding
- conversion to one of a set of standard formats
- optical character recognition (OCR) to capture text characters for searching
- OCR correction (since OCR is inherently error prone)
- creation and input of metadata and cataloging information
- special techniques for non-textual materials, such as music, images, videotape, etc.

Policy issues, discussed below, will determine the resources made available for digitization and how they will be allocated.

Paradigm Shift

If digital libraries are to become truly useful, they must assist users in making the transition from paper books to digital hypermedia. Many people report that they derive a high degree of comfort from books for the following reasons:

- *Portability*. Books can readily be carried, are compact, light in weight and comfortable to read. Anything you can't read in bed will never displace a book.
- *Reliability*. Reading books would still be possible even if every computer on earth were down.
- *Familiarity with the medium*. The pages of a book are easy to turn, the book can be opened to any page, and the linear hierarchical organization of the material is easy to grasp.
- *Low cost*.
- *Ability to annotate*. Comments and corrections can be written in a book; passages can be marked for emphasis or studying, and a book can be resold to recover costs.

Digital Librarians

A move to electronic libraries will alter the fundamental role of librarians. Far less attention will be paid to acquisitions, cataloging and circulation, and much more to systems, online assistance, navigation assistance and conversion issues. Unless graduate programs for digital librarians are established to create trained personnel, the real risk exists that users, particularly young adults who grew up in a digital generation, will outstrip the ability of librarians to assist them.

7. Adaptive Networkload Models to Cope with Multimedia Induced Network Dynamics[1]

S.V. Raghavan and N. Swaminathan

Department of Computer Science and Engineering
Indian Institute of Technology, Madras, India

The society around us is emerging as an Information (rich) society primarily due to the phenomenal growth in Information Technology. The information available around us is in multimedia form such as data, voice, video, images, and graphics. To make such multimedia information available to the teeming millions in this country requires development of an ultra high-speed network and associated technology. Modern network designs revolve around a single box called "switch", which is responsible for maintaining the Quality of Service (QoS) guaranteed flow between any two points in a network. There are four different approaches; viz., Cell Switches, IP Routers, IP switches, and learning Switches. Of these, the learning switches concentrate on the behaviour of the endusers, learns the usage pattern, predicts in short and long term, and reserves resources for ensuring variable QoS, and thus maximizes the utilization. To develop such a switch, new approaches to workload modeling become essential.

In this paper, a learning and a generative model that can adapt itself to the dynamics in the workload in a networked environment is presented. The operation of the proposed model is illustrated with two applications; the first one deals with virtual path bandwidth management in an ATM network and the second one deals with the efficient transport of compressed video traffic over ABR service of the ATM network.

Key Words: Adaptive Workload Model, Genetic Algorithm, Multimedia

1. Introduction

The phenomenal growth in information technology has made it possible for information to be available to anyone, anywhere, and at anytime in Multimedia form. Multimedia, large user population, bandwidth hungry applications and

[1] The talk given at INSA Meeting (15–16 March 1999) was on Multimedia and Emerging Information Society. This paper focuses on just one aspect, viz., the process of building adaptive networkload models as they are the central theme of R&D to cope with the increased use of multimedia and high-speed networks.

Metadata

This term is often used to mean information *about* an item, rather than the information in the item itself. Examples include the author, title, data of acquisition, price paid, donor, etc. It is particularly critical to capture metadata that is not present in or derivable from the item. For example, the author's data of birth is often not printed in a book but can be important in distinguishing among authors with similar names (Particularly parents and children). Libraries may "share" content by simply providing links, but uniform access to the content requires uniform metadata and a procedure for generating and storing it economically. It is of little avail to exchange documents at light speed if they must be held up for months until a human cataloger can prepare metadata.

Character Set Representations

This is not merely a question of different alphabets and writing systems, a major hurdle in itself, but also an issue of how characters are represented. For example, there are several widely differing mappings of Chinese characters into ASCII. There is some appeal to having a worldwide universal standard, such as Unicode, but the notion of attempting to list all of the world's glyphs and freeze them in a standard reduces flexibility and tends to overlook obscure or variant writing systems and restrict the development of new ones. Possibly a standard should be developed that permits new character sets so long as the definition of the glyphs and the representation mapping is maintained in an accessible location.

Scalability: The Billion-User Problem

A major problem encountered in digital library development is scalability—the expansion of system capabilities by many orders of magnitude. For example, a Web site, even one with huge capacity, may be choked if many people access it at the same time. Assuming that before long approximately a billion people will be able to connect to the Internet, if only one per cent of them are interested in a topic (a number that is far too low for subjects of global concern such as the death of Princess Diana), that is a collection of 10 million people. If a server requires 100 milliseconds to grant access to a Web page, then the population would have to wait 12 days for everyone to see the same page. Therefore, technology that seems instantaneous when used on a small scale may become impossibly cumbersome when expanded.

One can imagine speeding up access to a page by adding more servers in response to anticipated demand, but even this numerical solution does not scale. If the problem is delivery of an HDTV move (which takes 10 seconds to download at 10 gigabits per second), distributing the film to even one million people (a tenth of a per cent of anticipated net users and fewer than the attendance at a major film during its first weekend of release), would require 120 days. Increasing the number of servers by an order of magnitude would not make the delay even remotely tolerable.

Bandwidth scalability is largely a hardware and networking problem. Keyword searching presents a problem of an entirely different sort. The commercial Web searchers now index approximately 50 million documents. A search can easily return 1,000 hits. This is a number small enough that a user could consider glancing at all of them to find what he wants. If the corpus being searched contained 50 billion pages (less than the number of pages in all books), a search might return a million hits, which would instead require a lifetime of effort to review. Therefore, building a digital library index, particularly one to be shared among many libraries, is not simply a matter of building a large one. Access methods, screening and navigation tools must also be provided

Even if a library has a few million books, its staff members can be generally familiar with the nature and extent of its holdings. A library with a billion books and several billion other items would be qualitatively different and probably beyond the ability of any person to master. The sheer volume of transactions, catalog records, new acquistions and help requests would be overwhelming. This is particularly true if the library permits access by computer programs as well as humans. It is apparent then that new organizational concepts on a grand scale will be required if digital information systems are to scale properly.

Systems and Architecture

A digital library "system" is composed of multiple hardware and software components, including the following:

- scanners
- computers and servers
- storage devices
- media
- catalogs
- converters
- networks
- displays
- multimedia interfaces
- usage measurement software
- human processing procedures
- reference assistants

All of the above are interconnected through networks and gateway software and must be designed for scalability, interoperability and reliability. This is a daunting challenge, particularly since the technology in many of the areas is changing rapidly. A system that uses one speech recognition system for input must be designed so that the recognizer can be replaced with another very quickly. Otherwise, the digital library, potentially one of the most responsive and useful systems in the world, can become fossilized.

Search

The problem of locating items in libraries is frequently referred to as "search," although that word tends to imply that one knows in advance what one is looking for, and possesses/handles indicators or index terms to serve as finding aids. This narrow view ignores the activity of browsing or even the higher-level function of becoming acquainted in general with a library's holdings. Browsing in a traditional library is a physical activity—it involves scanning shelves on which related works have been placed in proximity, and occasionally withdrawing them from the shelves for examination. Browsing in a digital library is a logical activity mediated by a computer. It does not require physical proximity in any sense; indeed, two consecutive items examined maybe stored on different continents. The question, then, is how can a library user (not to say the library staff) become familiar with the whole of recorded human information in a way that makes it accessible and useful?

We adopt the term "navigation" to mean moving about in a digital collection. Search is a directed form of navigation in which the goal is defined in advance with reasonable clarity. The result of a search may be an item, a collection of items, or any part of an item, even down to a single glyph. Tools must be provided that enable users to move about at varying levels of granularity within the corpus.

The usual requirement for a search is that the user is looking for a specific piece of information or a summary of what is available about a certain topic. A common case is that the user wants the answer to a specific question, such as when the postcard was invented. Only rarely does such a question translate naturally into a keyword query. Such retrieval is indirect in the sense that the user wants to learn A, but formulates a query B, to which he receives a set of retrieved documents that must be scanned to determine whether the answer to A is among them. It would be far better simply to allow the user to ask question A instead of requiring him to convert it to some query language.

Non-Textual Matter

The existence of Web searchers proves that text can be searched without being indexed or cataloged. At least on a microscopic level, documents can be located purely by their content. Many documents consist of text plus other information such as mathematical equations, tables and drawings, that themselves cannot be searched directly but can often be located by the presence of relted text. Purely non-textual matter is very different. Although substantial progress is being made on video searching (through the use of extensive captioning cues, speech recognition and other aids), content searching of music and visual materials is non-existent or in its infancy. The problem is further complicated by the existence of work that combines media in various ways.

Translingual Issues

Most library items, particularly in non-English-speaking countries, are not in English. The central translingual library question is how users may navigate through materials in foreign languages and make effective use of them. Translingual search is currently a research problem for which obvious solutions do not work. A keyword search cannot be made multilingual merely by translating the keywords one at a time. The number of possible translations of each word may be very large, so an explosion in the number of hits may result. This approach also takes no account of idiomatic uses, untranslatable words such as particles, and numerous other language-related phenomena. An interim solution is the use of translation assistants—programs that offer dictionary entries or partial or suggested translations of text protions. These show great promise for users who are at least partially familiar with the language of the retrieved document.

Synthetic Text

A user who is looking for general information on a particular topic is constrained in traditional libraries to go to an encyclopedia (which may have no entry or an outdated one on the topic of interest) or to refer to books that are generally about the subject under consideration. The time necessary for the user to obtain an overview at the appropriate level may be large because of the volume of repetitive material obtained. Programs are needed that are able to scan hits with the particular query in mind and produce abstracts, summaries, translations or analyses of the retrieved material.

Information Reliability

It seems inevitable that the class of works available through digital libraries will include electronic-only publications, ephemeral and unreviewed materials and even fabricated or counterfeit matter. The ease of publishing on the Internet combined with the absence of traditional methods for evaluating reliability makes it likely that library users will be retrieving works of questionable authenticity and value. Issues concerning the Internet and digital materials include:

- *Reliability*. How can a user (or an automated agent) evaluate the reliability of digital materials? what information must be maintained about the source of the item and its creator to facilitate a decision?
- *Version control*. How can changes made to a document be tracked and the appropriate catalog entries updated?
- *Archiving*. What assurance can there be that the digital materials will be retained somehow in their original form for an indefinite period?
- *Authenticity*. How can the genuineness of materials be assured?
- *Reviews*. The system should allow the user to scan reviews of the retrieved work and then add his own reviews or comments to a database.

- *Citations*. How may a user readily learn which works have cited the retrieved work, either favourably or unfavorably?

Economics and Policy

while a huge amount of material is in the public domain and may be freely assimilated into a digital library, the most valuable items are recent and protected by copyright. In order to induce copyright owners to allow their content to be accessed or downloaded from digital libraries, mechanisms need to be developed to compensate them appropriately. In the most extreme case, an author might himself produce but a single electronic copy of a work. In order to justify his effort, he might have to sell it for $100,000. Such a sale would be impossible if the buyer were not able to charge for use of the material, and in fact charge enough to make a profit.

Fortunately, digital libraries theoretically permit precise measurement of the use made of content. A secure browser, for example, might prevent copying, printing or retransmission of material. Automated permission systems can be developed whereby users can pay directly for certain kinds of licenses. These in turn require metadata concerning the collection of rights the library has obtained for the item.

However, the implementation of charging requires another paradigm shift. The cost of building and maintaining traditional libraries is borne by governments, foundations and corporations, but hardly ever by individuals direcly. Usage of materials is free, despite the high cost of maintenance. Note that authors receive substantial money on account of libraries, because currently each library that wants a book must purchase a copy of it, and the authors of popular books receive large royalties. In the digital world, the following are necessary to preserve this revenue stream:

1. centralized organizations finance
2. subscription fees
3. fees for individual use

Policy

Digital library policy includes several areas:

- *International cooperation*. Many antiquities, national libraries and much television content are government-owned. Will governments share their materials with others? Can the world's nations cooperate to build a worldwide digital library?
- *Government vs. private funding*. How will digital libraries be funded?
- *National priorities*. Are digital libraries national priorities to be regarded as fundamental infrastructure such as roads, or must they compete with other projects for funds?
- *Allocation of resources*. Which works will be digitized first? Should

priority be given to items that are decaying? How are budgets to be divided between software/hardware and research/development?

- *Librarianship*. How will educational programs for digital librarians develop?
- *Copyright laws and conventions*. Are the laws of various countries conducive to digital exchange of information? Can content holders prevent the use of materials in a digital age?

Conclusion

The Universal Library can be accessed from the Web Site *www.ulib.org*. At present, there are close to 10,000 out of copyright classic books available at the web site. In addition, the site has links to newspapers, magazines, paintings, music, and video lectures. We are currently exploring the possibility of an initial "million volume digital library" project as phase 1 of getting "all authored works on line". While the ultimate vision is will take many decades and many billions of dollars, a global effort is create a universal information resource will lead to democratization of information, promote understanding and preserve our culture and heritage for generations to come.

7. Adaptive Networkload Models to Cope with Multimedia Induced Network Dynamics[1]

S.V. Raghavan and N. Swaminathan

Department of Computer Science and Engineering
Indian Institute of Technology, Madras, India

The society around us is emerging as an Information (rich) society primarily due to the phenomenal growth in Information Technology. The information available around us is in multimedia form such as data, voice, video, images, and graphics. To make such multimedia information available to the teeming millions in this country requires development of an ultra high-speed network and associated technology. Modern network designs revolve around a single box called "switch", which is responsible for maintaining the Quality of Service (QoS) guaranteed flow between any two points in a network. There are four different approaches; viz., Cell Switches, IP Routers, IP switches, and learning Switches. Of these, the learning switches concentrate on the behaviour of the endusers, learns the usage pattern, predicts in short and long term, and reserves resources for ensuring variable QoS, and thus maximizes the utilization. To develop such a switch, new approaches to workload modeling become essential.

In this paper, a learning and a generative model that can adapt itself to the dynamics in the workload in a networked environment is presented. The operation of the proposed model is illustrated with two applications; the first one deals with virtual path bandwidth management in an ATM network and the second one deals with the efficient transport of compressed video traffic over ABR service of the ATM network.

Key Words: Adaptive Workload Model, Genetic Algorithm, Multimedia

1. Introduction

The phenomenal growth in information technology has made it possible for information to be available to anyone, anywhere, and at anytime in Multimedia form. Multimedia, large user population, bandwidth hungry applications and

[1] The talk given at INSA Meeting (15–16 March 1999) was on Multimedia and Emerging Information Society. This paper focuses on just one aspect, viz., the process of building adaptive networkload models as they are the central theme of R&D to cope with the increased use of multimedia and high-speed networks.

high-speed networks are inherently complex and as such renders efficient management of resources difficult. Modern network designs revolve around a single box called "switch", which plays a major role in maintaining the *Quality of Service (QoS) guaranteed flow* between any two points in a network. Hence the design of an intelligent switch becomes extremely important. There are four different types of switches (some are called routers) and switch-related thinking. They are:

1. Switches which are technology oriented such as ATM switches.
2. Routers that are IP oriented and memoryless.
3. IP switches that have memory (based on IPV6 flow or Label or Tag).
4. Switches that can learn the characteristics of the traffic.

In general, the switches are designed with the assumption that the *contract for QoS* will be generated by the user and hence concentrates on mechanisms to enforce a contract. In the four approaches mentioned earlier, the switches or routers, view the source (of packets or cells) and the scope (only at the switch/router or end-to-end) of a flow differently. For example, the *IP based routers* concentrate on subnets (which are large aggregation of traffic) and not on endusers. The *IP switches* that have memory, expects the label to be affixed by the source, at the time of flow initiation. *The switches that learn* concentrate on the behaviour of the endusers, learn their usage pattern, predict in short and long term, and reserve resources for ensuring variable QoS.

In this paper, a learning and a generative model that can learn and adapt to the dynamics of the workload in a networked environment is presented. Two applications of the proposed model in high-speed network management are discussed. The *first* one deals with the *virtual path bandwidth management in an ATM network* and the *second* one deals with the *efficient transport of compressed video traffic over ABR service of an ATM network.*

The rest of the paper is organized as follows: section 2 discusses the nature of the ultra high-speed network. Section 3 proposes an adaptive workload model for efficient management of such ultra high-speed networks. Section 4 discusses applications of the proposed adaptive workload model, first in Virtual Path bandwidth management in an ATM network secondly in efficient transport of compressed video over explicit rate networks. Section 5 summarizes the essential contributions of this work.

2. The Emerging Context: Ultra High Speed Backbone Network

The movement of multimedia information across the country requires an ultra high-speed backbone network. Management of such a network requires a good estimate of the traffic matrix and response through appropriate resource allocation strategies, in order to maximize the utilization, yet maintaining the QoS guaranteed to each flow.

Traditionally, capacity planning was done with the explicit knowledge of the characteristics of the applications that are using the network, because the traffic was *static*. But today the applications that generate network traffic are *dynamic* as they are inherently bursty. Furthermore in a country-wide network, the number of users in the system at any given point of time keeps varying, adding to the burstiness. Therefore, the conventional workload models meant for performance tuning and capacity planning are no longer valid. Hence the need arises for an adaptive model for the workload that can learn. In this paper, a generative model for the workload in a network that can learn, adapt, and predict, is proposed.

3. Adaptive Workload Model

An adaptive model has the advantage of continuous optimization under dynamic workload variations. Such adaptive workload models learn the traffic pattern and generate future loads. Two questions that arise when learning is mentioned are:

- ➢ *What is the measure of their intelligence?*
- ➢ *How to quantify learning?*

We answer these questions by introducing two parameters: *r-value* as a metric to *measure the intelligence* in the adaptive workload model and the learning ability[2] as a metric to quantify the learning ability of the model. Example.1 illustrates the r-value of a sample.

Example: Let us consider a binary string of length 2 m. Let us define R as the number of occurrences of one or more 1's succeeding a 0 in the sequence. For instance, if the string is 010010, R has the value 2. Now consider the set of strings $S=\{w \mid w$ has equal number of 0's and 1's$\}$. In other words S represents all the strings of length 2 m which have equal number of 0's and 1's. The set S can be partitioned into $(m + 1)$ subsets by R values ranging from 0 to m. At the extremities, we have the sets 1^m0^m with $R = 0$ and $(01)^m$ with $R = m$. The ratio R/m serves as a measure of the presence of the pattern "01" in the string S.

We refer to the ratio R/m of the above Example as the *r-value* [SVR98] and the *pattern of interest* embedded within a string ("01" in the above Example) as a *concept* to be learned. The learning model primarily uses the r-value associated with a concept to learn the trend in the behaviour of a system. The concept chosen should be such that the information content is maximum. For instance, in an Unix environment, at the command level of the workload hierarchy, one may be interested in the number of times an user uses the editor command *vi* followed by the text formatter *latex* followed by a viewer

[2] By learnability (we coin a new word for reasons of brevity) we mean learning ability.

xdvi. In this case, the concept is defined as the order of usage of commands vi followed by latex followed by xdvi. Based on the history of command usage, the learning model can learn about the occurrence of the concept and can then predict the future occurrences (of the concept) based on what has been learned. Occurrences of a concept is the order in which symbols (representing the user commands) of the alphabet occur within a string. In other words, this order is equivalent to the presence of a substring in a given string. Hence, if we have a count of the number of occurrences of a particular substring in a given string, then the learning model has simply to learn that count. To compare the string the model has learned with the string that the model is attempting to learn, we use the normalized metric namely the r-value. The r-value serves as a metric for comparison even for strings with different lengths.

A learning model that produces strings with r-value very close to the r-value of the string it tries to learn, is characteric of a good learning model. This observation enables development of a methodology which can make a model adaptive. Let us consider the string to represent the workload in a networked environment. Whenever the workload characteristics change, the learning model must learn the new workload characteristics and adapt itself to generate strings with r-value closer to the new workload characteristics. Now we need to gauge the learnability of the model. We use the Learnability index (L) which is in effect the comparison of the r-values of the concept in the generated command sequence with those in the actual command sequence, to quantify learnability.

The key to the whole issue is the choice of the concept. The concept actually represents the substring of interest to the one who is using the learning model. In this paper we focus our attention on *how to learn the given concept(s)* rather than define methodologies for identifying the concept(s) itself for a given environment. However, we illustrate through case studies the choice of the concept for three different instances of networking.

In reality one may be interested in more than one concept. In the case of multiple concepts within a string, the r-value definition depends on the relationship shared by the concepts themselves. In a typical computing environment there will be C, Java, VB and other language programmers as shown in Fig. 1. In this situation, though the programmers may use different languages, yet their workload cycle remains the same. We consider the workload patterns generated by the different language programmers to share an OR relationship. When considering the workload pattern of programmers with that of document preparation group and managers we consider them to share an AND relationship since the workload cycles are different.

The essential components of an adaptive workload model are illustrated in Fig. 2. The actual input stream is the string representing the environment. The concept(s) the adaptive workload model is trying to learn appears as substrings

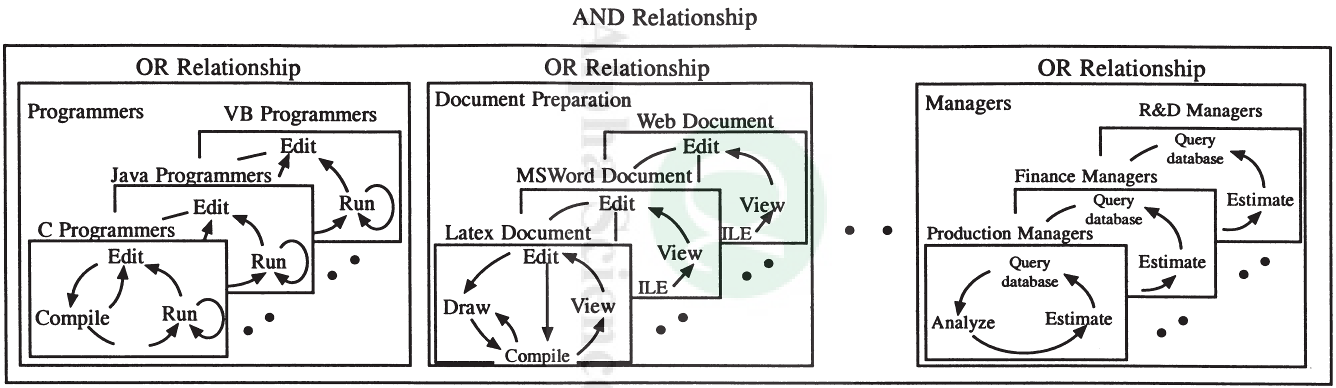


Fig. 1. Multiple Workload Cycles in a Computing Environment

embedded within the string. The learning module learns the concept and controls the generation of the predicted string. The predicted string represents the future behaviour of the system under observation.

Learning module shown in Fig. 2 could be implemented using learning techniques, such as hidden Markov models, artificial neural networks, genetic algorithms, or some regression methods. Genetic Algorithms have been used in a variety of optimization problem domains by Goldberg (1989), but they have rarely if at all, been used for dynamic problem domains such as workload behaviour learning.

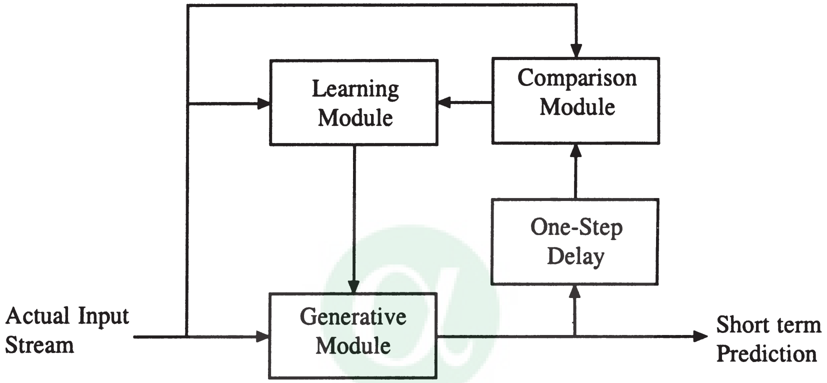


Fig. 2. Framework for the Adaptive Networkload Model

4. Applications of the Adaptive Workload Model

The adaptive networkload model finds application in many domains. As mentioned earlier, we look at the virtual path bandwidth management in an ATM network, in which we implement an adaptive workload model using genetic algorithms. The second application is the efficient transport of compressed video over explicit rate networks in which the adaptive workload model uses artificial neural networks and autoregression. The choice of a learning technique is based on the time scale within which they are applicable as in Natarajan (1991).

4.1 Virtual Path Bandwidth Management in ATM Networks

In an ATM network, transmission efficiency can be improved by dynamically changing the bandwidths of the virtual paths (VP) according to the changing traffic, see Burgin *et al.* (1991), Ohta *et al.* (1992). The bandwidth demand within a VP depends on the traffic pattern of the call arrivals experienced by that VP. The aggregate bandwidth demand of the individual calls within a VP can be mapped to discrete bandwidth ranges. These ranges in turn can be mapped to symbols of an alphabet as shown in Table 1. Thus, the characteristics of the VP bandwidth demand sampled at regular time intervals can be encoded

Table 1 Mapping of Bandwidth Ranges to Symbols

Bandwidth Range	Symbol
0–64 Kbps	1
64–128 Kbps	2
.	.
.	.
1984–2048 Kbps	24

as a string composed of the symbols of an alphabet. The string now represents the bandwidth demand pattern within a VP. We refer to the string as the demand string.

The call level throughput of an ATM network can be improved if the VP bandwidths are altered in accordance with the VP bandwidth demand patterns. The VP bandwidth control mechanism should respond to drastic changes in demand within a VP, which could be recognized by the adaptive workload model by setting a threshold value for the difference between two successive samples of bandwidth-demands. If the difference in bandwidth-demand between two successive samples is greater than the current value of the threshold, then the change in demand is deemed serious enough to warrant a contribution to the control information of the demand string. The threshold value is dynamically changed by setting it as a linear function of the minimum and maximum differences in successive samples present in the previously arrived demand string as given in eqn. 1.

$$T = \lfloor(\delta_{\min} + \delta_{\max})/2\rfloor \tag{1}$$

where T : The threshold value

$\delta_{\min}$: Minimum difference between successive bandwidth-demands

$\delta_{\max}$: Maximum difference between successive bandwidth-demands

Description of the Simulator: A call level ATM simulator was developed in C language for the purpose of simulating the call level traffic in a typical ATM network. Two physical network topologies one containing 4 nodes (Fig. 3 (left)) and the other containing 11 nodes (Fig. 3 (right)). All the VPs were assumed to be unidirectional. Call arrivals are assumed to be Poisson and the call holding times were assumed to be exponentially distributed. The calls were assumed to be of three types based on their bandwidth-demand and duration of calls, the details of which are given in Table.2. The proportion of calls of type A, B and C generated are 90%, 9% and 1% respectively. The traffic from one node to any other node in the network is assumed to be uniformly distributed. The traffic in the VPs named vp1–2, vp1–3 of the 4–node network and vp1–2, vp1–11 of the 11-node network were monitored. The physical links shared by these VPs are shown in thick line in Fig. 3. The

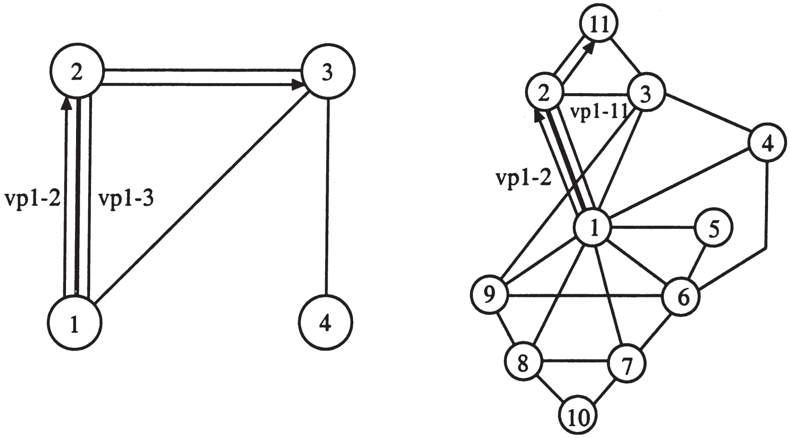


Fig. 3 Networks used in the simulation

bandwidth-demand in these VPs were sampled at intervals of one second and one day's worth of data was used for the analysis.

Table 2. Call Characteristics used in the Simulation

Call Type	Bandwidth Demand Per Call In 64 kbps	Mean Call Duration	Call Blocking Prob. Acceptable
A	1	3 min	0.03
B	2	30 min	0.03
C	4	15 sec	0.03

The maximum bandwidth of the physical link for the simulation purposes was intentionally chosen to be small in order to deal with a smaller alphabet set. However, the approach is scalable because a higher maximum bandwidth will only result in the cardinality of the alphabet set becoming larger. We consider T1 (1.5 Mbps) type of physical links, hence with every symbol of alphabet representing a multiple of 64 kbps the cardinality of the alphabet set was 24.

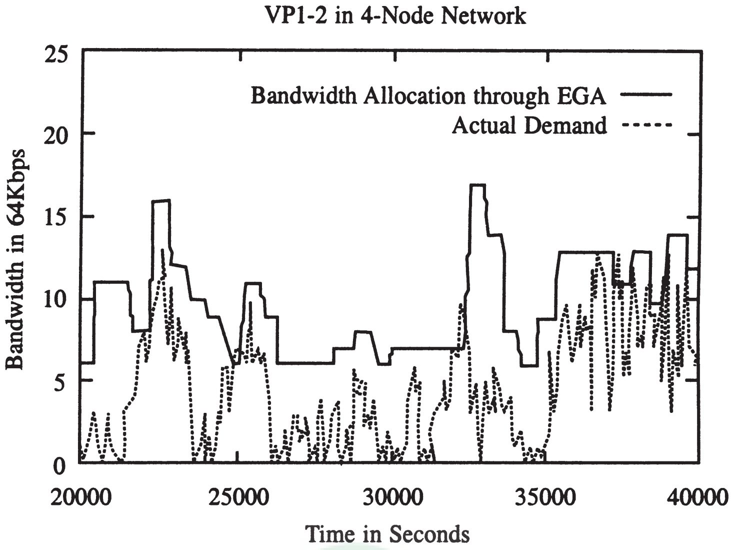
Prediction of Bandwidth Demand: The bandwidth demand in a VP is sampled at 1 second intervals. The bandwidth samples are collected for an observation period of 10 minutes. The difference in bandwidth demand between the samples is represented by symbols as given in Table 1. *The set of symbols for the entire 10 minutes is a demand string. The adaptive workload model is used to predict the next demand string on the basis of the previous three demand strings*. The three demand strings represent half an hour worth of data which is going to influence the predicted demand for the next 10 minutes. From the population, the demand string that has an r-value that lies between the minimum and maximum of the r-values of the previous three demand strings is predicted as the next demand string. The actual bandwidth demand is determined from the predicted demand string by considering the mean, standard deviation, maximum and minimum of the demands within the predicted demand string.

The effectiveness of the VP bandwidth management using the adaptive workload model is validated by means of simulation. The value of the bandwidth predicted by the proposed model for the 4-node and 11-node networks are shown in Fig. 4. The demand during the period from 36000 to 40000 was raised by 25% to test the adaptability of the proposed model. The bandwidth allocation is able to follow the demand for the 4-node network very closely while for the 11-node network during the period from 36000 to 40000 seconds the allocation is far below the demand. The low bandwidth allocation for VP1–11 in the 11-node network is due to the sharing of physical link bandwidth by many VPs flowing through it.

4.2 Efficient Transport of Compressed Video in Explicit Rate Networks

In this section we describe another application of the adaptive workload model in efficient transport of compressed video over explicit rate networks. Compressed video traffic is found to exhibit both short and long term dependency, see Beran *et al.* (1995) due to which the video frames are highly correlated. This correlation makes it possible to forecast the sizes of the upcoming frames based on the knowledge of the previous frame sizes. The forecast could be used to make short term reservations of resources in the network, thus improving the multiplexing gain.

The scheme for compressed video transport over an ABR service of an ATM network is given in Fig. 5. The unencoded raw video frames are given as input to the encoder. The encoded frames from the encoder are sent to the ATM interface via a source adaptation buffer for absorbing the short-term variabilities between the adjacent frames, see Duffield *et al.* (1998). The encoder provides the sizes of the encoded frames to the forecasting module. The forecasting module learns the trend and predicts the size of up-coming group of frames. The rate estimator determines the bandwidth to be requested from the network using the forecast of the *size of group of frames* and the backlog frames to be drained from the buffer. The network gives feedback regarding the available rate in the network to the video quality controller. Using the feedback, the video quality controller controls the quantization level of the encoder in order to reduce or increase the frame sizes to be accommodated within the available rate. If the forecasting module underestimates the actual size of the group of frames, then the quantization level is adjusted to make up for the error in estimate. Number of quality reductions by way of adjusting the quantizaiion level within a time period



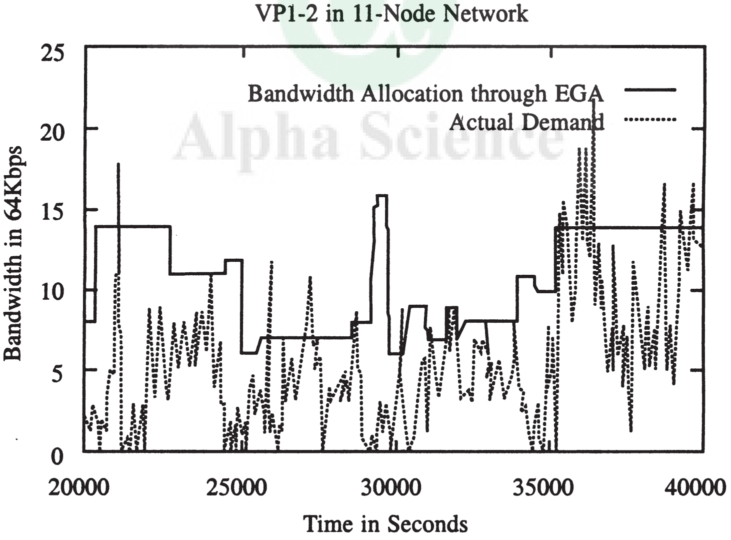


Fig. 4. Bandwidth Allocation for VP1–2 of the 4-Node and 11-Node Network through Evolutionary Genetic Approach (EGA)

depends on the accuracy of forecast of the upcoming size of group of frames. Hence, the number of quality reductions within a given period serves as a metric for assessing the performance of the learning model.

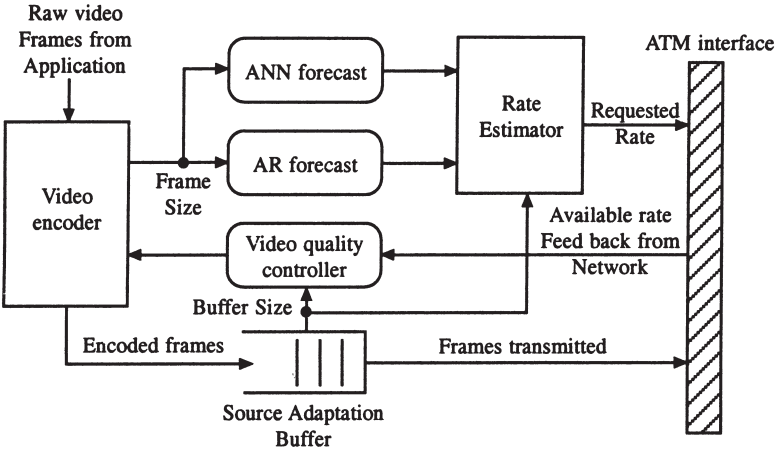


Fig. 5. The Video Source Scheme

Short Term Prediction Based on Estimates: The ABR source periodically submits required-rate requests to the network. The required-rate depends on the size of the frames that need to be transmitted. The required-rate is computed by the rate estimator based on the predicted size of the up-coming group of frames (In a typical MPEG-1 stream the group of frames correspond to the GOPs) and the backlog frames. The proposed prediction model uses a combination of the estimates by the Auto Regression (AR) and Artificial Neural Network (ANN) schemes for predicting the size of the up-coming frames as given below:

$$S = \alpha S_{ANN} + (1 - \alpha)\, S_{AR} \tag{2}$$

where S : Combined Estimate
S_{ANN} : Estimate by ANN
S_{AR} : Estimate by AR
α : Weighting Factor

The value of the weighting factor α used in eqn. 2 is closely related to the scene change rate of the video stream for which the prediction is being made. The performance of the ANN that is trained using one video stream did not perform well for other video streams. The performance difference is due to the large variation in the characteristics of the video streams that makes it difficult for a unique model to characterize all video types. The difficulty in having a unique model for all video types motivated us to use a combined estimate of up-coming video frame sizes using ANN and AR methods. The weightage given to one estimate over the other is controlled by the weighting factor α. In the next section along with the discussion on the simulation studies on ABR-Video transport, we also discusses the choice of right value for α for a given video type.

Simulation Configuration and Assumptions: The network considered for the simulation study consists of two ATM switches connected by a 4 Mbps link. There are 4 MPEG1 sources that are connected to the ATM switch through 10 Mbps links and their corresponding destinations are connected to the second ATM switch through 10 Mbps links. The service class used to carry the video stream is ABR. For the simulation, we assume the maximum delay incurred in the source to be 100 ms. A trace driven simulation approach is used where each video frame is converted into ATM cells and transmitted.

For the ANN based estimator, a *feed forward* neural network is used with *back propagation algorithm* for training the network. The input layer consists of three nodes, the hidden layer consists of six nodes and the output layer has one node. An asymmetric hyperbolic function, see Yegnanarayana (1999) was used as the output function. The AR estimator is the same as suggested in Duffield *et al.* (1998) including the frame quantization algorithm.

Effect of α on Quality of Video Transmission: The number of frame truncations experienced by MTv trace for different values of α was simulated and the results are shown as plots in Fig. 6. For a single connection all the three video traces show improvement in quality for higher values of α. When more and more sources are multiplexed the number of frame truncations reaches a minimum at α = 0.7. *Thus, the adaptive workload model works best with minimum quality reductions when α = 0.7 for all the video traces used in the simulation.*

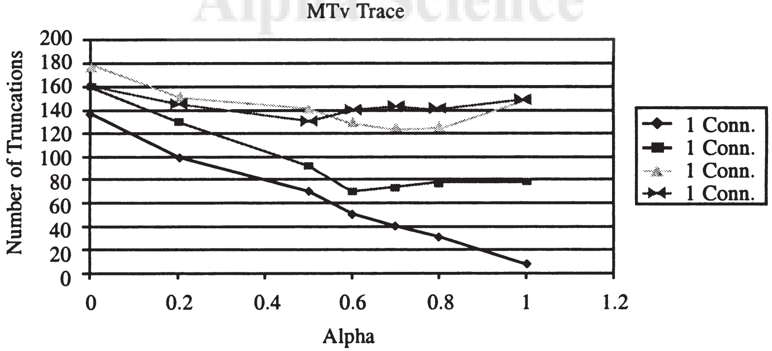


Fig. 6. The Number of Frame Truncations versus *a* for Different Number of Simultaneous Connections for MTv Trace.

Adaptation to Sudden Scene Changes: In order to study the effectiveness of the adaptive workload model in adapting to rapid scene changes, changes using the Learnability index (L) is shown in Fig. 7. The MTv trace has high scene change rate yet, the adaptive model is able to learn the sudden changes within 10 to 15 frames as seen in Fig. 7.

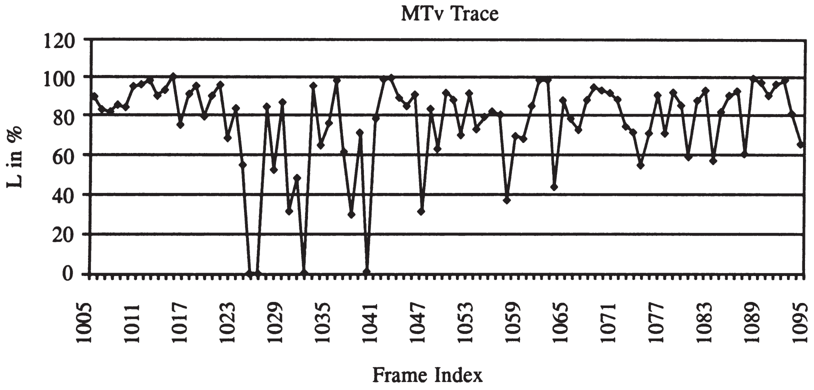


Fig. 7. Learnability index (L) versus Frame Index for MTv Trace

5. Conclusion

Information technology along with high-speed networks has made it possible for information to be available in multimedia form for anyone, anywhere, and at anytime. Such multimedia traffic, induce lot of dynamics in the network load. In this paper we proposed a learning and generative model that can adapt to the dynamics in the workload. Such learning generators enable continuous optimization of the system under consideration. We have illustrated the applicability of the learning generator in two areas viz., VP bandwidth management in an ATM network and efficient transport of compressed video over explicit rate networks. We presented sufficient results from the simulation studies that prove the effectiveness of the learning generators. The adaptive workload model applied in the VP bandwidth management context was able to learn the dynamics in the bandwidth demand pattern and make short term predictions that was within 15% of the actual demand. In the second instance the adaptive workload model was able to learn and adapt to the scene changes in a typical video stream and make short term predictions that improved the video transmission quality. The learning generators also find application in other domains such as security and high-end server performance engineering. These are currently being investigated.

References

1. Adas A.M., (1998), Using Adaptive Linear Prediction to Support Real-Time VBR Under RCBR Network Service Model, *IEEE/ACM Trans. On Networking*, **6(5)**, 635–644.
2. Agrawala Ashok K., Mohr J.M. and Bryant R.M., (1976), An Approach to the Workload Characterization Problem, *IEEE Computer*, **9(6)**, 18–32.
3. Beran J., Sherman R. Taqqu M. and Willinger W. (Oct 1998), Long-range Dependence in Varable-Bit-Rate Video. Traffic. *IEEE Communications Mag*.

4. Burgin J. and Dorman D. (1991), Broadband ISDN Resource Management: The Role of Virtual Paths, (*IEEE Communications Mag*., **29(9)**, 44–48.
5. Calzarossa M. Italiani. M. and Serazzi. G. A Workload, (1986), Model Representative of Static and Dynamic Characteristics. *Acta. Informatica*, **23**, 255–266.
6. Duffield N.G., Ramakrishnan. K.K. and Amy. R. Reibman, (April 1998), SAVE: An Algorithm for Smoothed Adaptive Video over Explicit Rate Networks, *IEEE INFOCOM*, San Francisco, Cal., U.S.A.
7. Ferrari D., (1972), Workload Characterization and Selection in Computer Performance Measurement, *Computer*, **5(4)**, 18–24.
8. ftp://ftp-info3.informatik. uni-wuerzburg. de/pub/mpeg.
9. Goldberg D.E., (1989), *Genetic Algorithm in Search, Optimization and Machine Learning*, Addison Wesley, ISBN: 0201157675.
10. Grossglauser M., Keshav S. and Tse. D., (1995), RCBR: A Simple and Efficient Service for Multiple Time Scale Traffic. *ACM SIGCOMM*, p. 219–230.
11. Haring G., (1982), On State-Dependent Workload Characterization of Software Resources, *Proc. of ACM SIGMETRICS Conf*., p. 51–57.
12. Natarajan B.K., (1991), *Machine Learning: A Theoritical Approach,* Academic Press, ISBN: 1558601481.
13. Ohta S. and Sato K., (1992), Dynamic Bandwidth Control of the Virtual Path in Asynchronous Transfer Mode Network, *IEEE Trans. On Commun*., **40(7)**, 1239–1247.
14. Ramakrishnan K.K. Lakshman T.V. and Partho. P. Mishra., (April 1997), Support for Commpressed Video with Explicit Rate Congestion Control in ATM Networks, *Proc. of NOSSDAV' 96*, Zushi, Japan.
15. Serazzi G., (1986), *Workload Characterization of Computer Systems and Computer Networks*, Amsterdam, The Netherlands: North Holland.
16. Raghavan S.V. and Kalyanakrishnan R., (1985), On The Classification of Interactive Users Based on User Behavior Indices, *ACM SIGMETRICS*, **13(2)**, 40–48.
17. Raghavan S.V., Joseph P.J. and Haring G., (Sept 1995), Workload Models for Multi-Window Distributed Environments, Proc. of 8th International Conference on Modeling Techniques and Tools for Computer Performance Evaluation, Heidelberg, Germany, p. 314–325.
18. Raghavan S.V., Prabhakaran B. and Satish K. Tripathi, (1996), Synchronization Representation and Traffic Source Modeling in Orchestrated Presentation, *IEEE J. of Selected Areas in Comm*., **14(1)**, 104–113.
19. Raghavan S.V., Guenter Haring, Srinivasan V. and Vishnu Priya N., (1998), Learning Generators for Workloads, *Proc. Workshop on Workload Characterization in High Performance Computing Environments, MASCOTS'98, Montreal, Canada.*
20. Yegnanarayana B., (1999), *Artificial Neural Networks*. Prentice Hall of India., 1999, ISBN: 81–203–1253–8.
21. Zhang H. and Knightly E., (April 1995), RED-VBR: A. New Approach to Support VBR Video in packet-switching Networks. *Proc. of NOSSDAV'95*, Durham, North Holland, p. 275–286.

8. The Net-University: An Indian Perspective

Mukul K Sinha

Expert Software Consultants Ltd.,C-17, Almora Bhavan,
NDSE-I, New Delhi 110049

The *virtual university* paradigm that has been developing as a form of the distance education, has now evolved into a full fledged new genre having completely distinct roles, and distinct activities, and which can be defined in its own terminology.

In a virtual university, education is imparted through *virtual classes* where Internet is the basic mechanism for communication, storage, and delivery. The major components of the virtual class are: the courseware and its delivery; the process of conducting classes, interactive participation, and collaborative learning; and the process of evaluation.

In a virtual class, the integrated role of a present day faculty gets decomposed into multiple roles to be performed cooperatively by a set of educational professionals. Experiences have confirmed that education through virtual classes with few face-to-face interactions is far more effective than the present face-to-face teaching.

In India, the societal demand of education has increased by an order of magnitude, both in quantity as well as in quality. It has been proposed that these distinctive societal demand can be satisfied better by a two-tier virtual university, called the *Net_University.*

A Net_University is a consortium of a *core university* along with a set of *associate institutes/universities.* The course is conducted by the core, which would be a centre of excellence, and the delivery is done through virtual class along with face-to-face classes parallely conducted at the associate institutes.

The paper first traces the evolution of the virtual class from the classical distance education, then it discusses the model of the virtual class, the strategy for conducting it, its effectiveness and limitations. The virtual university model is discussed next, and also its inadequacy in the present form in the Indian social and cultural milieu. In the end, it proposes the Net-University model, a two-tier virtual university, that may be more suitable for India.

Key Words: Asynchronous Learning, Computer Mediated Learning, Distance Education, Distance Learning, Courseware, Virtual University, Web Education.

1. Introduction

The university education system is expected to go through a revolutionary structural change in the coming decade or two. More than two decades ago, the *distance education* appeared as an evolutionary extension of the university to cover students who either cannot commit full time to studies or/and cannot co-locate themselves with the university town. In the recent past, with the integration of computer and communication technology, continuous decrease in the price of personal computers, along with the explosion of the world-wide-web, the nature of the distance education system is going through continuous qualitative improvement. Now, this process of improvement has gathered enough momentum to bring qualitative transformation not only to the distance education system but also to restructure the present university education system entirely, see Turoff (1997).

It is expected that in the coming decade or two, most of the universities will get restructured into *Virtual Universities*, a new genre, where the distinctions between *on-campus* and *off-campus*, full-time and part-time, and local and distance students will almost vanish. It is also expected that this restructuring will be neither easy, nor simple, nor uniform. In addition, it would have some distinct local customization.

In India, the education is primarily the responsibility of the state, and at present, it is not at all able to fulfill the societal demand of education, neither in quality nor in quantity. In the last couple of decades, the private sector has been encouraged to setup educational institutions, and a large number of such institutions for higher education have been setup.

In the beginning, it gave some reprieve, but later we realized that very few of these institutions could produce professionals up to the expected quality level. The *extreme shortage of quality staff* was realized the *primary* reason for this. And this shortage became the biggest impediments even for the private sector to further setup additional institutions for higher education.

In addition, India has to face continuous shortage of high quality of staff in the fields of information technology, medical sciences etc., as emigration of good faculty to the other countries in large number would remain unabated, or may even increase, in the near foreseeable future. It is not the shortage of fund, but the shortage of high quality of teaching staff that is the stumbling block. We strongly feel that the societal demand of higher education, in such a large quantity with appropriate quality, cannot be solved through the education delivery mode of the present university system.

In this paper, we assert that the concept of the *Net University*, a two-tier virtual university, can only satisfy the present societal demand of higher education. We also propose that each university of repute must cooperate and coordinate with few other universities and institutions to build a Net_University so that with the limited number of quality staff large number of quality professionals can be produced, and trained without any geographical constraint or relocation.

2. The Distance Education and its Evolution

It is the higher education institutions that took initiative first, to develop the distance education program almost two decades ago. It mostly flourished in USA and Europe. In India, for the higher technical education, Birla Institute of Technology and Sciences, Pilani (Rajasthan) was the first institution that took this initiatives by setting up a *Distance Learning Programmes Division* [BITS 99] for its postgraduate M.Tech/Ph.D. degrees way back in 1988.

In the pre-Internet era, one of the *primary* reason for a university to initiate the distance education program was to *cover new student community* (consumer base) previously untapped, viz., the working individuals who want to pursue their career further, but cannot take study as full-time, nor can relocate themselves either. As the distance education provides the asynchronous mode of learning it is quite natural that it will be extremely appealing for higher degrees, and that also mostly in technical and other professional fields.

If for a university getting sufficient number of full-time students for higher education is a problem (for postgraduate degrees many universities always have problem), and in that scenario the distance education program roping in off-campus students nicely satisfies the requirements of both, the university as well as the students. In addition, it *provides additional revenue* to the university with relatively little extra financial or manpower investment.

Earlier, most of the universities that took initiatives in distance education program focused their attention to cover students from areas restricted by their national boundaries. Later, when the Western and the US universities faced resource crunch, they took distance education program in a big way to *supplement* their finance. In addition, for this aim, they focused their attention to graduate level courses instead, as they attracted students in larger number. In addition, they started extending their *coverage areas beyond their national boundaries*, to the Middle-east, the South-East Asia, and off late to the Indian subcontinent as well.

Now, the distance education has become a big business that has the *global market*. As the English language is the mode of instructions for larger part of the world for higher education, the universities of the US, the UK, and Australia are entering and nurturing this global market. The Australian Government took the distance education as one of their key fields of export business, way back in late eighties, and in the year 1999 their foreign exchange earning from distance education is comparable to that of their wheat export.

In India, the distance education program is expanding but the speed is very slow. Mostly, it is restricted to information technology field, and that also for special certificate courses. Apart from BITS Pilani, Indira Gandhi Open University [IGNOU 1999] is the only other institution that plans to give postgraduate degrees through the distance education programs.

2.1 The Classical Distance Education

The *classical* distance education program did not use either computer, or the

Internet infrastructure for its operations. It provided printed booklets for the course material, a series of discussion sessions (optional), typically one per module by local instructors in few cities convenient to off-campus students, followed by assignments for each module, and finally evaluated by mid-term and the final examinations.

Limitations

It is perceived that education imparted by the classical distance education program is inferior to that imparted by *face-to-face* teaching by an instructor in a class.

It has been accepted satisfactory for the purpose of training where new skills to use/operate/maintain a new machine/package are to be learned.

It has been felt and argued that for the learning of concepts, the classical distance education is not at all good, as learning of concepts requires continuous interaction with the instructor as well as students.

It has been also argued that in the *synchronous face-to-face class*, the instructor carries students along with by continuous monitoring and authority. Depending on the progress of the class, the instructor can change his pace, may digress, or cover additional topics for the benefit of the class. These things are not possible in the *asynchronous* distance learning.

For these above cited reasons, most of the reputed Indian technical universities have consciously refrained to start the distance learning program. Perhaps, they felt that such a program would affect their social image adversely.

Distance Education Centres

Many universities realized these limitations, and tried to improve the classical distance education by associating few institutions/companies in many cities as its *Distance Education Centres* for that course. For each course conducted by the university, there is a part-time adjunct faculty, also called the *mentor* (may be an employee of the company) at the distance education centre for interactions with the students. The company provides infrastructure to the mentor to conduct at least one discussion per course module to the local off-campus students at pre-specified time. To assure high quality of the imparted education these universities demand that students must be attached with one of its distance education centres, see Subramanian (1999).

2.2 Evolution

Since the early nineties, we are seeing that the distance education is going through a continuous qualitative improvement. We feel that its evolution is being pushed both due to the technological push as well as the social push.

Technological Push

The *three* primary technological reasons that are giving qualitative push to the evolution of the distance education are: the *falling prices* of personal computers, the availability of *multimedia capabilities* in PCs, and the proliferation of *Internet with high bandwidth* sufficient for streaming media delivery.

Though these technological advancements are improving all aspects of our socio-economic life, they cannot increase the output capacity of the present university system where the process of imparting education is through face-to-face classes (called FtF *Classes*). On the other hand, incorporation of these advancements in the distance education programs are not only making it far efficient, it is showing a sharp improvement in the quality of the output, the most important metrics of any educational program, see Valianathan (1998).

Social Push

The *two* primary social reasons that are giving qualitative push to the evolution of the distance education are: the incorporation of information technology for the *restructuring of existing socio-economic processes* throughout the world, see Turoff (1997), Hiltz *et al.* (1993), and as a consequence to this, the explosive *increase in the demand of information technology professionals* of all kinds, at all levels.

As the present university system cannot at all fulfill this demand, it is realized that the distance education, with vast qualitative improvement, can perhaps be the only route, which may deliver sufficient professional of the required quality and number.

2.3 New Models of the Distance Education

At present, any search on Internet for the keyword *distance education* will give a large number of higher education schemes, offered by various departments of various universities, including the most prestigious university like Stanford University, USA. Again, different departments/universities have followed different strategies, different models to impart the distance education. The proliferation of various new terms, like *virtual classroom, Turoff* (1995), *asynchronous learning network Hiltz* (1997), *computer aided learning, Murray* (1997), *web-based courses, Merron* (1998), *virtual institute, virtual university Chellappa et al.* (1997), *Web University,* etc., shows that the distance education field is the most active growing area, which is yet to achieve its full potential.

Information technology industry has realized that *the future of the distance education has to be primarily based on Internet mediated communication*, and hence, many companies in the field of information technology, viz., IBM, Apple, Lucent, Macromedia, Microsoft, etc., have come up with a suit of tools/packages to conduct Internet mediated distance education. In addition, many R&D units active in this field have also come with commendable products, see Murray *et al.* (1997), and are offering services as well.

3. The Virtual Class

In this section, we present the most advanced model of the distance education, called *the virtual class*, its various components, and their experiences of the institutes that are following strategies closer to it.

In the virtual class environment, it is presumed that each off-campus student either owns a personal computer with multi-media capabilities, or has an access to such a system. The Internet is the basic delivery, communication and storage infrastructure for all aspects of the virtual class program. The Institute/University offering courses through virtual classes must provide easy access to the Internet to all its staff participating in the program. The Internet having sufficient bandwidth must be available, round the clock without fail, to staff and all students as well, as all interactions between students and instructors, among students, and between students and the university administrative staff is mediated through it. The postal delivery system that was the basic infrastructure for the efficient execution of the classical distance education program has very limited role to play for conducting virtual classes.

The students of a virtual class have to bear the extra expenditure of Internet access, and also the additional telephone bill to access the Internet.

3.1 Courseware

Existence of a courseware is a pre-requisite for conducting a course through virtual class. A *courseware* is the instructional material developed by the author of the course with multi-media content, to impart education through Internet mediated communication. As the instructional material is to be published on the Internet it has to be *structured* in such a way that a student of the virtual class should be able to *explore* and *navigate* through a browser (either standard or special) for its learning exercise. Structuring of a courseware is always distinct from that of a printed book as a book has strong influence of sequential access, and the presentation is limited to text and image media.

A courseware is not restricted to the original writing of the author, but has also *links* to all the relevant information available on the Internet.

Unlike a course book that is written with a strong intention to complement teaching, the structure of a courseware is strongly oriented towards *self-learning* and *collaborative learning* paradigm. Currently, this is very active research area, and researchers are aiming to identify semantic relations in the learning process for each specific knowledge domain so that author can design/structure their courseware to expedite the learning process, see Turoff (1995), Merron (1998), Picciano (1998), Subramanian (1999).

Again, a courseware is not a static instructional material like a book, the author can dynamically integrate new material to keep it up-to-date, revise and improve the content, prune links to outdated materials, set new links to recent publications, and refine the presentation based on the feedback of the students.

Thus, an author can continuously rejuvenate his courseware by incorporating current development and publications, and its content and the structure completely tuned for imparting education with strong orientation of self-learning and collaborative learning.

In FtF classes, an instructor tries to attract attention of the whole class either through oratory skill, dramatic gestures, humor, or coercion. On the other hand, for a passive courseware, the author must compensate his physical absence, at least partially, by his writing skill. The author should use written language through skillful ways, using humor and metaphor, raising inquisitiveness, stimulating questioning, and making opportunities for collaborative participation.

In FtF classes, to explain a concept, the instructor uses his oratory skills, writing/drawing on board, and the body gestures. On the other hand, an author of the courseware has multi-media mode of delivery. A media rich courseware not only makes the course more attractive, it makes concept learning much easier. A student can play repeatedly for better understanding.

To keep the interest intact, the courseware should be decomposed into smaller modules and subodules, and a sub-module should be made accessible only when the class has progressed to that point.

An author has to learn and use an authoring tool to design the courseware, place it on the Internet, and control the access rights. The authoring tool must be powerful enough to let the author easily modify the content dynamically, may be first on his personal computer, and later to be uploaded on the Internet host site.

Students participating in the virtual class are expected to download components of the courseware, annotate it, modify it as he/she wishes, and send to peer students or to the author/instructor for collaborative learning or as a feedback. The access tool must permit them to do that.

As a courseware is expected to have its own lifecycle, *courseware engineering* is emerging a new field that plans to cover all aspects of courseware development process: goal analysis, course planning, instructional design, production, testing, installation, and maintenance (evaluation-revision), see Bostock (1998).

3.2 Strategy for Conducting Virtual Classes

In this section, we elaborate the strategy for conducting virtual classes that have emerged out of experiences of various universities conducting Internet mediated distance learning programs giving comparatively far better result, see Hiltz (1995), Merron (1998).

Course Preparation

The courseware should be structured into a set of modules, and a module should last from one week to at most two weeks. The modular structure of the whole course (not the content details), and the amount of time each module will take should be made available to the students in the very beginning so that they can plan it better.

While the whole course might be already placed on an Internet site, but it's unrestricted access should be prohibited to students to start with. Access to its components (one component equivalent to a FtF lecture) should be permitted progressively, to give students the feeling of progress of the virtual class.

Delivery of e-lectures

To excite student's inquisitiveness, a stimulus material must be put every week. To move the virtual class from one virtual lecture (*e-lecture*) to another, the instructor should place new materials at least twice a week. Students are encouraged to asks questions to the instructor, or discuss with their peers through e-mails. Various tools are available for students to conduct e-conference among themselves, or with the instructor, on specific topics of their choice.

Each students should be given one assignment per week, if possible, different from others, with a strict deadline and a penalty for the delay. Assignment should be placed only after the student has read the new course materials placed that week. This enforces students to log to the course site at least thrice per week.

All students must be continuously monitored, and the authority to access new materials is given to a student only when the student has completed the assignment due satisfactorily, and within the deadline. A student after submitting his/her assignment is permitted to view, read and comment on assignments submitted by others.

The scheme permits a student to interact with the virtual class asynchronously, on its own choice of hours and days, with sufficient but limited flexibility in its duration, ensuring the virtual class to move from one module to the next, together. If a student cannot keep the pace with the class, he is forced to drop-out.

Controlling Collaborative Learning

As knowledge is a social construct, it is imperative that the virtual class must arouse interactions of students with instructor, and also among peer course-mates, in a focused and coordinated way. The amount of constructive interaction generated in the virtual class is the main challenge to the instructor, and the measure of the success of the virtual class.

For understanding of basic concepts, collaborative learning is supposed to be essential. Internet based seminars, debates, case study discussions, group projects, exchange of solutions of assignments, etc., are examples of collaborative learning practices that have been found extremely successful in virtual classes. These activities are also often employed in FtF classes, but their success in the virtual class environment is far more.

In the FtF classes, the collaborative learning is centered around the instructor, and students have to interact synchronously, with very little time to react. Hence, only a small subset of students is usually active participants. The main reason for the success of collaborative learning schemes mentioned above is the asynchronous nature of the virtual class environment where the students have enough time for reflexive thinking, and respond only after consolidating their thought. Neither they are under stress of their peers' scrutiny.

Like assignments and unit tests, even mid-term and final examination questions may be distinct for each student. Unlike the assignments and unit tests, the mid-term and final examinations are mostly conducted in controlled environment having enough flexibility for students to choose their own time, and site of the controlled environment within a limited but flexible duration.

3.3 Experiences of Conducting Virtual Classes

There are few reports viz. Hiltz (1997), Murray (1997), Picciano (1998) that have enumerated conclusions drawn after conducting virtual classes for few years.

In a virtual class, for students who are participating for the first time, the first e-lecture and the first assignment are very crucial. The first e-lecture and the first assignment must encourage/force students to respond/ask questions electronically, so that their inhibition is broken, and later, they should participate in the class according to their own natural mental process.

The response to student's e-mail question should be within twenty-four hours, and in no case it should exceed forty-eight hours. On an average, a student sends three to four e-mails per week to the instructor. It is quite obvious, that an instructor will need assistants to satisfy students' queries within a deadline time frame.

It is also advisable for the instructor should make himself/herself available in on-line chat mode, at least twice a week, for an hour or two, to conduct class conference for the students to moderate synchronous electronic interactions. It has been also realized that a meaningful class conference can be conducted if the size of active participants is at least ten, but not more than twenty-five. In case, the class size is large, it is advisable to split into sections, and different instructor should conduct each section.

The faculty conducting the virtual class felt that in relation to the FtF classes they have to work far harder for the development of the courseware, and to keep it up-to-date. Replying e-mail queries itself was taking at least an hour per day, making them feel like perpetual professor! Again, moderating the class conference, selecting and checking the assignments, monitoring student's participation in collaborative learning schemes need far more help from teaching assistants.

3.4 Effectiveness and Limitations of the Virtual Class

On completion of a virtual course, the students' felt that the quality of education was as good as, and even better than the FtF classes. The most distinguishing feelings they got in a virtual class were that they had better access to the instructor, worked harder, participated more actively (due to *asynchronous reflexive thinking*) in the class, learned more, and the articipation was more convenient (as they could choose their own time and pace). It gave them far higher satisfaction.

A virtual class is better suited for motivated, disciplined, and high ability students. For less motivated student, the asynchronous mode of interaction may encourage the tendency to postpone, and this leads to higher percentage of drop-outs.

Again, good reading, writing, reflexive thinking skills are required for better participation in the virtual class.

The virtual class does not provide room for the development of verbal expression, leadership quality, and interpersonal relationship.

4. Mixing Virtual Class with Face-to-Face Class

In the strategy of conducting the virtual class described in the previous section, the Internet is the basic medium for storage, delivery, and communication infrastructure. Lectures are through publication on the Internet, and interactions are through e-mails, and e-conferences. There is no visual interaction between the instructor and students, or among students.

It has been felt that the virtual class is more satisfying to students when it is supplemented by a couple of face-to-face classes with the instructor, one in the beginning, preferably one also in the middle, and finally one in the end.

Further, it has been felt that when a virtual class is augmented with video lectures it strengthens the imparting of education even more, see Hiltz (1995). The video lectures can either be broadcasted through television, or sent to students as video tapes by post periodically. The tapes contains the video of the instructor giving lectures, in a parallely (or a previously) conducted FtF class.

Similarly, it has also been felt that permitting the students of FtF classes to access the courseware of the virtual class strengthens their understanding process as well, see Murray (1997).

It is now being felt that the same course should be offered simultaneously through FtF class as well as through virtual class, and both modes of delivery should be synchronized. Students should be given choice to opt in either mode, and in some cases, may be even permitted to switch from one mode to another.

In addition, students of FtF classes are also permitted to participate in the collaborative learning schemes, assignments submission, and the evaluation process of the virtual class.

The experience of one such experience where various mode mixes were experimented has been reported in detail in Hilt (1995). The result shows that on the overall students' rating, the virtual class augmented with video is comparatively the best, followed by the virtual class having a parallel FtF class.

We strongly feel that, in future, the virtual class augmented with video, and supplemented with a couple of ft. meetings would emerge as the best mode of delivery for imparting education.

4.1 Decomposition of the Roles of a Faculty

In FtF classes, a faculty (instructor or teacher) has to perform multiple tasks associated with imparting education synchronously, and hence, he has to play multiple roles, one role for each task. Apart from the broad curriculum of the course, it is the faculty who first decides in detail, the whole course content and then designs the sequence of instructions to cover all aspects of knowledge. The faculty may prepare transparencies, lecture notes, etc., for conducting classes and/or for distribution among the students.

The faculty who conducts the FtF classes, gives lectures, interacts with the class synchronously, ask questions to various students to perceive the amount of absorption, and may do on-line modifications in the pace, and/or the sequence of the pre-planned instructions to carry the whole class together.

The faculty also answers questions raised by students, clarifies doubts, encourages collaborative learning among students, and monitors it. To cajole, control, and coerce the class (to make the imparting of education smooth and within the desired time frame) are also among the responsibilities of the faculty.

The faculty also sets test papers, question papers, assignments, group projects etc. He/she also evaluates, and finally gives grades to the students.

In addition, the faculty has to share a large administrative load of running the university (and hostels) as well.

In short, a faculty of an FtF class has to perform multiple roles in an integrated way. A faculty may be excellent in performing some roles, good in some other, and may be extremely bad in some other. Neither he is ever trained for such an integrated role, nor it is possible to expect one person to excel in all types of roles. In a FtF class, the mode of imparting education is synchronous, as well as face to face, and hence, it is not possible to decompose the various tasks so that different tasks are as signed to different individuals.

In a virtual class, as the education is imparted asynchronously, and in virtual mode (opposed to face-to-face mode), a group of faculty, each good in a specific role, can be involved to conduct a course. The integrated role of the present day university professor can be performed combinedly by a set of educational professionals viz., an *author* (who will develop the course in detail and publish on the Internet), a *lecturer* (who will deliver lectures to be broadcasted, or taped), an *instructor* (who will interact with students synchronously in parallel FtF classes), the facilitators (who will interact with students through e-mail/on-line chat), an *event coordinator* (who will coordinate all collaborative learning exercises like seminar, debate, case-study, etc.), a *monitor* (who will monitor each student's participation in the virtual class, and counsel the student accordingly), an *assigner* (who will set assignments, test and question papers), an *examiner* (who will examine the assignments, and the answers given by students, and give grades to them), and a *course-coordinator* who will coordinate all the activities associated with the conducting of the course.

In short, the virtual class provides objective basis for division of labour in imparting education, and we strongly feel that various professions associated with imparting education in virtual mode would emerge in near future. Again, these professionals need not be hired on full-time basis, nor they need to be co-located physically.

4.2 Decomposition of University Roles

At the highest level of abstraction, the *primary role* of a university is to impart education to students synchronously through face-to-face classes, and awards degrees to the successful students.

To impart education, the university has to perform multiple tasks, viz., to develop course curricula for each degree to be awarded, to hire faculty to deliver the various courses to the students, to provide learning environments and tools (lecture rooms, laboratories, software tools, libraries, canteens, etc.), to conduct tests and examinations, to arrange for evaluation of students, to award degrees, and to provide various other support activities.

As the education is imparted synchronously, and in face-to-face mode, the university has a very important *secondary role* to play, i.e., to provide physical environment, viz., land, building, and a score of other overhead activities to keep the physical environment conducive to academic activities, the primary role of the university.

Usually, for a university, the volume of work associated with its secondary role is far more than that associated with its primary role. And hence, it is natural to extrapolate that a virtual university, i.e., a university imparting education through asynchronous virtual mode, will be far efficient, give better quality of output, and will need lesser investment to fulfill the societal obligations of education.

5. The Virtual University

In this section, we describe the model of the virtual university, its distinguishing features with respect to a physical university, and the prerequisites to run a virtual university efficiently.

5.1 The Virtual University: The Model and Distinguishing Features

A virtual university imparts education through virtual classes sometime augmented with video, and supplemented with a couple of FtF meetings.

The distinguishing features of a virtual university in contrast to a physical university can be summarized as follows:

— It is a university in abstract with global coverage, whereas a physical university has to be located at a city,

— It imparts education asynchronously in multiple modes of delivery,

— Its education process flows from a courseware (not from course-books) that can be media-rich in its presentation,

— Its education process is learner centric as against faculty centric of the FtF classes,

— It provides better attention to each individual student,

— It provides better collaborative learning environments,

— It decomposes the multiple roles of a traditional faculty, facilitating multiple professionals to participate in a coordinated way in conducting a course,

— It is as good if not better than the physical university, as far as the quality of education and output.

5.2 Prerequisites for a Virtual University

It must have high bandwidth, high quality communication network accessible to all students, all professionals participating in education imparting, and university administrators.

It must have hardware and software tools to store and control access to the courseware, to facilitate interaction among students and academic professionals, and to administer university's other administrative activities.

It must acquire courseware already developed, or get developed courseware of its need.

It must hire services of faculty with good oratory skills for broadcasting (or video-taping) the lectures.

It must hire services of various kinds of professionals to conduct the various components of the virtual class.

It should control the process of evaluation, and the delivery of degrees, preferably electronically, to successful students.

It should hire staff for administrative activities to be done, mostly electronically.

6. The Net_University

We feel that restructuring the existing university system in the direction of the virtual university would take a specific route, the route of the Net_University. The route would depend mainly on the characteristics of the present societal demand of education, and also the specificity of the Indian socio-cultural milieu.

6.1 The Educational Demand: The Present Indian Scene

The present Indian educational demand has following distinguishing features:

- the demand for higher education, especially in information technology, is extremely large,
- the present number of institutes/universities are not at all sufficient to satisfy the societal needs,
- the number of well qualified staff is insufficient to provide quality of education in FtF mode,
- the number of well qualified staff would be in short supply even in future as better remuneration and/or quality of life encourage them to emigrate to other countries,
- the best educational institutes are mostly in state sector, and the state is not having sufficient fund to further expand them,
- there has not emerged any model to harness investment from private sector for the educational institute of repute so that its productivity of the limited number of well qualified staff can be increased exponentially.

6.2 The Educational Demand and The Indian Socio-cultural Specificity

As discussed earlier, the classical distance learning program emerged in the West to cater to the working people who want to acquire higher degree without committing full-time, and without co-locating themselves. In contrast, in India, we are looking to the virtual university model to satisfy the societal educational demand with specific constraints deliberated in the previous subsection.

The salient points of Indian socio-cultural specificity are followings:

- the major educational demand is emerging, not from the working people, but from fresh students who are ready to commit full-time,
- the desirous students are spread all-over India, and they are ready to co-locate themselves,
- the students would love to follow regular timings, and synchronous group teaching as far as possible (the concept of studying asynchronously, through a computer sitting alone at home is not socially acceptable at present),
- the students are very particular in getting quality of education, not just the degree,
- the competition for the admission into the institutes of repute is fierce whereas seats in many new/private institutes with ill-equipped laboratories, and lesser quality of staff, are consistently remaining unfilled.

6.3 The Net_University

The abstract model of the virtual university, neither is suitable in the present Indian social and cultural milieu, nor it can be structured afresh. We propose the Net_University model, a *two-tier virtual university*, that will be more suitable for India, and that can be structured with co-operation from existing universities and institutes of excellence. Such cooperative approach of institution partnership are also emerging in the west with good result, see Kroder *et al.* (1996).

The Net_University is not a new university, but *a consortium of universities*.

A Net_University has one university of excellence as its core (called *the core university*), and a set of other existing universities or institutes are associated (called *the associate institutes*) with it in a federal manner.

The course is conducted under the control of the core university, and the degree is awarded by the same.

The courseware is designed by the core university, but the development of media-rich content is done with participation of associate institutes. All courseware is associated with instructor's guide for conducting FtF classes.

The delivery of the education should be through virtual class along with the FtF classes conducted in parallel at multiple associate institutes.

The entire education process should be shared in a co-operative and federal manner by the faculties of the core university, and the associate institutes.

The course-coordinator, the author, the lecturer, the monitor, and the examiner must be the faculty of the core university, guaranteeing the quality of course content, its delivery, special attention to each student, and also the evaluation process.

All other educational activities, like conducting the FtF classes with video-tapes, the interactions with students, the laboratory activities, the collaborative learning activities, etc., have to be conducted by the professionals of associate institutes for their students. Any student will also be free to interact electronically with the faculty of the core university.

In addition, the author/lecture of the core university should pay one or two visits to each institute, at pre-specified point of the course, mainly to conduct special FtF meetings with the students of the associate institutes.

The private sector can participate in starting associate institutes, in the initial investment required to develop the courseware and the communication infrastructure, and also in providing administrative services to run the Net_University.

7. Prospects

In last couple of years, there are multiple western universities that are seriously trying to enter in the Indian education market through various types of distance education programs. It is well-established fact that almost all educational universities of repute in the English speaking advanced countries have a good number of Indians among their faculty. It is also well established fact that India is going to be the main source of highly educated technical manpower which is necessary for all the western advanced countries to keep their engine of development running.

In this scenario, we must realize that India having a stock of highly educated technical power should be the main exporter in the world education market, and not the importer. But this can only happen if we develop and master the technology of conducting virtual university indigenously. Again, due to lack of investment, and absence of sufficient quantity of highly qualified faculty existing at a single university, this can be done only through co-operation.

The Net_University model, not only provide a road map to develop indigenous technology to satisfy the educational demand of the home sector, it also gives opportunity to master it, perfect it, and improve it to a higher quality as the home market itself is very large.

Activities like authoring a course, media-rich courseware development, facilitating collaborative learning of the virtual class, evaluation etc., can be done in India even for the university of excellence of the most advanced countries.

Again, the Net_University can very easily integrate institutes of Middle-East Asia, South and South-East Asia, and English speaking African countries, making it a good source of foreign exchange earning.

Initiatives have to be taken from the Indian Universities that have established their international reputation over the years. I hope we do not miss this opportunity.

References

1. Off-Campus Programmes: Distance Learning and Collaborative Programmes in Part V, Bulletin 1998–99, Birla Institute of Technology and Sciences (BITS), Pilani, India.
2. Bostock, S., (March 1998), Courseware Engineering: An Overview of the Courseware Development Process, URL: http://www.keele.ac.uk/depts/cs/Stephen_Bostock/docs/atceng.htm
3. Chellappa, R., Barua, A. and Whinston, A.B., (1997), An Electronic Infrastructure for a Virtual University, Comm. ACM, **40** (9).
4. Hiltz, S.R. and Turoff, M., (1993), The Network Nations: Human Communications via Computer, MIT Press, Cambridge, Mass.
5. Hiltz, S.R., (1994), The Virtual Classroom: Learning Without Limits via Computer Networks, Norwood NJ: Ablex Publishing Corp.
6. Hiltz, S.R., (March 1995), Teaching in a Virtual Classroom", International Conference on Computer Assisted Instruction, ICCAI 95, Hsinchu, Taiwan.
7. Hiltz, S.R., (1997), Impact of College-level Courses via Asynchronous Learning Networks: Some Preliminary Results, Journal of Asynchronous Learning Networks.
8. Hiltz, S.R. and Wellman, B., (1997), Asynchronous Learning Networks as a Virtual Classroom, Communication of the ACM, **40** (9).
9. School of Computer and Information Sciences: Information brochure and application form, Indira Gandhi Open University (IGNOU), New Delhi, January 1999.
10. Kroder, S.L., Suess, J. and Sachs, D., (May 1996), Lessons in Launching Web-based Graduate Courses, T.H.E. Journal.
11. Merron, J.L., (1998), Managing a Web-based Literature Course for Undergraduates,

Online Journal of Distance Learning Administration, **1** (iv), State University of West Georgia, Distance Education.

12. Murray W. Goldberg and Sasan Salari, (June 1997), An Update on WebCT (World-Wide-Web Course Tools) — a Tool for the Creation of Sophisticated Web-Based Learning Environments, *Proc. NAUWeb '97—Current Practices in Web-Based Course Development,* Flagstaff, Arizona.
13. Murray W. Goldberg, (1997), CALOS: First Results From an Experiment in Computer-Aided Learning, *Proc. ACM's 28th SIGCSE Technical Symposium on Computer Science Education.*
14. Pant, M.M., (1999), Internet and Education: With Focus on Higher Education, Prospective in Education, **15**, Special Issue.
15. Picciano, A.G. (March 1998), Developing an Asynchronous Course Model at a Large Urban University", *Jour. Asynchronous Learning Networks,* **2** (1).
16. Rudenstein, N., (May 1996), The Internet is Changing Higher Education, Harvard Conference on the Internet and Society.
17. Subramanian, K.R.V., (1999), Web Based Courseware—Issues, Challenges and Prospects, Proc. Compu. Soc. of India.
18. Turoff, M., (March 1995), Designing a Virtual Classroom, International Conference on Computer Assisted Instruction, ICCAI 1995, Hsinchu, Taiwan.
19. Turoff, M., (April 1997), Alternative Futures for Distance Learning: The Force and the Darkside, Invited Keynote Presentation, UNESCO/Open University International Colloquium.
20. Turoff, M., (1997), Virtuality, Comm. ACM, **40** (9).
21. Valianathan Purnima, (1998), CBTs: Making the Best of Both Technology and Education, Proc. Compu. Soc. India, New Delhi.

9. CHOIS for Disseminating Infotainment/ Internet in Indian Society

Mohan Tambe

Innomedia Technologies Pvt. Ltd., 3278, 12th Main HALM, 2nd stage
Indira Nagar, Bangalore-560008

CHOIS technology (Cable TV based Home & Office Interactive Services), has been innovated with the end-objective being empowerment of Indian homes with infotainment services. The technology ushers "Interactive Media" as the fourth broadcast media in the world. Like other broadcast media it has the least cost of reaching to every home and can sustain itself through advertisement revenues. It piggy-backs on existing infrastructure such as cable-TV and telephones in order to reach the homes at the least additional cost. CHOIS services include information-on-demand, video-on-request, messaging, internet access and e-commerce applications. CHOIS services are already functional in many areas of Bangalore and Mysore on a pilot basis, and are now poised for deployment throughout the country. The paper elaborates on the nature of the CHOIS technology, its goals and the philosophies behind it.

Introduction

There has been more speculation about what would constitute Electronic Media and how it would transform society, than for any other media. Enough has been written about it at times when its implementation was still years away. Little did people realize that a new emerging paradigm might turn out to be something very different from what they could imagine.

Just like the Internet was not prophesied, but simply one day surrounded the world, the emerging layers are always much different from the ingredients they are made of. Internet is more like the roads which link all the cities together. What is of more importance are the roads which link up the homes and offices in each city: they are the ones which define the day to day transactions of a society. Interactive Media is more akin to the city roads, which have to connect each home.

In Bangalore, Innomedia Technologies, has been innovating Interactive Media solutions from 1996 hand-in-hand with cable-TV users. This is especially relevant for the developing countries such as India, since the social ethos needs infrastructure, economics etc. are very different than the western world. Like culture, solutions which will put people "interacting" have to originate right out here. These have been the needs, which have moulded the CHOIS technology.

Chois Technology

CHOIS technology has converted the passive one-way broadcast mode of television into a two-way communication medium. The technology ushers Interactive Media as the fourth broadcast media in the world. Like other broadcast media it has low cost of reaching to homes and can sustain itself through advertisement revenues. It piggy-backs on existing infrastructure such as cable TV and telephones in order to reach the homes at the least additional cost. CHOIS services include information on demand, Pay-per-view, Video-on-request, messaging, Internet access, shopping, surveys, polls and e-commerce applications. CHOIS services are already functional in many areas of Bangalore and Mysore on a pilot basis and are now poised for deployment throughout the country.

How Chois Works

A National Telecomputer based at Bangalore, which stores all the contents, is linked by satellite to different Telecomputers in the Metros. These in turn are linked to CHOIS servers at the Cable headend which are then linked to CHOIS pads at the users end. CHOIS pads are analog/digital set-top boxes, which come with a remote and are attached to the viewer's TV sets.

The National Telecomputers has the master repository of the contents and coordinates all the telecomputers through out the country. It is responsible for updating all the CHOIS servers country wide through a satellite data broadcast and keeps a record of usage statistics. The Telecomputers in the Metros control the CHOIS server in their region and allow addition of local content. They have connection to several online services and can also act as Internet Service Provider for the CHOIS servers. The CHOIS server installed at the Cable head end provides Interactive Media services. The CHOIS pad offers high-speed data reception through an internal cable modem, which offers greater bandwidth than the standard telephone modem.

Aim of Chois Services

Since the progress of the society depends on the quality of information that percolates into each home, Interactive Media has to be molded with considerable care. The short-term expediencies of getting audience hooked on the new media cannot vitiate the long-term goals of empowering and enlightening them. CHOIS services thus have been blended with considerable care. Acknowledging that entertainment has to be the primary focus, the aim is to inform people about the quality entertainment around them. This is done through reviews, synopsis, clips, movie trailers etc. The idea is to give enough preview information to allow a person to decide on what to see in more details. CHOIS services thus provide an index to the best choices in entertainment. It would be possible for the user to even see a full-length movie or hear an album, by booking it through the CHOIS Video-on-Request facility.

CHOIS services provide the essential general knowledge for effective interaction in society. Providing in depth information— Railway timetable, Bus schedules, Airline schedules, Entertainment and other events in the city for making better decisions remains an important goal of the CHOIS services. Finally although entertainment is substitutable, Information is not. Long term progress in dependent on the information inputs. CHOIS services would endeavor towards cultivating people towards use of information which can make difference in their lives.

Information Inundation

Human mind has finite capacity for information absorption. The more it is inundated with information the more it loses. What remain are fragments of different information, leading to a scatter-minded behaviour. Information which is useful, is that which can be applied directly or indirectly for achieving tangible results. Information, which is not sought for, but is thrust, doesn't get integrated with rest of the information in the brain—this results in more confusion and a feeling of ill at ease.

When faced with a deluge of information people turn off from the whole thing. Thus, they may stop visiting libraries, or watching an informative channel. They figure out that their life goes on fine, when they are not distracted by unnecessary information.

On the other hand, the capacity in people for entertaining material is much higher before it becomes surfeiting. This is because most of it doesn't contribute much to new information, but just consolidates further whatever a person knew before. Success of the Hindi movies rests on use of the stereo type formulae, with a few new twists.

Interactive Media, thus should not overwhelm people with encyclopedic nature of information, but rather should aim at consolidating and refreshing what they are already aware of, or would be of help in their day-to-day life.

Needs of People

A common method for gauging the need for newer products and services is through a Market Survey. This indeed was done at the initial stage. People opined about the nature of interactive services, which they would be interested in, and how much they would be willing to pay for them. Many of these input differ significantly from what we found later from the actual operations.

The response of the people to queries are more designed as ideal. Responses are also based on what they think is expected from them. People in fact hide their basic primitive motives while putting across a bold logical front. Unlike the earlier predictions, we have found a very poor response from people to pure information and educational material. Even for service such as timetables for airlines, trains and buses we found the usage to be of infrequent nature. This appears consistent to the average travel needs of the people. In contrast, we found heavy usage of glamorous contents dealing with films, fashions and celebrities. This was not surprising, as most of the TV channels were full of them. The success of a new service depends on satisfying the latent needs of the people. It is well known that success of pay TV and Internet services has been largely due to presence of pornographic material. In fact, in India and several other countries which have strict censorship laws the pay TV has not succeeded. The broadcast regulations and fear of erosion of our value system is also hindering the proliferation of DTH in India.

Success of interactive media then has to be through providing material of good quality, which also fits within the needs of the people. Entertainment and glamour material which people are already used to, fits within this requirement. Need for other types of contents will emerge after people have become familiar with the new medium.

TV and Interactive Media

There has been a debate going on for several years as to what will be the right medium for introducing Interactive Media in homes: the TV or the PC. By now the issue is more or less settled in favour of the TV simple because of its ubiquity and familiarity.

There are more important reasons though as to why TV is the more suitable medium. At the outset one should realise that if we consider the display aspects of TV and PC, both are basically equivalent. It is possible for a computer screen to be shown on a TV and a TV programme to be shown on a computer monitor. TV, however, is not suitable for browsing many pages of text, since reading text becomes strainful because of the flicker of the interlaced display. Moreover, we are not used to reading substantial information from a long distance. A PC screen in contrast has a higher resolution and a flicker free display, which is read, from a short distance. Browsing large amount of information on PC thus becomes as convenient as reading a book. Furthermore, everyone witnesses real life events as well as movies, in presence of other people. In fact, viewing an event together (in company) enhances its appeal, since it allows immediate sharing of the experiences. The spontaneous response of the audiences can continuously be compared to ones own, which adds another dimension to the perception of the event. All this is certainly not possible on a PC screen.

The TV and PC would over time overlap in their functions and complement each other. People at homes will first become comfortable with interactive information services through their TVs and would then venture into using PCs. TV would thus be the ideal medium for introducing Interactive Media into our homes. The contents of the Interactive Media have to appeal more to the heart, intuition and feelings and be a conduit for socially useful material. As Interactive Media has the power to reshape society it has to emanate from the TV. The PC would co-exist in lesser number and would primarily be used as a work tool and a front end for Internet.

Internet in India

The spread of Internet in India is restricted to the offices, academia and to a certain extent to businesses. It has made very limited inroads into the average homes. There are only about 2 lakh Internet subscribers under the government service provider VSNL. This is a very small miniscule of the educated population in our country. Only recently the regulations have permitted Internet spread through private operators. This is yet to percolate into the different parts of this vast country. Its spread is restricted to the populated areas amongst the larger cities where there is a considerable penetration of computers. The communication costs associated with usage of Internet acts as a deterrent to most of the vast middle class in our country.

The communication infrastructure in our country is not so efficient as to support extensive use of Internet and full potential of the net cannot be harnessed. Even with this limited usage and penetration of the Internet subscribers the bottle necks in the infrastructure cannot support meaningful access to the host of information and services available on the World Wide Web. Through high cost backbone infrastructure these services can of course be harnessed. Bringing these services to the common man at an affordable cost is possible through deployment of CHOIS technology at the cable operator.

Broadcast vs Point-to-Point

Multimedia information is known to guzzle up bandwidths of existing networks. The point-to-point networks naturally gets bottle-necked while catering to the diverse requirements of a relatively few people. In contrast, the broadcast networks allows the multimedia contents to be sent simultaneously to a large population. They, therefore, becomes a natural choice for developing countries with population problem. While point-to-point telephone networks have worked well for developed countries with sparse population, their reach has been poor in countries like ours. The media which have really worked in India are of broadcast nature such as newspaper, radio and TV.

Point-to-point network will become a bottle-neck even for developed countries, since the demand for contents will far exceed the network capacity. Field trials and pilot projects conducted in USA for Interactive Media services have come to the conclusion that it is important to re-orient the services so that they can be broadcast one-way.

Broadcast information becomes affordable to the end-user since the contents gets sponsored through advertisers. An analysis of the contents needed at home indicates that 95% of the requirement of people can be met with just 5% of the contents.

Conclusion

Information infrastructure is crucial for growth of a nation, however, the enormous costs for establishing it are beyond most of the nations, even the developed ones. The cable TV industry has already demonstrated how it can rapidly spread and consolidate itself in the country without requiring huge investments or initiatives of the Government.

CHOIS technology has moulded the Interactive Media in a manner that it can be spread in the most cost-effective manner through the cable-operators themselves. It aims first to provide the basic Interactive Media services to everyone, and then enrich it with Internet. The Interactive Media is designed to enlighten and empower people at homes, thus accelerating the ascent of the society.

References

1. Gunther Marc, (1996), The Cable Guy's Big Bet on the Net, Fortune, Nov. 25, 1996.
2. Tambe Mohan, (March 1998), CHOIS: Cable based Home & Office Interactive Services, Cable Quest.
3. Guy Kawaski, (1999), Rules for Revolutionaries, Harper Collins.