Lecture notes for

M.Sc. Data Communication Networks
and
Distributed Systems

D51 – Basic Communications and Networks

Saleem N. Bhatti [1]
Department of Computer Science
University College London

October 1995

[1] These notes are based on a script originally produced by Dr. Duncan Wingham, formerly of the Department of Electrical and Electronic Engineering, UCL

1 Introduction and Overview of Communication Systems

1.1 Classification of communication networks

Communication networks are usually defined by their size and complexity. We can distinguish four main types:

- Small. These networks are for the connection of computer subassemblies. They are usually contained within a single piece of equipment.
- **Local area networks (LAN).** These networks connect computer equipment and other terminals distributed in a localised area, e.g. a university campus, factory, office. The connection is usually a cable or fibre, and the extent of the cable defines the LAN.
- **Metropolitan area networks (MAN).** These networks are used to interconnect LANs that are spread around, say, a town or city. This kind of network is a high speed network using optical fibre connections.
- **Wide area networks (WAN).** These networks connect computers and other terminals over large distances. They often require multiple communication connections, including microwave radio links and satellite.

LANs may have a number of different physical configurations, i.e. the manner in which workstations on the LAN are physically connected. The physical configuration will often reflect the **media access control (MAC)** method used to allow the workstation to gain access to the connection media. Most LANs are **shared medium** networks, i.e. there is effectively one link between all the workstations on the LAN and each must wait its turn for the use of the media. There are various methods of controlling how and when a workstation gets its turn to use the media, e.g. **carrier sense multiple access with collision detect (CSMA/CD)** or the use of **token** passing.

One of the most popular configurations is the **bus** arrangement as used by **Ethernet** and **token-bus** (Figure 1).

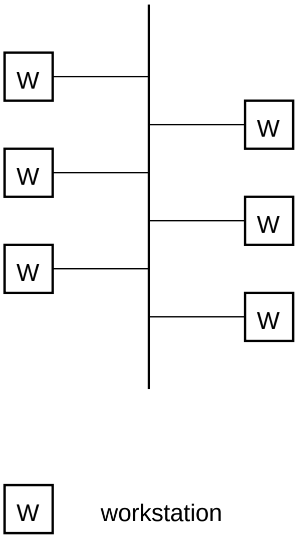


Figure 1: LAN bus arrangement

In this configuration, all the workstations effectively hang off from one piece of wire. Another common configuration is a **ring**, for example **token-ring** (Figure 2).

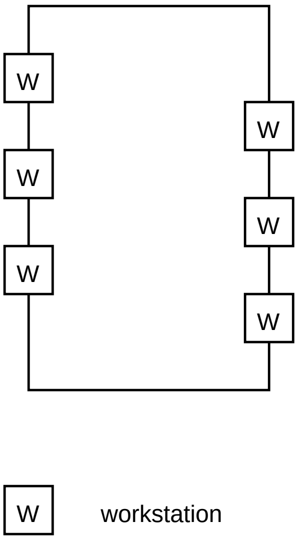


Figure 2: LAN ring arrangement

A **star** configuration is not seldom used to interconnect workstations on a LAN. Instead it is used to interconnect LANs via an **exchange** or **switch**. However, recent renewed interest in the use of **unshielded twisted pair (UTP)** cabling and also due to the interest in using **asynchronous transfer mode (ATM)** technology (and other high-speed networks), we will see the use of star configurations a more common occurrence (Figure 3).

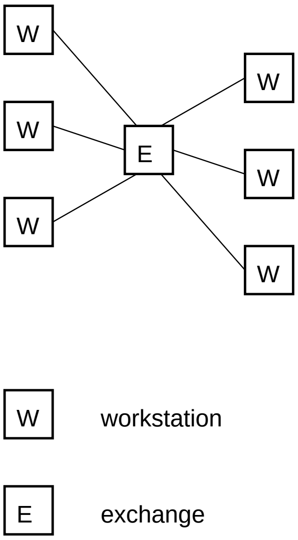


Figure 3: LAN star arrangement

You may hear these configurations referred to as **LAN topologies**. LAN **segments** may be interconnected by use of a **backbone** LAN that allows communication between the segments. The segments help to localise traffic, e.g. within an office or a single floor in a building. Each segment is a LAN in itself, but is connected to the backbone via a **bridge** (Figure 4).

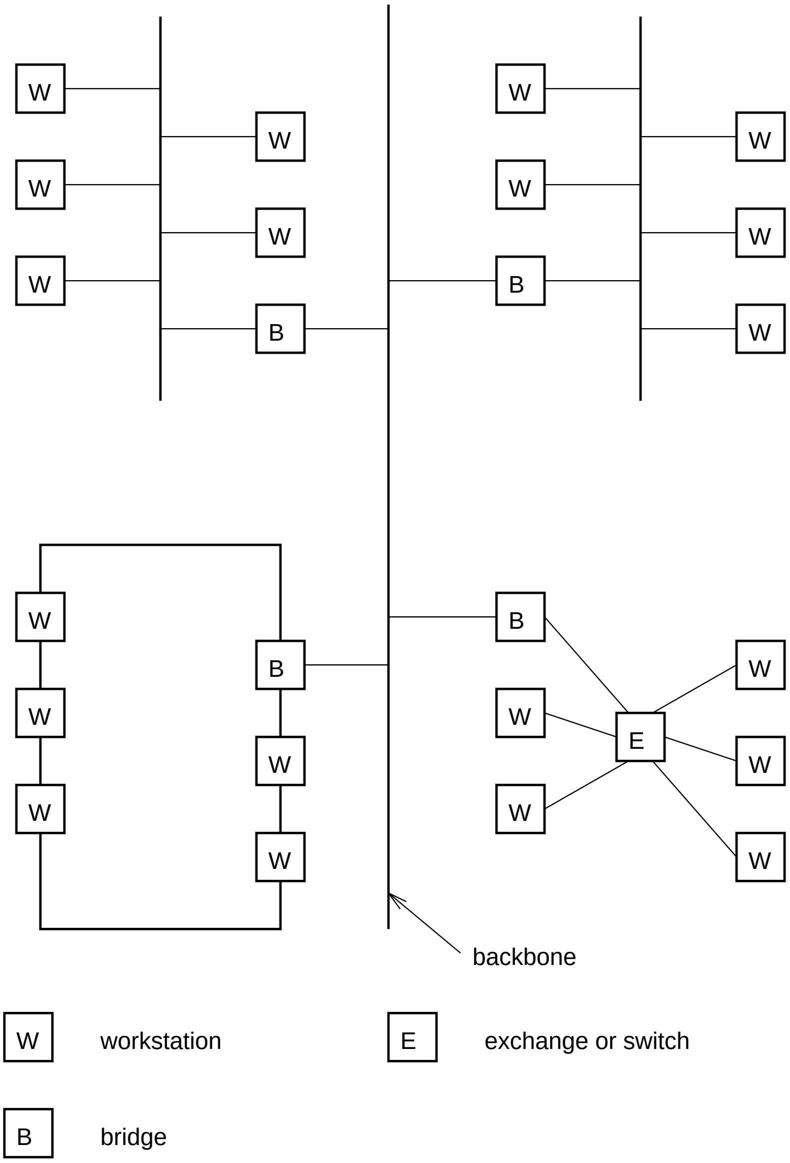


Figure 4: Backbone interconnecting LAN segments

WANs often exploit public networks such as the public telephone system (Figure 5). There are two types of public network:

- **Public switched telephone network (PSTN).** This is the ordinary telephone system. This system exists in all countries of the world. It was designed specifically for the transmission of voice communication. Digital systems employing this network must produce a 'voice-like' signal, and accept the low transmission rates.

- **Public switched data network (PSDN).** This is a public network designed specifically for the transmission of digital data. They arose from privately owned WANs that required higher performance than the PSTN. Many countries of the world are introducing PSDN services. They can support much higher transmission rates. An **integrated services digital network (ISDN)** is the term given to all-digital networks that can carry simultaneously voice and data communication, and offer additionally a variety of teletex services. ISDN services are being introduced all round the world and are indeed on offer in the UK.

MANs are now taking some of the roles that WANs once had in particular environments for LAN ↔ LAN interconnection, due to the high speeds that they offer compared to most PSTNs and PSDNs, e.g **fibre distributed data interface (FDDI)** and **distributed queue dual bus (DQDB)** offer 100Mb/s, whereas a PSTN line may offer, say, 19.2Kb/s with a fast modem.

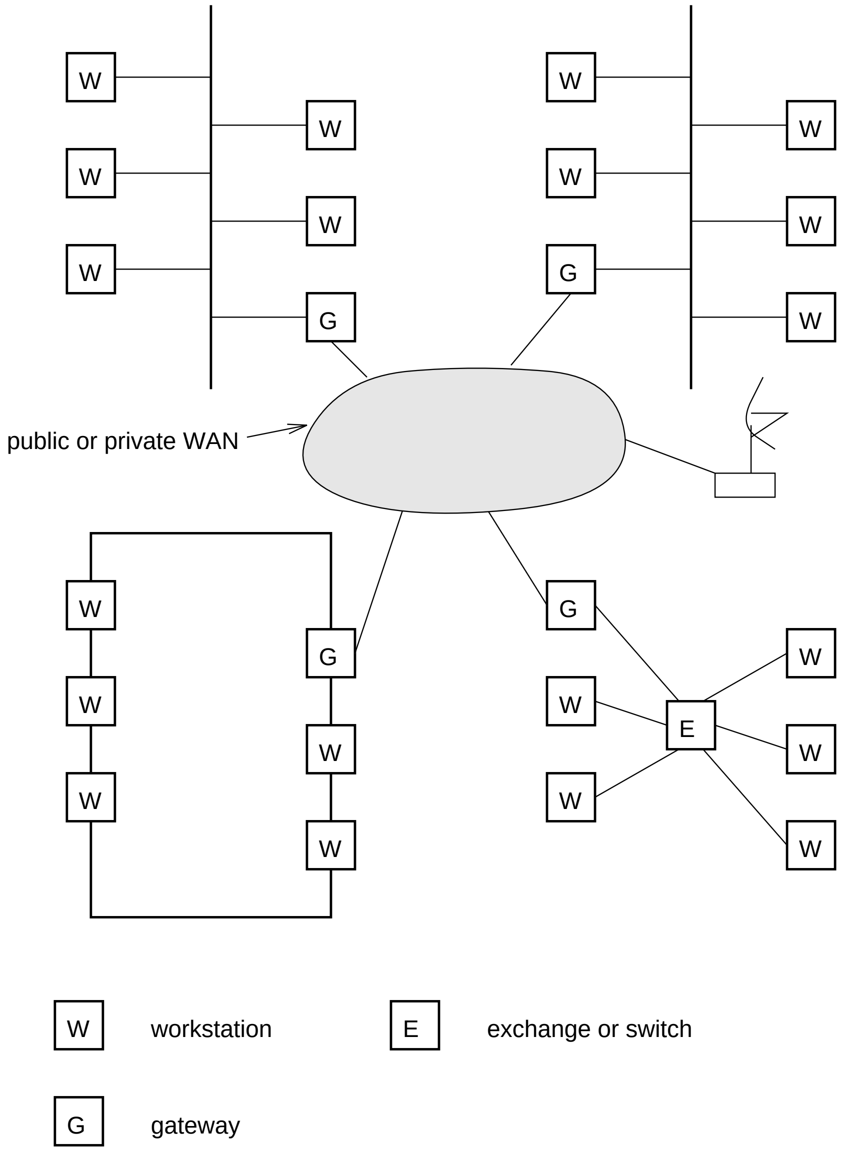


Figure 5: WANs to interconnect LANs

At the present time the world's communication systems are in a state of flux. Historically, digital computing equipment conformed to the requirements of the PSTN. Now, increasingly, analogue traffic such as voice and video is conforming to the requirements of the PSDN.

Public networks employ two types of switching. Switching describes the method by which the corresponders are connected. A **circuit switched network (CSN)** establishes a connection through the network that is then used exclusively by the two correspondents. (Of course, only in 19th century telephone exchanges is the switching actually current driven.) The PSTN is a circuit switched network. A **packet switched network (PSN)** divides the message into packets, that are addressed to the recipient. The packets are then forwarded through the network, together with many other packets. These are locally distributed on arrival. The Post Office is a packet switched network. More relevantly, LAN communication is exclusively via a PSN. The outstanding advantage of the PSN is that the two correspondents can communicate at different rates, permitting much more efficient use of the communication channel.

1.2 The variety and description of telecommunications traffic

Telecommunications traffic is characterised by great diversity. A non-exclusive list is the following:

- Digital data is universally represented by strings of 1s or 0s. Each one or zero is referred to as a **bit**. Often, but not always, these bit strings are interpreted as numbers in a binary number system. Thus $101001_2 = 41_{10}$. The information content of a digital signal is equal to the number of bits required to represent it. Thus a signal that may vary between 0 and 7 has an information content of 3 bits. Written as an equation this relationship is:

  $$I = \log_2(n) \;\text{ bits} \tag{1}$$

  where n is the number of levels a signal may take. It is important to appreciate that information is a measure of the number of different outcomes a value may take. Thus, the **individual values** 0 or 7 each have an information content of 1 bit; the **range of integer values** between 0 and 7 have an information content of 3 bits.

  The **information rate** is a measure of the speed with which information is transferred. It is measured in **bits/second or b/s** (sometimes also **bytes/second or B/s**).

- Audio signals. An audio signal is an example of an analogue signal. It is different in kind from a digital signal. It occupies a frequency range from about 200Hz to about 15KHz. Speech signals occupy a smaller range of frequencies, and telephone speech typically occupies the range 300Hz to 3300Hz. The range of frequencies occupied by the signal is called its **bandwidth** (Figure 6).

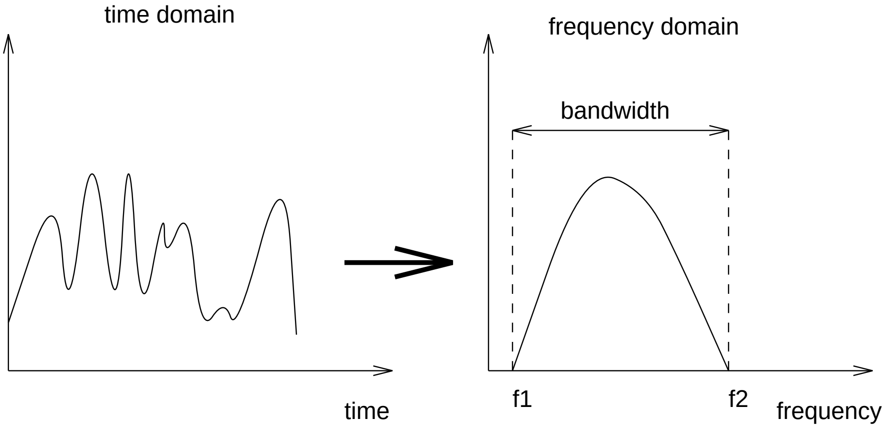


Figure 6: Time domain and frequency domain representation of a signal

- Television. A television signal is an analogue signal created by linearly scanning a two-dimensional image. Typically the signal occupies a bandwidth of about 6MHz.

- Teletext is written (or drawn) communications that are interpreted visually. Telex describes a message limited to a predetermined set of alphanumeric characters. Facsimile describes the transmission of documents that have been converted into a discrete two-tone image.

1.3 The conversion of analogue and digital signals

In order to send analogue signals over a digital communication system, or process them on a digital computer, we need to convert analogue signals to digital ones. This process is performed by an **analogue-to-digital converter (ADC)**. The analogue signal is **sampled** (i.e. measured at regularly spaced instants) (Figure 7) and then **quantised** (i.e. converted to discrete numeric values) (Figure 8). The greater the number of quantisation levels, the lesser the **quantisation error**. The converse operation to the ADC is performed by a **digital-to-analogue converter (DAC)**.

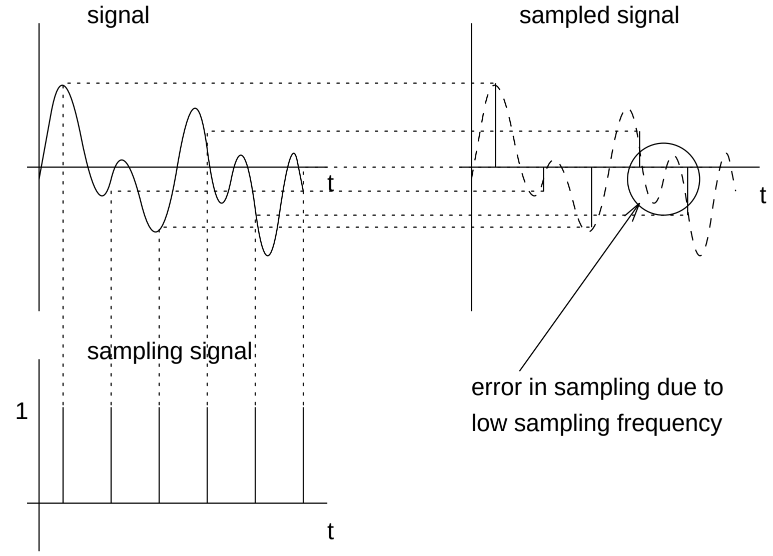


Figure 7: Periodic sampling of an analogue signal

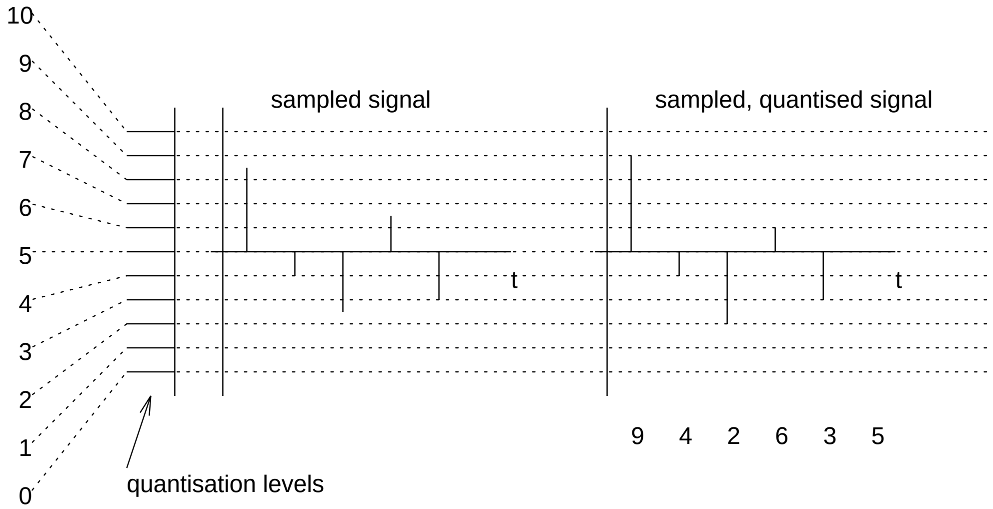


Figure 8: Quantisation of a sampled signal

The ADC process is governed by an important law. The **Nyquist-Shannon Theorem** states that an analogue signal of bandwidth B can be completely recreated from its sampled form provided it is sampled at a rate s equal to at least twice its bandwidth. That is:

$$s \geq 2B \tag{2}$$

The rate at which an ADC generates bits depends on how many bits are used in the converter. For example, a speech signal has an approximate bandwidth of 4KHz. If this is sampled by an 8-bit ADC at the Nyquist sampling rate, the bit rate R is:

$$R = 8\text{bits} \times 2B = 64000 \ \text{b/s} \tag{3}$$

1.4 The transmission of information

The communications system is responsible for the transmission of information from the sender to the recipient. At its simplest, the system contains (Figure 9):

- a **transmission channel** that is the physical link between the communicating parties.
- a **modulator** that takes the source signal and transforms it so that it is physically suitable for the transmission channel.
- a **transmitter** that actually introduces the modulated signal into the channel, usually amplifying the signal as it does so.
- a **receiver** that detects the transmitted signal on the channel and usually amplifies it (as it will have been attenuated by its journey through the channel).
- a **demodulator** that receives the original source signal from the received signal and passes it to the sink.

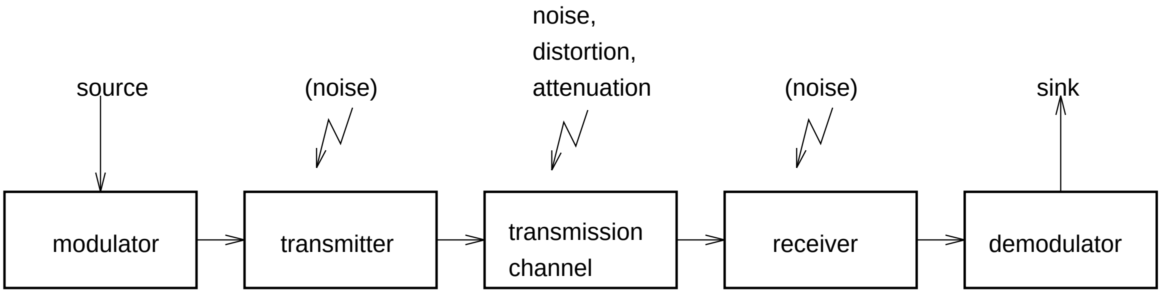


Figure 9: A communications system

Systems are distinguished by the type of signal presented to the modulator. An analogue system is designed for transmitting analogue signals. The PSTN is an analogue system. A digital system is designed to transmit digital signals. The PSDN is a digital system.

There is considerable scope for confusion if the use of the words analogue and digital are not qualified. An analogue system can transmit digital signals, so long as they are made to conform to the analogue expectations of the system. This is the function of the modem (modulator/demodulator) used to connect computers to the PSTN. On the other hand, digital systems cannot transmit analogue signals, because their greater speed relies on the digital form of the message. If the message source is analogue, like speech, then it must first be converted to a digital message by an ADC.

When used to describe the modulation, the words *analogue* and *digital* refer to the message. All signals transmitted by the system are analogue, since they are continuously varying voltages or fields. The signals transmitted over the channel are called the **carrier**. **Analogue modulation** describes the modulation of an analogue signal onto the analogue carrier. Radio, for example, uses analogue modulation. **Digital modulation** describes the modulation of a digital signal onto the analogue carrier. Thus analogue systems may use analogue or digital modulation; digital systems use digital modulation.

The use of digital signals and modulation has great advantages over analogue systems. These are:

- **High fidelity.** The discrete nature of digital signals makes their distinction in the presence of noise easy. Very high fidelity transmission and regeneration are possible.
- **Time independence.** A digitised signal is a stream of numbers. Once digitised a signal may be transmitted at a rate unconnected with its recording rate.
- **Source independence.** The digital signals may be transmitted using the same format irrespective of the source of the communication. Voice, video and data may be transmitted using the same channel.
- **Signals may be coded.** The same transmitted message has an infinite number of meanings according to the rule used to interpret it.

The one disadvantage of digital communication is the increased expense of transmitters and receivers. This is particularly true of of real-time communication of analogue signals. Complex and expensive circuitry is needed to perform real-time ADC and DAC. Radio and television broadcasts are analogue because the terminal cost must be kept low. However, digital manipulation of television images is now common prior to transmission, where equipment expense is less relevant. This situation is changing. Audio and video frequency digital players (compact disc (CD), digital audio tape (DAT), digital compact cassette (DCC) and LaserDisc) have been commercially available for some time.

The price to be paid for these gains is the greatly increased complexity of the encoding and decoding procedures. In particular, timing control and the correct identification of the digital structure are crucial. These problems do not exist with analogue communication.

1.5 The relationship between information, bandwidth and noise

The most important question associated with a communication channel is the maximum rate at which it can transfer information. Information can only be transferred by a signal if the signal is permitted to change. Analogue signals passing through physical channels may not change arbitrarily fast. The rate at which a signal may change is determined by the bandwidth. In fact it is governed by the same Nyquist-Shannon law (equation 2) as governs sampling; a signal of bandwidth B may change at a maximum rate of $2B$. If each change is used to signify a bit, the maximum information rate is $2B$.

The Nyquist-Shannon theorem makes no observation concerning the magnitude of the change. If changes of differing magnitude are each associated with a separate bit, the information rate may be increased. Thus, if each time the signal changes it can take one of n levels, the information rate is increased to:

$$R = 2B \log_2(n) \ \text{ b/s} \tag{4}$$

This formula states that as n tends to infinity, so does the information rate.

Is there a limit on the number of levels? The limit is set by the presence of **noise**. If we continue to subdivide the magnitude of the changes into ever decreasing intervals, we reach a point where we cannot distinguish the individual levels because of the presence of noise. Noise therefore places a limit on the maximum rate at which we can transfer information. Obviously, what really matters is the **signal-to-noise ratio (SNR)**. This is defined by the ratio of signal power S to noise power N, and is often expressed in **deciBels (dB)**;

$$SNR = 10\log_{10}(S/N) \text{ dB} \tag{5}$$

Also note that it is common to see following expressions for power in many texts:

$$P_{dBW} = 10\log_{10}(S/1) \text{ dBW} \tag{6}$$

$$P_{dBm} = 10\log_{10}(S/0.001) \text{ dBm} \tag{7}$$

i.e. equation 6 expresses power as a ratio to 1 Watt and equation 7 expresses power as a ratio to 1 milliWatt. These are expressions of **power** and should not be confused with SNR.

There is a theoretical maximum to the rate at which information passes error free over the channel. This maximum is called the **channel capacity,** C. The famous **Hartley-Shannon Law** states that the channel capacity, C, is given by:

$$C = B\log_2(1 + (S/N)) \text{ b/s} \tag{8}$$

The Hartley-Shannon Law comes from considering the maximum number of signal levels that can be used within a communication channel. If the peak signal value is V_s and the peak noise value is V_n then the power in each of these two signals is $S = V_s^2$ and $N = V_n^2$ respectively. Because of the noise, we must separate our information signal levels by (at least) $\sqrt{N}$, so the maximum number of levels we can have is $\sqrt((S + N)/N)$. From equation 4 for the information rate:

$$\begin{aligned} R &= 2B\log_2\sqrt{\frac{S+N}{N}} \\ &= B\log_2(1 + S/N) \text{ b/s} \end{aligned}$$

Note that S/N is linear in this expression. For example, a 10KHz channel operating in a SNR of 15dB has a theoretical maximum information rate of $10000\log_2(31.623) = 49828$b/s.

The theorem makes no statement as to **how** the channel capacity is achieved. In fact, in practise channels only approach this limit. The task of providing high channel efficiency is the goal of coding techniques. The failure to meet perfect performance is measured by **the bit-error-rate (BER)**. Typically BER values are of the order 10^{-6}.

1.6 The description and types of communication channels

The communication channel is a crucial part of the communication network. It is an analogue part of the system, and is described in terms of the analogue – quantities bandwidth, absorption etc. It limits the bandwidth and noise power, which we have already identified as factors determining the maximum information rate of the channel. Additionally, the channel can introduce various kinds of distortion into the signal, and also has a delay associated with it.

When a signal is introduced into a channel of bandwidth B, the signal emerging from the end of a channel has its bandwidth reduced to the bandwidth B of the channel. Signals lying outside the bandwidth of the channel are removed. The channel places a maximum value of the bandwidth of the system, irrespective of the signal bandwidth. According to the type and length of the channel, the bandwidth will be determined (Figure 10).

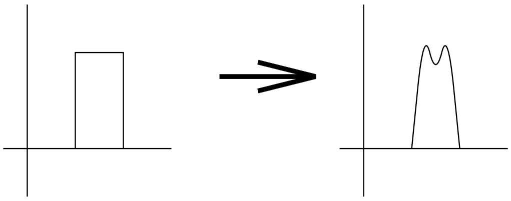


Figure 10: The effects of limited bandwidth

Absorption is the term used to describe the loss of signal power as the signal moves through the channel (Figure 11). The longer the channel, the higher the absorption. It is usually specified in **dB/m**. Absorption is usually frequency dependent. It reduces the available bandwidth. **Equalisers** are frequency dependent amplifiers that restore the spectral balance of the signal.

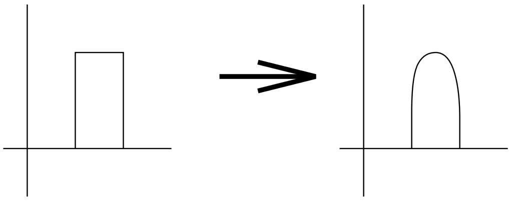


Figure 11: The effects of channel absorption

Attenuation is the decrease of signal strength as it propagates along a channel (Figure 12). **Repeaters** are amplifiers placed along the channel that restore the signal power. The distance between the repeaters is determined so as to minimise any errors in the signal due to attenuation effects, for instance to maintain a minimum SNR along the channel. Some attenuation affects are due to absorption. The amount of attenuation is measured in **dB/m**. Some typical values are shown in Table 1.

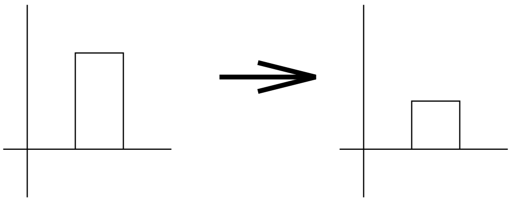


Figure 12: The effects of signal attenuation

Dispersion (sometimes called **delay distortion**) describes the effect of different frequencies moving at different speeds along the channel (Figure 13). This introduces changes in shape of the signal. Dispersion is not easy to correct for, and channels are sometimes limited by its effect.

Propagation delay is the term used to describe the time taken for the signal to pass along the channel. Electromagnetic signals travel at velocities near that of light in vacuo of 3×10^8m/s. A more typical value is 2.5×10^8m/s. The importance of propagation delay varies widely with the length of the channel.

Noise is the term used to describe all signals present at the receiver that are not part of the message signal (Figure 14). The source of these noise signals vary widely.

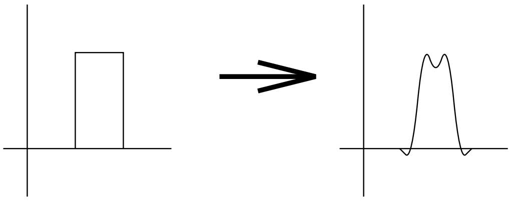


Figure 13: The effects of dispersion (delay distortion)

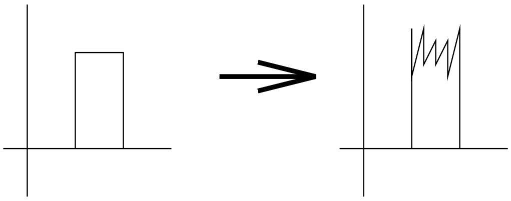


Figure 14: The effects of noise on a signal

Impulse noise is common in low-frequency circuits and arises from electric fields generated by electrical switching. It appears as bursts at the receiver, and when present can have a catastrophic effect due to its large power. Other peoples signals can generate noise: **cross-talk** is the term given to the pick-up of radiated signals from adjacent cabling. When radio links are used, **interference** from other transmitters can be problematic.

Thermal noise is always present. It is due to the random motion of electric charges present in all media. It can be generated externally, or internally at the receiver. A common property of all thermal noise is that it may be described by the relation:

$$N = kTB \text{ Watts} \tag{9}$$

where $k = 1.38 \times 10^{-23}$J/K (known as **Boltzman's Constant**), T is the absolute temperature in degrees Kelvin, and B is the receiver bandwidth. Equation 9 says that hot systems generate more noise.

The combined effects of attenuation, absorption, dispersion, and noise may result in **bit errors** occurring in the transmitted signal as shown in Figure 15.

Communications channels can be a physical connection between correspondents. Wires, coaxial cables and optical fibres are used for this purpose. Roughly speaking, the bandwidth and expense increases along this list. When distances, or the cost of cable laying become very large, it can

Transmission medium	**Frequency**	**Loss** $[dB/Km]$
Copper wire pair (0.3cm diameter)	1KHz	0.05
Twisted pair (16 gauge)	10KHz	2
	100KHz	3
	300KHz	6
Coaxial cable	100KHz	1
	1MHz	2
	3MHz	4
Fibre-optic cable	4 $\times 10^{14}$Hz	5

Table 1: Some typical media loss values

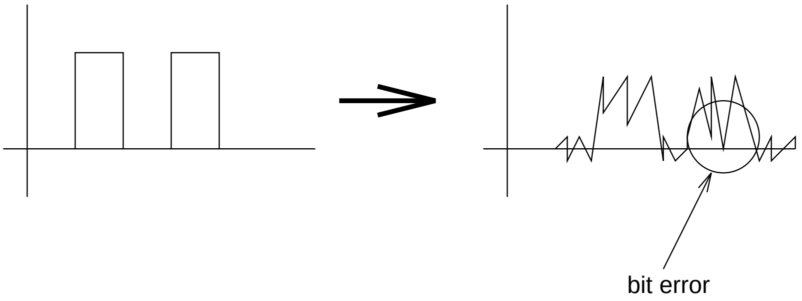


Figure 15: A bit error in a transmitted signal due to channel effects

be advantageous to use radio waves as the channel. Microwave links are high bandwidth **line--of-sight** links. Satellite links (usually) use **geostationary** satellites to provide inter-continental high-bandwidth communications. There is now also considerable interest in the use of networks of **low earth orbiting satellites (LEOs)** for providing global communication. Optical links (i.e. the transmission through free-space of light waves) can be used for short, high bandwidth transmission, but are fairly rare in use.

Radio and television **broadcasting** is the natural method for communicating unselectively with large numbers of correspondents, or when one or both of the correspondents are moving. Mobile communications employ networks of radio transmitter/receiver stations.

1.7 Digital transmission and switching

The demand for services such as multimedia, video and real-time applications bring new requirements and challenges. Much of the successful evolution of these services will depend on the deployment of high speed networks, based on optical fibre technology. Some of the requirements for this have already been addressed by standards bodies in the description of **Broadband ISDN (B-ISDN)** services. It seems most likely that the provision of the B-ISDN services will be made possible by the use of **asynchronous transfer mode (ATM)** as the carrier at the subnetwork level. ATM will be used in conjunction with SONET/SDH to offer network high speed network services, with rates of 155Mb/s to 622Mb/s (and higher) being planned. **Synchronous Optical Network (SONET)** and **Synchronous Digital Hierarchy (SDH)** are two similar architectures proposed for providing high-speed networks, the latter (a European standard) being based heavily on the former (a US standard).

With the real-time applications that need to run over large, diverse networks, issues arise concerning the **Quality of Service (QoS)** guarantees required from the network in order to maintain the required network resources for the needs of a particular application. Most of the current activity in this increasingly popular area of research focus around providing models for QoS in terms of various parameters such as speed, delay, jitter and various error rates.

To fully exploit the huge bandwidth offered by optical fibre technology we require electronic equipment that will be able to operate at speeds of at least the rate of optical transmission. Currently, our use of the optical bandwidths is limited by the electronic technology that is used for transmitting and receiving the signals, as the electronic components used approach the operational limits of their fabrication techniques, e.g. CMOS. The electronic components not only restrict the way in which the information is actually transmitted into the fibre (the type of signal modulation used), but also the speed at which information is switched between the links that make up the network.

Network switches are crucial elements in the provision of high speed networks. Switches link the physical paths which the data takes through the network. Although it may be possible to build on known switching techniques and improve their throughput, with the increasing speed of transmission of the data, the switches must also become faster if they are not to become performance bottlenecks in the evolution of high speed networks.

1.8 Standards

International bodies concerned with the interconnection of telecommunications equipment have for decades provided standards for the connection of terminals to these lines. These systems were initially analogue, and the associated standards were primarily concerned with defining the hardware and electrical interfaces. V-series standards relate to the connection of terminals to PSTNs, D-series to PSDN's and I-series to ISDNs. The X-series recommendations relate to network services and protocols that are independent of the underlying subnetwork technology. This in marked contrast to computer manufacturers, who in the past have tended to introduce internal standards that make different manufactures equipment incompatible. Such systems are described as **closed systems**, in contrast to the **open systems** of the telephone networks.

The recognition that compatibility was potentially advantageous led to the formation of the **International Organisation for Standardisation (ISO)**. This body has generated a range of standards concerning the connection of computer equipment, the connection of LANs, the connection of equipment and LANs to PSDNs, and more recently the connection to the higher level functions of the ISDNs. ISO works in close collaboration with the **ITU-TSS (International Telecommunications Union-Telecommunication Standardization Sector)**, which was formerly the **CCITT (International Consultative Committee for Telephony and Telegraphy)**. These groups receive contributions and suggestions from various national bodies, such as BSI (British Standards Institute) and **ANSI (American National Standards Institute)**, **IEEE (Institute of Electrical and Electronic Engineers)**, as well as from commercial companies with interest in the areas to be standardised. ANSI and IEEE also produce their own documents which are often adopted as ISO or ITU standards.

While these international standards bodies are important, they are not the only people involved with the promotion of the the use of data communications technologies. Indeed, the largest network in the world, the **Internet** uses standards that were developed within its own community, consisting originally of many universities and research establishments. The **Internet Engineering Task Force (IETF)** continues to encourage the publishing of **Request For Comments (RFC)** documents that are available freely to anyone with Internet access.

Many other groups are concerned with the use of standards or producing **profiles** for architectures based on existing standards for instance the **Network Management Forum (NMF)** and the **Open Software Foundation (OSF)**. These two organisations are particularly interested in promoting agreements that will encourage use and interoperability of 'open' standards.

2 Communication Techniques

2.1 Time, frequency and bandwidth

Previously, the ideas of the frequency content and bandwidth of an analogue signal were introduced. In this section, we wish to examine these ideas more closely.

Most signals carried by communication channels are modulated forms of sine waves. A sine wave is described mathematically by the expression:

$$s(t) = A\cos(\omega t + \phi) \tag{10}$$

The quantities A, ω, and ϕ are termed the **amplitude**, **frequency** and **phase** of the sine wave, respectively. We can describe this signal in two ways. One way is to describe its evolution in time (**time domain model**), and equation 10 is a mathematical representation of this evolution. The second way is to describe its frequency content (**frequency domain model**). The cosine wave, $s(t)$, has a single frequency, $\omega = 2\pi f$.

This representation is quite general. In fact we have the following theorem due to **Fourier**:

> **Any signal $s(t)$ of time may be represented as the sum of a set of cosinusoidal and sinusoidal waves of different frequencies and phases.**

Mathematically:

$$x(t) = A_0 + \sum_{n=1}^{\infty} A_n \cos(\omega n t) + \sum_{n=1}^{\infty} B_n \sin(\omega n t) \tag{11}$$

where:

$$\begin{aligned}
A_0 &= \frac{1}{T}\int_{-T/2}^{T/2} x(t)dt && (12)\\
A_n &= \frac{2}{T}\int_{-T/2}^{T/2} x(t)\cos(\omega n t)dt && (13)\\
B_n &= \frac{2}{T}\int_{-T/2}^{T/2} x(t)\sin(\omega n t)dt && (14)\\
\omega &= \frac{2\pi}{T} && (15)
\end{aligned}$$

where A_0 is the d.c. terms (i.e. $f = 0$) and T is the period of the waveform. The description of a signal in terms of its constituent frequencies is called its **frequency spectrum**.

As an example, consider the square wave (Figure 16):

$$\begin{aligned}
s(t) &= 1; \quad 0 < t < \pi, 2\pi < t < 3\pi, \cdots \\
&= 0; \quad \pi < t < 2\pi, 3\pi < t < 4\pi, \cdots
\end{aligned}$$

This has the Fourier series:

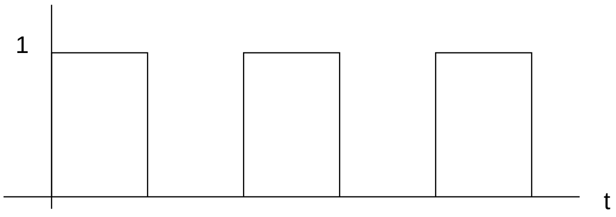


Figure 16: A square wave

$$s(t) = \frac{1}{2} + \frac{2}{\pi}[\sin(t) + \frac{1}{3}\sin(3t) + \frac{1}{5}\sin(5t) + \frac{1}{9}\sin(9t)\cdots] \tag{16}$$

A graph of the spectrum has a line at the odd harmonic frequencies, 1, 3, 5, 9, ..., whose respective amplitudes decay as $2/\pi,\ 2/3\pi,\ 2/5\pi,\ 2/9\pi,\cdots$. The spectrum of a signal is usually shown as a two-sided spectrum with **positive** and **negative** frequency components (as shown in Figure 17). An explanation of this is given later.

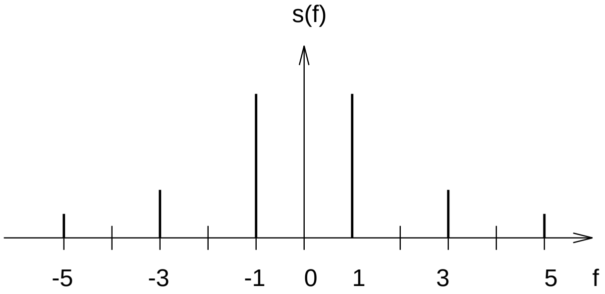


Figure 17: Frequency spectrum for the square wave

So, for our square wave of Figure 16, we have (using equations 12, 13 and 14):

$$\begin{aligned} A_0 &= \frac{1}{2\pi}\int_0^{\pi} 1dt \\ &= \frac{1}{2} \end{aligned}$$

$$\begin{aligned} A_n &= \frac{2}{2\pi}\int_0^{\pi} 1.\cos(nt) \\ &= \frac{1}{\pi}\left[\frac{\sin(nt)}{n}\right]_0^{\pi} \\ &= \frac{1}{n\pi}\sin(n\pi) \\ &= 0 \end{aligned}$$

$$\begin{aligned} B_n &= \frac{2}{2\pi}\int_0^{\pi} 1.\sin(nt) \\ &= \frac{1}{\pi}\left[\frac{-\cos(nt)}{n}\right]_0^{\pi} \\ &= \frac{1}{n\pi}(1-\cos(n\pi)) \end{aligned}$$

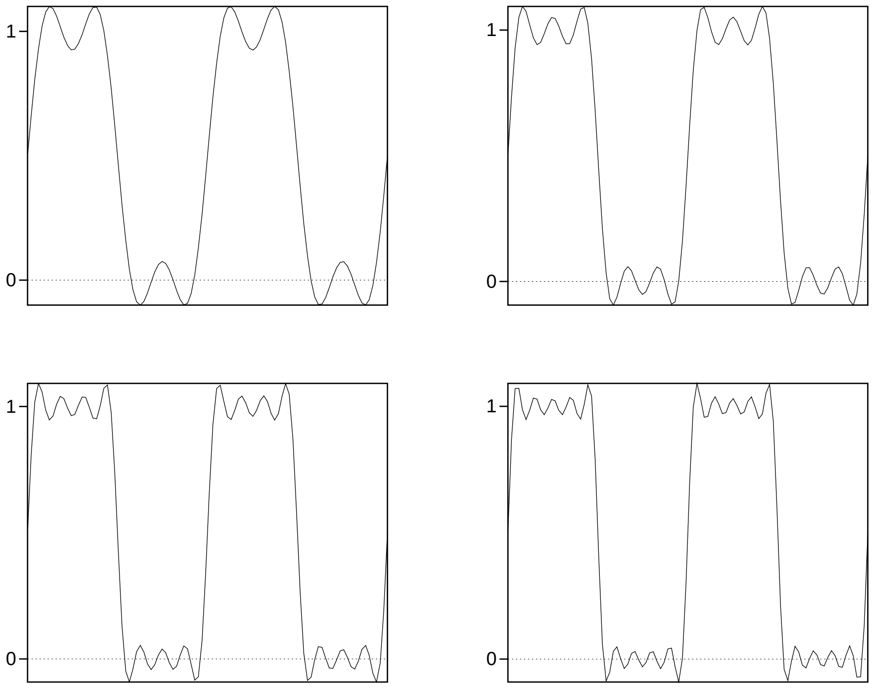


Figure 18: A square wave with 3, 4, 5 and 6 of its Fourier terms

$$\begin{aligned}
B_1 &= \frac{2}{\pi} \\
B_2 &= 0 \\
B_3 &= \frac{2}{3\pi} \\
B_4 &= 0 \\
B_5 &= \frac{2}{5\pi} \\
\ldots
\end{aligned}$$

and hence our expression for the square wave as given in equation 16. ($A_n = 0$ is visible by inspection, as $s(t)$ is an odd function.)

For an arbitrary square pulse train, the frequency spectrum is a continuous function, $\sin(x)/x$ or synch(x) (Figure 19).

In fact, the expression 11 is a simplification, but sufficient for our needs. In general, any signal $x(t)$ can be represented as follows:

$$x(t) = C_0 + \sum_{n=1}^{\infty} C_n \cos(\omega n t + \phi_n) \tag{17}$$

where C_0 is the d.c term. The frequency description of this signal would describe the amplitudes C_n as a function of the frequency harmonics ω_n with C_0 as the d.c. term.

Now, if we use the following identity in equation 17:

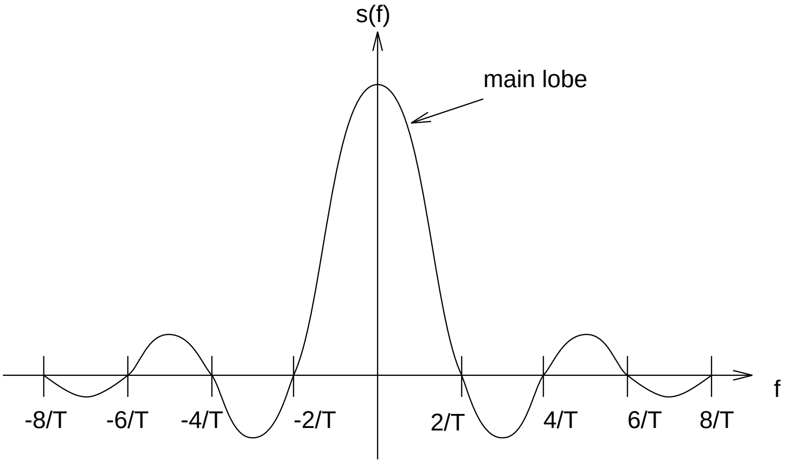


Figure 19: Frequency spectrum for a square pulse train

$$\cos\theta = \frac{\mathrm{e}^{j\theta} + \mathrm{e}^{-j\theta}}{2} \tag{18}$$

and let $\theta = \omega t$ in the expression for c_n, we get:

$$\begin{aligned} C_n \cos(n\theta + \phi_n) &= \frac{C_n}{2}(\mathrm{e}^{j(n\theta+\phi_n)} + \mathrm{e}^{-j(n\theta+\phi_n)}) \\ &= \frac{C_n}{2}(\mathrm{e}^{jn\theta}\mathrm{e}^{j\phi_n} + \mathrm{e}^{-jn\theta}\mathrm{e}^{-j\phi_n}) \\ &= c_n\mathrm{e}^{jn\theta} + c_{-n}\mathrm{e}^{-jn\theta} \end{aligned} \tag{19}$$

where:

$$c_n = \frac{C_n}{2}\mathrm{e}^{jn\phi} \tag{20}$$

$$c_{-n} = \frac{C_n}{2}\mathrm{e}^{-jn\phi} \tag{21}$$

and we note that $c_{-n} = c_n^*$ the complex conjugate of c_n and also we assume that all C_n are real. Further:

$$\mathrm{e}^{jn\theta}|_{n=0} = \mathrm{e}^0 = 1 \tag{22}$$

and we can define $c_0 = C_0$, so that we can rewrite our expression for $x(t)$ in equation 17 as:

$$x(t) = \sum_{n=-\infty}^{\infty} c_n \mathrm{e}^{j\omega nt} \tag{23}$$

which is the exponential form of the Fourier series. In this expression, the values c_n are complex and so $|c_n|$ and $\arg(c_n)$ are the magnitude and the phase of the spectral component respectively:

$$c_n = \frac{1}{T} \int_{-T/2}^{T/2} x(t) \mathrm{e}^{-j\omega nt} dt \tag{24}$$

where $w = 2\pi/T$, T is the period of the waveform.

Signals whose spectra consist of isolated lines are periodic, i.e. they repeat themselves indefinitely. The lines in this spectrum are infinitely thin; they have zero bandwidth. The Hartley-Shannon law (equation 8) tells us that the maximum information rate of a zero bandwidth channel is zero. Thus, zero bandwidth signals carry no information.

To permit the signal to carry information we must introduce the capacity for aperiodic change. The consequence of an aperiodic change is to introduce a spread of frequencies into the signal. Detailed theoretical consideration leads us to the Nyquist-Shannon law that a signal of bandwidth B may change in an aperiodic fashion at a maximum rate of $2B$ (equation 2).

If the square-wave signal discussed in the previous example is replaced with an aperiodic sequence of 1s and 0s, the spectrum changes substantially. The discrete harmonic components are replaced by a continuous range of frequencies whose shape is a **synch** ($\sin(x)/x$) function. There a number of features to note:

- The bandwidth of the signal is only approximately finite. Most of the energy is contained in a limited region called the **main-lobe**. However, some energy is found at all frequencies.
- If there are two changes in the period T (in the time domain), the width of the main-lobe (in the frequency domain) is $4/T$. The Nyquist-Shannon law (equation 2) states that if a signal has bandwidth $4/T$, then the maximum number of changes that may take place is $8/T$ per second, or 8 changes in the period T. If each change is taken to represent 1 bit, it is apparent that this signal does not carry as much information as the bandwidth permits.
- The spectrum has positive and negative frequencies. These are symmetric about the origin. This may seem non-intuitive but can be seen from equation 23. However, note that a signal has components at a certain **frequency** and the notion of **pairs** of positive and negative frequency components is a mathematical convenience in examining the signal.

The bandwidth of a communication channel is limited by the physical construction of the channel. The Nyquist-Shannon law places a limit on the maximum number of aperiodic changes per unit time the signal may perform. It is the task of the modulator to make best use of the available bandwidth. However, the modulator is itself constrained by its physical design and construction, and it may or may not approach the Nyquist limit in its operation. To ease calculations, this method of analysis is used often with the approximations that the signal being examined is periodic and is band-limited.

2.2 Digital modulation, ASK, FSK and PSK

There are three ways in which the bandwidth of the channel carrier may be altered simply. It is worth emphasising that these methods are chosen because they are practically simple, not because they are theoretically desirable. These are the altering of the amplitude, frequency and phase of the carrier wave. These techniques give rise to **amplitude-shift-keying (ASK)**, **frequency--shift-keying (FSK)** and **phase-shift-keying (PSK)**, respectively.

ASK describes the technique by which a carrier wave is multiplied by the digital signal $f(t)$. Mathematically, the modulated carrier signal $s(t)$ is:

$$s(t) = f(t) \cos(\omega_c t + \phi) \tag{25}$$

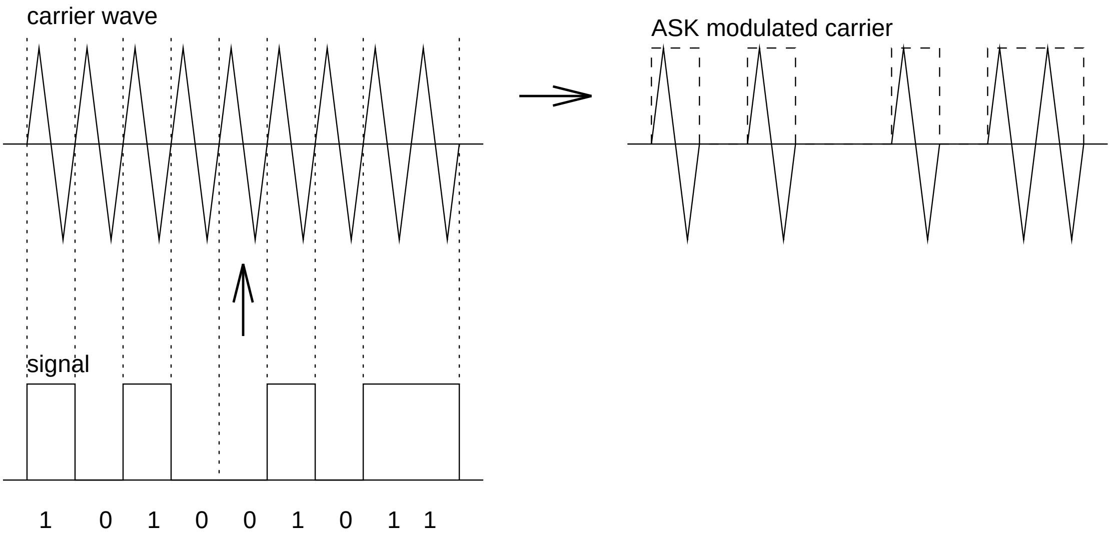


Figure 20: Amplitude shift keying

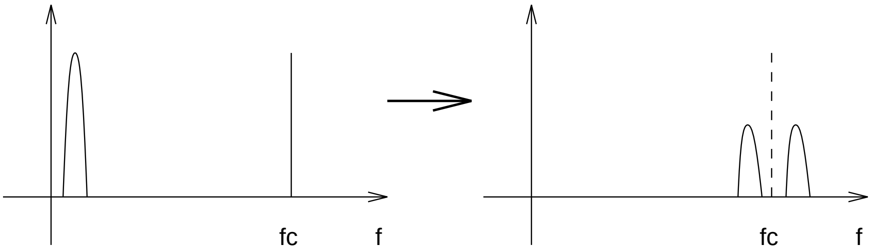


Figure 21: Amplitude shift keying – frequency domain

It is a special case of **amplitude modulation (AM)** (Figures 20 and 21). Amplitude modulation has the property of translating the spectrum of the modulation $f(t)$ to the carrier frequency. The bandwidth of the signal remains unchanged. This can be seen if we examine a simple case when $f(t) = \cos(\omega_m t)$ and we use the identities:

$$\cos(A+B) = \cos(A)\cos(B) - \sin(A)\sin(B) \tag{26}$$
$$\cos(A-B) = \cos(A)\cos(B) + \sin(A)\sin(B) \tag{27}$$

then:

$$\begin{aligned} s(t) &= \cos(\omega_m t)\cos(\omega_c t) \\ &= \frac{1}{2}[\cos((\omega_c + \omega_m)t) + \cos((\omega_c - \omega_m)t)] \end{aligned} \tag{28}$$

This method is called **double side-band supressed carrier (DSB-SC)** AM.

The fact that AM simply shifts the signal spectrum is often used to convert the carrier frequency to a more suitable value without altering the modulation. This process is known variously as **mixing**, **up-conversion** or **down-conversion**. Some form of conversion will always be present when the channel carrier occupies a frequency range outside the modulation frequency range.

As a matter of practical fact, the conventional form for performing AM is **envelope modulation** or **envelope AM**, which is usually performed as follows:

$$s(t) = (1 + f(t))\cos(\omega_c t + \phi) \tag{29}$$

which, using our simple modulation of $f(t) = \cos(\omega_m t)$, expands to:

$$\begin{aligned} s(t) &= (1 + \cos(\omega_m t))\cos(\omega_c t) && (30)\\ &= \cos(\omega_c t) + \frac{1}{2}[\cos((\omega_c + \omega_m)t) + \cos((\omega_c - \omega_m)t)] && (31) \end{aligned}$$

the difference now being that the carrier signal is present in $s(t)$. The significance of this will be explained later. (However, both forms can be used for ASK.)

FSK describes the modulation of a carrier (or two carriers) by using a different frequency for a 1 or 0. The resultant modulated signal may be regarded as the sum of two amplitude modulated signals of different carrier frequency (Figures 22 and 23):

$$s(t) = f_0(t)\cos(\omega_{c0} t + \phi) + f_1(t)\cos(\omega_{c1} t + \phi) \tag{32}$$

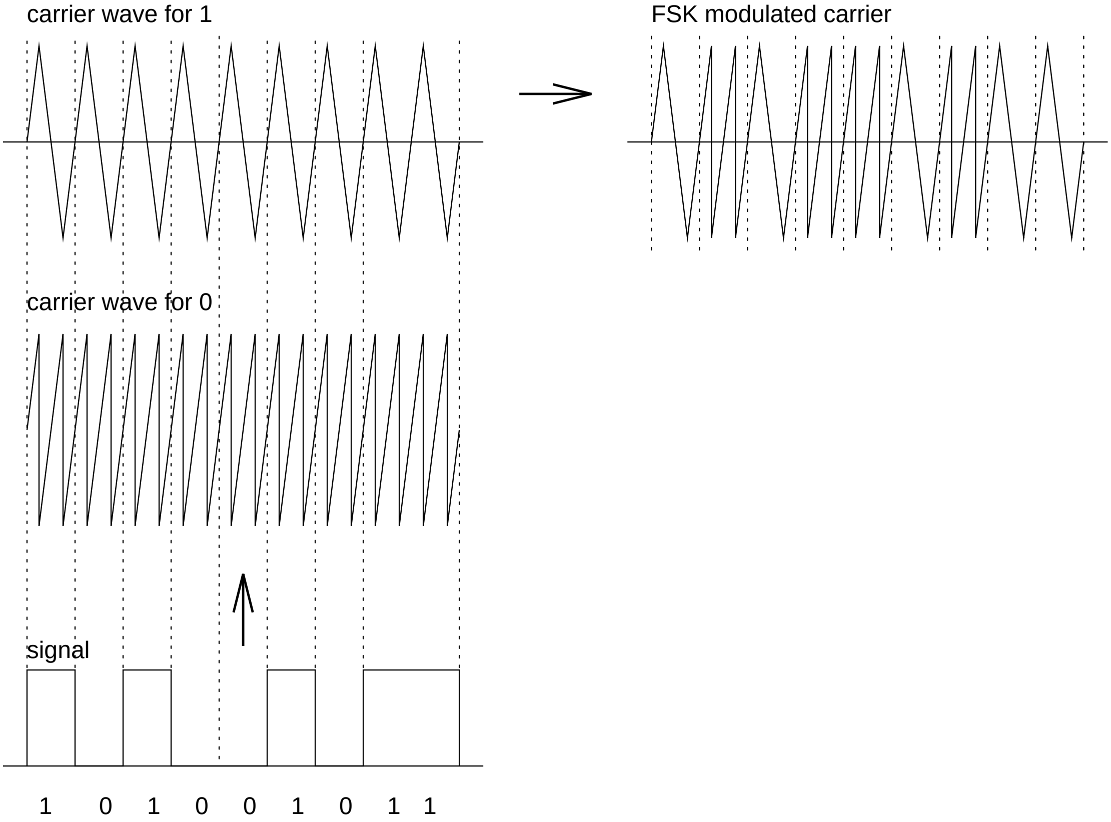


Figure 22: Frequency shift keying

FSK is classified as wide-band if the separation between the two carrier frequencies is larger than the bandwidth of the spectrums of $f_0(t)$ and $f_0(t)$. In this case the spectrum of the modulated signal appears as two separate ASK signals. Narrow-band FSK is the term used to describe an

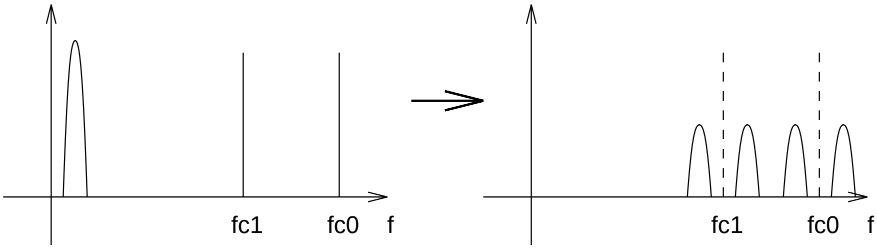


Figure 23: Frequency shift keying – frequency domain

FSK signal whose carrier frequencies are separated by less than the width of the spectrum $f_0(t)$ and $f_1(t)$. Generally the bandwidth of FSK is greater than ASK for the same modulation.

PSK describes the modulation technique that alters the phase of the carrier. Mathematically:

$$s(t) = \cos(\omega_c + \phi(t)) \tag{33}$$

Binary phase-shift-keying (BPSK) has only two phases, 0 and π. It is therefore a type of ASK with $f(t)$ taking the values -1 or 1 (Figure 24), and its bandwidth is the same as that of ASK. Phase-shift-keying offers a simple way of increasing the number of levels in the transmission without increasing the bandwidth by introducing smaller phase shifts. **Quadrature phase-shift-keying (QPSK)** has four phases, $0, \pi/2, \pi, 3\pi/2$. **M-ary PSK** has M phases, $2\pi m/M; \quad m = 0, 1, \cdots M-1$. For a given bit-rate, QPSK requires half the bandwidth of PSK and is widely used for this reason.

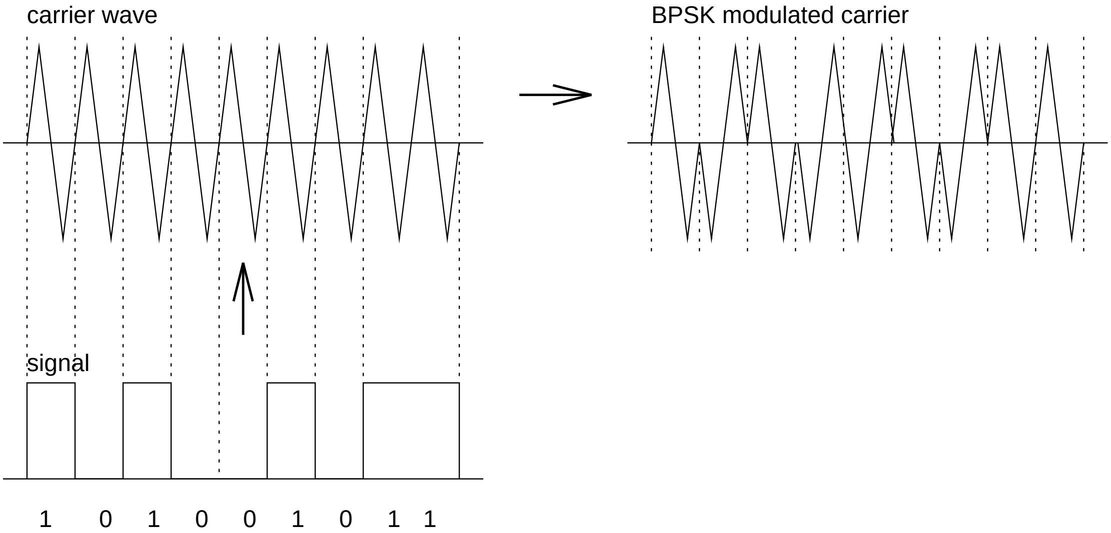


Figure 24: Binary phase shift keying

The number of times the signal parameter (amplitude, frequency, phase) is changed per second is called the **signaling rate**. It is measured in **baud**. 1 baud = 1 change per second. With binary modulations such as ASK, FSK and BPSK, the signaling rate equals the bit-rate. With QPSK and M-ary PSK, the bit-rate may exceed the baud rate.

2.3 Spread spectrum techniques

We know from the Hartley-Shannon Law (equation 8) that we can make a trade-off between signal power and bandwidth in order to increase the channel capacity:

$$C = B\log_2(1 + (S/N)) \text{ b/s}$$

However, due to the logarithmic relationship, a significant increase in the signal power, S, does not produce a similarly proportionate increase in the channel capacity. So, frequency allocation constraints permitting, we may increase the bandwidth to achieve the desired result:

$$\begin{aligned} C &= B\log_2(1 + (S/N)) \\ \frac{C}{B} &= \frac{\log(1 + (S/N))}{\log 2} \\ &\approx 1.44\log(1 + (S/N)) \\ &\approx 1.44\frac{S}{N} \end{aligned} \tag{34}$$

with the last step in the above derivation is made by using the logarithmic expansion:

$$\log(x) = x - \frac{x^2}{2} + \frac{x^3}{3} - \frac{x^5}{5} + \cdots$$

with $x = 1 + S/N$. Since S/N is typically of the order of 0.1 or less in spread spectrum techniques, we consider all terms in the expansion except the first as negligible. So, for instance, for a 32Kb/s channel with a SNR of 0.001 (-30dB), B $\approx$ 22MHz. The two main criteria for using spread spectrum are:

- The transmitted bandwidth is much greater than the minimum bandwidth that the signal requires (as given by the Hartley-Shannon Law).
- Some function other than the information being sent determines the radio frequency bandwidth.

The first of these criteria has been demonstrated. The second criteria is determined by the type of modulation technique used. Clearly, incoherent modulation techniques by themselves will not be useful as the noise power is higher than the signal power! We shall not delve deeply into mechanisms, but shall look at one particular technique that is used called **frequency hopping**. In frequency hoping, the large bandwidth is effectively split into frequency channels. The signal is then spread across the channels as shown in Figure 25.

The **hop set** (channel hopping sequence) is not arbitrary, but determined by the use of a pseudo random sequence. The receiver can reproduce the identical hop set and so decode the signal. The **hop rate** (the rate at which the signal switches channels) can be thousands of times a second, so the **dwell time** (time spent on one channel) is very short. If the hop set is generated by a pseudo random number generator then the **seed** to to that generator is effectively a key decoding the transmitted message, and so this technique has obvious security applications, for instance military use (anti-jamming devices or secure transmission), or (a more domestic but certainly useful example) in mobile phone systems.

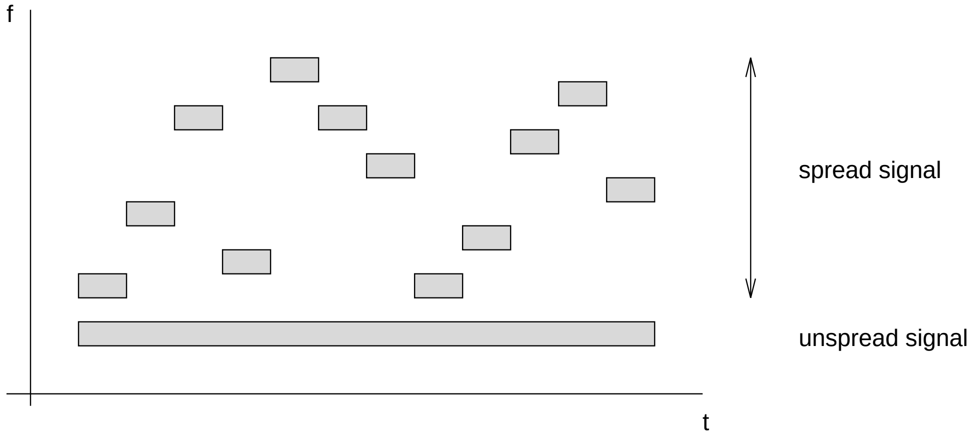


Figure 25: Frequency hopping spread spectrum technique

2.4 Digital demodulation, DPSK and MSK

From the discussion above it might appear that QPSK offers advantages over both ASK, FSK and PSK. However, the demodulation of these signals requires various degrees of difficulty and hence expense. The method of demodulation is an important factor in determining the selection of a modulation scheme. There are two types of demodulation which are distinguished by the need to provide knowledge of the phase of the carrier. Demodulation schemes requiring the carrier phase are termed **coherent**. Those that do not need knowledge of the carrier phase are termed **incoherent**. Incoherent demodulation can be applied to ASK and wide-band FSK. It describes demodulation schemes that are sensitive only to the power in the signal. With ASK, the power is either present, or it is not. With wide-band FSK, the power is either present at one frequency, or the other. Incoherent modulation is inexpensive but has poorer performance. Coherent demodulation requires more complex circuitry, but has better performance.

In ASK incoherent demodulation, the signal is passed to an **envelope detector**. This a device that outputs the 'outline' of the signal. A decision is made as to whether the signal is present or not. Envelope detection is the simplest and cheapest method of demodulation. In optical communications, phase modulation is technically very difficult, and ASK is the only option. In the electrical and microwave context, however, it is considered crude. In addition, systems where the signal amplitude may vary unpredictably, such as microwave links, are not suitable for ASK modulation.

Incoherent demodulation can also be used for wide-band FSK. Here the signals are passed to two circuits, each sensitive to one of the two carrier frequencies. Circuits whose output depends on the frequency of the input are called **discriminators** or **filters**. The outputs of the two discriminators are interrogated to determine the signal. Incoherent FSK demodulation is simple and cheap, but very wasteful of bandwidth. The signal must be wide-band FSK to ensure the two signals $f_0(t)$ and $f_1(t)$ are distinguished. It is used in circumstances where bandwidth is not the primary constraint.

With coherent demodulation systems, the incoming signal is compared with a replica of the carrier wave. This is obviously necessary with PSK signals, because here the power in the signal is constant. With binary PSK the comparison is performed by multiplying the incoming signal with a replica of the carrier. If the output of this process is $h(t)$, using the identity in expression 26 with $A = B$, we have that:

$$h(t) \quad = \quad f(t)\cos(\omega_c t)\cos(\omega_c t)$$

$$
\begin{aligned}
&= \frac{f(t)}{2}[1+\cos(2\omega_c t)] \\
&= \frac{f(t)}{2} + \frac{f(t)}{2}\cos(2\omega_c t)
\end{aligned}
$$

i.e. the original signal plus a term a twice the carrier frequency. By removing, or **filtering** out, the harmonic term, the output of the demodulator is the modulation $f(t)$. With QPSK, the processing is more complicated, and two separate demodulators are required. The demodulator complexity increases rapidly for M-ary PSK; for this reason it is rarely used.

The difficulty with coherent detection is the need to keep the phase of the replica signal, termed **local oscillator**, 'locked' to the carrier. This is not easy to do. Oscillators are sensitive to (among other things) temperature, and a 'free-running' oscillator will gradually drift in frequency and phase. Suppose there is some phase error ϕ present in the local oscillator signal. After filtering, the output of a BPSK demodulator will be:

$$
\begin{aligned}
h(t) &= f(t)\cos(\omega_c t)\cos(\omega_c t+\phi) \\
&= \frac{f(t)}{2}[\cos(\phi)+\cos(2\omega_c t+\phi)] \\
&= \frac{f(t)}{2}\cos(\phi) + \frac{f(t)}{2}\cos(2\omega_c t)
\end{aligned}
$$

According to the value of ϕ, $h(t)$ may take the value -1 to 1 with every value in-between. Clearly, the consequence for the correct interpretation of the demodulated signal is catastrophic.

There are two methods to prevent such an occurrence. In one, a pilot carrier signal is sent in addition to the modulated carrier. This pilot carrier is used to synchronise the local oscillator phase. This original carrier is present when using envelope AM, for instance. The alternative is to employ another form of modulation, **differential phase-shift-keying (DSPK)**. Differential PSK is actually a simple form of coding. The modulating signal is not the binary code itself, but a code that records **changes** in the binary code. This way, the demodulator only needs to determine changes in the incoming signal phase. Because the drifts associated with local oscillators occur slowly, this is not difficult to arrange.

The PSK signal is converted to a DPSK signal with two rules:

- a 1 in the PSK signal is denoted by no change in the DPSK
- a 0 in the PSK signal is denoted by a change in the DPSK signal

The sequence is initialised with a leading 1. An example of the pattern is thus:

```
 PSK    0 1 0 0 1 1 0 1
DPSK  1 0 0 1 0 0 0 1 1
```

Coherent detection is also necessary for the correct demodulation of narrow-band FSK. We have already noted that the form of the modulation signal gives rise to a bandwidth that is larger than that required by the Nyquist limit, and that has appreciable energy outside this bandwidth. This is in part due to the abrupt changes in the phase of the signal in FSK and PSK. By combining coherent detection with appropriate smoothing of the phase change, narrow-band FSK can become very bandwidth efficient, approaching the Nyquist limit, and have low side-lobe levels. This form of modulation, which is a combination of FSK and PSK, is termed **minimum-shift-keying (MSK)**,

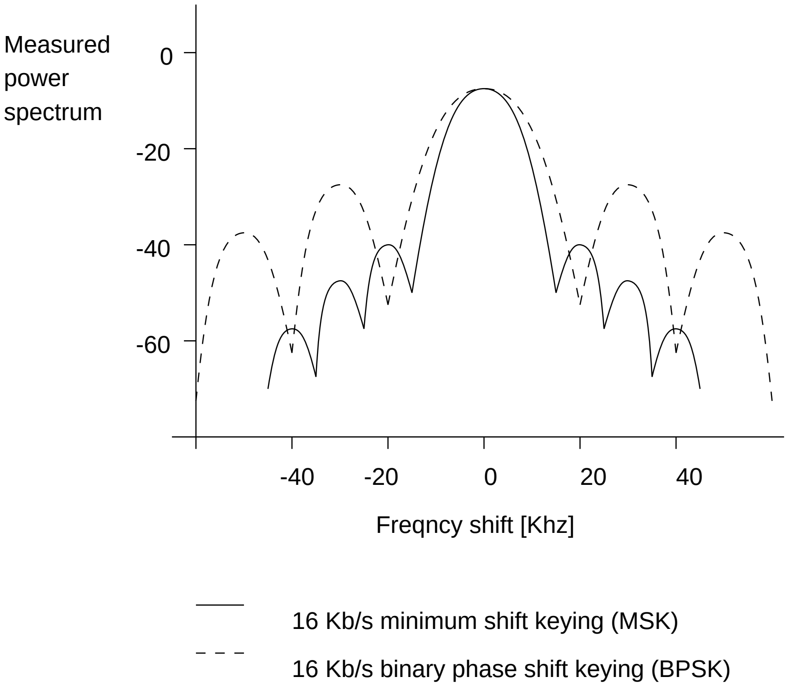


Figure 26: A sketch graph of the measured spectrum power of MSK and BPSK

or **fast-FSK (FFSK)**. Figure 26 shows a sketch graph that compares the power spectrum of MSK and BPSK.

To illustrate the way in which the complexity of demodulation determines the method used we consider two examples.

A **modem** (**mo**dulator/**dem**odulator) is used to connect computer equipment to the PSTN to communicate with other equipment. The PSTN has a **pass-band** from 400Hz to 3400Hz; its bandwidth is 3000 Hz. Modems must be inexpensive if they are to be widely available. Variations in amplitude are common on telephone lines and ASK is not suitable. Incoherent FSK is widely used for low-rate modems. In all, four frequencies are used, two to transmit and two to receive. The bandwidth available per frequency is 750Hz, and with some simple smoothing baud rates up to 1200b/s are possible. For rates above this value PSK is necessary, and rates up to 2400b/s are possible.

For the second example, consider the problem of modulation choice for the down-link of a satellite microwave link. Here the requirement is for maximum data-rate, coupled with the need to avoid saturation of the available bandwidths close to the link bandwidth. In this case, MSK would be used, due to its good side-lobe performance and low bandwidth use. The additional cost of the demodulation processor is immaterial in comparison with the cost of the satellite.

2.5 Noise in communication systems; probability and random signals

Noise plays a crucial role in communication systems. In theory, it determines the theoretical capacity of the channel. In practise it determines the number of errors occurring in a digital communication. We shall consider how the noise determines the error rates in the next subsection. In this subsection we shall provide a description of noise.

Noise is a random signal. By this we mean that we cannot predict its value. We can only make statements about the probability of it taking a particular value, or range of values. The

probability density function (PDF) $p(x)$ of a random signal, or random variable x, is defined to be the probability that the random variable x takes a value between x_0 and $x_0 + \delta x$. We write this as follows:

$$p(x) = P\{x_0 < x < x_0 + \delta x\} \tag{35}$$

The probability that the random variable will take a value lying between x_1 and x_2 is then the integral of the pdf over the interval $x_2 - x_1$:

$$P\{x_1 < x < x_2\} = \int_{x_1}^{x_2} p(x)dx \tag{36}$$

The probability $P\{-\infty < x < \infty\}$ is unity. Thus:

$$\int_{-\infty}^{\infty} p(x)dx = 1 \tag{37}$$

A density satisfying equation 37 is termed **normalised**.

The **cumulative distribution function (CDF)** $P(x)$ is defined to be the probability that a random variable, x is less than x_0:

$$P(x_0) = P\{x < x_0\} = \int_{-\infty}^{x_0} p(x)dx \tag{38}$$

From the rules of integration:

$$P\{x_1 < x < x_2\} = P(x_2) - P(x_1) \tag{39}$$

Also, from equations 37 and 38:

$$P(\infty) = 1, \quad P(-\infty) = 0 \tag{40}$$

The meaning of the two functions $p(x)$ and $P(x)$ become clearer with the aid of examples.

- **Continuous distribution.** An example of a continuous distribution is the **Normal**, or **Gaussian** distribution:

  $$p(x) = \frac{1}{\sigma\sqrt{2\pi}} \mathrm{e}^{\frac{-(x-m)^2}{2\sigma^2}} \tag{41}$$

  where m is the **mean** value of $p(x)$. The constant term ensures that the distribution is normalised. This expression is important as many naturally occurring noise sources can be described by it, e.g. **white noise**. Further, we can simplify the expression by considering the source to be a zero mean random variable, i.e. $m = 0$. σ is the standard deviation of the distribution.

  How would this be used? If we want to know the probability of, say, the noise signal, $n(t)$, having the value $\pm v_1$, we would evaluate:

  $$P\{-v_1 < x < +v_1\} = \int_{x=-v_1}^{x=+v_1} p(x)dx$$

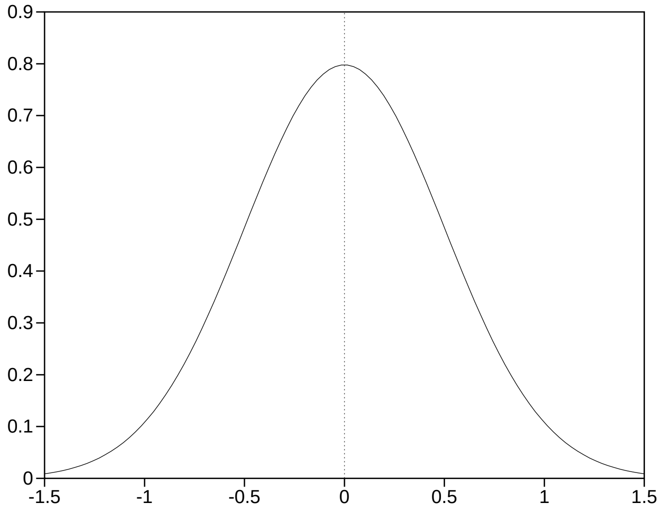


Figure 27: A zero mean Gaussian distribution with $\sigma = 0.5$

In general, to evaluate $P(-x_1 < x < +x_1)$, if we use:

$$\begin{aligned} u &= \frac{x}{\sigma\sqrt{2}} && (42) \\ dx &= du\,\sigma\sqrt{2} && (43) \end{aligned}$$

then from equation 38 and equation 41 with $m = 0$ we have:

$$\begin{aligned} P\{-x_1 < x < +x_1\} &= \frac{1}{\sqrt{\pi}} \int_{x=-x_1}^{x=+x_1} \mathrm{e}^{-u^2} du \\ &= \frac{2}{\sqrt{\pi}} \int_{x=0}^{x=x_1} \mathrm{e}^{-u^2} du && (44) \end{aligned}$$

The distribution function $P(x)$ is usually written in terms of the **error function** $\mathrm{erf}(x)$:

$$P(x) = \mathrm{erf}(x) \tag{45}$$

where $\mathrm{erf}(x)$ is given by:

$$\mathrm{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x \mathrm{e}^{-u^2} du \tag{46}$$

which has the property:

$$\mathrm{erf}(x) = -\mathrm{erf}(-x) \tag{47}$$

The integral is difficult to evaluate and is approximated by use of tables which exist for various value of x. Some values of $\mathrm{erf}(x)$ are shown in Table 2. Note the extremely low probability of the value falling outside ±2.

A more practically useful measure is to consider $P\{-k\sigma < x < +k\sigma\}$ and use equations 42 and 45, which gives us:

x	$P(x)$	x	$P(x)$
0.5	0.5205	2.0	0.9953
1.0	0.8427	2.5	0.9995
1.5	0.9661	3.0	0.99997

Table 2: Some values for the error function $P(x)$

$$P_k(k\sigma) = \mathrm{erf}(\frac{k}{\sqrt{2}}) \tag{48}$$

This is useful in that it gives us a feeling of how likely it is that the value of the noise signal will exceed certain thresholds as given by k. A similar table to that of Table 2 can be evaluated for P_k and is given in Table 3.

k	P_k	k	P_k
0.5	0.383	2.5	0.988
1.0	0.683	3.0	0.997
1.5	0.866	3.5	0.9995
2.0	.955	4.0	0.99994

Table 3: Some values for the error function, P_k

So we see that it is highly unlikely for our noise signal to take values outside the range $\pm 3\sigma$.

The **complementary error function**, erfc(x), is defined as:

$$\mathrm{erfc}(x) = 1 - \mathrm{erf}(x) = \frac{2}{\sqrt{\pi}} \int_x^\infty \mathrm{e}^{-u^2} du \tag{49}$$

- **Discrete distribution.** Probability density functions need not be continuous. If a random variable can only take discrete values, its PDF takes the forms of lines. An example of a discrete distribution is the **Poisson** distribution:

  $$p(n) = P\{x = n\} = \frac{\alpha^n}{n!}\mathrm{e}^{-\alpha} \tag{50}$$

  where $n = 0, 1, 2, \cdots$. The most celebrated example of the Poisson distribution is its accurate description of the number of men in the Austro-Hungarian army killed by kicks from a horse in each year.

  Poisson distributions are normally plotted as line charts as shown in Figure 28.

We cannot predict the value a random variable may take on a particular occasion. We can introduce measures that summarise what we expect to happen on average. The two most important measures are the **mean** (or **expectation**) and the **standard deviation**.

The mean η of a random variable x is defined to be:

$$\eta = \int_{-\infty}^{\infty} x \, p(x) dx \tag{51}$$

or, for a discrete distribution:

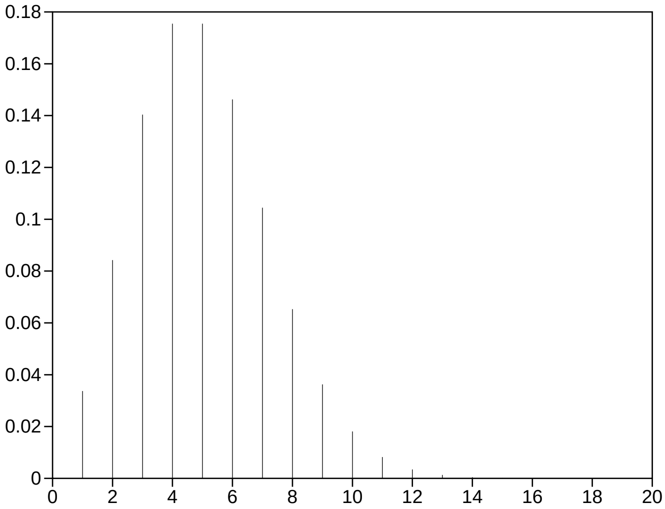


Figure 28: An example Poisson distribution with $\alpha = 5.0, \;\; n = 1, \cdots, 20$

$$\eta = \sum_{n=-\infty}^{n=\infty} n \; p(n) \tag{52}$$

(If n can take only a limited range of values we adopt the convention that $p(n) = 0$ outside this range).

In the examples above we have assumed that the mean of the Gaussian distribution to be 0, the mean of the Poisson distribution to is found to be α. The mean of a distribution is, in common parlance, the average value. On average then, the Austro-Hungarian army lost α men per year as a result from horse kicks.

The standard deviation is a measure of the **spread** of the probability distribution around the mean. A small standard deviation means the distribution (and hence occurrences) are close to the mean. A large value indicates a wide range of possible outcomes. The standard deviation σ is defined to be:

$$\sigma^2 = \int_{-\infty}^{\infty} (x - \eta)^2 p(x) dx \tag{53}$$

or, for a discrete distribution:

$$\sigma^2 = \sum_{n=-\infty}^{n=\infty} (n - \eta)^2 p(n) \tag{54}$$

The square of the standard deviation is called the **variance**. The Gaussian distribution (equation 41) contains the standard deviation within its definition. The Poisson distribution (equation 50) has a standard deviation of α^2.

In many cases the noise present in communication signals can be modeled as a zero-mean, Gaussian random variable. This means that its amplitude at a particular time has a PDF given by equation 41. The statement that noise is zero-mean says that, on average, the noise signal takes the value zero. The mean power in the noise signal is equal to the variance of the PDF. We have already seen that the signal-to-noise ratio (SNR) is an important quantity in determining the performance of a communication channel. The noise power referred to in the definition of SNR (equation 5) is the mean noise power. It can therefore be rewritten as:

$$SNR = 10\log_{10}(S/\sigma^2) \ \text{dB} \tag{55}$$

Moreover, if the source of this noise is thermal noise we have from equation 9 that $\sigma^2 = kTB$. Irrespective of the source of noise, the variance is sometimes written in the terms of E_n the power per unit Hz, so that:

$$\sigma^2 = E_n B \ \text{Watts} \tag{56}$$

Note that E_n is an energy, because Energy = Power × Time = Power/Frequency.

2.6 Errors in digital communication

We noted earlier that one of the most important advantages of digital communications is that it permits very high fidelity. In this sub-section we shall investigate this more closely. We shall consider in detail only BPSK systems, and comment on the alternative modulations.

In the absence of noise, the signal, V, from a BPSK system can take one of two values, $\pm v_b$. In the ideal case, if the signal is greater than 0, the value that is read is assigned 1. If the signal is less than 0, then value that is read is assigned 0. When noise is present, this distinction between $\pm v_b$ (with the threshold at 0) becomes blurred. There is a finite probability of the signal dropping below 0, and thus being assigned 0, even though a 1 was transmitted. When this happens, we say that a **bit-error** has occurred. The probability that a bit-error will occur in a given time is referred to as the **bit-error rate (BER)**. In actuality, we may decide that our threshold of deciding whether the signal is interpreted as a 0 or a 1 is set at $v_b/2$ such that any signal detected between a 0 is read if $-v_b < V < 0$ and a 1 is read if $v_b < V < 0$.

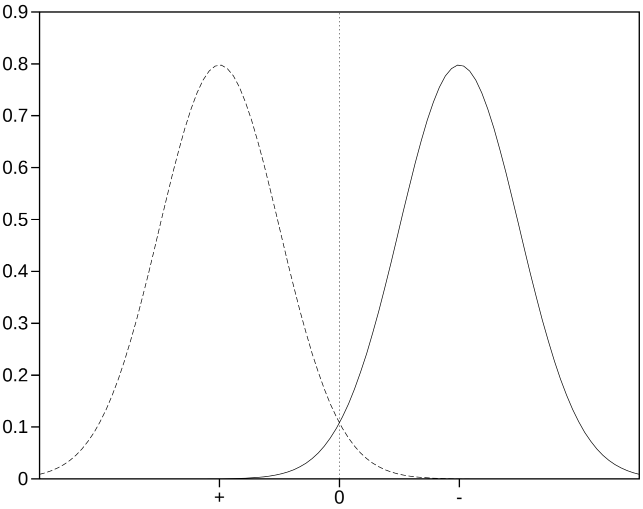


Figure 29: Schematic of noise on a three level line

We suppose (without loss) that the signal V, which has the signal levels $\pm v_b$ with (Gaussian) noise N of variance σ^2 (Figure 29. The probability that an error will occur in the transmission of a 1 is:

$$P\{N + v_b < 0\} \quad = \quad P\{N < -v_b\}$$

$$\begin{aligned} &= \frac{2}{\sqrt{\pi}}\int_{x=-\infty}^{x=-v_b} \mathrm{e}^{-u^2}du \\ &= \frac{1}{2}-\frac{1}{2}\mathrm{erf}(x=-v_b) \\ &= \frac{1}{2}\mathrm{erfc}(x=v_b) \end{aligned} \tag{57}$$

Similarly, the probability that an error will occur in the transmission of a 0 is:

$$\begin{aligned} P\{N+v_b>0\} &= P\{N>v_b\} \\ &= \frac{2}{\sqrt{\pi}}\int_{x=+v_b}^{x=+\infty} \mathrm{e}^{-u^2}du \\ &= \frac{1}{2}-\frac{1}{2}\mathrm{erf}(x=v_b) \\ &= \frac{1}{2}\mathrm{erfc}(x=v_b) \end{aligned} \tag{58}$$

Note that 57 and 58 are the same. This is an important result as it gives us an expression for the probability of error without reference to which value (a 1 or a 0) is transmitted.

It is usual to write these expressions in terms of the ratio of (energy per bit) to (noise power per unit Hz), E_n. The power S in the signal is, on average v_b^2, and the total energy in the signaling period T is v_b^2T. The average energy per bit is therefore:

$$E_b = (v_b^2T+v_b^2T)/2 = v_b^2T \tag{59}$$

Using the expressions 42, 58, 56, 59 we have:

$$\begin{aligned} \frac{1}{2}\mathrm{erfc}(x=v_b) &= \frac{1}{2}\mathrm{erfc}(\sqrt{\frac{v_b^2}{2\sigma^2}}) \\ &= \frac{1}{2}\mathrm{erfc}(\sqrt{\frac{E_b}{2TE_nB}}) \end{aligned}$$

For BPSK, the signaling period T is half the reciprocal of the bandwidth B, i.e. $T=1/2B$; thus:

$$P\{\mathrm{error}\} = \frac{1}{2}\mathrm{erfc}(\sqrt{\frac{E_b}{E_n}}) \tag{60}$$

All coherent detection schemes give rise to error rates of the form in equation 60. For example, QPSK has twice the error probability of BPSK, reflecting the fact that with a quadrature scheme, there are more ways an error can occur. Narrow-band FSK has an error probability rather worse than QPSK, although its numerical value depends on the exact scheme used. Figure 30 shows graphs of $P\{\mathrm{error}\}$ for incoherent ASK, incoherent FSK, BPSK and DPSK; the expressions for $P\{\mathrm{error}\}$ are given in Table 4.

Incoherent demodulation schemes always have a higher probability of error than coherent schemes. Incoherent schemes are forms of power detection, i.e. produce an output proportional to the **square** of the input. Power detection always decreases the SNR. It is quite easy to see why this is so. Suppose the input, X, is of the form $X=v+N$, as before. The input SNR is:

Modulation scheme	$P\{\text{error}\}$
Incoherent ASK	$\approx \frac{1}{2}\mathrm{e}^{-\frac{E_b}{4E_n}}$
Incoherent FSK	$\frac{1}{2}\mathrm{e}^{-\frac{E_b}{2E_n}}$
BPSK	$\frac{1}{2}\mathrm{erfc}(\sqrt{\frac{E_b}{E_n}})$
DPSK	$\frac{1}{2}\mathrm{e}^{-\frac{E_b}{E_n}}$

Table 4: Expressions for error rates in some modulation schemes

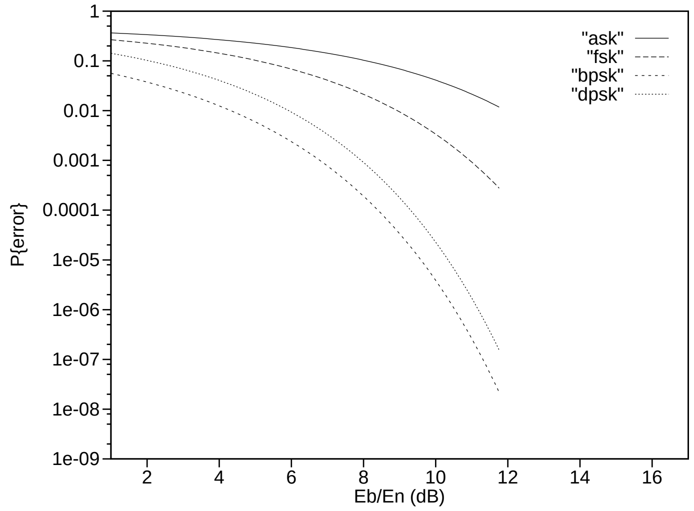


Figure 30: Comparison of error rates in some modulation schemes

$$SNR_{in} = \frac{v^2}{N^2} \tag{61}$$

If we square the input, the output is:

$$X^2 = (v+N)^2 = v^2 + 2vN + N^2 \tag{62}$$

Assuming the SNR is high, $vN \gg N^2$, and the SNR of the output is:

$$\begin{aligned} SNR_{out} &\approx \frac{(v^2)^2}{(2vN)^2} \\ &\approx \frac{v^2}{4N^2} \\ &\approx \frac{SNR_{in}}{4} \end{aligned} \tag{63}$$

This decrease in the signal-to-noise ratio causes an increase in the error probability. The detailed analysis is beyond our scope. Although poorer, however, their performance is good nonetheless. This explains the widespread use of incoherent ASK and FSK.

Error rates are usually quoted as bit error rates (BER). The conversion from error probability to BER is numerically simple: BER $= P\{\text{error}\}$. However, this conversion assumes that the probabilities of errors from bit-to-bit are independent. This may or may not be a reasonable assumption. In particular, loss of timing can cause multiple bit failures that can dramatically increase the BER.

When signals travel along the channel, they are being attenuated. As the signal is losing power, the BER increases with the length of the channel. Regenerators, placed at regular intervals, can dramatically reduce the error rate over long channels. To determine the BER of the channel with N regenerators, it is simplest to calculate first the probability of no error. This probability is the probability of no error over one regenerator, raised to the Nth power:

$$P\{\text{No error over } N \text{ regenerators}\} = (1 - P\{\text{error}\})^N \tag{64}$$

assuming the regenerators are regularly spaced and the probabilities are independent. The BER is then determined simply by:

$$P\{\text{error over } N \text{ regenerators}\} = 1 - P\{\text{No error over } N \text{ regenerators}\} \tag{65}$$

This avoids having to enumerate all the ways in which the multiple system can fail.

2.7 Timing control in digital communication

In addition to providing the analogue modulation and demodulation functions, digital communication also requires timing control. Timing control is required to identify the rate at which bits are transmitted, and to identify the start and end of each bit. This permits the receiver to correctly identify each bit in the transmitted message. Bits are never sent individually. They are grouped together in segments, called **blocks**. A block is the minimum segment of data that can be sent with each transmission. (Usually, a message will contain many such blocks.) Each block is **framed** by binary characters identifying the start and end of the block. Sometimes, blocks are also referred to as **frames**. In addition to bit synchronisation, the receiver must also provide frame synchronisation, permitting the correct identification of the start and end of each block.

The type of method used depends on the source of the timing information. If the timing in the receiver is generated by the receiver, separately from the transmitter, the transmission is termed **asynchronous**. If the timing is generated, directly or indirectly, from the transmitter clock, the transmission is termed **synchronous**.

Asynchronous transmission is used for low data-rate transmission and stand-alone equipment. The block length is only 8 data-bits, permitting different clocks with only approximate synchronism to be used. Very commonly, the 8 bits represent an ASCII character. For this reason, frame synchronisation is often referred to as character synchronisation in asynchronous systems.

Synchronous transmission is used for high data rate transmission. The timing is generated by sending a separate clock signal, or embedding the timing information into the transmission. This information is used to synchronise the receiver circuitry to the transmitter clock. Because the clocks are synchronised, much longer block lengths are possible.

The type of framing, character or block, is sometimes used to indicate whether a system uses asynchronous or synchronous transmission. This terminology does obscure the important difference between the different methods of timing control.

A typical asynchronous character frame consists of a **start-bit**, 8 **data-bits** and one or two stop bits. The start bit is chosen to have an opposite polarity to the **line-idle** state. Its arrival causes an idle/not-idle transition, that indicates to the receiver that a character is arriving. The 8 data bits are then sampled (or **clocked**) through the receiver. The stop-bit(s) return the line to the idle state (Figure 31).

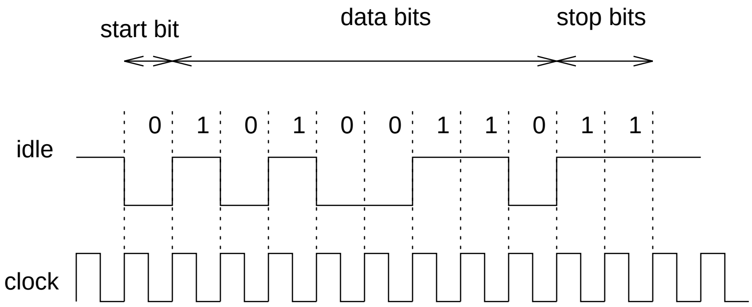


Figure 31: Asynchronous character receiver principle

The clocking is achieved by providing the receiver with a clock running several times faster than the bit-rate (Figure 32). In this way, having identified the idle/start-bit transition, the receiver can easily identify the approximate centre of each bit. Because only 8 data-bits are transmitted, the clock need only be approximate; a 5% difference between transmitter and receiver clocks can be tolerated. (It should be recognised that with modern clock circuitry, a 5% error in clock rate is an enormous error.)

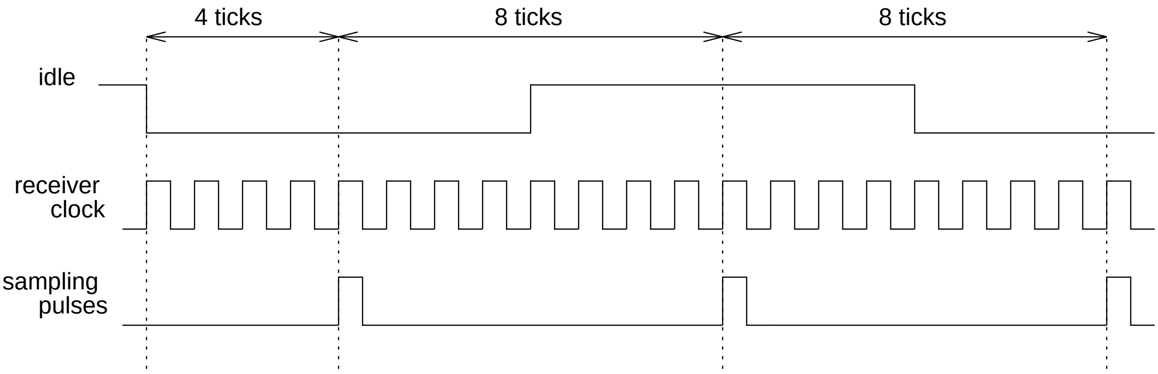


Figure 32: Asynchronous receiver operation with ×8 clock

The addition of start and stop bits reduces the information rate of the channel. If one start and stop bit are used, the data-rate is 8/10 times the signaling rate. This is a general result. The addition of timing information always reduces the information rate of the channel.

Asynchronous transmission is used for transmissions up to 20Kb/s. When the data rate is high, the addition of framing characters around each byte becomes very inefficient. It is natural to wish to increase the block length, to maximise the data-rate. However, as the block length becomes longer, so errors in receiver clock rate less easy to tolerate, because they are accumulative. For example, if a 10000 bit frame is used, and we need to maintain the sampling point to within the central 50% of the bit period, the maximum clock rate error we can tolerate is 5 parts in 2×10^5. In consequence, asynchronous transmission is inappropriate for high data rates, and synchronous transmission must be used.

When using synchronous transmission, special procedures must be adopted to permit the unique identification of the frame. This is achieved by setting the **frame header** to a predetermined pattern (Figure 33). On transmission, any repeat of this pattern in the data is destroyed by the addition of binary zeroes, that are removed again on reception. This process is known as **bit-stuffing**.

Synchronous receivers require a timing signal from the transmitter. An additional channel may be used in the system to transmit the clock signal. This is wasteful of bandwidth, and it is

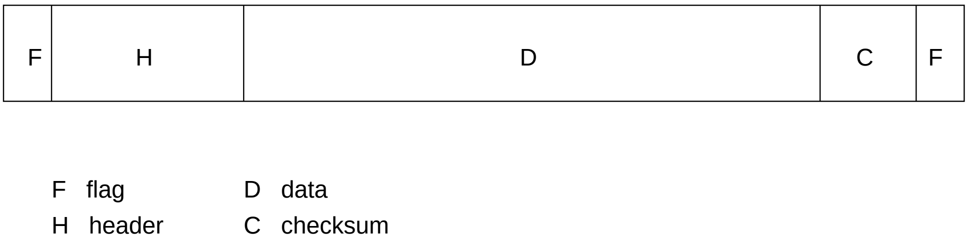


Figure 33: General strcuture of a frame used in synchronous transmission

more customary to embed the timing signal within the transmitted data stream by use of suitable encoding.

One way of embedding the timing signal is to encode the binary signal in such a way as to ensure a 1/0 transition with each bit. This transition is used to generate the clock signal used by the receiver. **Bipolar encoding** (Figure 34) is an example of such a code. This uses a three level line; +V volts to represent a 1 and -V volts to represent a 0, and 0 volts between transmitting a 1 or 0. Bipolar encoding is an example of a **return-to-zero (RZ)** code.

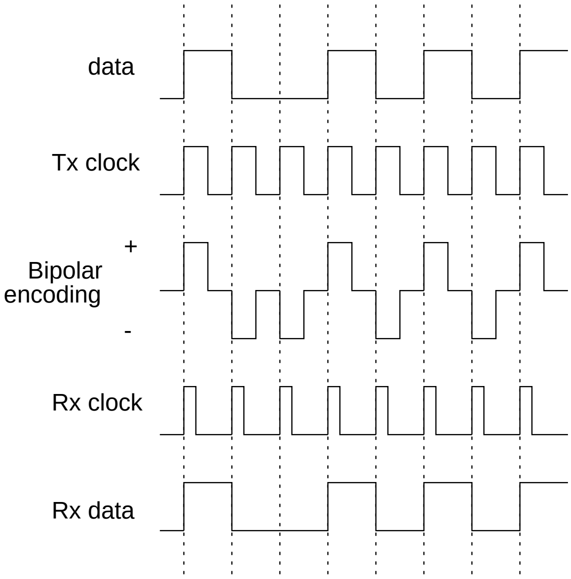


Figure 34: Bipolar encoding

Another example of a timing code scheme is **phase** or **Manchester encoding** (Figure 35). In this scheme, a **transition** in voltage level is used to represent a 1 or zero, rather than a voltage level itself, so only a two level line is needed. A 1 is represented by a low-high transition and 0 by a high-low transition. A transition occurs in the middle of each bit. This is an example of a **non-return-to-zero (NRZ)** code and requires a two level line.

These codes are very inefficient in their use of bandwidth. Both codes require a doubling of bandwidth over the original bit-stream, with no increase in data rate. RZ additionally requires a third signal level with no increase in data rate.

This increase can be avoided by using a device called a **digital phase locked loop (DPLL)**. This uses a **non-return-zero-inverted (NRZI)** coding (Figure 36). In NRZI, the signal is first NRZ encoded. Then, the NRZI signal is produced by the rule that a 1 in the NRZ signal produces no change in the NRZI signal, while a 0 in the NRZ signal produces a change in the NRZI signal.

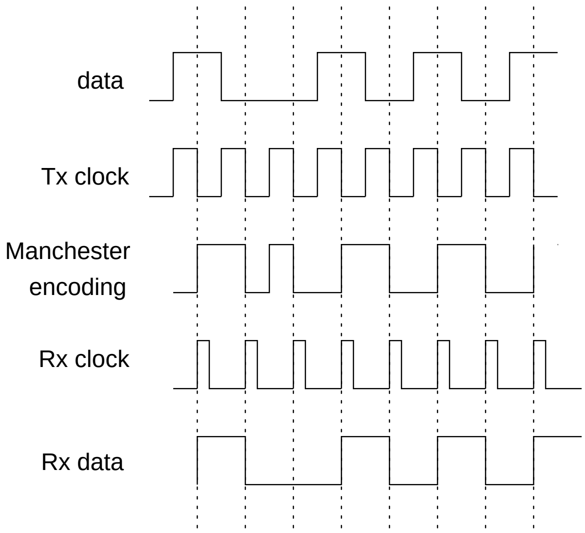


Figure 35: Manchester encoding

The details of the DPLL need not concern us. All we need to understand is that a DPLL has the property that, when presented with a timing signal, it can generate a second timing signal in synchronism with the first. This second signal is used to clock the data through the receiver. The DPLL has a settling time; the correct timing signal is not generated instantaneously. Conversely, the DPLL has 'inertia'; it will continue to generate accurate timing after the 'driving' signal has turned off. These properties permit the bit-stream itself to be used as a driver. On average, the bit stream will present sufficient 1/0 and 0/1 changes to ensure stable output from the DPLL. The practise of bit-stuffing often has the second function of ensuring sufficient 1/0 transitions for the DPLL. When using a DPLL, it is necessary to precede the frame with a **preamble** consisting of synchronisation characters whose only function is to permit the DPLL to settle before the **frame header** or **start-of-frame** bytes are read.

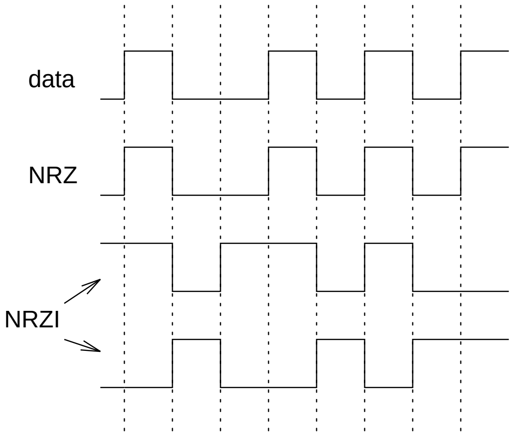


Figure 36: NRZI encoding for use with DPLL

2.8 Design limitations on maximum data-rate and channel capacity

In the methods described in this section, we have found that the choice of modulation technique places a limit on the data rate. This limit is imposed by the number of levels in each signaling interval. For a binary system, if there is two levels per signaling interval, and the maximum data-rate equals the signaling rate. We have also seen that the addition of framing and timing information will degrade the data-rate, so that in general, for a binary system, the data-rate will fall below the signaling rate. These limitations exist by virtue of practise, not principle; by introducing more complex circuitry and signal processing we may increase the data-rate.

An additional limitation on data-rate in some systems arises due to **inter-symbol interference (ISI)**. ISI is the name given to the mistaken identification of one bit due to the presence of a bit in a neighboring time-slot (Figure 37). The importance of ISI varies according to the system. Its effects can be reduced by signal processing, although simple techniques to perform this may require an increase in bandwidth. Its effects can dominate systems relying on crude incoherent detection. The data-rate of optical fibres is usually determined by the need to separate bit-intervals sufficiently to avoid ISI.

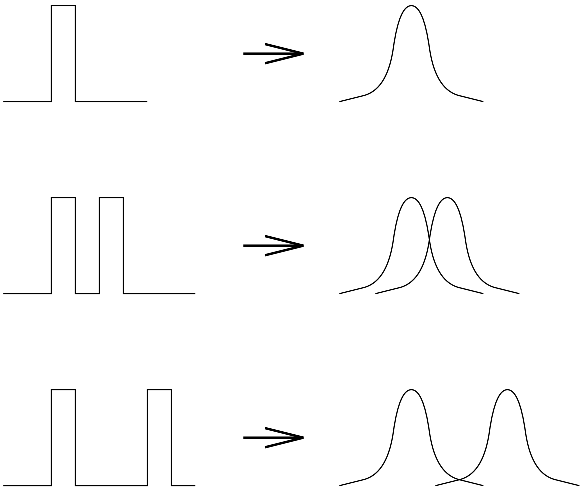


Figure 37: Intersymbol interference

Earlier, we introduced the Hartley-Shannon law that states that the maximum rate at which information can be transmitted error-free over a communication channel is the channel capacity C, given by equation 8:

$$C = B \log_2(1 + (S/N))$$

For a given bandwidth, this relation states that as S/N tends to infinity, so does the capacity. It is apparent that the methods we have just described will not satisfy this law. To take BPSK, for example, the maximum capacity in the absence of noise is equal to the signaling rate.

This inability to satisfy the Hartley-Shannon equation is often explained away by stating that it is a theoretical relation, and that we have to deal with the practical world. This is neither a good nor correct explanation. The law is valid for systems whose input and outputs may take any value. Since this is not true of our systems, the equation is not satisfied, but there is no reason why it should be. (We may derive an alternative relation for binary channels.) The remarkable feature of the Hartley-Shannon law is not that our channels do not obey it but that the message may be transmitted **error-free**. This is an astonishing result. If we consider our results of above we can see that, for a finite signal-to-noise ratio, we always have a finite probability of error. It is the existence of a channel capacity, not its functional form, that makes equation 8 such an important relationship.

It is important to understand the distinction between maximum data-rate and capacity. The maximum data-rate is determined by the system design, and is associated with a given error rate.

The channel capacity is the maximum rate that the channel can transmit data error-free, given its design. Clearly, the channel capacity cannot exceed the maximum-data rate. However, we have (as yet) said nothing concerning how error-free transmission might be achieved.

3 Communication channels

Earlier, we have described the terminology and general properties of communication channels. In this section, the properties of particular channels are described and contrasted. Broadly speaking, we may distinguish between channels that are physically connected with cable or fibre, and channels that have no physical connection, such as microwave links. Cable systems can be divided into **transmission lines**, that carry an electrical voltage between two conductors, and **waveguides**, that carry an electromagnetic wave.

3.1 Transmission lines

A transmission line is a pair a conducting wires held apart by an **insulator** or **dielectric**. They come in a variety of construction geometries. The simplest and least expensive form is **two-wire cable**. **Unshielded twisted pair (UTP)** cable consists of (often sets of) two wires sheathed in an insulator and twisted together. **Shielded twisted pair (STP)** cable contains (often sets of) two wires surrounded and separated by a solid **dielectric**. The dielectric is contained within a copper braid, that shields the conductors from external noise sources. The entire construction is housed in a flexible, waterproof cover.

Ribbon cable individual copper wires held together by an insulator in such a way that the individual wires are insulated from each other but the insulator effectively forms a flat ribbon (e.g. cable used to connect components in PCs such as disc drives).

The use of these kinds of cable is limited by two factors: **attenuation** and **cross-talk**. There are three principle sources of attenuation. **Resistance** (or **impedance**) **losses** are simply the loss resulting from the resistance of the wires. This loss is minimised by the choice of a metal with low resistivity. Copper is chosen for this reason. (Gold is even better, and is in fact used on satellites to reduce losses.) **Dielectric losses** are caused by the heating effects when a varying electric field passes through a dielectric. **Radiation losses** occur because the cable acts as an antenna. All these losses increase with frequency.

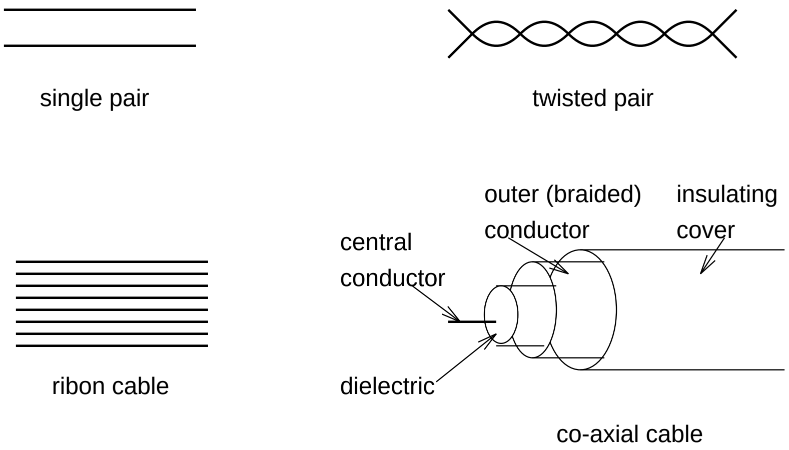


Figure 38: Some different types of transmission lines

When a transmission line can act as an antenna, it can also act as a receiver. Lines prone to radiation loss are also susceptable to **pick-up**, or **cross-talk**. UTP and ribbon cable are particularly prone to this fault. The STP is designed to reduce this pick-up.

All these lines have strong attenuation at frequencies above 1MHz. They are generally used for for low bit-rate communication. Two-wire or four-wire cable is standard for the connection of individual telephone receivers. UTP is the normal method of connection for computer terminals and short high bit-rate connections.

Attenuation increases with both frequency and length (it is usually specified in dB/m at a particular frequency). Because of this fact, it is not possible to give hard-and-fast rules concerning the bandwidth availability of transmission lines. A twisted pair can support rates of many Mb/s over short distances (metres), but over long distances (kilometres) will be completely unsuitable at these data-rates. Using sophisticated modulation techniques and coding schemes it is possible to achieve many Mb/s over reasonable distances using UTP or STP, e.g. 100Mb/s over UTP in a LAN environment.

For long distances, or more generally for rates in excess of several Mb/s, **coaxial cable** is used. Coaxial cable has a central wire, surrounded by a dielectric, in turn concentrically sheathed in a braided conductor. The cable is finally surrounded in a water-proof, flexible sheath. Coaxial cable is familiar to you – it is the cable used to connect your television ariel. The supreme advantage of this method of construction is its resistance to radiation losses. The outer conductor acts to shield out any external fields, whist preventing any internal fields escaping.

Until the advent of optical fibre, coaxial lines were the standard method of long-distance, high bit-rate communication. They are expensive (but getting cheaper, especially as demand rises), and only used where necessary. Typical attenuation values for coaxial cable are 10dB/Km at 10KHz, 50dB/Km at 500MHz. For very long-haul routes, repeaters and equalisers are necessary.

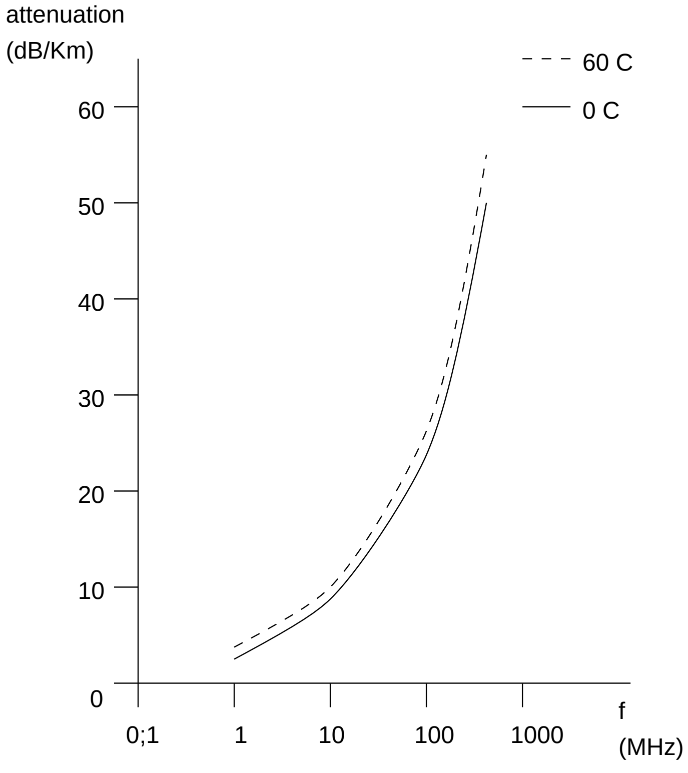


Figure 39: A sketch graph of attenuation in 0.375 inch coaxial cable

For short runs (metres), very high bit-rates are sustainable on coaxial cable – up to 1Gb/s. At rates above this, the cables develop a waveguide action, and dispersion becomes a problem. Although it is possible to design coaxial cable that has still improved performance, these developments have been largely superceded by optical fibre.

3.2 Optical fibre waveguide

For many years it has been appreciated that the use of optical (light) waves as a carrier wave provides an enormous potential bandwidth. Optical carriers are in the region of 10^{13}Hz to 10^{16}Hz, i.e. three to six orders of magnitude higher than microwave frequencies. However, the atmosphere is a poor transmission medium for light waves. Optical communication only became a widespread option with the development of low-loss dielectric waveguide. In addition to the potential bandwidth, optical fibre communication offers a number of benefits:

- **Size, weight, flexibility.** Optical fibres have very small diameters. A very large number of fibres can be carried in a cable the thickness of a coaxial cable.
- **Electrical isolation.** Optical fibres are almost completely immune from external fields. They do not suffer from cross-talk, radio interference, etc.
- **Security.** It is difficult to tap into an optical line. It is extremely difficult to tap into an optical line unnoticed.
- **Low transmission loss.** Modern optical fibre now has better loss characteristics than coaxial cable. Fibres have been fabricated with losses as low as 0.2dB/Km.

The primary disadvantage of optical fibre are the technical difficulties associated with reliable and cheap connections, and the development of an optical circuit technology that can match the potential data-rates of the cables. The speed of these circuits, which are electronically controlled, is usually the limiting factor on the bit-rate. The difficulty of connection and high-cost of associated circuitry result in optical fibres being used only in very high bit-rate communication. There is considerable current debate as to whether optics will ever completely replace electronic technology. In addition, good phase control of an optical signal is extremely difficult. Optical communications are forced to use the comparatively crude method of ASK modulation.

Optical fibre is a waveguide. The fibre (in its simplest form) consists of a core of glass of one refractive index, and a cladding of a slightly lower refractive index (Figure 40). The fibre is then surrounded by a refractive sheath. Typical fibre dimensions are 100μm to 500μm diameter.

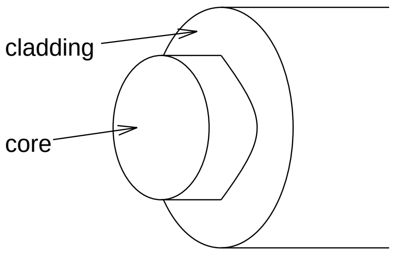
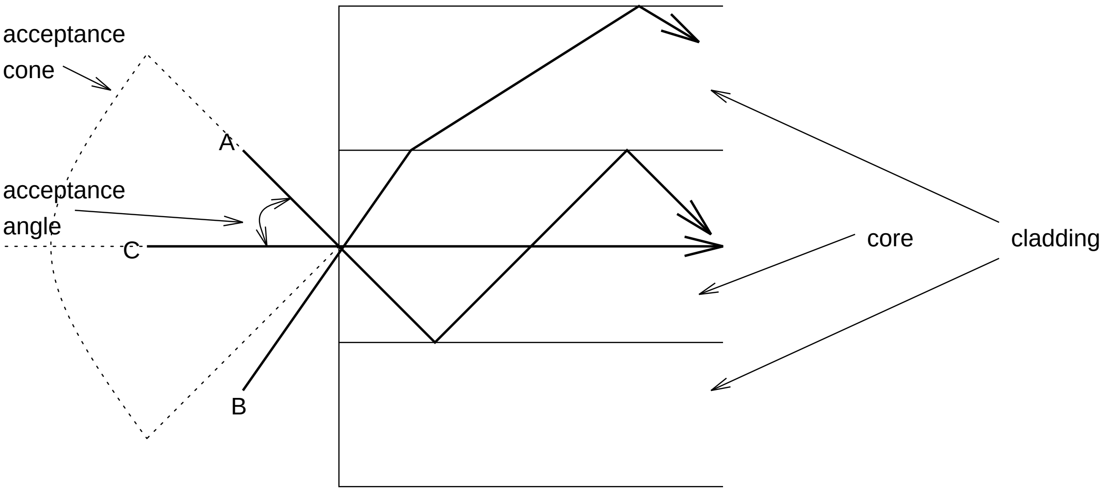


Figure 40: The basic structure of a fibre optic waveguide

In simple terms, the action of a waveguide can be partially understood by considering the rays down the fibre. A light-wave entering the fibre is either refracted into the cladding, and attenuated, or is totally internally reflected at the core/cladding boundary. In this manner it travels along the length of the fibre. The maximum angle at which it may enter the guide and travel by total internal reflection is termed the **acceptance angle** (Figure 41). It is also possible for the wave to follow a helical path down the guide. These rays are called **skew-rays**.

However, this view is too simple to explain all features of waveguide behaviour. In fact, it is not possible for the wave to take any ray down the guide. Only certain rays can be taken. These rays are called **modes**. For any particular frequency, there is a different ray. The modal action of a waveguide is a consequence of the wave nature of the radiation. A **mono-mode fibre** is a



A propogated by total internal reflection

B evetually lost by radiation

C shortest path along fibre

Figure 41: Waveguide action of an optical fibre

fibre that only has one acceptable ray-path per frequency. A **multi-mode** fibre has a number of possible rays that light of a particular frequency may take.

The attenuation of light in the guide has a number of sources. Absorption of light occurs in the glass and this decreases with frequency. Scattering of light from internal imperfections within the glass – **Rayleigh scattering** – increases with frequency. Waveguide imperfections account for low-level loss that is approximately constant with wavelength. Bending the waveguide changes the local angle of total internal reflection and loss increases through the walls. A combination of these effects results in a minimum absorption of about 2dB/Km to 5dB/Km in the 0.8μm to 1.8μm wavelength region (see Figure 42). It is these wavelengths that are used for transmission.

In addition to attenuation, optical waveguides also suffer from dispersion. The dispersion has two sources. Due to the modal behaviour, a waveguide is an intrinsically dispersive device. Put simply, rays of different frequency travel on different paths having different lengths. Because the different frequencies travel different lengths they take different times. In addition to the waveguide dispersion, however, is the material dispersion. Glass is an intrinsically dispersive media. In single mode fibres the material dispersion dominates the waveguide dispersion.

The bandwidth of optical fibres is dominated by the dispersion. In fact, the bandwidth of individual fibres is actually much the same as high quality co-axial cable. It is ironic that the principle justification for optical communication, very large bandwidth, has not in practise been realised. However, it is possible to lay many hundreds of optical fibres in the same cable cross-section as a single co-axial cable.

Fibre optic cable is available in three basic forms:

1. **Stepped-index fibre.** In this type of fibre, the core has a uniform refractive index throughout. This generally has a core diameter of 100μm to 500μm. This is a multi-mode fibre. (Figure 43.)

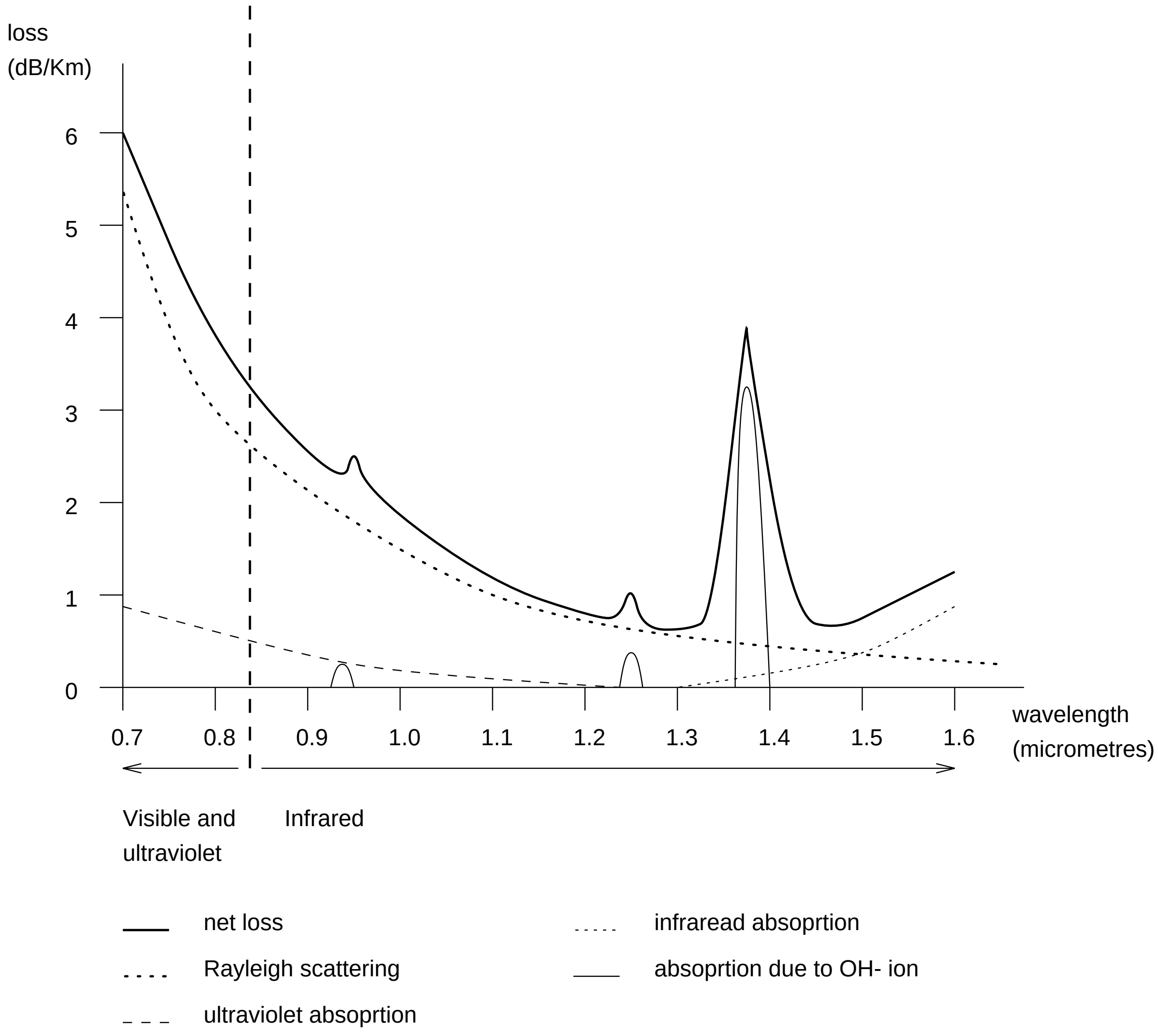


Figure 42: Sketch graph showing contributions to net spectral loss is a glass core

2. **Graded-index fibre.** In this type of fibre, the core has a refractive index that gradually decreases as the distance from the centre of the fibre increases. This generally has a core diameter of 50μm. This is a multi-mode fibre. (Figure 44.)

3. **Mono-mode fibre.** As the name suggests, the distinguishing characteristic of this fibre is that allows only a single ray path. The radius of the core of this type of fibre is much less than that of the other two, however it does have a uniform refractive index. (Figure 45.)

From, 1 to 3, we find that the cost of production increases, the complexity of transmitter and receiver increases, while the dispersion decreases. This latter property change means that the mono-mode fibre also has the potential to provide greater bandwidth. As it becomes cheaper to produce mono-mode fibre technology, we will see an increased use of this type of optical fibre. Figure 46 gives typical operational information for a mono-mode fibre.

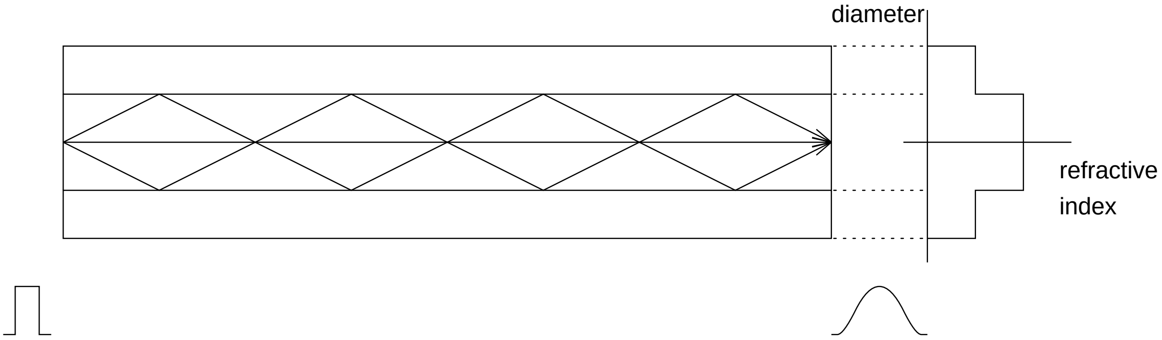


Figure 43: Stepped-index fibre

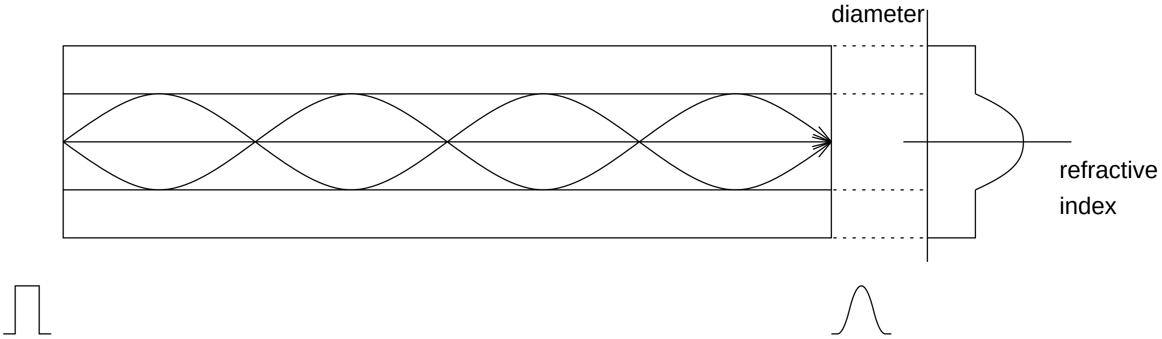


Figure 44: Graded-index fibre

3.3 The electromagnetic spectrum; propagation in free-space and the atmosphere; noise in free-space

In this section we shall consider the physical properties of free-space electromagnetic waves, and how the atmosphere influences the propagation of electromagnetic waves. In the following sections, we shall describe how these properties have determined the selection of frequencies for communication.

The electromagnetic spectrum is divided up into a number of bands, as shown in Table 5.

Description	**Frequency**	**Wavelength**
High frequency	3 - 30MHz	100 - 10m
VHF	50 - 100MHz	6 - 3m
UHF	400 -1000MHz	75 - 30cm
Microwaves	3×10^9 - 10^{11}Hz	10cm - 3mm
Millimetre waves	$10^{11} - 10^{12}$Hz	3mm - 0.3mm
Infrared	$10^{12} - 6 \times 10^{14}$Hz	0.3mm - 0.5μm
Light	$6 \times 10^{14} - 8 \times 10^{14}$Hz	0.5 μm - 0.4μm
Ultra-violet	$8 \times 10^{14} - 10^{17}$Hz	0.4μm - 10^{-9}m
X-rays	$10^{17} - 10^{19}$Hz	10^{-9}m - 10^{-13}m
Gamma rays	$> 10^{19}$Hz	$< 10^{-13}$m

Table 5: The higher frequencies of the electro-magnetic spectrum

Propagation of waves in free-space is different from that in cable or waveguides. With respect to signal propagation, these latter are one-dimensional systems, and a wave does not lose energy

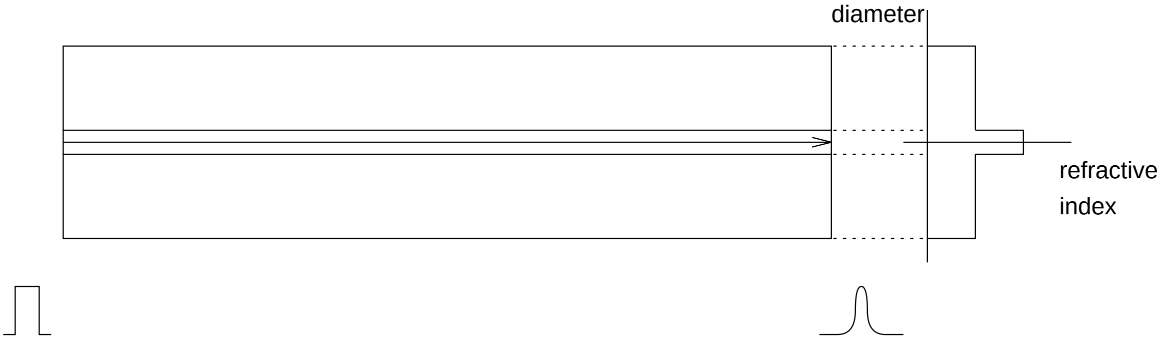
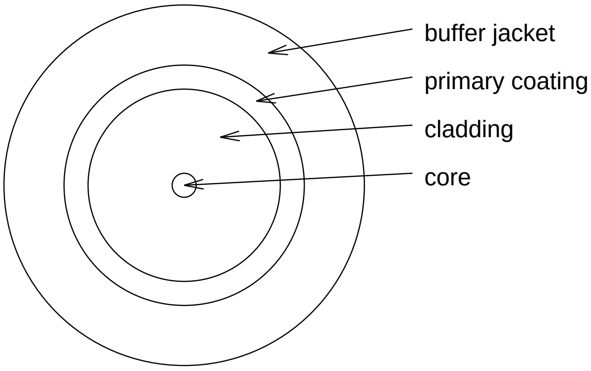


Figure 45: Mono-mode fibre



core diameter	$3 - 10\mu m$
cladding diameter	$50 - 125\mu m$
buffer jacket diameter	$50 - 125\mu m$
acceptance angle	$8 - 10°$

Figure 46: Operational information for a mono-mode fibre

as it travels, except that due to absorption or scattering. In three-dimensions waves radiate spherically. As they travel, the surface area they occupy increases as the square of the distance traveled. However, since energy is conserved, the energy per unit surface area must decrease as the square of the distance. Thus the power of free-space waves obey an inverse square law. For each doubling of the distance between the source and receiver, a 6dB loss is experienced. For all frequencies up to millimetre-wave frequencies, this **free-space loss** is the most important source of loss. Because of it, free-space systems usually require much more power than cable or fibre systems.

When waves traveling in free-space are obstructed, new waves result from the interaction. There are four types of interaction:

1. **Reflection.** This occurs when a wave meets a plane object. The wave is reflected back without distortion.
2. **Refraction.** This occurs when a wave encounters a medium with a different wave speed. The direction and speed of the wave is altered.

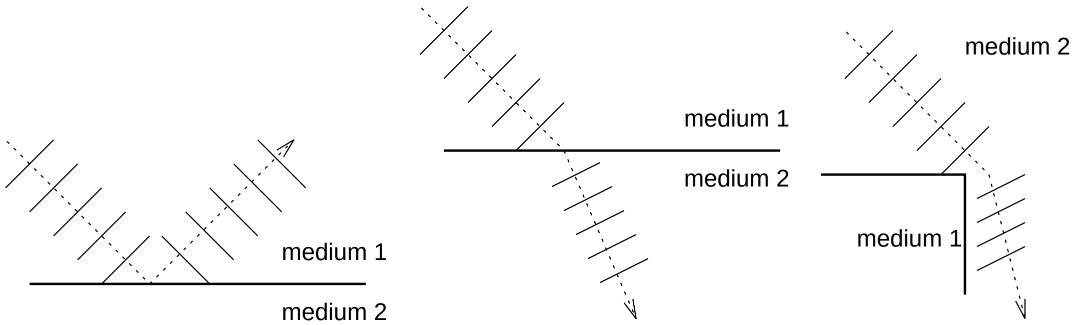


Figure 47: Reflection, refraction and diffraction

3. **Diffraction.** This occurs when the wave encounters an edge. The wave has the ability to turn the corner of the edge. This ability of waves to turn corners is called diffraction. It is markedly dependent on frequency – the higher the frequency, the less diffraction. Very high frequencies (light) hardly diffract at all; 'light travels in straight lines'.

4. **Scattering.** Catch-all description of wave interactions that are too complex to be described as reflection, refraction or diffraction. Typically the result of scattering is to remove radiation of the wave, and re-radiate it over a wide range of directions. Scattering too is strongly frequency dependent. Usually it will increase with frequency.

Figure 47 depicts reflection, refraction and diffraction of an electromagnetic wave.

The propagation of electromagnetic waves through the atmosphere is strongly influenced by the atmosphere (Figure 48). From the point of view of wave propagation, there are two layers. The **troposphere** is the lowest layer of the atmosphere. It extends (typically) from the surface to a height of 50Km. It contains all the Earth's weather, all the liquid water, most of the water vapour, most of the gaseous atmosphere, and most of the pollution. The **ionosphere** extends from the top of the troposphere into outer space. The ionosphere plays a crucial historical role in radio communication. It consists of oxygen molecules that are ionised by the action of the sun. During the day, the quantity of ions rises. At night, the ions recombine to form uncharged oxygen molecules. The ionisation of the atmosphere converts the ionosphere into a **plasma**: an electrically neutral gas of positive and negative charges.

A plasma has a wave-speed that is a strong function of frequency. The consequence of this is that, to a low frequency wave, the ionosphere behaves as a mirror. Waves are simply reflected. This permits a mode of propagation in which the wave bounces forwards and backwards between the ionosphere and the Earth (Figure 49). It was this mode of propagation that permitted Marconi, against the experts' advice, to achieve cross-Atlantic radio communication, and, in the process, discover the ionosphere. The Earth itself also acts as a mirror for electromagnetic waves. As the frequency increases to around 50MHz, the ionospheric effect reduces, and at higher frequencies becomes invisible.

High Frequency (HF) propagation. The propagation of HF consists of two waves. The **sky-wave**, that bounces from the ionosphere, and the **ground-wave**, that propagates along the ground. Because of the long wavelengths, HF communications are not effected by the troposphere, and can usually bend around even large objects such as hills. The ionosphere permits communication over great distances: these are the frequencies of the **radio ham**. The ionosphere is not dependable, however, and signal strengths can vary considerably – effect is maximum in early evening, explaining the overcrowding of the medium wave-band in Britain with European transmissions in the evening.

Very High Frequency (VHF) and **Ultra High Frequency (UHF) propagation.** These

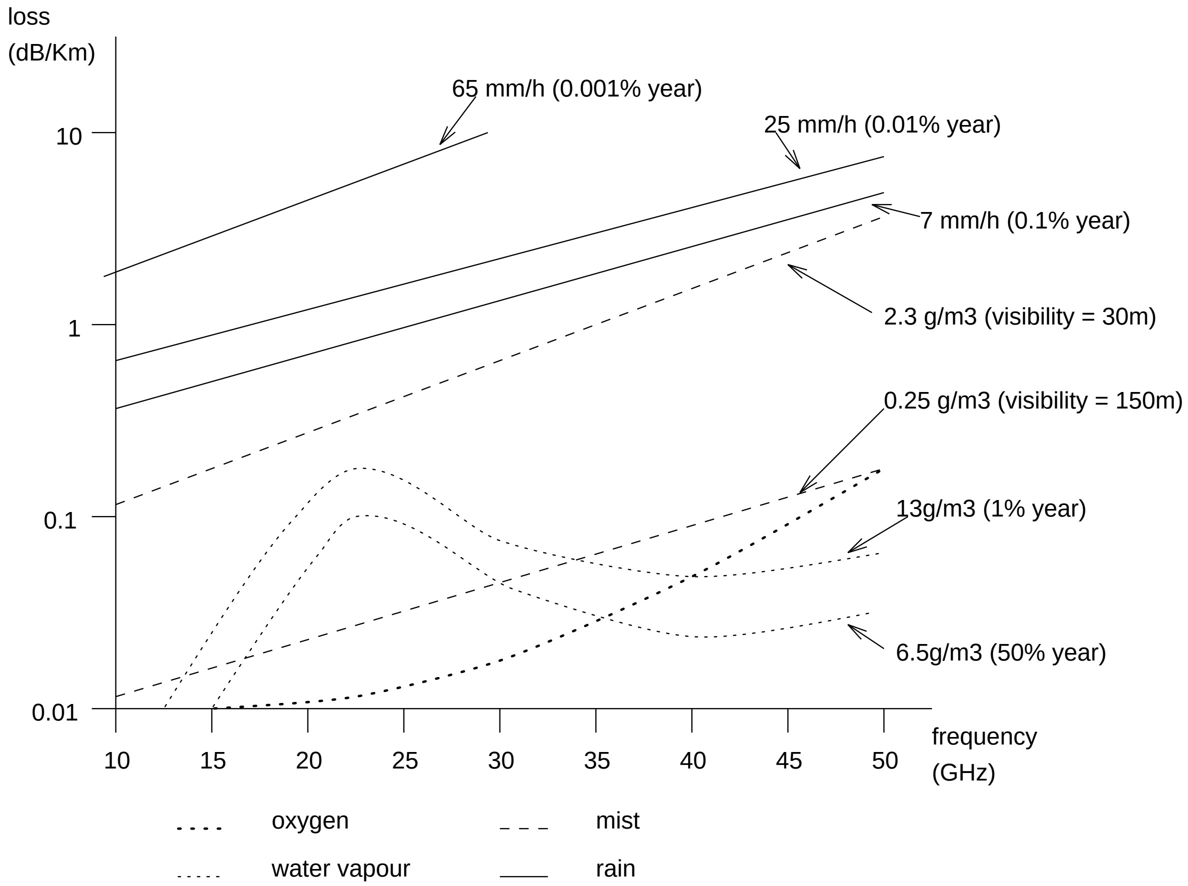


Figure 48: Sketch graph of attenuation due to rain, mist, water vapour and oxygen (temperate climate at ground level)

frequencies are not effected by the troposphere, but are too high to exploit the ionosphere. In addition their decreased wavelength makes them increasingly less able to diffract around obstacles. At UHF in particular, areas in deep radio shadow can experience very poor reception, for example TV signal reception in the prescence when surrounded by large buildings. In urban areas, most reception is due to scattered arrivals. Experiments have shown that the free-space loss is increased in urban areas, to the extent of following a fourth power law. These effects are noticeable under motorway bridges: medium wavelength transmission is usually better than VHF. VHF is used for FM radio stations and UHF for TV transmissions.

When the received signal consists of a number of scattered arrivals, an effect called **fading** occurs. According to the location, the many arrivals may interfere constructively or destructively. The phase relationship between the arrivals changes as the location of receiver changes. This causes a moving receiver to experience fluctuations between strong and weak signals. This effect is known as fading.

Microwave and millimetre wave propagation. Microwave frequencies are usually completely blocked by obstructions. They need **line-of-sight** geometries. In addition, the wavelengths are small enough to become effected by water vapour and rain. Above 10GHz these effects become important. Above 25GHz, the effects of individual molecules becomes important. Water and oxygen are the most important gases. These have resonant absorption lines at 23GHz, 69GHz and 120GHz (Figure 50). Above these frequencies, absorption is increasingly effected by the Earth's **black-body** spectrum. These peaks leave 'windows' that may be used for communication, notably

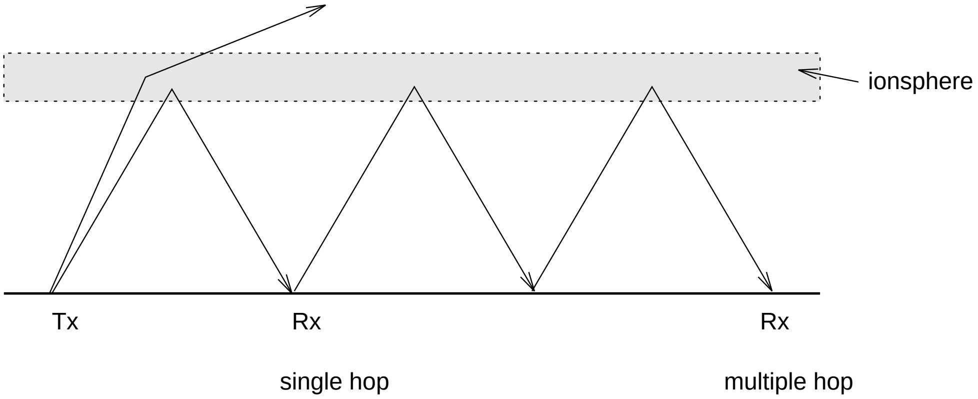


Figure 49: Using the ionosphere for radio wave propagation

at 38GHz and 98GHz.

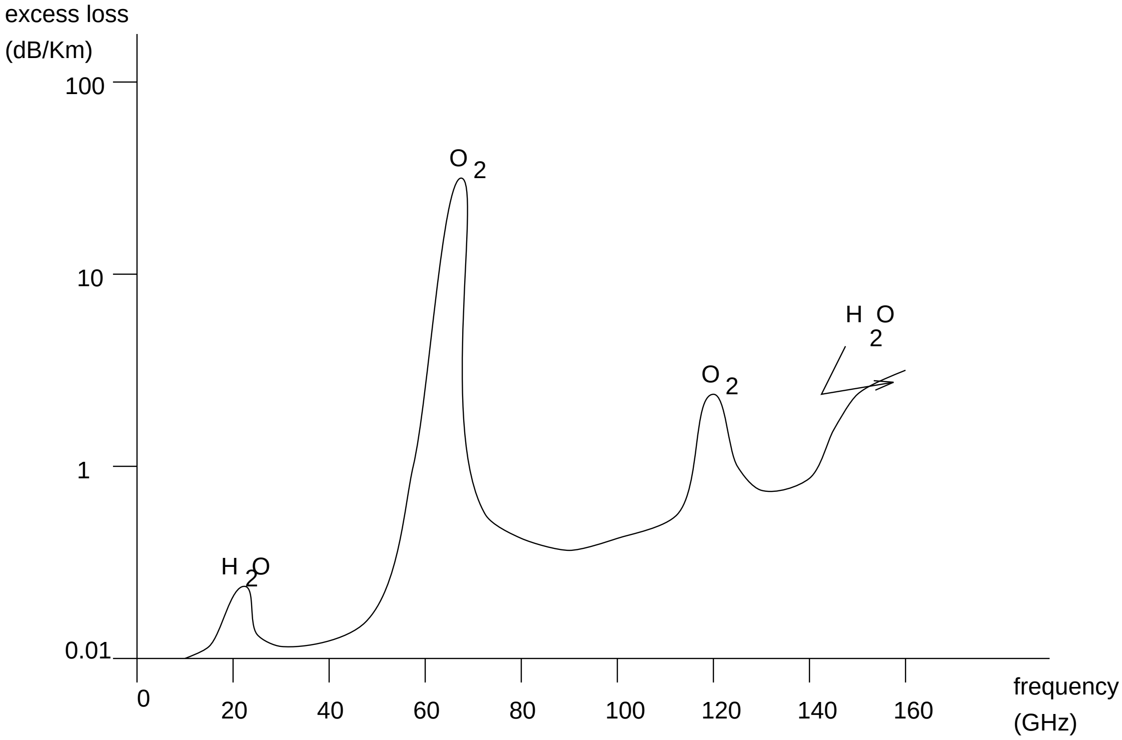


Figure 50: Sketch graph of attenuation due to water and oxygen for microwave and millimetre wave

The use of the atmosphere as a transmission medium means that control over the noise environment is not possible. There are a number of sources of noise. Natural noise is classified by its source. **Atmospheric noise** dominates the low frequencies, up to 2MHz. The primary cause of this noise is electrical discharges in the atmosphere: lightning. So-called **galactic noise** is radio noise from outside the solar system. It is important up to 200GHz. The sun also generates a small noise contribution at these frequencies.

In urban environments these natural sources of noise are dominated by **man-made noise**. In the HF and VHF bands these are mostly other signals, although the very high power early warning radars can generate noise in these bands. At higher frequencies car ignition systems generate high noise levels. In comparison, receiver noise typically is small in the HF, UHF and VHF. At microwave frequencies, however, the thermal noise generated by the receiver (as given by equation 9) is usually the dominant source of noise in the system.

3.4 Microwave link communication

The maturity of radio frequency (RF) technology has permitted the use of microwave links as the major trunk channel for long distance communication. The use of microwave links has major advantages over cabling systems:

- **Freedom from land acquisition rights.** The acquisition of rights to lay cabling, repair cabling, and have permanent access to repeater stations is a major cost in the provision of cable communications. The use of radio links, that require only the acquisition of the transmitter/receiver station, removes this requirement. It also simplifies the maintenance and repair of the link.
- **Ease of communication over difficult terrain.** Some terrains make cable laying extremely difficult and expensive, even if the land acquisition cost is negligible.

The use of microwave links has a number of disadvantages, that mainly arise from the use of free-space communication:

- **Bandwidth allocation is extremely limited.** The competition for RF bandwidth from various competing users leads to very strict allocations of bandwidth. Unlike cabling systems, that can increase bandwidth by laying more cables, the radio frequency (RF) bandwidth allocation is finite and limitied. In practise, bandwidth allocations of 50MHz in the carrier range 300MHz to 1GHz are typical.
- **Atmospheric effects.** The use of free-space communication results in susceptibility to weather effects, particularly rain.
- **Transmission path needs to be clear.** Microwave communication requires line-of-sight, point-to-point communication. The frequency of repeater stations is determined by the terrain. Care must be taken in the system design to ensure freedom from obstacles. In addition, links must be kept free of future constructions that could obstruct the link.
- **Interference.** The microwave system is open to RF interference.
- **Restrictive Costs.** The cost of design, implementation and maintenance of microwave links is high. Many countries are not well equipped with good technical resources to provide efficient and continuous operation.

The modern urban environment presents a particular challenge, in that bandwidth allocation, RF interference, link obstruction and atmospheric pollution place maximum constraints on the system simultaneously. However, urban environments also have the highest land acquisition values too. Many modern cities have found it cost effective to build a single, very high tower to house an entire city's trunk communication microwave dishes. These towers are now a common feature of the modern urban landscape.

As the demand for bandwidth increases, microwave links will become increasingly unable to deliver. The use of increased carrier frequencies in the millimetre wave region would be advantageous. However, for technical reasons, no efficient method of producing large quantities of millimetre power have been found. This is a necessity, given the increase in atmospheric attenuation at millimetre wave frequencies.

The tremendous growth of optical fibre technology and its bandwidth capacity has led to a gradual replacement of microwave links with optical fibre cable for point-to-point communication. The rate of growth has depended on the communication policy of various nations. Europe has seen a rather more rapid growth than the US, reflecting the increased investment by government, however, there are many cable TV companies in the US that are laying fibre based networks. The primary disadvantage of optical cable is that it is largely restricted to point-to-point communications. Microwave and satellite systems retain an advantage in the ease with which they may be extended to include additional users.

Cable laying for intercontinental use is an expensive operation that, like satellites, can be justified only if the additional bandwidth is fully used. Optical fibre links now connect Britain and the US, and areas of Japan. Within the national network in Britain, the replacement of microwave systems by optical fibre is well advanced, especially in heavily populated areas of the country. Microwave systems will continue for some time to come, however, if only because they are already paid for. In addition, microwave links retain a considerable cost advantage in sparsely populated areas.

3.5 Satellite communication

Satellite communication became a possibility when it was realised (by the science fiction writer, Arthur C. Clarke) that a satellite orbiting at a distance of 36,000Km from the Earth would be **geostationary**, i.e. would have an angular orbital velocity equal to the Earth's own orbital velocity. It would thus appear to remain stationary relative to the Earth if placed in an equatorial orbit. This is a consequence of Kepler's law that the period of rotation T of a satellite around the Earth was given by:

$$T = \frac{2\pi r^{3/2}}{\sqrt{g_s R^2}} \tag{66}$$

where r is the orbit radius, R is the Earth's radius and $g_s = 9.81ms^{-2}$ is the acceleration due to gravity at the Earth's surface. As the orbit increases in radius, the angular velocity reduces, until it is coincident with the Earth's at a radius of $36,000Km$. In principle, three geostationary satellites correctly placed can provide complete coverage of the Earth's surface (Figure 51).

For intercontinental communication, satellite radio links become a commercially attractive proposition. Space communication showed phenomenal growth in the 1970s, and will continue to grow for some years to come. The growth has been so rapid that there is now danger of overcrowding the geostationary orbit. Satellite communication has a number of advantages:

- The laying and maintenance of intercontinental cable is difficult and expensive.
- The heavy usage of intercontinental traffic makes the satellite commercially attractive.
- Satellites can cover large areas of the Earth. This is particularly useful for sparsely populated areas.

Satellite communication is limited by four factors:

- Technological limitations preventing the deployment of large, high gain antennas on the satellite platform.
- Over-crowding of available bandwidths due to low antenna gains.
- The high investment cost and insurance cost associated with significant probability of failure.

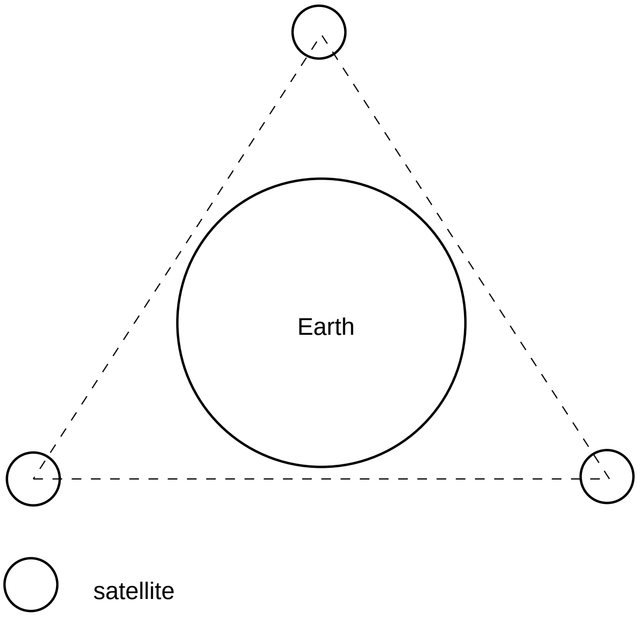


Figure 51: Geostationary satellites providing global coverage

- High atmospheric losses above 30GHz limit carrier frequencies.

A microwave antenna has two functions. It provides **gain** (i.e. amplification). It also directs the radiation into confined regions of space: the **antenna beam**. These properties are largely dependent on the antenna size. For a circular, dish antenna, the gain G is related to the antenna area A by the formula:

$$G = \frac{4\pi A}{\lambda^2} \tag{67}$$

where λ is the wavelength of the transmitted carrier. The angular width of the antenna beam Θ is related to the antenna radius R by:

$$\Theta = 0.61\frac{\lambda}{R} \tag{68}$$

Thus, large antennas have high gains and narrow beams. The profile of a typical antenna beam is shown in Figure 52.

The cost of constructing an antenna is a strong function of its size. A rough rule of thumb is the the cost is proportional to the diameter cubed. Thus a doubling of the antenna size will result in the satellite cost increasing eight times. As a result, antenna sizes are limited. The limitation in antenna size means that the satellite beam is wide. In order to prevent electromagnetic interference with terrestrial stations, the power radiated by the satellite is limited by international convention. In any event power, is severely limited on a satellite platform.

Because the radiated power is low, large receiving antennas are required. The larger the receiver antenna, the larger the antenna gain, and hence the better the receiver SNR. The SNR is a function of the bandwidth, and the atmospheric attenuation. Ground stations close to the poles of the Earth have low elevation **look angles**, and signals have to pass through a thicker section of atmosphere. The size of receiver antenna is determined by the two requirements; 500MHz receive bandwidth and full capability at ±80° of latitude.

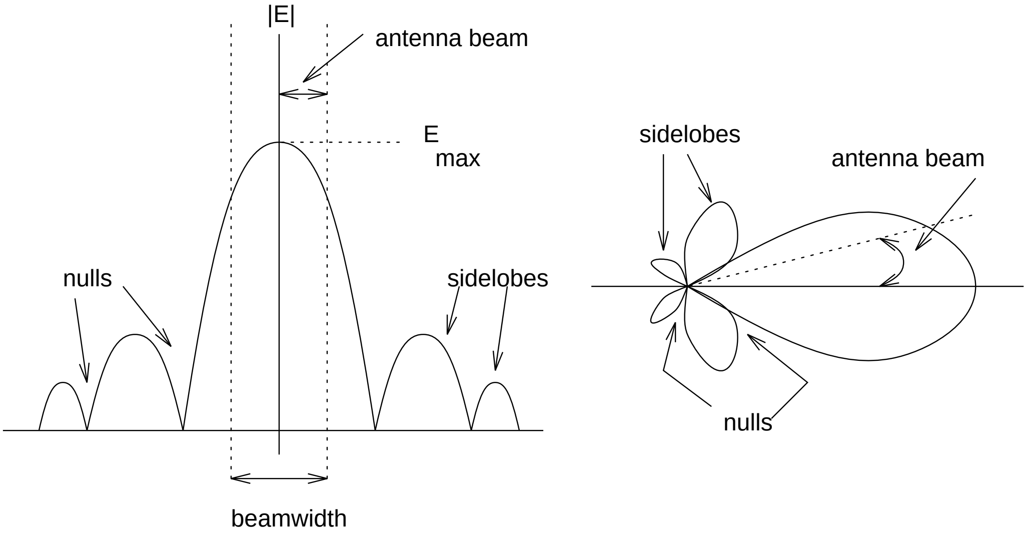


Figure 52: A typical antenna beam profile of a dish antenna

A standard INTELSAT receiver is 30m in diameter. An antenna this large has a very narrow beam, typically 0.01°. A geostationary satellite is not truly stationary, it wanders slightly in the sky. The very narrow beam width of the receiver requires automatic tracking of the satellite, and continuous pointing of the receiver antenna. An INTELSAT ground station is thus a large, expensive piece of equipment.

Satellite systems are extremely expensive. As an example, the break down for a particular British satellite is as shown in Table 6.

Item	Cost [$Million]
Satellite construction	300
Investment finance	300
Insurance	300
Launch	100
	1000

Table 6: Example costs for a satellite system

The use of satellites for regional communication is possible if there is sufficient demand for traffic. By reducing the range of latitudes down to ±60°, and reducing the bandwidth down to 50MHz, large reductions in satellite and ground station receiver costs are possible. One such **direct--to-user (DTU)** system is the Satellite Business System (SBS) covering a range of business and governments users with a demand for high speed data links in the US. The region is split into areas, roughly coincident with the satellite antenna gain contours, denoting increased cost of receiver technology. It is important to realise that the economies of satellite communication only make this regional communication possible if the system is heavily used (Figure 53).

Improvements in satellite receiver technology have permitted smaller antennas to be used as ground station receivers. However, antennas are reciprocal. They have the same directional characteristics in transmit and receive. The use of low gain, wide beam earth stations for DTU systems has contributed considerably to the bandwidth overcrowding problem, particularly in the US.

Recently there has been interest in **low-earth orbiting (LEO)** satellites. Here, a satellite placed in a 1000Km orbit has an orbital time of 1 hour. These satellites can be operated in a **store-and-**

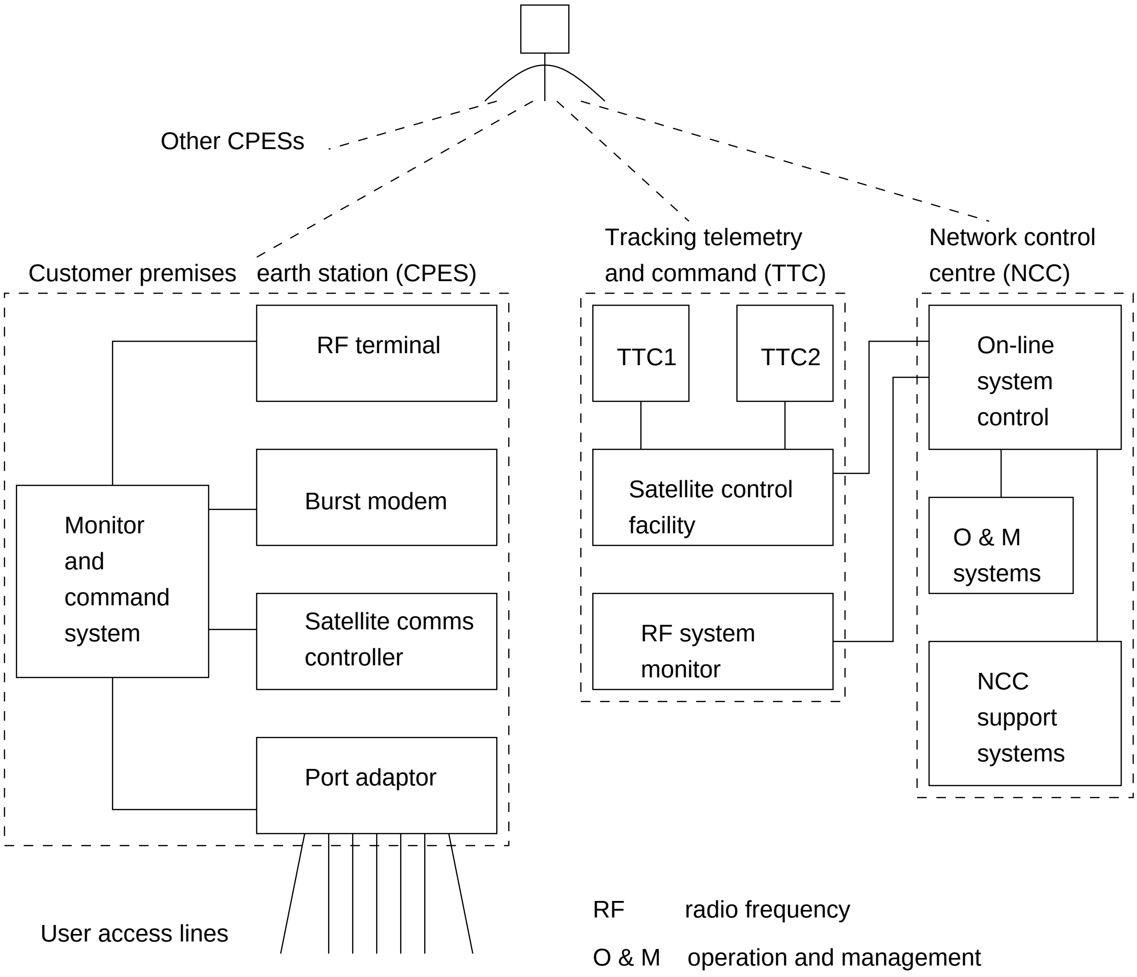

Other CPESs
Customer premises earth station (CPES)
Tracking telemetry and command (TTC)
Network control centre (NCC)
RF terminal
Burst modem
Monitor and command system
Satellite comms controller
Port adaptor
TTC1
TTC2
Satellite control facility
RF system monitor
On-line system control
O & M systems
NCC support systems
User access lines
RF radio frequency
O & M operation and management

Figure 53: The Satellite Business System operational schematic

forward mode, picking up data at one part of the globe and physically transferring it to another. Because the data-rates and orbit radius are greatly reduced, small, low-cost satellites and ground stations are possible. However, such satellites have yet to demonstrate any commercial success.

3.6 Mobile communications

The use of mobile radio-telephones has seen an enormous boost in the 1980s and 1990s. Previous to this time, citizen band (CB) radio had served a limited market. However, the bandwidth assignation for CB radio was very limited and rapidly saturated. Even in the U.S., a total of only 40 10KHz channels were available around 27MHz. The use of digital mobile telephones has a number of advantages over CB radio:

- Access to national and international telephone system.
- Privacy of communication.
- Data independent transmission.
- An infinitely extendable number of channels.

Mobile communications are usually allocated bands in the 500MHz to 1GHz band. At these frequencies the effects of scattering and shadowing are significant. Lower frequencies would improve this performance, but HF bandwidth is not available for this purpose. The primary problems associated with mobile communication at these frequencies are:

- Maintaining transmission in the fading circumstances of mobile communication.
- The extensive investigation of propagation characteristics required prior to installation.

Mobile communication work by limiting transmitter powers. This restricts the range of communication to a small region. Outside this region, other transmitters can operated independently. Each region is termed a **cell**. These cells are often represented in diagrams as hexagons. In practise the cell shape is determined by local propagation characteristics. Together the cells will completely cover the area supplied with mobile communication coverage (Figure 55). The cells are grouped into **areas** with the cells in one area interconnected by a terrestrial links between a **base station (BS)** in each cell.

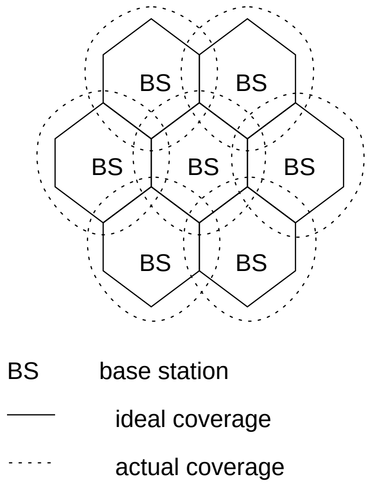


Figure 54: Use of cells to provide geographical coverage for mobile phone service

The total bandwidth available is divided into **channels**. These channels are then grouped in sets and the sets allocated for use in cells in such a way as to ensure that the same set is not used in adjacent channels (Figure 55). Provided that the separation between the cells is sufficient, the **co-channel interference** should be negligible. A useful measure is the ratio d/r which may have the value or around 3 or more.

Within each cell, the mobile user communicates with a transmitter within the cell. As the mobile user approaches a cell boundary, the signal strength fades, and the user is passed on to a transmitter from the new cell. Each cell is equipped with a base station that transmit/receive to/from the mobile within the cell. Within a single cell, a number of channels are available. These channels are (usually) separated by frequency. When a mobile initiates a call, it is assigned an idle channel within the current cell by the **mobile switching centre (MSC)**. He/she uses the channel within the cell until he/she reaches the boundary. He/she is then allocated a new idle channel within the next cell.

For example, the American **advanced mobile phone service (AMPS)** makes use of a 40MHz bands in the range 800MHz – 900MHz. Each band is split into a 20MHz transmit and 20MHz receive bandwidth. These bands are split into 666 two-way channels, each having a bandwidth of 30KHz. These channels are subdivided into 21 sets of channels, arranged in 7 groups of 3. The nominally hexagonal pattern contains 7 cells, a central one and its 6 nearest neighbours. Each cell

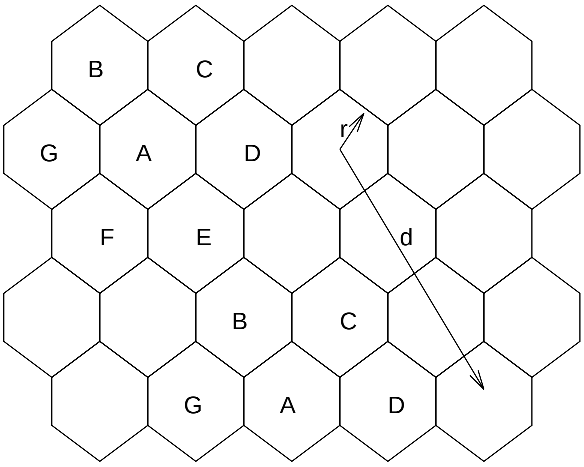


Figure 55: Frequency re-use in cells

is assigned a different group in such a way that at least two cells lie between it and the next block using that set. With a total of 666 channels, it is possible to assign three sets of 31 channels per cell.

The great strength of this type of network is the ease with which more channels may be introduced. As demand rises, one simply reduces the cell size. Then the same number of channels is available in a smaller area, increasing the total number of channels per unit area. In a well planned system, the density of cells would reflect the user density.

AMPS is a **first generation** mobile phone system. It uses analogue modulation. It is one of six incompatible first generation systems that exist around the world. Currently, **second generation** systems are being introduced. These are digital in nature. One aim of the second generation mobile systems was to try and develop one global standard, allowing use of the same mobile phone anywhere in the world. However, there are are currently three digital standards in use, so this seems unlikely. The pan-European standard is known as **Global System for Mobile communication (GSM)**, and is now available in the UK. The services planned for the GSM are similar to those for ISDN (e.g. call forwarding, charge advice, etc.). Full GSM will have 200KHz physical channels offering 270Kb/s. Currently, one physical channel is split between 8 users, each having use of 13Kb/s (the rest is used for channel overhead). The aims of the GSM system are:

- Good speech quality
- Low terminal cost
- Low service cost
- International roaming
- Ability to support hand-held portables
- A range of new services and facilities (ISDN!)

The heart of the mobile telephone network is the **mobile switching centre (MSC)** (Figure 56). Its task is to acknowledge the presence of the mobile user, assign him/her a channel, broadcast his/her dialed request, and then return the call. In addition, it automatically monitors the signal strength of both transmitter and receiver, and allocates new channels as required. This latter process, known as **hand-off**, is completely hidden to the user, although it is a major technical problem. In addition, the MSC is responsible for charging the call. The decision making ability of the MSC relies to a great extent on modern digital technology. It is the maturity of this technology that has permitted the rapid growth of mobile communications.

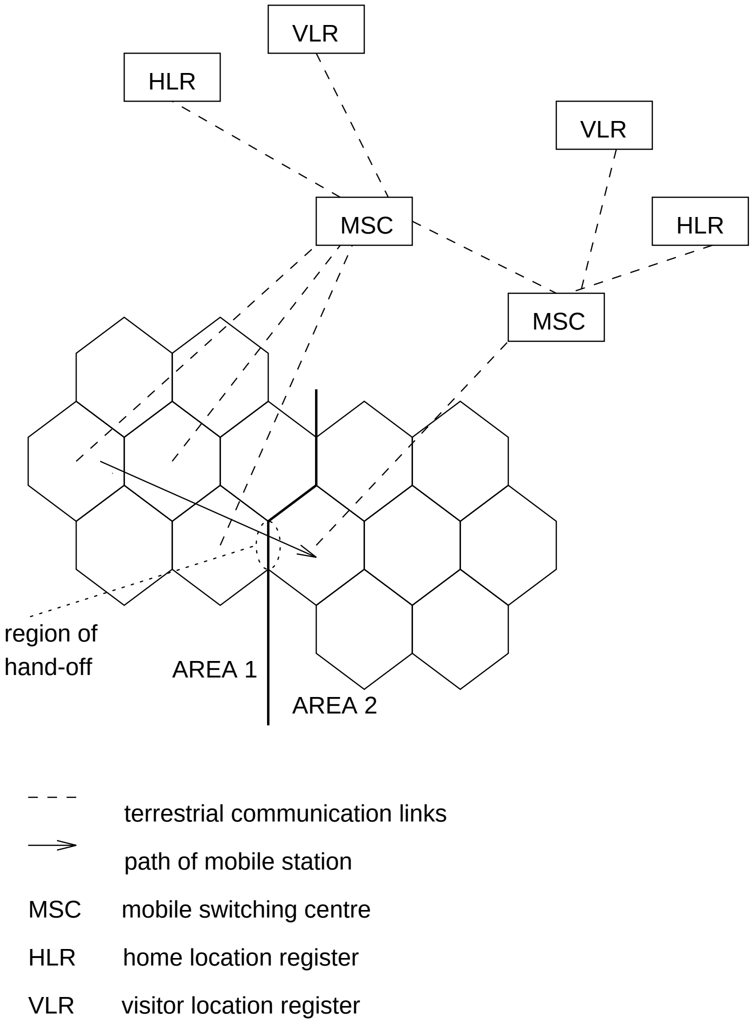

VLR
HLR
VLR
HLR
MSC
MSC
region of hand-off
AREA 1
AREA 2
- - - terrestrial communication links
—> path of mobile station
MSC mobile switching centre
HLR home location register
VLR visitor location register

Figure 56: A mobile network

When a **mobile station (MS)** is switched on, it starts a **beacon** signal that it broadcasts constantly to be received by the nearest base station. Each mobile network has at least one **home location register (HLR)** and at least one **visitor location register (VLR)**. The HLR stores information about MSs that subscribe directly to the service provider providing the mobile network, i.e. home MSs. The VLR contains information about MSs from other subscribers, i.e. visitor MSs. The VLR is updated by detecting the prescence of the beacon from the MS as it roams.

The principle problem with mobile communication is the variation in signal strength as the communicating parties move. This variation is due to the varying interference of scattered radiation – **fading**. Fading causes rapid variation in signal strength. The normal solution to fading, increasing the transmitter power, is not available in mobile communication where transmitter power is limited.

There is considerable research into suitable modulation schemes to cope with fading. The presence of fading has a severe effect on the BER. It can reduce the BER by several orders of magnitude over the non-fading case. In addition, some modulation schemes cope better than others. Surprisingly, coherent AM modulation is better than coherent FM in the fading environment. In any event BERs of 10^{-3} are typical, in comparison with BERs of 10^{-6} for non-fading channels. There should be no surprise if your mobile phone has a worse performance than that in your home.

The installation of a mobile telephone system requires a large initial effort in determining the propagation behaviour in the area covered by the network. Propagation planning, by a mixture of observation and computer simulation, is necessary if the system is to work properly. At UHF and VHF frequencies, the effects of obstructions is significant. Some of the effects that need to be considered are:

- **Free space loss.** This significantly increases in urban environments. Studies have indicated that a $(\text{distance})^{-4}$ relationship is more often followed than a $(\text{distance})^{-2}$ law.
- **Effect of street orientation.** Streets have a significant waveguide effect. Variations of up to 20dB have been measured in urban environments as a result of street direction.
- **Effects of foliage.** Propagation in rural areas is significantly effected by the presence of leaves. Variations of 18dB between summer and winter have been observed in forested areas.
- **Effect of tunnels.** Tunnels can introduce signal attenuation of up to 30dB according to the tunnel length and transmission frequency in use.

3.7 Quality of service (QoS) parameters

While knowledge of the physical nature of the channel is important, in order to determine if it is suitable for use within a particular communication system, we may need a more formal description of the channel in order that we may evaluate if it is suitable for the transporting a particular type of data. To do this, we need some metrics for determining the characteristics of the channel, the network and the data.

This subsection is intended as a introduction to some of these characteristics, which are commonly referred to as **quality of service (QoS)** parameters.

A note on terminology: In most literature, the use of the word **bandwidth** takes a different sense to that which has been used so far within this text. In this text, bandwidth is a measure of the range of frequencies that make up a signal or that can be passed through a communication channel, and has the units **cycles per second** or **Hertz (Hz)**. In most (computer science based) data communication literature, 'bandwidth' is used to refer to **capacity** or **throughput** and takes the units **bits per second (bps or b/s)** or **bytes per second (Bps or B/s)**. In order to conform to the terminology used in the popular literature, in this subsection the word 'bandwidth' is used in the latter sense. It could be argued that this usage of the word is incorrect. However, it is widely used in this manner and its meaning in a particular usage is (usually) obvious through context. So, the term 'bandwidth' is used to express the measure of the network throughput or data rate, e.g. 10Mb/s is the (theoretical) maximum bandwidth of Ethernet.

It is useful to think of different streams of data as individual **flows**, each of which may have its on requirements in order that it may be transported suitably across the network. For instance, **audio** flows and **video** flows may have different data transfer rates and be willing to tolerate different error rates. (There may also be need for **synchronisation** between flows, but that is beyond our scope.)

The nature of the traffic generated by applications is sometimes used to catagorise the types of application. The three mains types are:

- **Constant bit rate (CBR)** applications.

- **Variable bit rate (VBR)** applications.
- **Available bit rate (ABR)** applications.

CBR applications are characterised by the fact that they generate data traffic at a constant bit rate. Typically, such applications will specify the data rate at which they operate and expect that the network will provide this rate without fluctuation. An example of a CBR application is a N-ISDN videophone.

VBR applications generate data whose rate may vary during the lifetime of the application. Such applications try vary their data rate in order to make efficient use of the network resources offered to them. VBR sources may be characterised by a number of QoS parameters:

- **Minimum bandwidth,** D_{min}**:** the minimum network throughput required.
- **Peak/Maximum bandwidth,** D_{max}**:** the maximum network throughput required.
- **Mean bandwidth,** D_{av}**:** the average network throughput required.
- **Peak sustain,** N_p**:** a period for which the peak throughput may be sustained. This measure may be expressed in units of time or as a number of packets.
- **Burstiness,** N_b**:** this is the ratio of D_{max}/D_{min}.

To try and give these terms more meaning, let us consider an example where data is being sent as variable length packets with a maximum packet size of B_{max} – this is a very common scenario.

D_{max} is then the maximum data rate achieved when sending a packet of the maximum size:

$$D_{max} = \frac{B_{max}}{T_{max}} \tag{69}$$

where T_{max} is the time taken to send (transmit) a packet of size B_{max}. D_{max} is normally known to the user of the network service.

D_{av} not often be an exact measure. It can be hard to assess the average throughput of a VBR application as the exact nature of usage of the applications may change this value. With CBR systems, $D_{max} = D_{av} = D_{min}$. A VBR system may have a value of $D_{min} > 0$.

For example; an audio conferencing application may offer several audio encoding schemes (e.g. 78Kb/s 8-bit μ-law encoded 8KHz PCM, 17Kb/s GSM, 9Kb/s LPC), so this applications has minimum throughput requirement of 9Kb/s and a maximum of 78Kb/s. However, this can be considered as a CBR application with a choice encodings. Other applications may not have this choice of data rates, but truly vary their rate of data transfer (e.g. H.261 video applications).

N_p is a measure of the time for which it is expected that a VBR application will maintain its peak rate. This value may again be variable in real use, but it is possible to evaluate the expected value. For example, a VBR video application using a frame-to-frame difference coding scheme may have its peak data rate when there is a scene change in the video stream and this may be expected to last for the time required to transfer the first frame of the new scene, i.e. it will be a known quantity.

N_b has no formal quantatative definition, but it is taken to be the ratio D_{max}/D_{min}. So CBR applications will have $N_b = 1$, but VBR applications can have high busrtiness, e.g. high quality VBR video may have $N_b = 40$. Considering the above example of a variable length packet based system, this measure can be considered to reflect the length of the time between packets being sent – if there are many large packets generated for transmission at the same time, then the greater the time between the tranmission of the packets, the greater the burstiness.

ABR applications are applications that can make use of any bit rate. For example file transfer, electronic mail and Network File System (NFS). These applications differ from VBR applications as they can make use of arbitrarily small and arbitrarily large network capacities during the normal course of their operation. However VBR applications (such as VBR video) may have a minimum rate of operation below which they can not operate.

We can also consider the characteristics of the communication channel and the network as a whole.

When transmitting information, there will always be a **delay** from when the information is transmitted to when it is received. For a simple point-to-point channel, there are two main delay parameters we need to consider; the **propagation delay,** T_p and the **transmission delay,** T_x:

$$\begin{aligned} T_p &= \frac{L}{V_p} \text{ s} && (70) \\ T_x &= \frac{B_n}{D} \text{ s} && (71) \\ T_{rtt} &= 2(T_p + T_x) \text{ s} && (72) \\ a &= \frac{T_p}{T_x} && (73) \end{aligned}$$

L is the distance between transmitter and receiver [m], V_p is the speed of the signal [m/s], B_n is the number of bits to be transmitted and D is the data rate [b/s].

The **round trip time,** T_{rtt} is a useful measure – it is the minimum time that a transmitting station can expect to wait before it receives a a reply to a message from the receiving station.

If $a >> 1$ then this tells us that the propagation delay is the dominant delay factor for the channel – at high data rates, in such a channel there will be much data buffered in the channel itself. This measure is also commonly expressed as the **bandwidth-delay product**:

$$B_d = RT_p \text{ bits} \qquad (74)$$

This factor is important as it quantifies the amount of data that is in the channel and hence is not controllable. This can give important information to designers of the higher layers of the communication infrastructure. For example, a geostationary satellite link may have a propagation delay of 250ms and a data rate of 10Mb/s. This means that the link itself holds 2.5Mb of data (5Mb if you consider a duplex connection). Hence, an **automatic repeat request (ARQ)** protocol for error control would not be suitable for such a link because of the large amount buffering that would be required at the transmitter and receiver.

These are not the only delays inherent in a practical network. In a real network, there will be many links connected by switches, bridges and routers. As messages pass through these devices they will be subject to **switching delays** (the time taken for a switch to transfer messages between different links) and **queuing delays** within the same devices (caused by the buffering of messages awaiting to be processed).

The switching and queuing delays are not constant, and will depend on how heavily loaded the network is. Also, in a packet switched network, different packets for the same destination will not necessarily take the same route at all times, leading to out of order delivery (and there may also be loss). This means that the **end-to-end** delay experienced by a flow at the application level will not always be constant. This variation in the delay is referred to as **delay jitter** or just **jitter**. Delay and (sometimes more significantly) delay jitter must be carefully handled, especially in the context of real-time interactive applications that involve multi-media information flows.

While parameters related to throughput will provide useful information, we also need to know how reliable the channel is in determining if it is useful for a particular application. Parameters exists that quantify the expected **error rates** for a particular channel. We have already looked at some detail at one such measure – the **bit error rate (BER)**. The BER gives information about the likelyhood of single bit errors occurring on the channel. This in turn can be used to determine the performance expected of higher level protocols, and any **error detection** and/or **error correction** methods required. For example, if a channel has a BER of 10^{-6} and our higher level protocol uses frame sizes of 10,000 bits, then, on average, we can expect one frame in every 100 to have an error in it. If this is unacceptable, then a reduction in frame size to 1,000 bits will result in an error for one packet in every 1000.

Another measure that is used, but that is not strictly associated with the physical characteristics of the channel, is the **frames loss rate**. This may also be called **packet loss rate** or **cell loss rate** in other contexts. This measures the probability of individual frames (or packets or cells) not being delivered to their final destination. The frame may have been dropped somewhere within the network due to insufficient resources available at a particular switch or router somewhere along the path. Such measures are expressed in a similar way to BERs, e.g. a frame loss rate of 10^{-5} means that one frame in 100,000 will be lost.

4 Information and coding theory

Information theory is concerned with the description of information sources, the representation of the information from a source, and the transmission of this information over a channel.

4.1 Information sources and entropy

We start our examintaion of information theory by way of an example.

Consider a rain forecast in the Sahara desert. This forecast is an information source. The information source has two outcomes: rain or no-rain. Clearly, the outcome no-rain contains little information; it is a highly probable outcome. The outcome rain, however, contains considerable information; it is a highly improbable event.

In information theory, an **information source** is a **probability distribution**, i.e. a set of probabilities assigned to a set of outcomes. This reflects the fact that the information contained in an outcome is determined not only by the outcome, but by how uncertain it is. An almost certain outcome contains little information.

A measure of the information contained in an outcome was introduced by **Hartley** in 1927. He defined the **information** (sometimes called **self-information**) contained in an outcome x_i as:

$$I(x_i) = \log_2(\frac{1}{P\{x_i\}}) = -\log_2(P\{x_i\}) \tag{75}$$

This measure satisfies our requirement that the information contained in an outcome is proportional to its uncertainty. If $P\{x_i\} = 1$, then $I(x_i) = 0$, telling us that a certain event contains no information.

The definition 75 also satisfies the requirement that the total information in independent events should add. Clearly, our rain forecast for two days contains twice as much information as for one day. From equation 75, for two independent outcomes x_i and x_j:

$$\begin{aligned} I(x_i \text{ and } x_j) &= \log_2(\frac{1}{P\{x_i \text{ and } x_j\}}) \\ &= \log_2(\frac{1}{P\{x_j\}P\{x_j\}}) \\ &= \log_2(\frac{1}{P\{x_i\}}) + \log_2(\frac{1}{P\{x_j\}}) \\ &= I(x_i) + I(x_j) \end{aligned} \tag{76}$$

Hartley's measure defines the information in a single outcome. The measure **entropy** $H(X)$, (sometimes **absolute entropy**), defines the information content of the source X as a whole. It is the mean information provided by the source per source output or symbol. We have from equation 52:

$$H(X) = \sum_i P\{x_i\}I\{x_i\} = \sum_i -P\{x_i\}\log_2(P\{x_i\}) \tag{77}$$

A **binary symmetric source (BSS)** is a source with two outputs whose probabilities are p and $1-p$ respectively. The rain forecast discussed is a BSS. The entropy of the source is:

$$H(X) = -p\log_2(p) - (1-p)\log_2(1-p) \tag{78}$$

This function (Figure 57) takes the value zero when $p = 0$. When one outcome is certain, so is the other, and the entropy is zero. As p increases, so too does the entropy, until it reaches a maximum when $p = 1 - p = 0.5$. When p is greater than 0.5, the curve declines symmetrically to zero, reached when $p = 1$. We conclude that the average information in the BSS is maximised when both outcomes are equally likely. The entropy is measuring the average **uncertainty** of the source. (The term *entropy* is borrowed from thermodynamics. There too it is a measure of the uncertainty, or disorder of a system.)

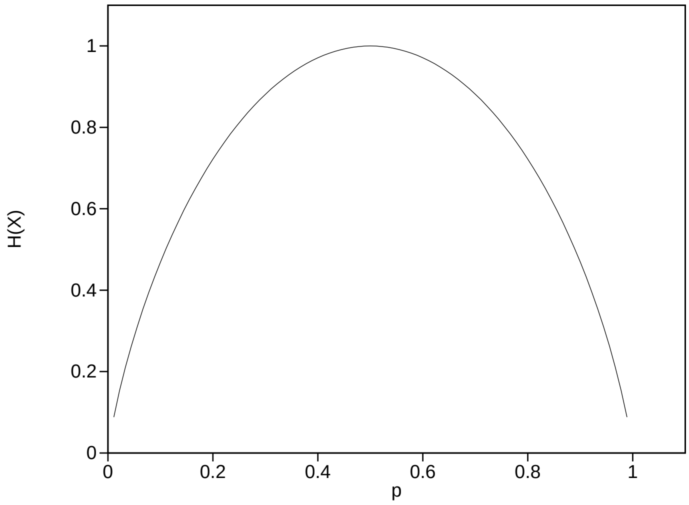


Figure 57: Entropy of the binary symmetric source

When $p = 0.5$, $H(X) = 1$. The unit of entropy is bits/symbol. An equally probable BSS has an entropy, or average information content per symbol, of 1 bit per symbol.

By long tradition, engineers have used the word *bit* to describe both the symbol, and its information content. A BSS whose outputs are 1 or 0 has an output we describe as a bit. The entropy of the source is also measured in bits, so that we might say the equi-probable ($p = 0.5$) BSS has an information rate of 1 bit/bit. The *numerator* bit refers to the information content. The *denominator* bit refers to the symbol 1 or 0. We can avoid this by writing it as 1 bit/symbol. When $p \neq 0.5$, the BSS information rate falls. When $p = 0.1$, $H(X) = 0.47$ bits/symbol. This means that, on average, each symbol (1 or 0) of source output is providing 0.47 bits of information.

For a BSS, the entropy is maximised when both outcomes are equally likely. This property is generally true. If an information source X has J symbols, its maximum entropy is $\log_2(J)$ and this is obtained when all J outcomes are equally likely. Thus, for a J symbol source:

$$0 \leq H(X) \leq \log_2(J) \tag{79}$$

4.2 Information source coding

It seems intuitively reasonable that an information source of entropy H needs on average only H binary bits to represent each symbol. Indeed, the equi-probable BSS generates on average 1 information bit per symbol bit. However, consider the rain forecast again. Suppose the probability of rain is 0.1 and that of no-rain 0.9. We have already noted that this source has an entropy of

0.47 bits/symbol. Suppose we identify rain with 1 and no-rain with zero. This representation uses 1 binary bit per symbol, and is using more binary bits per symbol than the entropy suggests is necessary.

The replacement of the symbols rain/no-rain with a binary representation is termed **source coding**. In any coding operation we replace the symbol with a **codeword**. The purpose of source coding is to reduce the number of bits required to convey the information provided by the information source.

Central to source coding is the use of **sequences**. By this, we mean that codewords are not simply associated to a single outcome, but to a sequence of outcomes. To see why this is useful, let us return to the problem of the rain forecast. Suppose we group the outcomes in threes, according to their probability, and assign binary codewords to these grouped outcomes. Table 7 shows such a code, and the probability of each codeword occurring. It is easy to compute that this code will on average use 1.2 bits/sequence. Each sequence contains three symbols, so the code uses 0.4 bits/symbol.

This example shows how using sequences permits us to decrease the average number of bits per symbol. Moreover, without difficulty, we have found a code that has an average bit usage less than the source entropy. However, there is a difficulty with the code in Table 7. Before a codeword can be decoded, it must be parsed. Parsing describes the activity of breaking the message string into its component codewords. After parsing, each codeword can be decoded into its symbol sequence. An **instantaneously parseable code** is one that can be parsed as soon as the last bit of a codeword is received. An instantaneous code must satisfy the **prefix condition**. That is, no codeword may be a prefix of any other codeword. This condition is not satisfied by the code in Table 7.

The code in Table 8, however, is an instantaneously parseable code. It satisfies the prefix condition.

Sequence	**Probability**	**Codeword**
NNN	0.729	0
NNR	0.081	1
NRN	0.081	01
RNN	0.081	10
RRN	0.009	11
RNR	0.009	00
NRR	0.009	000
RRR	0.001	111

Table 7: Variable length coding

Sequence	**Probability**	**Codeword**	**Letter**
NNN	0.729	1	A
NNR	0.081	011	B
NRN	0.081	010	C
RNN	0.081	001	D
RRN	0.009	00011	E
RNR	0.009	00010	F
NRR	0.009	00001	G
RRR	0.001	00000	H

Table 8: Instantly parseable variable length coding

The code in Table 8 uses on average 1.6 bits per sequence, or 0.53 bits/symbol. This is a 47%

improvement on identifying each symbol with a bit. In fact this is the **Huffman code** for the sequence set. The code for each sequence is found by generating the **Huffman code tree** for the sequence. A Huffman code tree is an unbalanced binary tree. The derivation of the Huffman code tree is shown in Figure 58 and the tree itself is shown in Figure 59. In both these figures, the letters A to H have been used in place of the sequences in Table 8 to make them easier to read.

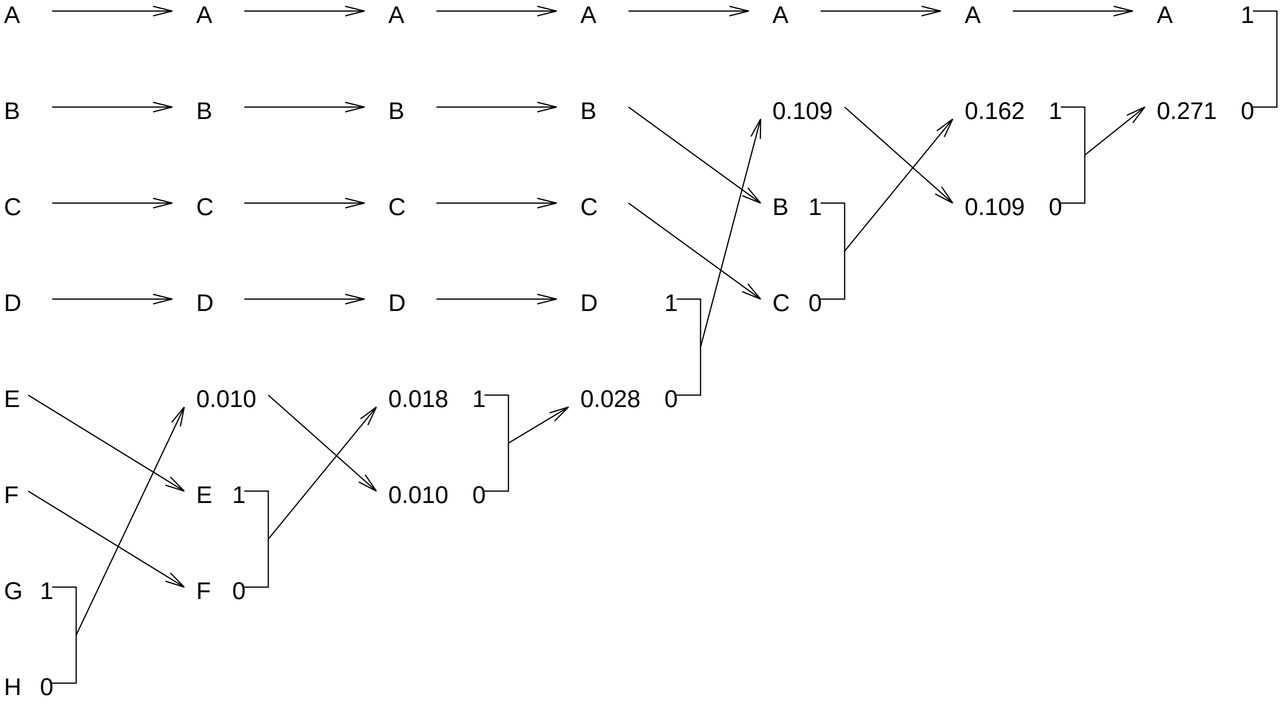


Figure 58: Example derivation of a Huffman code tree

In Figure 58, the sequences are ordered with respect to the probability of the sequence occurring, highest probability at the top of the list. (The probabilities are effectiveley used as **weights** in the process to be described.) The tree is derived bottom up, in terms of **branch nodes** and **leaf nodes** by combining weights and removing leaf nodes in progressive stages. As shown in Figure 58, the two lowest leaf nodes G and H have their weights added, and the topmost node is labelled with a 1 and the lower one with a 0. In the next stage the symbols G and H are represented by that weight, and the list is rewritten, again in order of the weights. The two lowest leaf nodes are now E and F, and they are labelled 1 and 0, respectiveley, and their weights are added to be taken on to the next stage. This continues until only two nodes remain. The Huffman tree shown in Figure 59 is then produced by following backwards along the arrows in Figure 58. To derived the codewords from the tree, descend from the top node (the root node), and list the 1s and 0s in the order they appear until you reach the leaf node for one of the letters.

Note that Huffman coding relies on the use of bit patterns of varialble length. In most data communication systems, the data symbols are encoded as bit patterns of a fixed length, e.g. 8-bits. This is done for technical simplicity. Often, coding schemes (such as Huffman coding) are used on a source symbol set to produce variable length bit length coding and refreed to as **compression algorithms**.

Huffman coding relies on the fact that both the transmitter and the receiver know the sequence set (or data set) before communicating and can build the code table. Where this is not possible **dynamic Huffman coding** can be used to build up the code table as the data is transmitted or received.

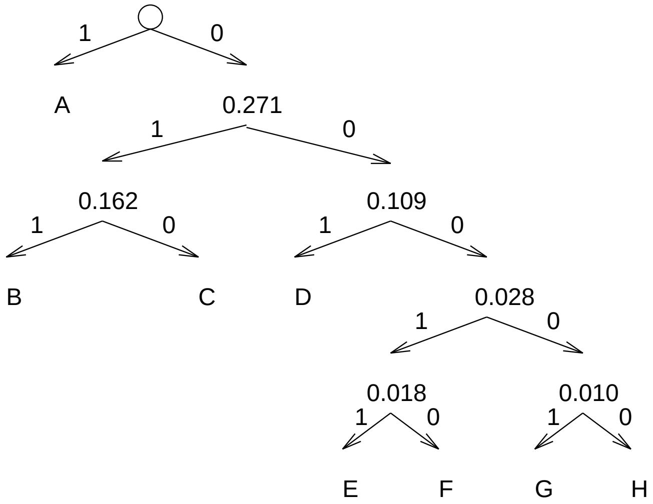


Figure 59: Example Huffman code tree

The **noiseless source coding theorem** (also called **Shannon's first theorem**) states that an instantaneous code can be found that encodes a source of entropy $H(X)$ with an average number of bits per symbol B_s such that:

$$B_s \geq H(X) \tag{80}$$

Ordinarily, the longer the sequences of symbols, the closer B_s will be to $H(X)$.

Like many theorems of information theory, the theorem tells us nothing of how to find the code. However, it is a useful result. For example, the code in Table 8 uses 0.53 bits/symbol which is only 11% more bits per symbol than the theorem tells us is the best we can do (0.47 bits/symbol). We might conclude that there is little point in expending the effort in finding a code closer to satisfying the inequality 80.

4.3 Channel coding; Hamming distance

The task of source coding is to represent the source information with the minimum of symbols. When a code is transmitted over a channel in the presence of noise, errors will occur. The task of channel coding is to represent the source information in a manner that minimises the error probability in decoding.

It is apparent that channel coding requires the use of redundancy. If all possible outputs of the channel correspond uniquely to a source input, there is no possibility of detecting errors in the transmission. To detect, and possibly correct errors, the channel code sequence must be longer than the source sequence. The rate R of a channel code is the average ratio of the source sequence length to the channel code length. Thus, $R < 1$.

A good channel code is designed so that, if a few errors occur in transmission, the output can still be identified with the correct input. This is possible because although incorrect, the output is sufficiently similar to the input to be recognisable. The idea of similarity is made more firm by the definition of a **Hamming distance**. Let x and y be two binary sequences of the same length. The Hamming distance between these two codes is the number of symbols that disagree. Suppose the code x is transmitted over the channel. Due to errors, y is received. The decoder will assign to y the code x that minimises the Hamming distance between x and y. For example, consider the codewords:

a	100000
b	011000
c	000111

If the transmitter sends `10000` but there is a single bit error and the receiver gets `10001`, it can be seen that the 'nearest' codeword is in fact 10000 and so the correct codeword is found.

It can be shown that to detect n bit errors, a coding scheme requires the use of codewords with a Hamming distance of at least $n+1$. It can also be shown that to correct n bit errors requires a coding scheme with at least a Hamming distnace of $2n+1$ between the codewords.

By designing a good code, we try to ensure that the Hamming distance between possible codewords x is larger than the Hamming distance arising from errors.

4.4 Channel capacity

One of the most famous of all results of information theory is **Shannon's channel coding theorem**. For a given channel there exists a code that will permit the error-free transmission across the channel at a rate R, provided $R \leq C$, the channel capacity. Equality is achieved only when the SNR is infinite.

As we have already noted, the astonishing part of this theory is the existence of a channel capacity. Shannon's theorem is both tantalising and frustrating. It offers error-free transmission, but it makes no statement as to what code is required. In fact, all we may deduce from the proof of the theorem is that it must be a long one. No one has yet found a code that permits the use of a channel at its capacity. However, Shannon has thrown down the gauntlet, in as much as he has proved that the code exists.

We shall not give a description of how the capacity is calculated. However, an example is instructive. The **binary channel (BC)** is a channel with a binary input and output. Associated with each output is a probability p that the output is correct, and a probability $(1-p)$ it is not. For such a channel, the channel capacity turns out to be:

$$C = 1 + p\log_2(p) + (1-p)\log_2(1-p) \tag{81}$$

Here (Figure 60), p is the bit error probability. If $p=0$ then $C=1$. If $p=0.5$ the $C=0$. Thus, if there is equal probability of receiving a 1 or 0, irrespective of the signal sent, the channel is completely unreliable and no message can be sent across it.

So defined, the channel capacity is a non-dimensional number. We normally quote the capacity as a rate, in bits/second. To do this we relate each output to a change in the signal. For a channel of bandwidth B, we can transmit at most $2B$ changes per second. Thus, the capacity in bits/second is $2BC$. For the binary channel, we have:

$$C = 2B(1 + p\log_2(p) + (1-p)\log_2(1-p)) \tag{82}$$

For the binary channel the maximum bit rate W is $2B$. We note that $C < W$, i.e. the capacity is always less than the bit rate. The data rate D, or information rate describes the rate of transfer of data bits across the channel. In theory, we have:

$$W \geq C \geq D \tag{83}$$

As a matter of practical fact:

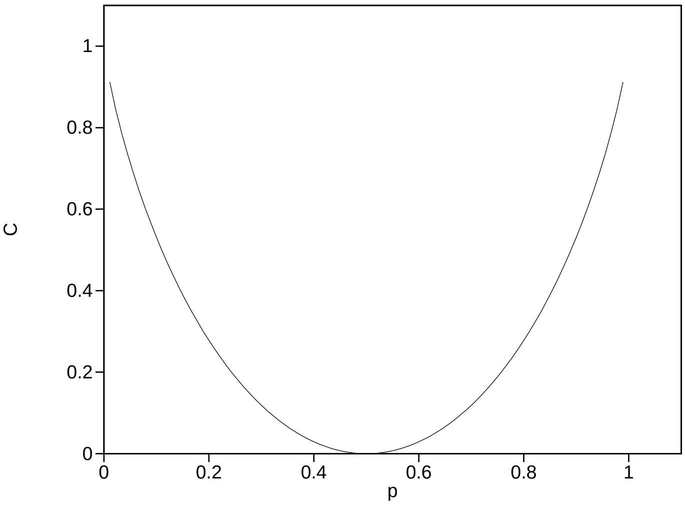


Figure 60: The capacity of a binary channel

$$W > C > D \tag{84}$$

Shannon's channel coding theorem applies to the channel, not to the source. If the source is optimally coded, we can rephrase the channel coding theorem: A source of information with entropy $H(X)$ can be transmitted error free over a channel provided $H(X) \leq C$.

All the modulations described earlier are binary channels. For equal BER, all these schemes have the same capacity. We have noted, however, that QPSK only uses half the bandwidth of PSK for the same bit rate. We might suppose that for the same bandwidth, QPSK would have twice the capacity. This is not so. We have noted that PSK modulation is far from optimum in terms of bandwidth use. QPSK makes better use of the bandwidth. The increase in bit rate provided by QPSK does not reflect an increase in capacity; merely a better use of bandwidth.

The capacity of the binary channel is much less than that calculated from the Hartley-Shannon Law (equation 8). Why so? The answer is that equation 8 applies to systems whose outputs may take any values. We use systems obeying equation 82 because they are technically convenient, not because they are desirable.

4.5 Error detection coding

The theoretical limitations of coding are placed by the results of information theory. These results are frustrating in that they offer little clue as to how the coding should be performed. Errors occur – they must, at the very least, be detected.

Error detection coding is designed to permit the detection of errors. Once detected, the receiver may ask for a re-transmission of the erroneous bits, or it may simply inform the recipient that the transmission was corrupted. In a binary channel, error checking codes are called **parity check** codes.

Practical codes are normally block codes. A block code converts a fixed length of K data bits to a fixed length N codeword, where $N > K$. The rate of the code is the ratio K/N, and the redundancy of the code is $1 - (K/N)$. Our ability to detect errors depends on the rate. A low rate has a high detection probability, but a high redundancy.

The receiver will assign to the received codeword the preassigned codeword that minimises the

Hamming distance between the two words. If we wish to identify any pattern of n or less errors, the Hamming distance between the preassigned codewords must be $n + 1$ or greater.

A very common code is the **single parity check code**. This code appends to each K data bits an additional bit whose value is taken to make the $K + 1$ word even (or odd). Such a choice is said to have even (odd) parity. With even (odd) parity, a single bit error will make the received word odd (even). The preassigned code words are always even (odd), and hence are separated by a Hamming distance of 2 or more.

To see how the addition of a parity bit can improve error performance, consider the following example. A common choice of K is eight. Suppose the BER is $p = 10^{-4}$. Then:

$$\begin{aligned}
P\{\text{single bit error}\} &= p \\
P\{\text{no error in single bit}\} &= (1 - p) \\
P\{\text{no error in 8 bits}\} &= (1 - p)^8 \\
P\{\text{unseen error in 8 bits}\} &= 1 - (1 - p)^8 \\
&= 7.9 \times 10^{-4}
\end{aligned}$$

So, the probability of a transmission with an error is 7.9×10^{-4}. With the addition of a parity error bit we can detect any single bit error. So:

$$\begin{aligned}
P\{\text{no error in single bit}\} &= (1 - p) \\
P\{\text{no error in 9 bits}\} &= (1 - p)^9 \\
P\{\text{single error in 9 bits}\} &= 9(P\{\text{single bit error}\}P\{\text{no error in other 8 bits}\}) \\
&= 9p(1 - p)^8 \\
P\{\text{unseen error in 9 bits}\} &= 1 - P\{\text{no error in 9 bits}\} - P\{\text{single error in 9 bits}\} \\
&= 1 - (1 - p)^9 - 9p(1 - p)^8 \\
&= 3.6 \times 10^{-7}
\end{aligned}$$

As can be seen, the addition of a parity bit has reduced the uncorrected error rate by three orders of magnitude.

Single parity bits are common in asynchronous, character oriented transmission. Where synchronous transmission is used, additional parity symbols are added that checks not only the parity of each 8 bit **row**, but also the parity of each 8 bit **column**. The column is formed by listing each successive 8 bit word one beneath the other. This type of parity checking is called **block sum checking**, and it can correct any single 2 bit error in the transmitted block of rows and columns. However, there are some combinations of errors that will go undetected in such a scheme (Figure 61)

Parity checking in this way provides good protection against single and multiple bit errors when the probability of the errors are independent. However, in many circumstances, errors occur in groups, or bursts. Parity checking of the kind just described then provides little protection. In these circumstances, a polynomial code is used.

The mechanism of **polynomial codes** is beyond the scope of this course. We shall content ourselves with a description. Polynomial codes work on each frame. Additional digits are added to the end of each frame. These digits depend on the contents of the frame. The number of added digits depends on the length of the expected error burst. Typically 16 or 32 digits are added. The computed digits are called the **frame check sequence (FCS)** or **cyclic redundancy check (CRC)**. Before transmission, each frame is divided by a generator polynomial. The remainder of

	P1	B6	B5	B4	B3	B2	B1	B0
	0	0	0	0	0	0	0	0
	1	0	1	0	1	0	0	0
	0	1	0*	0	0	1*	1	0
	0	0	1	0	0	0	0	0
	1	0	1	0	1	1	0	1
	0	1	0	0	0	0	0	0
	1	1	1*	0	0	0*	1	1
	1	0	0	0	0	0	1	1
P2	1	1	0	0	0	0	0	1

```
P1 is odd parity for rows
P2 is even parity for columns
* mark undetected error combination
```

Figure 61: Example of block sum check showing undetected errors

this division is added to the frame. On reception, the division is repeated. Since the remainder has been added, the result should be zero. A non-zero result indicates that an error has occurred.

A polynomial code can detect any error burst of length less than or equal to the length of the generator polynomial. The technique requires the addition of hardware to perform the division. However, with modern integrated circuitry, this hardware is now available inexpensively. CRC error checking is now quite common, and its use will increase.

4.6 Error correction coding

Error correction coding is more sophisticated than error detection coding. Its aim is to detect and locate errors in transmission. Once located, the correction is trivial: the bit is inverted. Error correction coding requires lower rate codes than error detection, often markedly so. It is therefore uncommon in terrestrial communication, where better performance is usually obtained with error detection and retransmission. However, in satellite communications, the propagation delay often means that many frames are transmitted before an instruction to retransmit is received. This can make the task of data handling very complex. Real-time transmission often precludes retransmission. It is necessary to get it right first time. In these special circumstances, the additional bandwidth required for the redundant check-bits is an acceptable price. There are two principle types: **Hamming codes** and **convolutional codes**.

A Hamming code is a block code capable of identifying and correcting any single bit error occurring within the block. It is identified by the numbers K and N; we talk of an (N, K) Hamming code. Hamming codes employ modulo 2 arithmetic. Addition in modulo 2 arithmetic is replaced by the **exclusive OR** (often written **XOR** or **EOR**) logic operation. This operation has a truth-table shown in Table 9.

A	B	A $\oplus$ B
0	0	0
0	1	1
1	0	1
1	1	0

Table 9: Truth Table for exclusive OR

The theory of Hamming codes is beyond the scope of our course. An example of a (11, 7) code is given.

Consider a Hamming code to detect and correct for single-bit errors assuming each codeword contains a seven-bit data field, e.g. an ASCII character. for example. Such a coding scheme requires four check bits since, with this scheme, the check bits occupy all bit positions that are powers of 2. Such a code is thus known as an (11, 7) block code with a rate of 7/11 and a redundancy of $1 - 7/11$. For example, the bit positions of the value 1001101 are:

```
Bit Position   11  10  9  8  7  6  5  4  3  2  1
Bit value       1   0  0  x  1  1  0  x  1  x  x
```

The four bit positions marked with `x` are used for the check bits, which are derived as follows. The four-bit binary numbers corresponding to those bit positions having a binary 1 are added together using modulo 2 arithmetic and the four check bits are then the four-bit sum:

```
11  =  1011
 7  =  0111
 6  =  0110
 3  =  0011
       1001
```

The transmitted codeword is thus:

```
Bit Position   11  10  9  8  7  6  5  4  3  2  1
Bit value       1   0  0  1  1  1  0  0  1  0  1
```

Similarly, at the receiver, the four-bit binary numbers corresponding to those bit positions having a binary 1, including the check bits, are again added together and, if no errors have occurred, the modulo 2 sum should be zero:

```
11  =  1101
 8  =  1100
 7  =  0111
 6  =  0110
 3  =  0011
 1  =  0001
       0000
```

Now consider a single-bit error; say bit 11 is corrupted from 1 to 0. The new modulo 2 sum would now be:

```
 8  =  1100
 7  =  0111
 6  =  0110
 3  =  0011
 1  =  0001
       1011
```

Firstly, the sum is non-zero, which indicates an error, and secondly the modulo 2 sum, equivalent to decimal 11, indicates that bit 11 is the erroneous bit. The latter would therefore be inverted to obtain the corrected codeword and hence data bits.

Hamming codes suffer from the same difficulty as block-codes. They offer protection against single-bit errors. They offer little protection against burst errors. Convolutional codes are designed to deal with this circumstances. Convolutional codes are different from previous codes we have examined in that they work in a statistical sense. By this we mean that we cannot say that, for example, every single-bit error will be corrected. We can only say that, on average, the use of the convolutional code will improve the error rate. We shall not examine these codes in detail. They are widely used. Their use can typically provide an error rate improvement of 3 orders of magnitude, with a code rate of 0.5.

4.7 Encryption

In all our discussion of coding, we have not mentioned what is popularly supposed to be the purpose of coding: security. We have only considered coding as a mechanism for improving the integrity of the communication system in the presence of noise. The use of coding for security has a different name: encryption. The use of digital computers has made highly secure communication a normal occurrence. The basis for key based encryption is that it is very much easier to encrypt with knowledge of the key than it is to decipher without knowledge of the key. The principle is just that of a combination lock. With a computer, the number of the digits in the lock can be very large. Of course, one still has to keep the combination secure!

The most commonly used encryptions algorithms are **block ciphers**. This means that the algorithm splits the **plaintext** (message to be encrypted) into (usually) fixed size blocks which are then subjected to various functions to produce a block of **ciphertext**. The most common functions are **permutations** based on **expansion** and **compression** and **straight** 'shuffling' transformations. In a straight permutation, the bits of an n bit block are simply reordered. In expansion, as well as being reordered, the group of n bits is converted to m bits (m > n), with some bits being duplicated. In compression, the n bit block in converted to a p bit block (p < n), with some of the original bits unused (Figure 62).

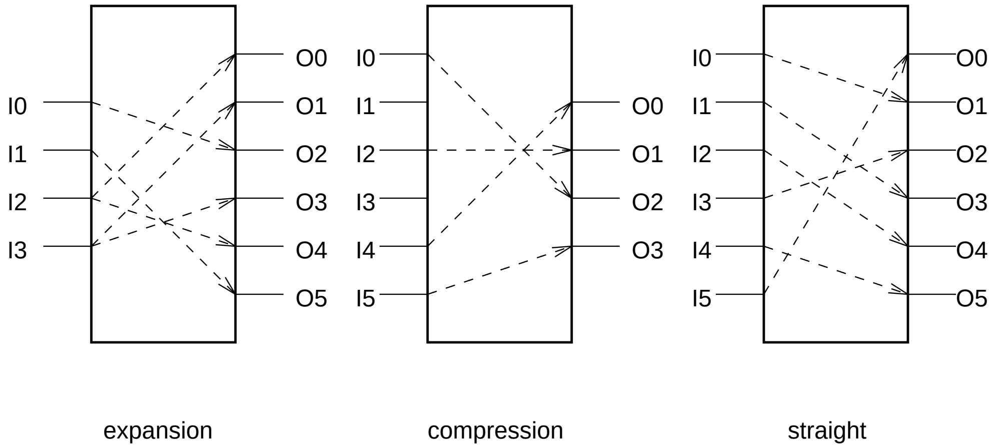


Figure 62: Examples of block cipher permutations

The most widely used form of encyrption is defined by the National Bureau of Standards and is known as the **data encryption standard (DES)**. The DES is a block cipher, splitting the data stream into 64-bit blocks that are enciphered separately. A (probably) unique key of 56 bits is then used to perform a succession of transposition and substitution operations. A 56 bit key has 7.2×10^{16} possible combinations. Assuming a powerful computer could attempt 10^8 combinations per second, it would still take over 20 years to break the code. If the code is changed once per year, there is little possibility of it being broken, unless the code breaker has additional information. The DES converts 64 bits of plaintext into 64 bits of ciphertext. The receiver uses the same key to decipher the ciphertext into plaintext.

The difficulty with this method is that each block is independent. This permits an interceptor in possession of the key to introduce additional blocks without the recipient being aware of this fact. In addition, the same plaintext will generate the same ciphertext, a fact of great value to someone attempting to break the code. These disadvantages are removed by chaining. **Chaining** describes the process of EOR-ing the plaintext of one block with the ciphertext of the previous block. In this way it is not possible to introduce blocks in a transparent fashion, and repetitions of the same plaintext generates different ciphertexts.

Like the combination of a lock, the system is only secure if the key is secure. If the key is changed often, the security of the key becomes a problem, because the transfer of the key between sender and receiver may not be secure. This is avoided by the use of matched keys. In a matched key scheme, the encryption is not reversible with the same key. The message is encrypted using one key, and decrypted with a second, matched key. The receiver makes available the first, public key. This key is used by the sender to encrypt the message. This message is unintelligible to anyone not in possession of the second, private key. In this way the private key need not be transferred. The most famous of such schemes is the **Public Key** mechanism using the work of **Rivest, Shamir and Adleman (RSA)**. It is based on the use of multiplying extremely large numbers and, with current technology, is computationally very expensive.

5 Broadband ISDN

The original specifications for the **intergrated services digital network (ISDN)**, were based around voice and non-voice telephone-type services: telephony, data, telex, facsimile, as it was hoped that the ISDN would evolve from the (then) emerging digital telephone networks. Indeed, this is one of the reasons that the fundamental element of an ISDN link is the 64Kb/s **B-Channel**. However, the planning for ISDN was started around 1976, and as technology evolved, so did the requirements of the users that wanted to use this technology. In 1988, the CCITT released a document that described a new set of **Broadband ISDN (B-ISDN)** services. To distinguish this new concept from the original ISDN service, we now refer to the latter as **Narrowband ISDN (N-ISDN)**.

Since then the CCITT has become the ITU. The B-ISDN work is far from complete, and some of the factors influencing the emergence of the B-ISDN from the ITU are:

- **Demand.** Users (both commercial and residential) are showing interest in receiving high speed, reliable services.
- **Technology.** This has been one of the biggest factors. Advances in technology have increased demand, as well as the ability to supply it. As data processing technologies available to the user have become more sophisticated so has his/her demands, while high speed transmission (based on the use of optical fibre), high speed switching and increased processing power make the realisation of these demands possible.
- **Service integration.** There is a need to integrate both circuit switched and packet switched services into one network that can provide interactive and distribution services.
- **Flexibility.** The resulting network must be able to satisfy the needs of the wide variety of users as well as the network operators in terms of its functionality and usability.

5.1 Broadband ISDN services

The need for a Broadband ISDN service sprung from the growing needs of the customers. The planned Broadband ISDN services can broadly be categorised as follows:

- **Interactive services.** These are services allowing information flow between two end users of the network, or between the user and a service provider. Such services can be subdivided:
  - **Conversational services.** These are basically end-to-end, real-time communications, between users or between a user and a service provider, e.g. telephone-like services. Indeed, B-ISDN will support N-ISDN type services. Also the additional bandwidth offered will allow such services as video telephony, video conferencing and high volume, high speed data transfer.
  - **Messaging services.** This differs from conversational services in that it is mainly a store-and-forward type of service. Applications could include voice and video mail, as well as multi-media mail and traditional electronic mail.
  - **Retrieval services.** This service provides access to (public) information stores, and information is sent to the user on demand only. This includes applications such as teleshopping, videotex services, still and moving pictures, telesoftware and entertainment.
- **Distribution services.** These are mainly broadcast services, intended for one way interaction from a service provider to a user or many users:

- **No user control of presentation.** This would be for instance, a TV broadcast, where the user can choose simply either to view or not. It is expected that cable TV companies will become interested in Broadband ISDN as a carrier for normal TV services as well as the high definition TV (HDTV) services that are forseen for the future, and **near video-on-demand (nVoD)** services.
- **User controlled presentation.** This would apply to broadcast information that the user can partially control, in that the user can decide which part of it he/she accesses, e.g. teletext and news retrieval services and full **video-on-demand (VoD)** services.

(Note also that the user-to-user signaling, user-to-network signaling, and inter-exchange (network-to-network) signaling are also provided but outside our scope.)

However, many of these services have very high throughput requirements, as shown in Table 10. The burstiness is the ratio of the peak bit rate to average bit rate.

Service	**Bit Rate [Mb/s]**	**Burstiness**
Data	1.5 to 130	1 – 50
Document transfer	1.5 to 45	1 – 20
Videoconference or videotelephony	1.5 to 130	1 – 5
Broadband video (nVoD and VoD)	1.5 to 130	1 –40
TV	30 to 130	1
HDTV	130	1

Table 10: Some Broadband service throughput needs

It is clear that high network capacity is required if this kind of service is to be offered to many users simultaneously. The N-ISDN can currently offer interfaces which aggregate B-Channels to give additional throughput, as shown in Table 11 and Table 12. However, these are not sufficient for our Broadband service requirements.

Channel	**Bit Rate [Kb/s]**	**Interface**
B	64	Basic rate
H0	384	Primary rate
H11	1536	Primary rate
H12	1920	Primary rate
D16	16	Basic rate
D64	64	Primary rate

Table 11: Narrowband ISDN channels

5.2 Network architecture

The B-ISDN needs to provide:

- Broadband services, as described in the last subsection.
- Narrowband services (for backwards compatibility).
- User-to-network signaling, to allow the B-ISDN user to initiate and control communication.
- Inter-exchange (network-to-network) signaling within the network, to allow the network to provide and control resources as requested by the B-ISDN user or by another network exchange.

- User-to-user signaling, to allow B-ISDN users to send control, operation and maintenance information to each other.
- Management facilities for controlling and operating the network.

It is intended that the B-ISDN will offer both **connection oriented (CO)** and **connectionless (CL)** services, however, the CO mode of operation is receiving the greatest attention at the moment, while CL service definitions mature. The broadband information transfer is provided by the use of **asynchronous transfer mode (ATM)**, in both cases, using end-to-end logical connections. ATM makes use of small, fixed size **cells** (53 octets) in which the information is transferred, along the logical connections. Each cell contains a header (5 octets) which is used to identify the logical connection to which the cell belongs.

Each logical connection is accessed as a **virtual channel (VC)**. User-to-user data VCs are unidirectional. Many VCs may be used to a single destination and they may be associated by use of a **virtual path (VP)**. There relationship between VCs and VPs with respect to the **transmission path** is depicted in Figure 63. The transmission path is the logical connection between the two end-points, and consists in reality of many **links** between network exchanges and switches.

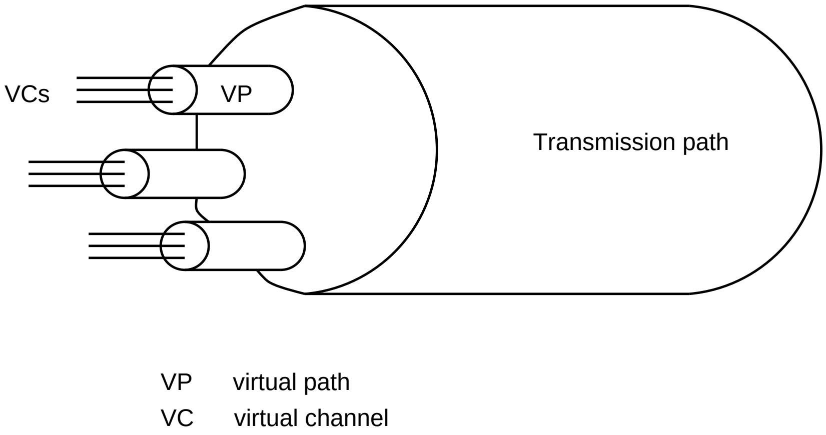


Figure 63: Transmission path model for Broadband ISDN

The VCs are identified at each end of the connection by a **virtual channel identifier (VCI)**. Similarly, the VP is identified by a **virtual path identifier (VPI)**. The VCI/VPI pair uniquely identifies a user-to-user information flow and is carried in each ATM cell header. Both VCIs and VPIs in general have only local significance. The concept of the **link** can be applied to both VCs and VPs in explaining the use of VCIs and VPIs, and we can say that the VCI/VPI pair identifies

Interface	**Bit Rate [Kb/s]**	**Structure**
Basic rate access	144	2B + D
Primary rate access	1544	23B + D64
		3H0 + D64
		H11
		etc
Primary rate access	2048	30B + D64
		5H0 + D64
		H12 + D64
		etc

Table 12: Narrowband ISDN interfaces

a particular link. VCIs and VPIs are used within the network for switching purposes, with **virtual channel links** and **virtual path links** being defined as the connection between two points where the either the VC or the VP is **switched**, respectively, i.e. the link is defined to exist between the two points where the VCI or VPI value is removed or translated (switched). There will be many virtual channel links comprising a **virtual channel connection (VCC)** and, similarly, many virtual path links in a **virtual path connection (VPC)**. This relationship is shown in Figure 64.

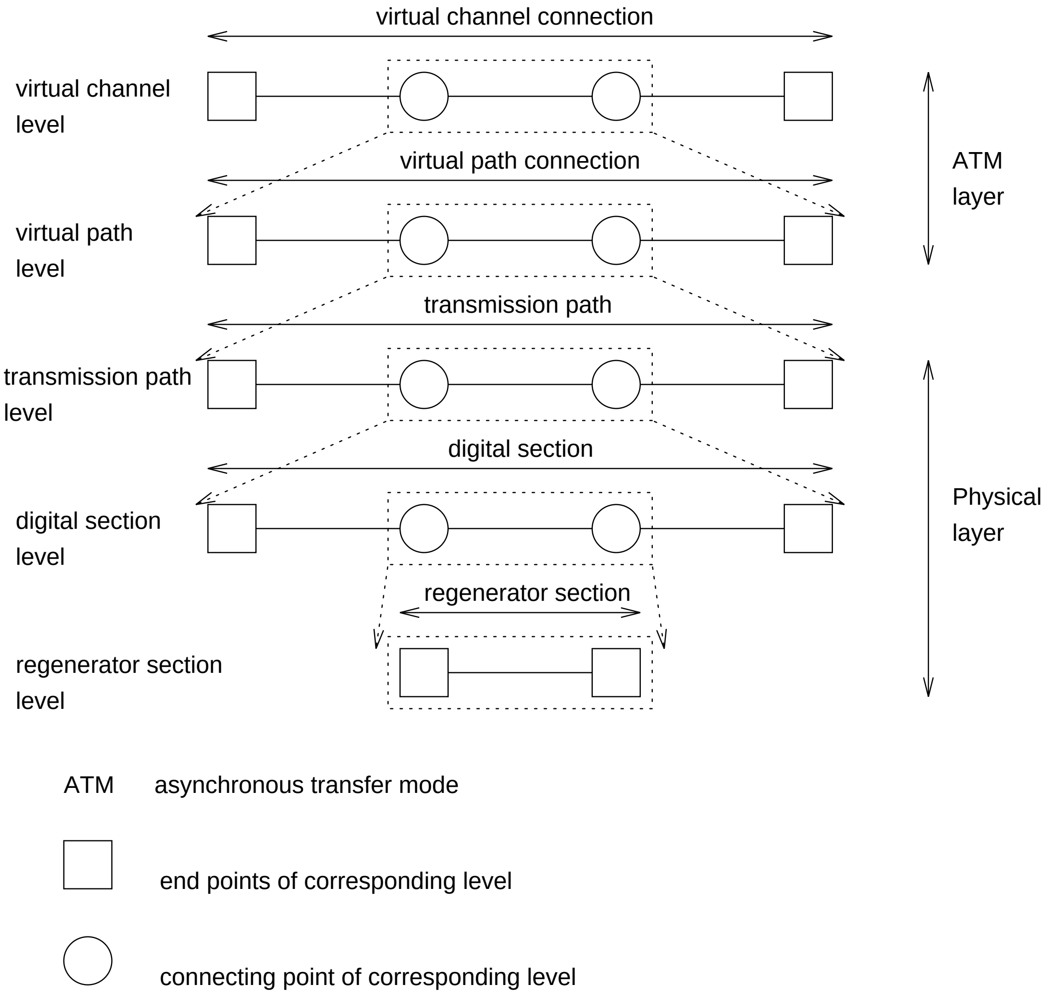


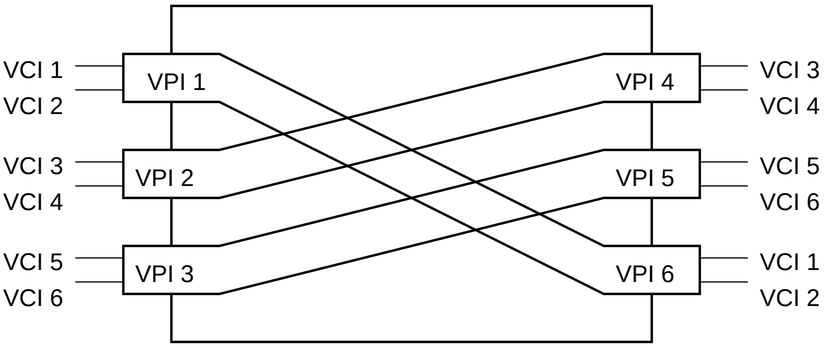
Figure 64: Hierarchical layer to layer relationship in the ATM layer and Physical layer

The **transmission path** consists of **digital sections** that are responsible for assembling bit/octet streams. In turn, each digital section may consist of **regenerator sections** which are sections between two regenerators (repeaters) on a physical link.

The switching of the VCs and VPs is depicted in Figure 65 and Figure 66.

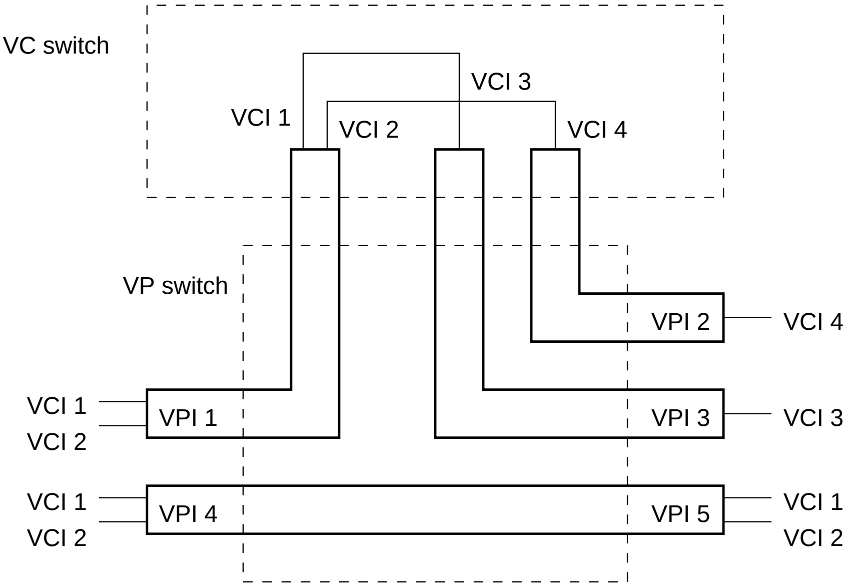
Figure 65 depicts a scenario which may result in the same VCI being used to identify the VCC at both ends of the transmission path. In this case the VC link may be the same as as the VC connection. However, in general, both the VCI and VPI values may be translated in a switch as shown in Figure 66.

The setting up of VCCs and VPCs between B-ISDN users may be done in a number of ways. For VCCs, one of the following procedures can be used:



VCI	virtual channel identifier
VPI	virtula path identifier

Figure 65: Virtual path switch



VC	virtual channel	VP	virtual path
VCI	virtual channel identifier	VPI	virtual path identifier

Figure 66: Virtual path and virtual circuit switch

- A semi-permanent or permanent VCC is established by the network provider as the result of a subscription procedure. This requires no signaling, and the VCCs behave like leased lines.
- A switched VCC is established and/or release by the use of user-to-network signaling procedure.
- If a VPC already exists between two B-ISDN users, then they may employ user-to-user signaling procedures on one of the the existing VCCs to establish and/or release other VCCs.
- A signaling VCC is established by the use of a **meta-signaling** procedure, using a special meta-signaling VCC.

The cell sequence integrity is guaranteed on any VCC. The QoS for a VCC will have been negotiated at subscription, but may be renegotiated at VCC establishment.

For VPCs, one of the following procedures can be used:

- A VPC is established by the network provider as the result of a subscription procedure. This requires no signaling.
- A VPC may be established by the B-ISDN user employing network management procedures.
- A VPC may be established by the network itself using network management procedures.

Again, cell sequence integrity is guaranteed for each VCC in the VPC but also for the VPC as a whole. This is an improvement over N-ISDN where the separate logical channels (B-Channels) must be used; amongst other things use of several B-Channels may result in skew in the data sent from one user to another. The QoS required for the VPC can be selected from a set of known classes that have been negotiated with the network provider. Some of the VCIs that could be used for the VPC may be reserved for special purposes, such as **operation and management (OAM)** activities.

5.3 Signaling

Broadband ISDN uses out-of-band signaling (as does N-ISDN). Instead of using a D Channel as in N-ISDN, a special VCC can be used for signaling. This also means that the B-ISDN user can have a much more flexible and powerful signaling capability, as further VCCs for signaling can be established as required. As the B-ISDN is meant to backwards compatible with N-ISDN applications, the B-ISDN also provides N-ISDN signaling. Also, the B-ISDN services bring their own requirements for more powerful signaling:

- Establish, release and maintain VCCs as required.
- Negotiate and/or renegotiate the QoS for VCCs.
- Allow mutli-connection calls, i.e. composite calls that require several different types of information flow to the same destination. For instance, a multimedia call that carries voice, video and data may use one VC each for the voice, video and data, each with its own QoS requirements and synchronisation requirements.
- Allow multi-party calls, i.e. calls between more than two end-points. For instance a conference call. This type of call requires facilities to allow users to leave or join the call/conference. If the call is a multi-media conference, then the multi-connection signaling facility will also be required and should operate in harmony.

SVC type	**Directionality**	**SVCs per user interface**
Meta-signaling	Bidirectional	1
General broadcast	Unidirectional	1
Selective broadcast	Unidirectional	Several possible
Point-to-point	Bidirectional	1 per signaling endpoint

Table 13: SVCs at the B-ISDN user-to-network interface

B-ISDN uses dedicated **signaling virtual channels (SVCs)**. There are four types as shown in Table 13.

There is only one **meta-signaling channel** per interface and it is permanent. It is used to set up instances of the point-to-point SVC as required. It is identified by the use of a specially allocated VCI/VPI.

The broadcast SVCs are directed from network to user. The **general broadcast SVC** is permanent and provides signaling information to all network users. The **selective broadcast SVC** is optionally set up by the network provider to provide additional signaling services to only certain terminals.

The **point-to-point SVC** exists only while it is required and is used to establish, maintain, control and release data VCCs between to end points.

5.4 Protocol Reference Model

The network is described in terms of a **protocol reference model (PRM)** (Figure 67). Not all of the PRM is fully defined. The main aspects of the model are that it can be viewed in terms of the three **planes** – the **user plane**, the **control plane** and and **management plane** – and in terms of the 3 **layers** – the **ATM adaptation layer**, the **ATM layer** and the **Physical layer**.

The functions of the layers are as follows:

- **ATM adaptation layer (AAL).** This layer is responsible for mapping the service offered by ATM to the service expected by the higher layers. It has two sublayers.
  - **Convergence sublayer (CS).** Responsible for presenting the ATM service to the higher layers. The functionality of this sublayer is very much dependent on the higher layer service.
  - **Segmentation and reassembly (SAR).** This layer is responsible for, at the transmitter, splitting the higher level PDU into 48 octet chunks, and at the receiving side, to reassemble the 48 octet chunks back into the original PDU.
- **ATM Layer.** This layer is independent of the physical medium over which transmission is to take place. It has four functions:
  - **Generic flow control (GFC) function.** This can be used to alleviate short term overload conditions above the ATM layer, as it is accessible by the user.
  - **Cell header generation and extraction.** At the transmitter, adds header information to a cell, and at the receiver removes it.
  - **Cell VPI/VCI translation.** Switching of VCs and VPs.
  - **Cell multiplex and demultiplex.** At the transmitter, multiplex cells into one continuous stream, and at the receiver demultiplex the cells according to VPI and VCI values.

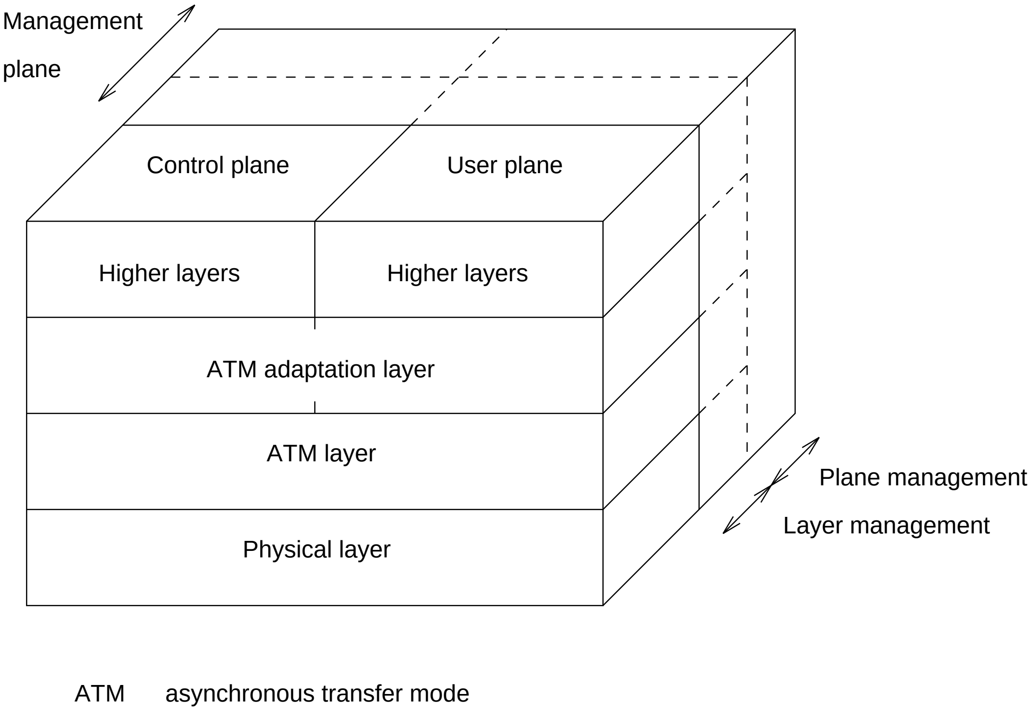


Figure 67: Broadband ISDN protocol reference model

- **Physical layer.** This consists of two sublayers:
  - **Transport Convergence (TC).** This sublayer has five functions:
    - **Cell rate decoupling.** Insertion and extraction of idle cells.
    - **Header error control (HEC) generation and verification.** In the transmitter, generation of the HEC, and in the receiver checking of the HEC. The HEC that is used can detect and correct a 1 bit error and can further detect certain multiple bit errors.
    - **Cell delineation.** In the receiver, detection of cell boundaries.
    - **Transmission frame adaptation.** Adapts cell flow according to the payload of the Physical level frame being used, e.g. for SDH.
    - **Transmission frame generation and recovery.** At the transmitter, generates Physical level frames, and at the receiver, extracts the ATM cells from the Physical level frame.
  - **Physical medium (PM).** This contains two sublayers:
    - **Bit timing.** Insertion and extraction of bit timing information and generation and reception of waveforms.
    - **Physical medium.** Bit transmission, bit alignment and optical ↔ electrical conversion, if required. (The physical medium need not be optical, at least for transmission rates of 155Mb/s and lower.)

This (sub-)layering of the PRM is depicted in Figure 68.

The **management plane** consists of two functions to perform **layer management** and **plane management**. The plane management is not layered as the other layers are. This is because it needs information on all aspects of the system to provide management facilities for the system as a whole. The **layer management** provides information and control facilities for the protocol

FUNCTION	(SUB) LAYER	
Higher layer functions	Higher layers	
Convergence	CS	AAL
Segmentation and reassembly	SAR	
Generic flow control Cell header generation/extraction Cell VPI/VCI translation Cell multiplex and demultiplex		ATM
Cell rate decoupling HEC sequence generation/verification Cell delineation Trsnamission frame adaption Transmission frame generation/recovery	TC	Physical layer
Bit timing Physical medium	PM	

ATM	asynchronous transfer mode	VPI	virtual path identifier
AAL	ATM adaptation layer	VCI	virtual channel identifier
CS	convergence sublayer	TC	transmission convergence
SAR	segmentation and reassembly	PM	physical medium
HEC	header error control		

Figure 68: B-ISDN layer functionality

entities that exists in each individual layer. This includes **operation and maintenance (OAM)** functions for each layer.

The **control plane** is responsible for the supervision of connections, including call set-up, call release and maintenance.

The **user plane** provides for the transfer of user information. It also includes mechanisms to perform error recovery, flow control etc.

Broadband ISDN intends to offer many Mb/s to the user, but intends to remaim backwards compatible with N-ISDN. Indeed, the Narrowband services will eventually need to be offered over the future global Broadband network to come. To this extent the user interface to B-ISDN is very similar to that for N-ISDN. Figure 69 shows the position of the **user to network interface (UNI)**, as well as the internal **network to network interface (NNI)** for B-ISDN. We will concentrate our attention mainly on the UNI. The UNI is accessed at the reference point T_B or S_B (Figure 70).

Note that in Figure 70, it is expected that N-ISDN (or even other PSTN) equipment will be able to connect to the Broadband network via a suitable **terminal adaptor (TA)**. The various functional groups are now described:

- **B-NT1.** This group contains functions that are considered to be part of OSI layer 1. It represents the physical connection point to the network, i.e. the socket on the wall. It includes functions such as:
  - **Line transmission termination.** Provision of the physical connection.
  - **Transmission interface handling.** The interface to the transmission channel, be it electrical or optical.
  - **Operation and Maintenance (OAM).** This is not normally associated with the socket in the wall. However, it is expected that for B-ISDN, more sophisticated management capabilities will be required than at present.
- **B-NT2.** This group contains OSI layer 1 and higher OSI layer functions:
  - **Adaptation functions.** For different physical media and network topologies.
  - **Multiplexing and demultiplexing.** The user data may be sent and received on several VCCs and VPCs.
  - **Buffering.** User data may be sent and/or received at varying rates with respect to the B-ISDN user and the network.
  - **Signaling.** VCCs/VPCs must be established, controlled and released.
  - **Interface.** Interaction with the B-ISDN user.
- **B-TE1.** Equipment requiring B-ISDN access.
- **B-TA.** Equipment allowing connection of other B-ISDN, N-ISDN and non-ISDN equipment.
- **B-TE2.** B-ISDN with special interface needs or N-ISDN equipment.
- **TE2.** Non-ISDN equipment.

Note that these are logical units. The physical implementation may be quite different. For instance, it may be common to find the following in the same physical unit, depending on need: B-NT1 and B-NT2; B-TE1 and B-NT2; B-TA and B-TE2 etc. Further, the way in which the terminal equipment is connected to the user-to-network interface via B-NT1/B-NT2 is not restricted with respect to local topologies (Figure 71).

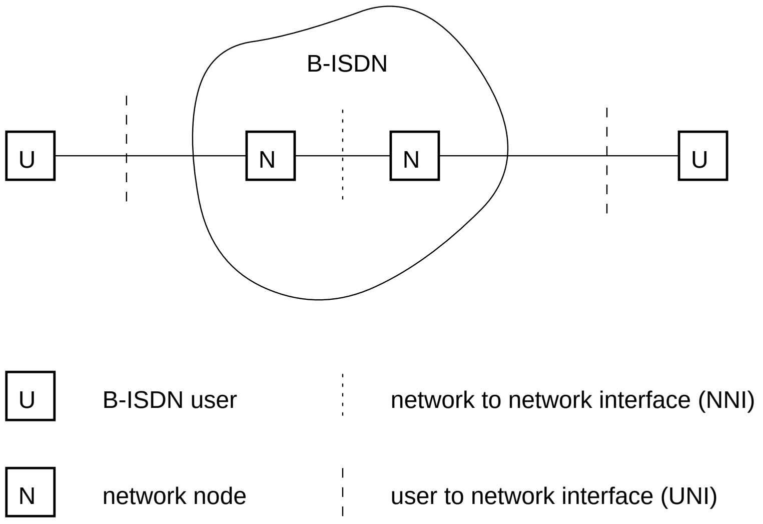

B-ISDN
U
N
N
U
U B-ISDN user
network to network interface (NNI)
N network node
user to network interface (UNI)

Figure 69: B-ISDN user and network interfaces

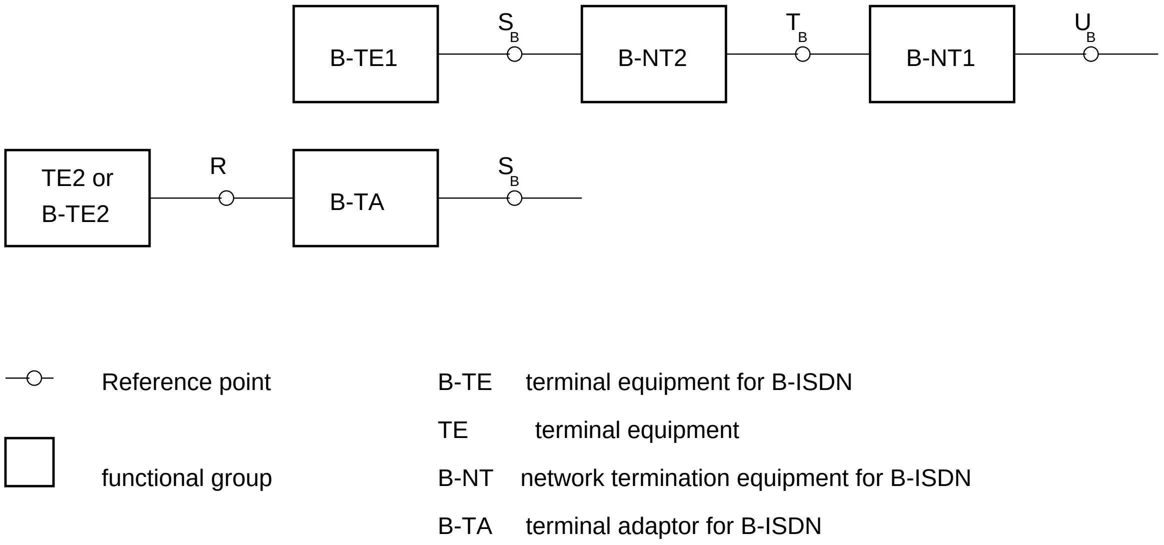

B-TE1
S_B
B-NT2
T_B
B-NT1
U_B
TE2 or B-TE2
R
B-TA
S_B
Reference point
functional group
B-TE terminal equipment for B-ISDN
TE terminal equipment
B-NT network termination equipment for B-ISDN
B-TA terminal adaptor for B-ISDN

Figure 70: B-ISDN UNI configuration reference points

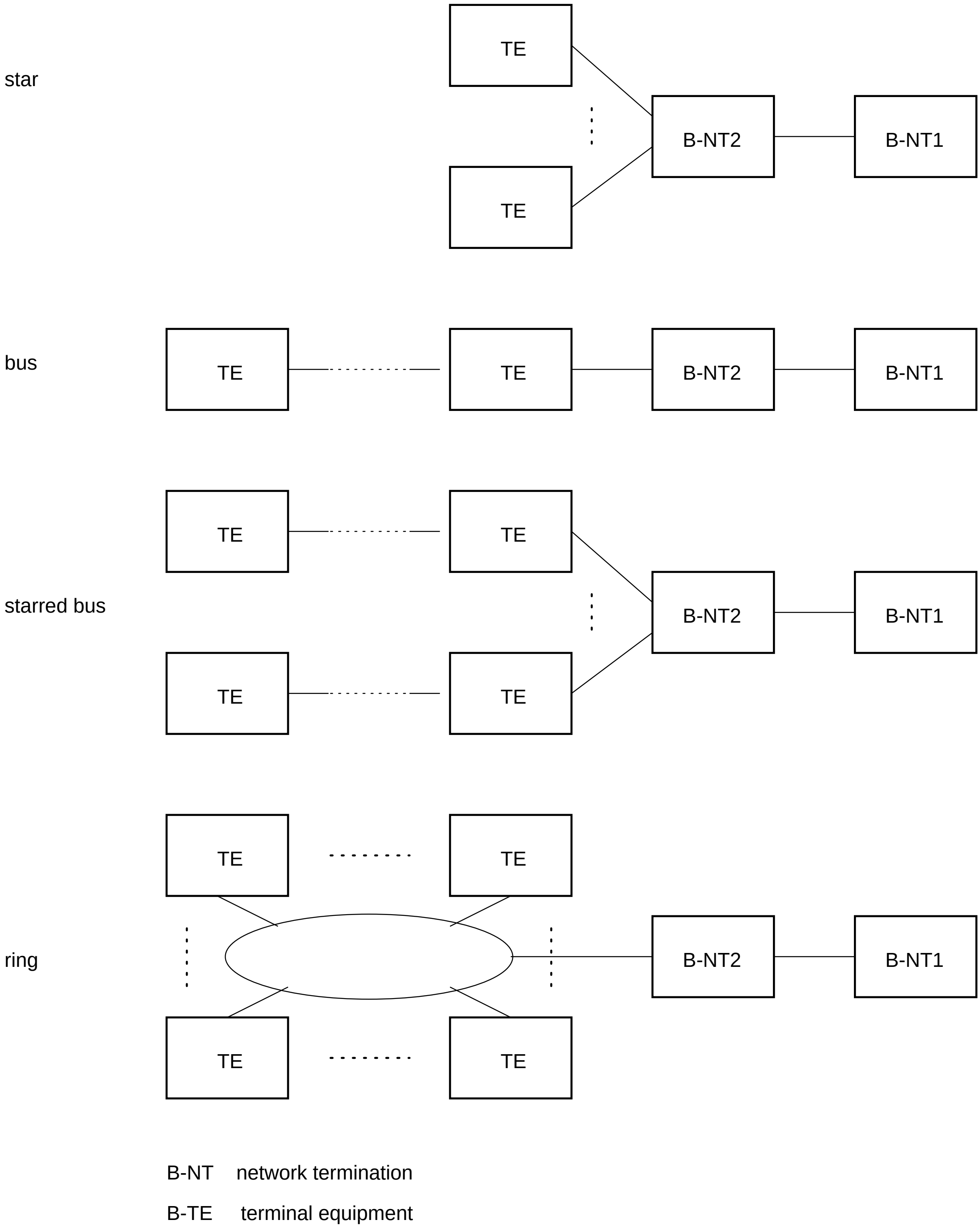
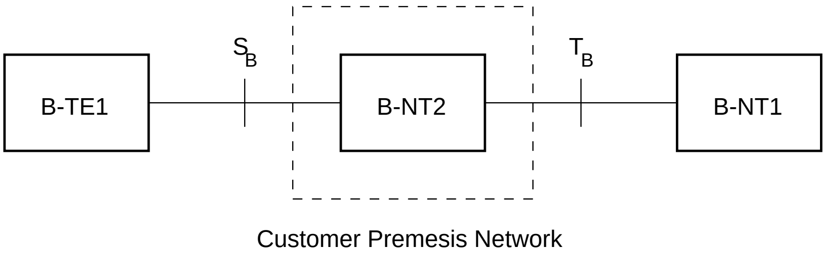


Figure 71: B-ISDN multiple interface configurations

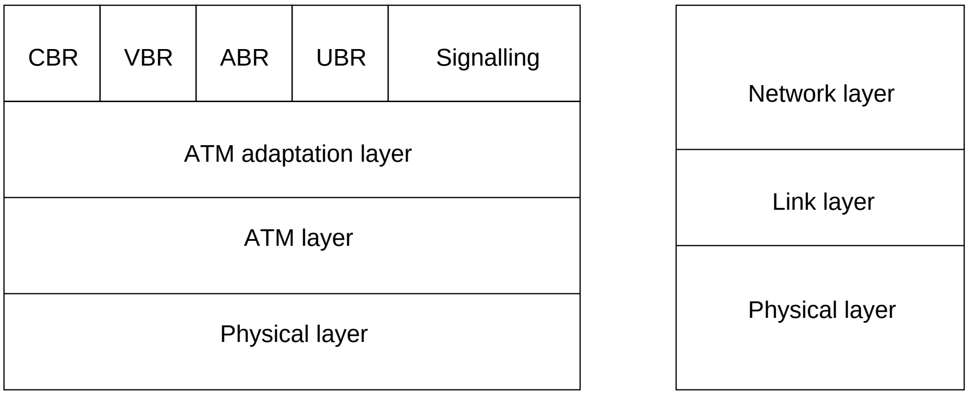


B-TE	terminal equipment
B-NT	network termination

Figure 72: B-ISDN customer premisis network configuration

The B-NT2 equipment is considered to be the **customer premisis equipment (CPN)** (Figure 72). This could in real terms be a **private branch exchange (PBX)** or other local switch.

The discussion above has mentioned the OSI reference model. This was developed in collaboration between the ISO and the (then) CCITT. It seems surprising therefore that there is no defined relationship between the B-ISDN PRM and the OSI reference model. Figure 73 is the author's view of the relationship between the two.



ATM	asynchronous transfer mode	ABR	available bit rate
CBR	constant bit rate	UBR	unspecified bit rate
VBR	variable bit rate		

Figure 73: B-ISDN PRM compared with the OSI model

As there is unlikely to be a user interface directly to the AAL, included in this figure are the interfaces to the service classes defined by the ATM Forum for the UNI:

- **Constant bit rate (CBR).** The CBR service offers a very simple, reliable guaranteed channel that effectively acts as a circuit emulation. The QoS of this service must be maintained throughout the lifetime of a CBR connection, as the data rate is expected to be constant. It is intended for use by applications with stringent real-time constraints on delay and jitter, e.g. real-time video.
- **Variable bit rate (VBR).** This service is also intended for use by by real-time applications.

However, it differs from CBR in that it does not expect the data rate to be constant, i.e. the sources may use variable bit rate coding for efficiency and also be statistically multiplexed.

- **Available bit rate (ABR).** This service class offers the B-ISDN user some degree of fairness, and also control of loss or delay with respect to QoS, but is intended for non real-time applications. It is likely that ABR QoS statements will specify that there are minimum acceptable parameters, but that if better QoS should become available then it will be used. ABR is intended for use by unit-oriented applications such as database access and electronic mail.

- **Unspecified bit rate (UBR).** UBR is intended for applications that send data very sporadically and the use of CBR, VBR or ABR would be wasteful of resources. In fact, this service class is effectively a best-effort approach which is similar to today's Internet (using IP). Applications that use this service would have non-real time requirements and not be too sensitive to loss, e.g. file transfers.

5.5 Operation and Maintenance

The operation and maintenance (OAM) of the B-ISDN network has five main actions:

- **Performance monitoring.** Managed entities are continuously or periodically monitored in order to generate maintenance event information.

- **Defect and failure detection.** Errors or malfunctions in the managed entities are detected or predicted resulting in the generation of maintenance event information or service alarms.

- **System protection.** To offer some degree of fault tolerance, the effect of failures in managed entities are minimised by the use of backups, standbys or other resources, and the failed entity is excluded form the normal operation of the system.

- **Failure or performance information.** Communication of failure and performance information as alarms to other managed entities in the management plane. Also acts as a response service to status report requests.

- **Fault localisation.** Use of test systems, both internal and external, to determine whether information about faults is complete or sufficient for other actions to take place.

These actions are supported by the use of OAM information flows in the ATM layer and Physical layer (Figure 74). Other management capabilities such as use of the **TMN (Telecommunications Management Network)** may also be employed in conjunction with (or instead of!) the OAM.

5.6 Asynchronous Transfer Mode (ATM)

To provide the new B-ISDN services, use of a technology called **asynchronous transfer mode (ATM)** is specified by ITU. ATM is a connection-oriented technique based on the use of fixed size packets termed **cells**. These cells are 53 octets in size, with 5 octets used for the **cell header**, and the remaining 48 octets for data (Figure 75).

The term **asynchronous transfer mode** needs some explanation. The words **transfer mode** say that this technology is specific way of transmitting and switching through the network. The term **asynchronous** refers to the fact that the cells are transmitted using asynchronous techniques, and the two end-points need not have synchronised clocks. Also, the use and allocation of cells and their subsequent multiplexing and transmission through the network is determined in an

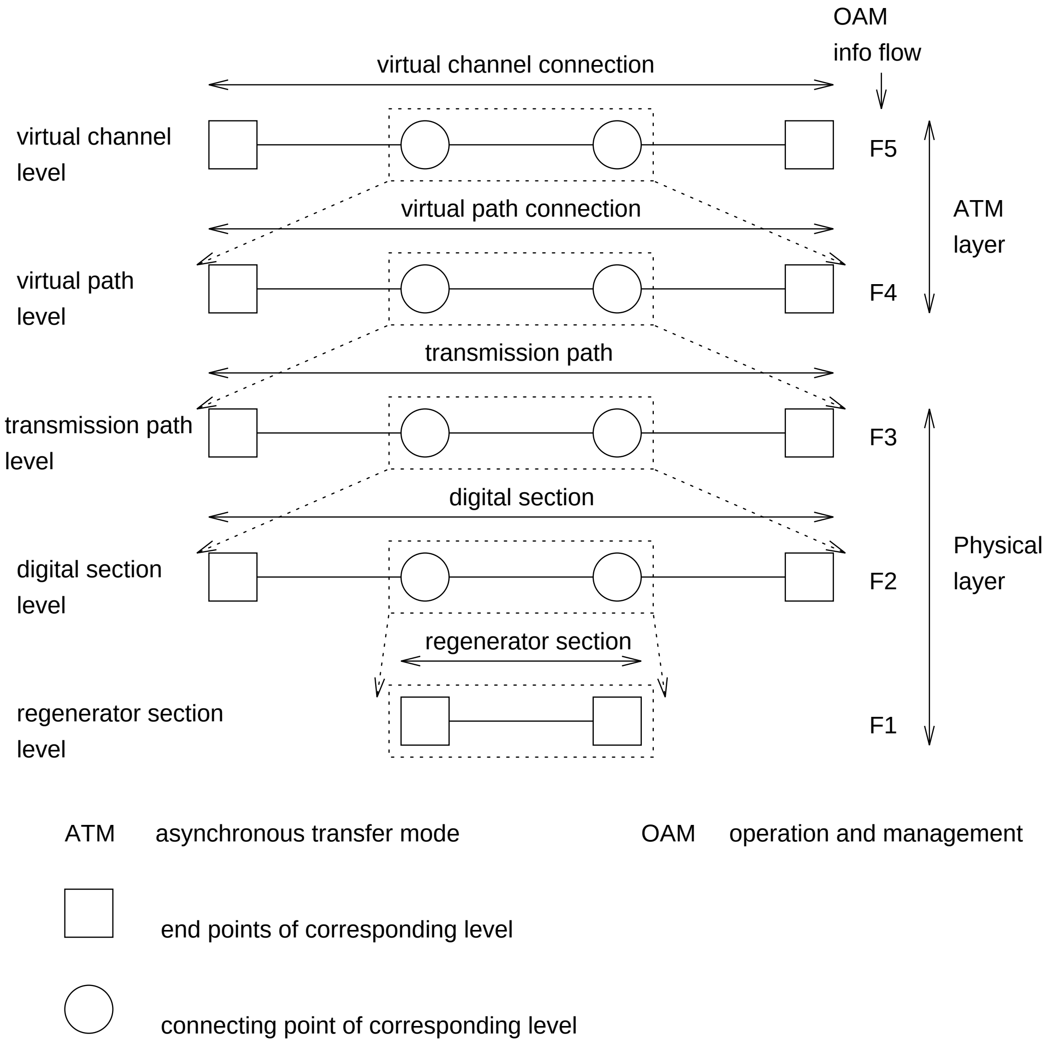


Figure 74: OAM information flows in the ATM layer and Physical layer

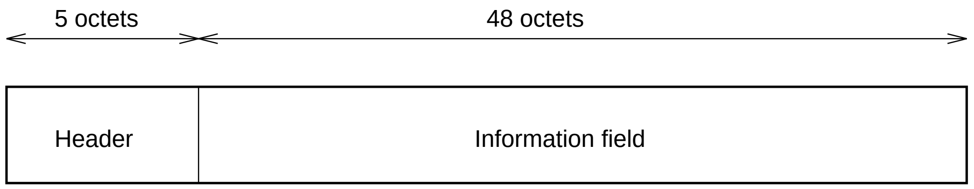


Figure 75: An ATM cell

asynchronous fashion, e.g. on demand, and is independent of the user (Figure 76 and Figure 77). ATM will support both circuit switched and packet switched (sometimes referred to as **circuit mode** and **packet mode**, respectively) services.

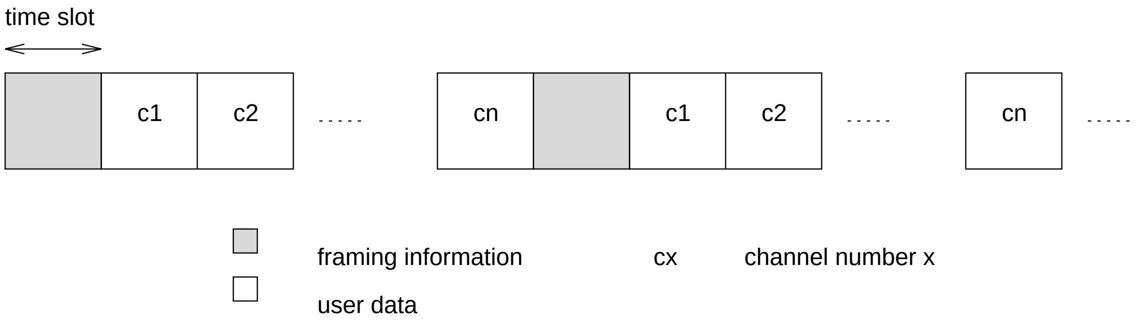


Figure 76: The principle of synchronous transfer mode

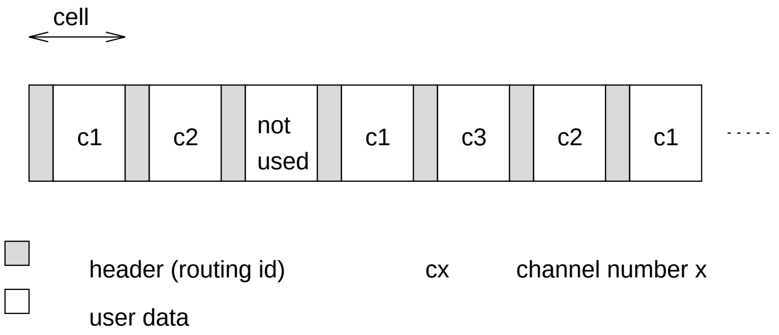


Figure 77: The principle of asynchronous transfer mode

There is much to know about ATM, however, here we will consider only the transmission aspects, by looking at the Physical layer and the ATM layer. There are several types of cells to consider, some are passed all the way through from the Physical layer to the ATM layer, while others never proceed beyond the Physical layer:

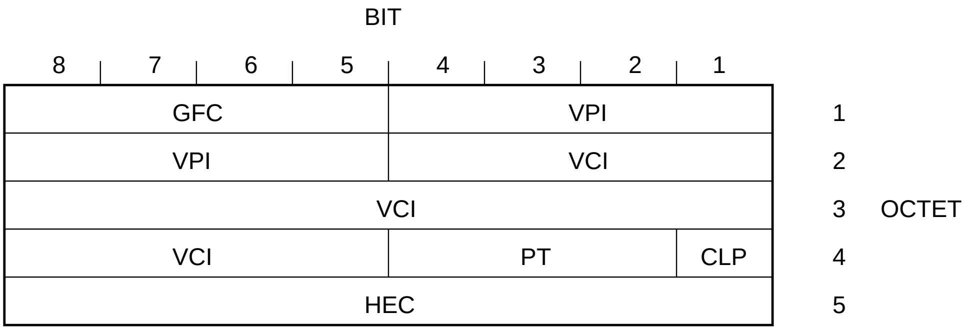
- **Assigned Cell.** (ATM Layer) A cell which is used to provide the service to the higher layers. This is the cell type that is used in transporting the higher level PDU.
- **Unassigned cell.** (ATM Layer) An ATM layer cell which is not an assigned cell.
- **Valid cell.** (Physical layer) Any cell that does not contain header errors, either through successful transmission or after being corrected by the physical layer.
- **Invalid cell.** (Physical layer) A cell with an error in its header that can not be corrected. It will be discarded by the physical layer.

- **Idle cell.** (Physical layer) A cell that is inserted/extracted by the physical layer for the purposes of payload capacity adaptation.

As explained earlier, there are two interfaces to the B-ISDN to consider – the user-to-network interface (UNI) and the network-to-network interface (NNI). For these two interfaces the internal structure of the ATM cell header is slightly different, as shown in Figure 78 and Figure 79.

The various parts of the cell header are:

- **Generic flow control (GFC).** (4 bits) Default value 0000. The exact use of this field is, as yet, not fully defined. This is used only in assigned or unassigned (i.e. ATM layer cells).
- **Virtual path identifier (VPI).** (8 or 12 bits) This is used for physical layer routing, together with the;
- **Virtual circuit identifier (VCI).** (16 bits) This is used in conjunction with the VPI field to provide physical level routing. Many cells with different VCIs may have the same VPI. Some VCI/VPI values are pre-assigned for special purposes (Table 14).
- **Payload type (PT).** (3 bits) Some PT values – **PT identifiers (PTIs)** – are pre-assigned (Table 15) for special purposes.
- **Cell loss priority (CLP).** (1 bit) Flag stating whether or not this cell can be dropped in the prescence of network congestion or other network difficulties. A value of 0 means that this cell has high priority and should not be discarded.
- **Header error control (HEC).** (8 bits) This is used by the physical layer for detecting and correcting errors in the cell header.



VPI	virtual path identifier	PT	payload type
VCI	virtual channel indentifer	CLP	cell loss priority
HEC	header error control	GFC	generic flow control

Figure 78: ATM cell header for UNI

5.7 ATM Adaptation Layer

The purpose of the ATM Adaptation Layer (AAL) is to adapt the PDUs passed down from the higher layer onto ATM cells. As the higher level PDUs may in general be of an arbitrary size, so one of the two sublayers in the AAL is responsible for **segmentation and reassembly (SAR)** of the higher layer PDUs. The other sublayer, the **convergence sublayer (CS)**, is responsible

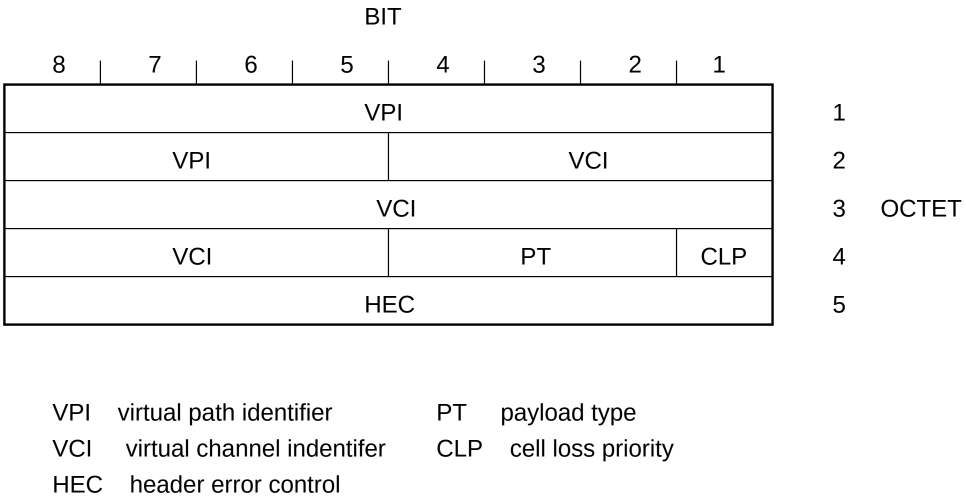


Figure 79: ATM cell header for NNI

VCI	VPI	Use
0	0 only	Unassigned cell
1	All	Meta-signaling
3	All	VP link network management (i.e. loopback the VP segment)
4	All	VP end-to-end management (i.e. loopback the VP end-to-end)
5	All	Signaling in point-to-point access
16	0 only	UNI (SNMP) network management (ATM Forum)

Table 14: Pre-assigned VCI/VPI values for ATM cells

for packaging the higher layer PDU with any additional information required for the adaptation necessary and offering an interface to the B-ISDN user. As mentioned previously, there are many and varied requirements for applications using the B-ISDN and as there are several AAL protocols defined. The classification of these protocols was guided by considering the following parameters:

- **Timing relationship.** The (non-)requirement for synchronisation between the receiver and sender.
- **Bit rate.** Constant or variable.
- **Connection mode.** Connection oriented or connectionless.

A summary of the (sensible) combinations of these parameters is given in Figure 80. Each of these classes may be loosely associated with the the ATM Forum classes described previously: Class A, CBR; Class B, VBR; Class C, ABR(ish); Class D, UBR. This mapping also give example use of the various classes. The various AAL classes are provided by several AAL protocols identified by a type number. The use of a particular AAL type for the provision of a particular service is not wholy well defined.

AAL Type 0. This is effectively a NULL AAL. It is not really an official AAL type but is mentioned for completeness.

AAL Type 1. This AAL type is normally used by Class A (CBR) services. The function performed by this AAL are:

- Segmentation and reassembly of user information.
- Handling of cell delay variation (jitter).
- Handling of cell reassembly variation.
- Handling of lost and misinserted cells.
- Source clock frequency recovery at receiver.
- Source data structure recovery at receiver.
- Monitoring and handling of AAL-PCI (protocol control information) bit errors.
- Monitoring and (possibly) correcting the bit errors in the user information field.
- For circuit emulation, monitoring and maintenance of end-to-end QoS.

AAL Type 2. This AAL type would be used with Class B (VBR). This type is not well defined and it seems possible that the it may be merged with AAL Type 1 in the future. Some of its functions are similar to AAL 1:

- Segmentation and reassembly of user information.
- Handling of cell delay variation (jitter).
- Handling of lost and misinserted cells.
- Source clock frequency recovery at receiver.
- Monitoring and handling of AAL-PCI bit errors.
- Monitoring and (possibly) correcting the bit errors in the user information field.

It also has these additional functions:

- Handle SDUs from a variable bit rate source.
- Transfer timing information between source and destination.
- Notify the higher layers of uncorrectable errors in AAL.

PTI	**Cell usage**	**ATM user to ATM user signaling**	**Congestion experienced**
000	User data	no	0
001	User data	no	1
010	User data	yes	0
011	User data	yes	1
100	OAM F5 link	–	–
101	OAM F5 end-to-end	–	–
110	Resource management	–	–
111	Reserved	–	–

Table 15: Pre-assigned PTI values for ATM cells

AAL Type 3/4. There was once separate Type 3 and Type 4 AALs, but they have now been merged. This AAL is now intended to support both Class C (ABR) and Class D (UBR) services.

In this AAL, the convergence sublayer is split into two (Figure 81), the **service specific convergence subslayer (SSCS)** and the **common part convergence sublayer (CPCS)**. The SSCS is application dependent, i.e. it could be for a VBR video application. The CPCS is responsible for constructing PDUs that can be sent to the other end user. There are two **modes** of operation of AAL 3/4; **message mode** and **streaming mode**.

The message mode is intended for use with framed data where the AAL-SDU is a logical unit of data with respect to the B-ISDN user (Figure 82). It allows the the transport of a single AAL-SDU in one or (optionally) more than one CS-PDU. The CS-PDU may be then further spilt into several SAR-PDUs. The AAL-SDU can be of an arbitrary size.

In streaming mode, the AAL-SDUs are of fixed size and one or more of them may be transported in a single CS-PDU (Figure 83). Each AAL-SDU is delivered in a separate SAR-PDU.

In both cases, the SAR sublayer provides error detection and both these modes can offer the following **operational procedures**:

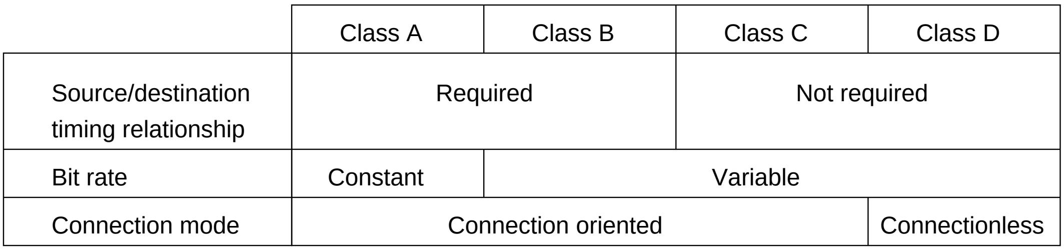
- **Assured operation.** Flow control and retransmission of missing or erroneous AAL--SDUs. Flow control restricted to point-to-point connections at the ATM layer and point--to-multipoint flow control possible.
- **Non-assured operation.** No retransmission of missing or erroneous SAR-PDUs. Optionally deliver erroneous PDUs to user. Allow flow control for point-to-point connections but not point-to-multipoint.

This AAL type also provides multiplexing at the SAR sublayer.

AAL Type 5.

AAL 3/4 was intended for use where there was a connection-orinted interface to the user and contains several control fields in the header that could be eliminated to provide a similar yet simplified service. This is the purpose of AAL5 which is also called the **simple and efficient adaptation layer (SEAL)**. Figure 84 shows how the AAL-SDU for AAL5 is broken down into ATM cells. AAL 5 has the following important features:

- The CS-PDU can be of the same size as AAL 3/4.
- The CS-PDU must be 48-byte aligned.
- There are only 8 bytes of PCI for the CS-PDU – the AAL 5 **tail**.
- The SAR layer is effectively a NULL function.



	Class A	Class B	Class C	Class D
Source/destination timing relationship	Required		Not required	
Bit rate	Constant	Variable		
Connection mode	Connection oriented			Connectionless

Figure 80: AAL service classification

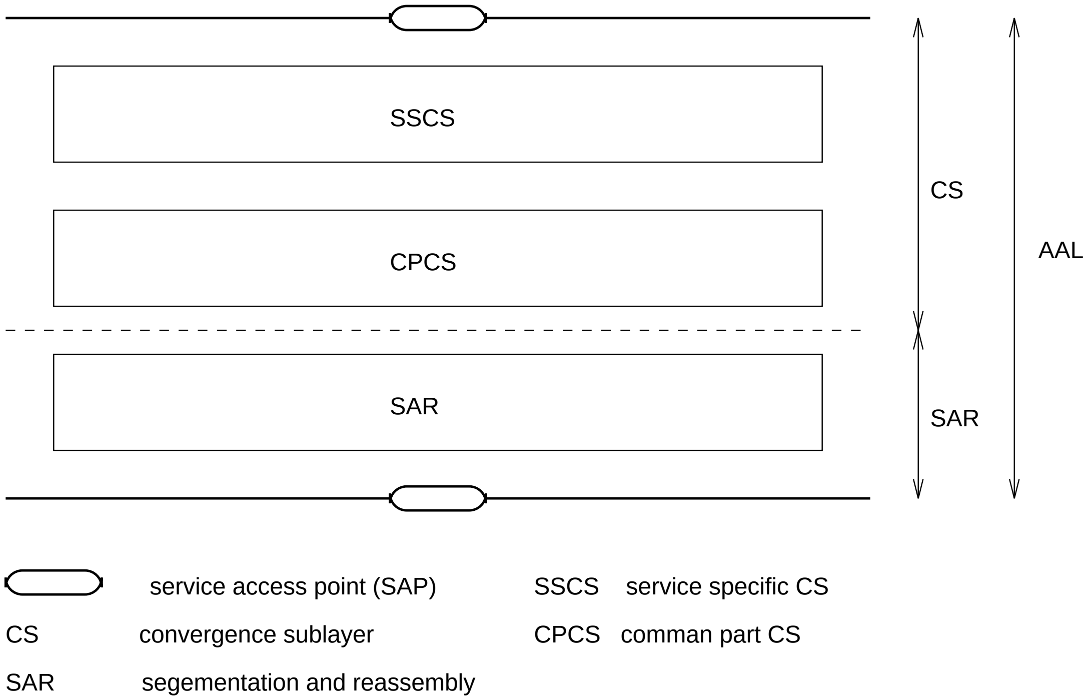


Figure 81: AAL 3/4 sub-layering

- Unlike AAL 3/4, AAL 5 provides no multiplexing at its user interface.

AAL 5 only add a very simple **tail** as a trailer to the AAL-SDU. The structure of this tail is shown in Figure 85.

The various fields perform the following functions:

- **AAL layer user-to-user identifier (UU).** This effectively identifies a **service access point (SAP)** or **port** at the AAL user interface.
- **Common part identifier (CPI).** The use of this field is undefined at present.
- **Length.** An integer value between 0 and 65,535.
- **Cyclic redundancy check (CRC).** For error detection – notifications of errors are sent to the AAL user.

AAL5 is intended for use by VBR sources with timing relationship between source and destination.

AAL 1 and AAL 2 are not used much.

AAL 3/4 has been chosen for the provision of the Bellcore **switched multi-megabit data service (SMDS)**, the European version of which is the **connectionless broadband data service (CBDS)**.

AAL 5 has been selected by the IETF to provide IP services over ATM. Also, it is used for signaling purposes in the control plane, where it is known as the **signaling AAL (SAAL)**.

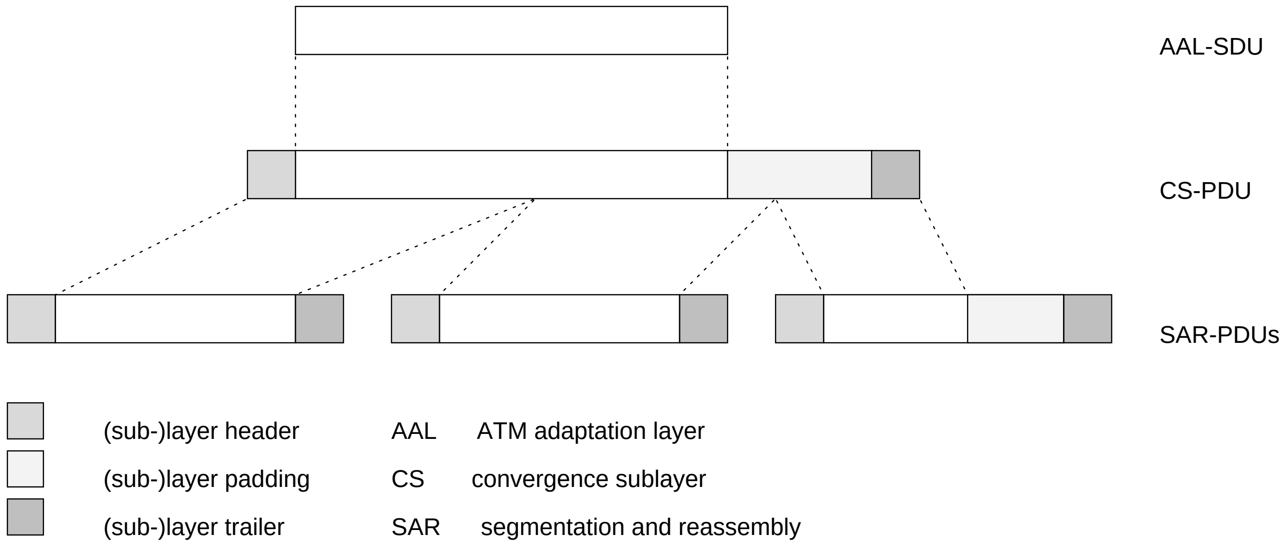


Figure 82: AAL 3/4 message mode service

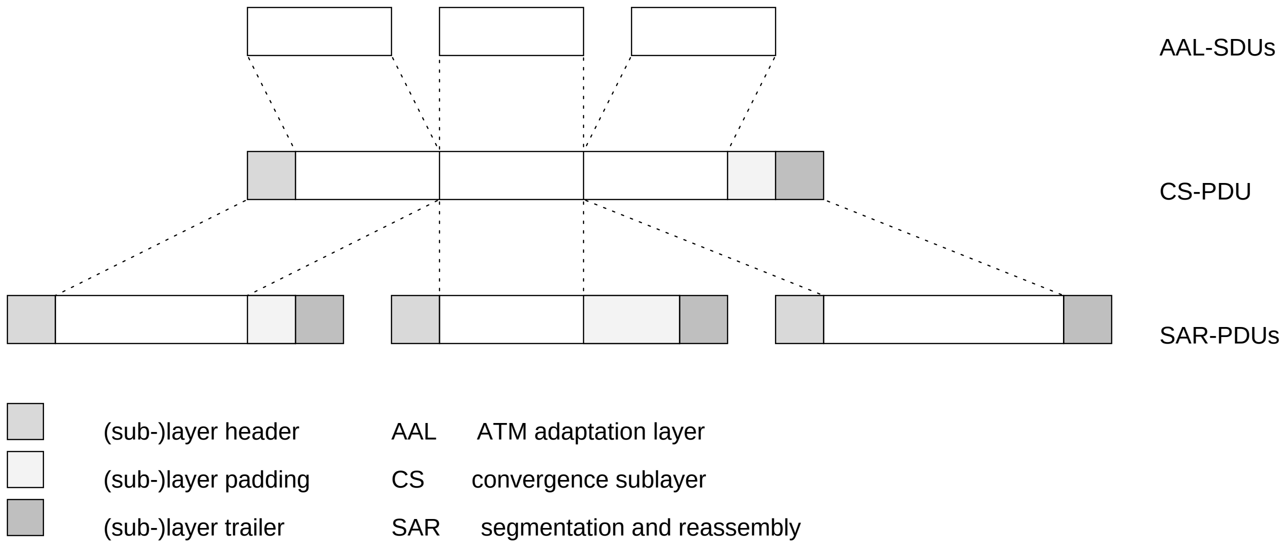


Figure 83: AAL 3/4 streaming mode service

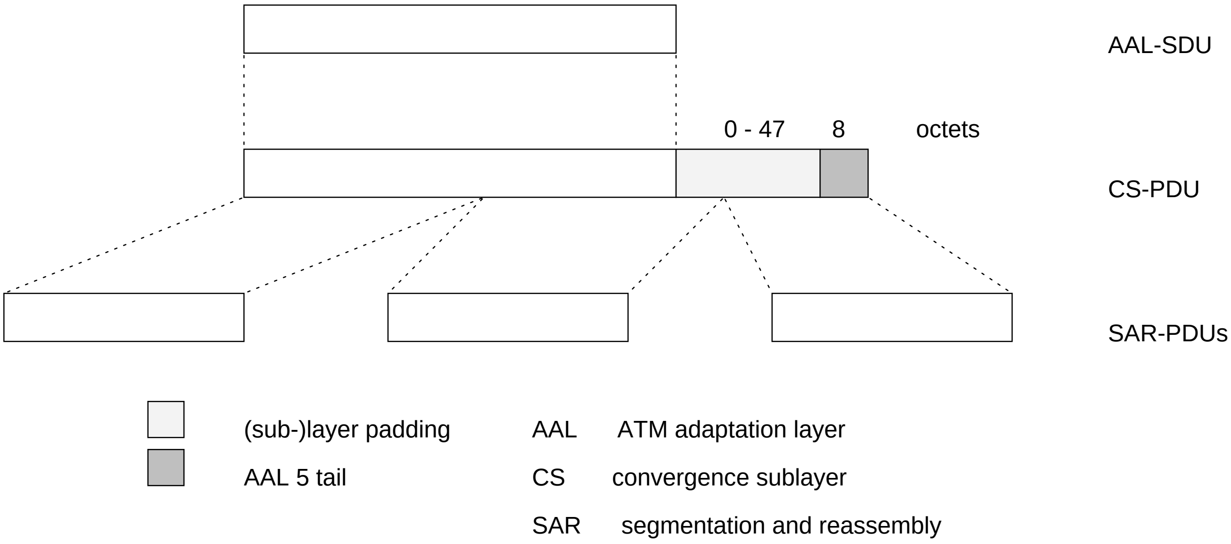


Figure 84: AAL 5 service

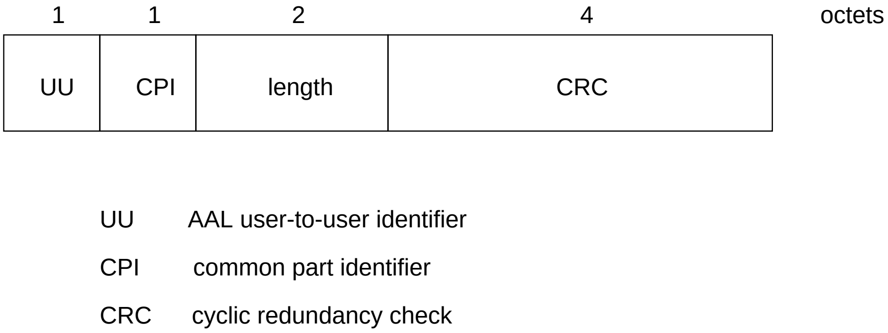

1
1
2
4
octets
UU
CPI
length
CRC
UU AAL user-to-user identifier
CPI common part identifier
CRC cyclic redundancy check

Figure 85: AAL 5 tail structure

5.8 Physical Layer; SONET and SDH

For B-ISDN, two transmission rates are currently specified; 155Mb/s and 622Mb/s, although it is highly likely that other higher rates will be standardised (Table 16). To achieve these rates, the physical layer uses a structuring called **synchronous digital hierarchy (SDH)**. This is an ITU standard based heavily on earlier work by ANSI called **synchronous optical network (SONET)**. For our interests the two are practically identical, and so further, only reference to SDH will be made, with specific differences in SONET being highlighted where necessary.

The labeling of the various rates in Table 16 are associated with the structuring of the physical level data: in SONET uses the concept of the **synchronous transport module (STS)** or **optical signal (carrier) (OC)** and SDH uses the **synchronous transport module (STM)**.

SDH is a point-to-point physical level transport system whose structure is depicted in Figure 86. The structure is split into **sections**, **lines** and **paths**, with logical manageable entities delimiting parts of the structure. Single sections consist of (fibre) cable that is terminated by equipment such as repeaters that are termed **section termination equipment (STE)**. The various sections of cable link together switches or multiplexors – **line termination equipment (LTE)** – to form a line. The end-to-end transmission path is defined by the the presence of **path termination equipment (PTE)** at the ends of the path.

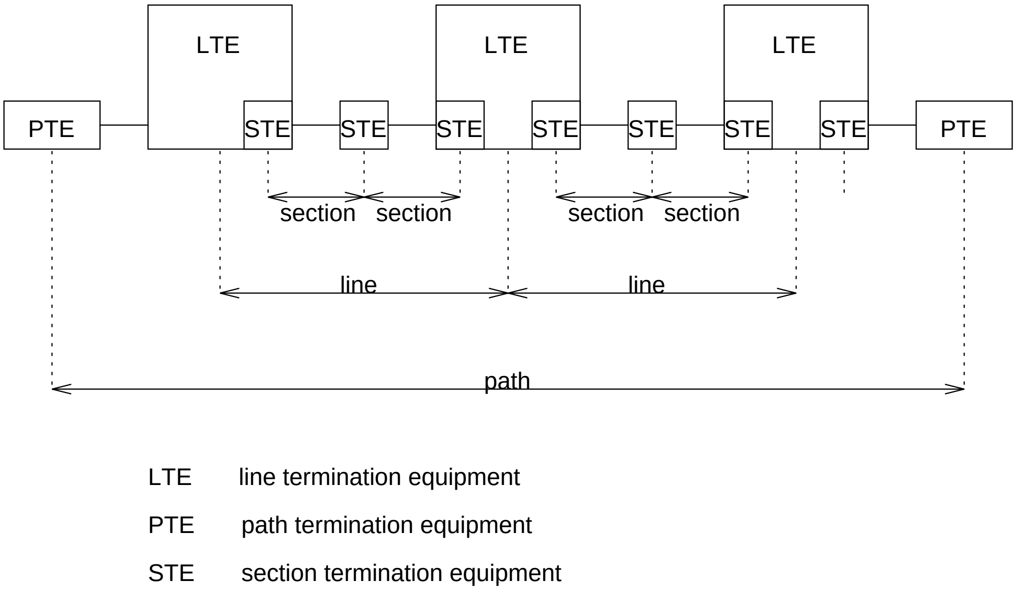

LTE
LTE
LTE
PTE
STE
STE
STE
STE
STE
STE
STE
PTE
section
section
section
section
line
line
path
LTE line termination equipment
PTE path termination equipment
STE section termination equipment

Figure 86: The structure of a SDH transmission path

SDH specifies a hierarchy of signals that are multiples of 155.52Mb/s to achieve higher data rates. For instance, the 155Mb/s rate, SDH uses a 9 × 270 octet frame. This frame is repeated with a frequency of 8KHz, so achieving a rate of 155Mb/s (Figure 87). The frame consists of 81 octets of **section overhead (SOH)** and the rest of the frame, which we will call the **payload**, is filled with ATM cells. The capacity available to the network user is then the **payload rate**. The **path overhead (POH)** are special control octets. The AU-4 pointer give the location of the VC-4 container which holds the ATM cells. There may not be an integral number of cells in the VC-4.

The 622Mb/s rate can be achieved simply extending the 155Mb/s SDH frame by increasing the field sizes (Figure 88).

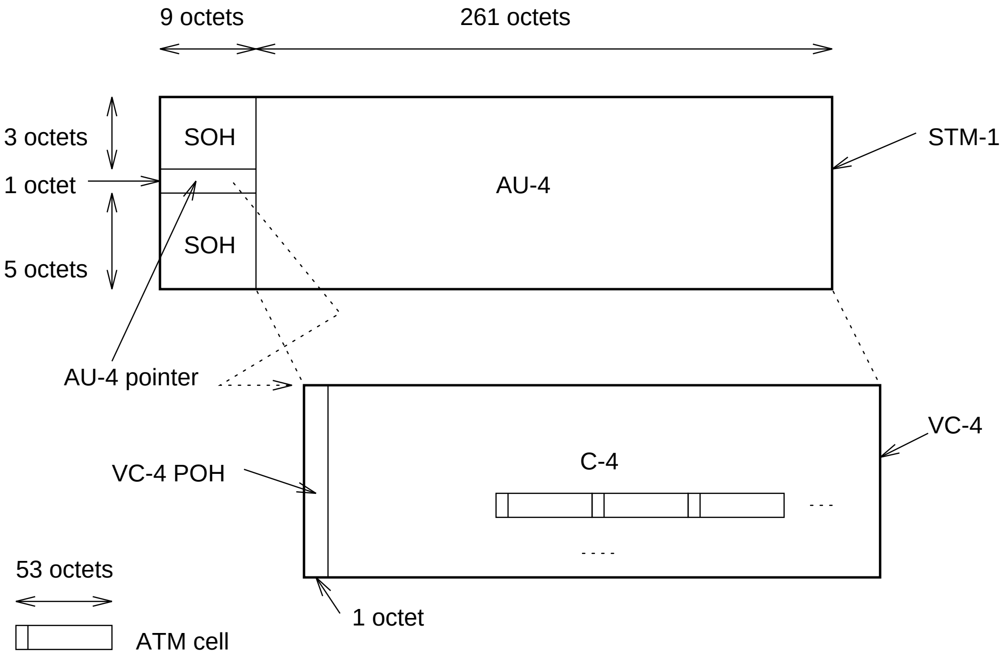
Some of the Physical layer specifications for the optical transmission of the signal are given in Table 17.

SONET	**SDH**	**Data Rate** [Mb/s]	**Payload Rate** [Mb/s]
STS-1/OC1	–	51.84	50.11
STS-3/OC3	STM-1	155.52	150.34
STS-12/OC12	STM-4	622.08	601.34
STS-24/OC24	STM-8	1244.16	1202.69
STS-48/OC48	STM-16	2488.32	2405.38
STS-192/OC192	STM-64	9953.28	9621.50

Table 16: SONET and SDH data rates and payload rates

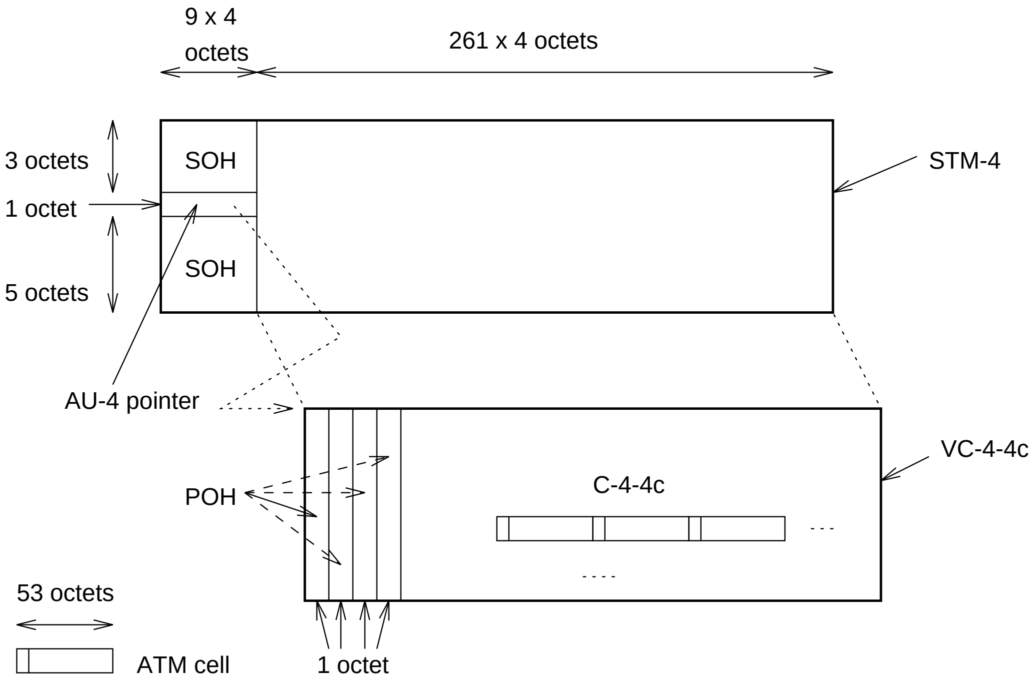
Item	**Specified solution**
Attenuation range	0 – 7dB
Transmission medium	2 mono-mode fibres, one for each direction
Operating wavelength	1310nm
Channel coding	non-return to zero (NRZ)

Table 17: Optical interface characteristics for SDH



SOH	section overhead	ATM	asynchronous transfer mode
POH	path overhead	C	container
AU	administrative unit	VC	virtual container
STM	syncronous transport module		

Figure 87: Frame structure for 155Mb/s SDH-based interface



SOH	section overhead	ATM	asynchronous transfer mode
POH	path overhead	C-4-4c	container 4 concatenated
AU	administrative unit	VC-4-4c	virtual container 4 concatenated
STM	syncronous transport module		

Figure 88: Frame structure for 622Mb/s SDH-based interface

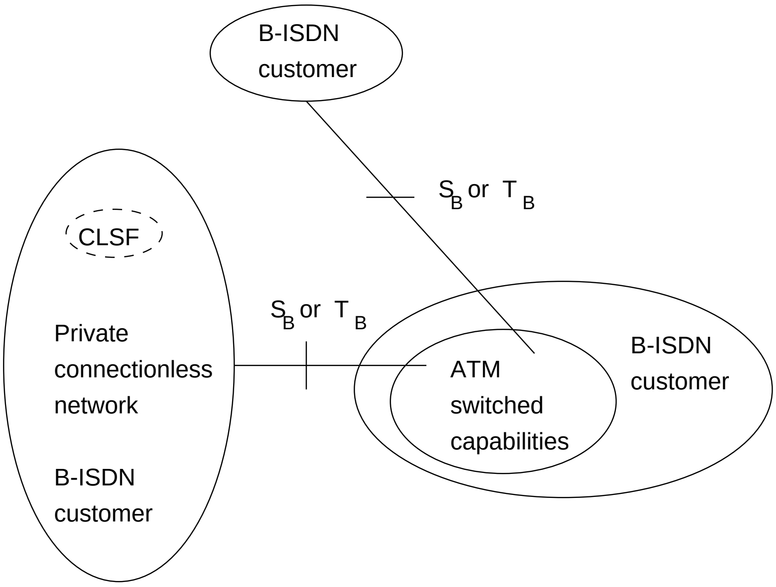
5.9 Connectionless Service

B-ISDN can also provide a connectionless (CL) service. This type of service is likely to be important as one of the first uses of B-ISDN is likely to be LAN/MAN interconnection. Most LANs and MANs use CL services and protocols at the subnetwork level, e.g. Ethernet.

There is always the possibility of **indirect** provision of CL services by putting the **CL service functions (CLSF)** outside the B-ISDN network and using ATM connections, i.e. a form of simulated CL (Figure 89). The CLSF would use ATM connections to the required destinations. These connections could be (semi-)permanent VCCs, pre-configured services specially supplied by the network provider or switched connections set up on demand. In the latter case, the user-to-network signaling would need to be very fast and efficient and the VCC set-up time would have to be minimised in order that the characteristics of the CL service expected by the B-ISDN user are maintained. Use of the permanent or reserved VCCs is obviously wasteful of resources.

A connectionless service can also be provided by the B-ISDN directly. **Connectionless service functions (CLSFs)** would be enabled either by a specialised connectionless service provider or by the B-ISDN itself. The CL service would use special connections between the user and the CLSF, which could be permanent, semi-permanent or switched circuits. Semi-permanent connections could be either VPCs with several VCCs offering CL services, or a single VCC. With switched connections would only allow the use of a single VCC at a time. In the case of a switched connection service, if the connection can not be established when the data transfer is requested, the data will be rejected by the CLSF.

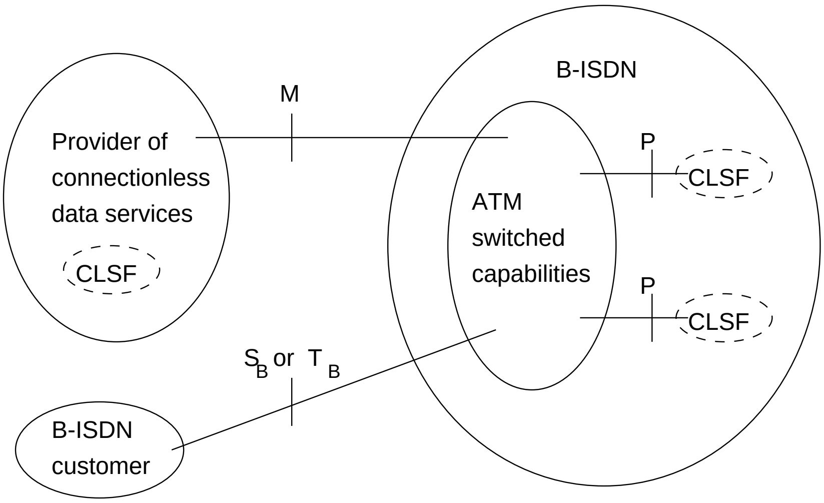
The general protocol structure for connectionless broadband service is depicted in Figure 91. The connectionless protocol used would provide mechanisms for allowing the user to specify addressing and QoS information , as well as providing routing functions. The AAL 3/4 or AAL 5 could be used to map the upper layer (connectionless) protocol on to ATM. A gateway would then map between the **connectionless network access protocol (CLNAP)** and the actual protocol used at the B-ISDN interface – **connectionless network interface protocol (CLNIP)**. However, while there is a logical separation made in the model between CLNAP and CLNIP, in reality these may be very similar and so the mapping would be simple.



ATM asynchronous trsanfer mode

CLSF connectionless service functions

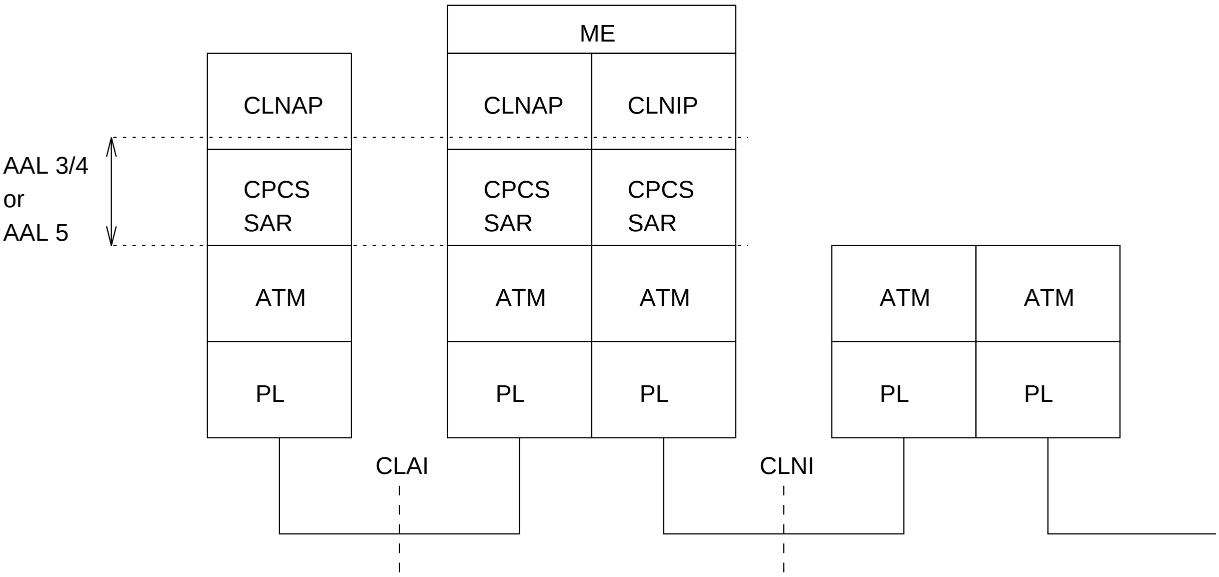
Figure 89: Indirect provision of connectionless service on B-ISDN



ATM asynchronous transfer mode

CLSF connectioneless service functions

Figure 90: Direct provision of connectionless service on B-ISDN



ATM	asynchronous transfer mode	CLNIP	connectionless network interface protocol
AAL	ATM adaptation layer	CPCS	common part convergence sub-layer
CLAI	connectionless access interface	ME	mapping entitiy
CLNI	connectionless network interface	SAR	segmentation and reassembly
CLNAP	connectionless network access protocol	PL	physical layer

Figure 91: Connectionless service protocol architecture for B-ISDN

6 Switching

Switching is concerned with connection of communication end-points within a network with out the need for direct, dedicated connections between all the possible end-points in the network. It's purpose is to efficiently determine the transmission links that are used in transporting information between two end-points. It is essentially a Physical Layer (Layer 1) and Data-Link Layer (Layer 2) function, and although its description will involve some description of routing type problems, it should not be confused with Network level (Layer 3) routing functions.

6.1 Transfer Modes

The *type* of switching used can be heavily dependent on the type of subnetwork, specifically if the subnetwork is connection of a **connection oriented (CO)** or a **connection less (CL)** nature. This can be described by its **transfer mode**, the way in which information is passed along the transmission path, multiplexed as well as switched.

Circuit Switching (CS). This is used in traditional telephone networks and also in N-ISDN. This involves the two end-points first setting up a (logical) connection between them (*connection establishment phase*). Then, their follows the required information exchange (*data transfer phase*), after which the connection is dropped (*connection release phase*). Any network resources that are required for the connection must be made available at connection establishment, maintained during data transfer and then freed at connection release. This kind of technology uses **time domain multiplexing** and **synchronous transfer mode** to allocate available capacity to users. An example of a circuit switched network is the normal telephone service.

Multi-rate circuit switching (MRCS). CS is inflexible and can be wasteful of resources. Although similar to CS, MRCS allows different basic rates by defining different channel types, and a user can choose which channel type to use. For instance, in with N-ISDN, H0, H11, etc. aggregate channels can be used when required (see also Table 11 and Table 12). The basic channel rate must be carefully chose in a trade-off between flexibility and efficiency. This type of transfer mode has a limited number of channels of each type that can be used, so is not very future-proof. For example, it may only be possible to offer a given number of H0 channels at a single exchange.

Fast circuit switching (FCS). Both CS and MRCS are inefficient for bursty data. FCS attempts to overcome this problem by caching some information about the type of connection and resources required, but does not actually set up the connection. When data transfer is requested, the cached information is used to set up a connection, and then the connection is torn down again after data transfer. Because there is no resource reservation, as with CS or MRCS, the **Quaility of Service (QoS)** may not always be guaranteed. There are also descriptions of mechanisms combining MRCS and FCS, but they become complex to design and implement, and one of the biggest problems is how to handle the high signaling rate involved.

Asynchronous transfer mode (ATM). This uses a CO based approach, setting up a connection, and multiplexing and switching fixed size cells onto the the connection. The connection has reduced functionality from the traditional CS type connection and is made possible by making assumptions about the reliability and capacity of the subnetwork. The fixed size cells and the asynchronous allocation of available network capacity means that this transfer mode is flexible and future-proof. ATM has been selected as the technology for use in B-ISDN.

Frame switching. This is similar to frame relay, differing only in the way it allows error control at the data-link level. Frame relay has no end-to-end error control or flow control, while frame switching has both of these.

Frame relay. This is similar to (and precursor to) ATM. Again, it is CS with reduced functionality, however the frames do not have to be of a fixed size. Commercial offerings are available form

Function	**X.25**	**Frame switching**	**Frame relay**	**ATM**
Framing	yes	yes	yes	yes
Bit stuffing	yes	yes	yes	no (Note 1)
Data CRC	yes	yes	yes	no (Note 2)
Error control	yes	yes	no	no
Flow control	yes	yes	no	yes
Multiplexing	yes	no	no	no (Note 3)

- **Note 1**: As fixed-size cells and fixed cell formats are used in ATM and SDH, there is no need for bit stuffing.
- **Notes 2**: Some AALs may provide error checking for the data portions of the AAL-SDU, however this is not part of the ATM or B-ISDN physical layer.
- **Note 3**: The assignments of VCI/VPI values that are used for multiplexing are for to the AAL or B-ISDN user.

Table 18: The functions of some transfer modes

BT in the UK (and by many national providers in other countries).

Packet switching (PS). In PS there is no prearranged connection between the end-points, and the information flow between them is contained in packets, each with its own header giving relevant information to facilitate switching, error control, flow control etc.

ATM, frame relay and frame switching seem very similar. Indeed, sometimes frame relay and ATM are considered synonymous, ATM sometimes being referred to as **cell relay**. However, this is inaccurate, and we will make the distinction as there are two different sets of ITU recommendations describing each.

This spectrum of transfer modes is summarised in Figure 92 and Table 18 gives a short summary of the main functional differences between the some of the transfer modes mentioned above, and the following Notes are referred to:

6.2 Switches

The purpose of switching is to provide interconnection between all the nodes on a network without the need for single connections between each pair of nodes (Figure 93). In this full mess interconnection, for N users, the complexity of interconnections required increases $\mathcal{O}(N(N-1)/2)$. This full mesh interconnection is not necessary and not desirable. Ideally, for N users we would like a complexity of $\mathcal{O}(N)$, however, we will always need some redundancy in the network for fault tolerance.

We require switching technology that provides a mechanism to establish connectivity as and when required. As we saw in the last subsection, the nature of the switching required depends on the transfer mode. In this subsection, we discuss something of the physical nature of the switches themselves. Before we go on to describe the types of switches, it is worth defining some terminology:

Switching element: This is a basic building block of a switch. A switching element will consist of a **input controllers** and **output controllers** which provide the input and outputs, respectively, to an **interconnection network** (Figure 94).

Switching fabric: This is an interconnection of switching elements to facilitate a particular

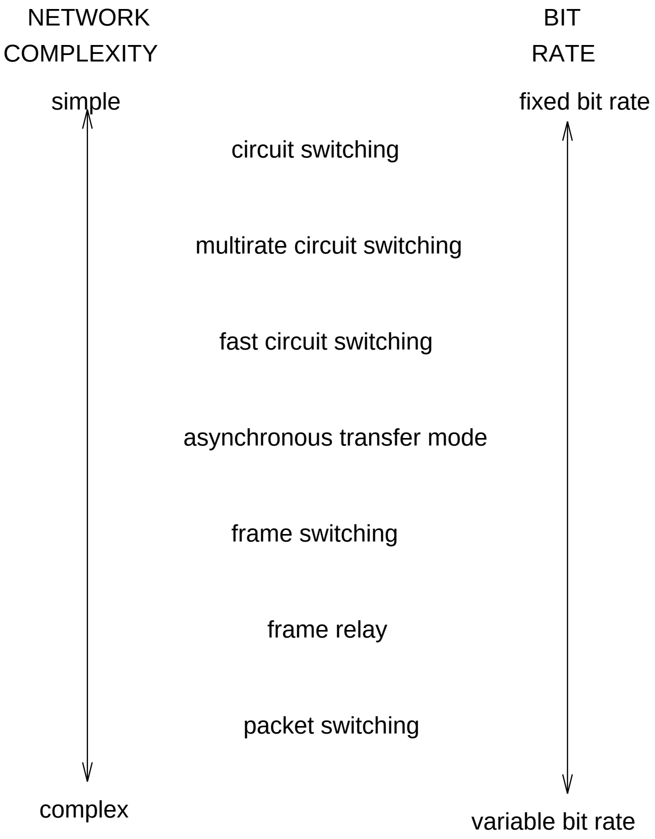


Figure 92: The range of transfer modes

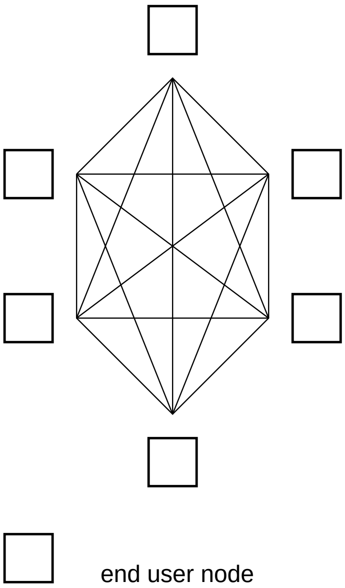


Figure 93: Direct connection between nodes is in efficient

switching mechanism. The switching fabric is characterised by the types and interconnection of its switching elements.

Switching elements have have many forms, however they have the following two common problems that must be resolved:

- **Collisions**: This occurs when more than one inputs are destined for the same output.
- **Blocking**: This occurs when the progress of one message through the interconnection network is stopped by a message that is not destined for the same output.

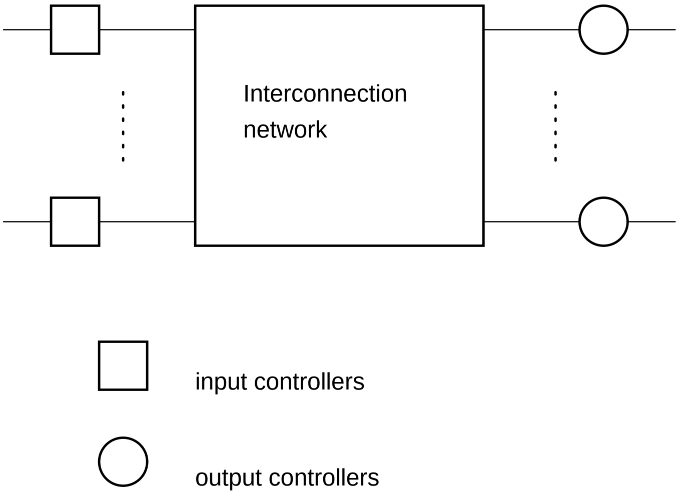


Figure 94: The main components of a switching element

In a **matrix switch** (or **space division switch**), the interconnection of network consists of a rectangular matrix of cross-points (Figure 95). This is the simplest form of switch. Buffering can be used at the inputs, at the outputs, at the cross-points or at any three of these in combination.

Buffers placed at the inputs (Figure 96) can be used to prevent collisions. If FIFOs are used, collisions will still occur if the messages at the head of the FIFOs are destined for the same port. Further, this will have the effect of blocking other messages in the queue. This problem can be overcome by the use of RAM buffers, but this may give rise to sequencing problems, and increase the complexity of the switching element.

Buffers placed at the outputs (Figure 97) can produce a non-blocking switching element if the switch is speeded up by a factor of N for a $N \times N$ switch. If the speed up is not possible, then buffers must additionally be used at the inputs.

Buffers can be placed at the cross-points of the interconnection matrix (Figure 98), resulting in a **Butterfly** switching element. This requires logic to prevent collisions, based on some fairness and timeliness metric. A Butterfly switch may require a large increase in the complexity of the switching element, as each cross-point now requires memory.

In the **time slot interchange (TSI)** technique (Figure 99), the line is logically partioned, with respect to time, into regular **time slots**. Users on each side of the TSI switch are allocated a time slot. For two users (one on each side of the switch) to communicate requires only that as the slot

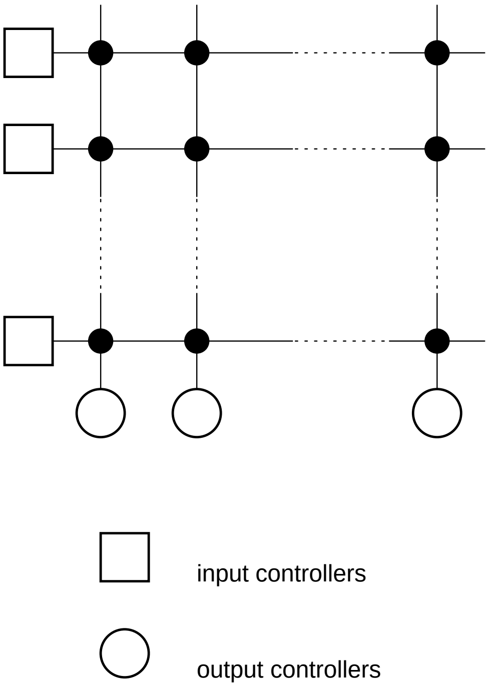


Figure 95: A matrix switching element

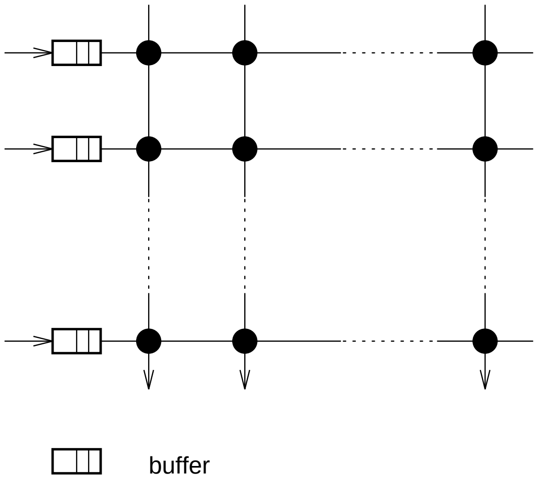


Figure 96: A matrix switching element with input buffers

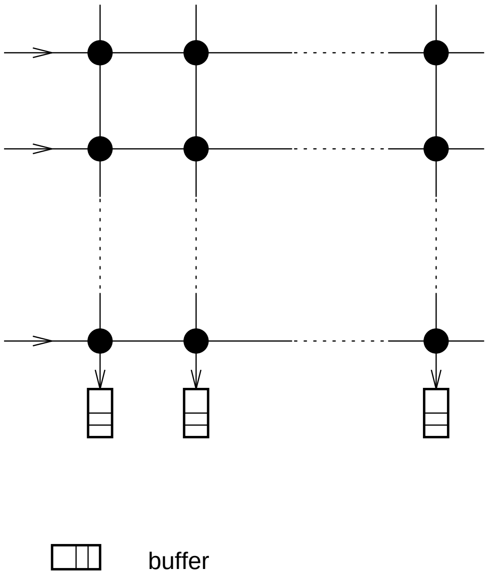


Figure 97: A matrix switching element with output buffers

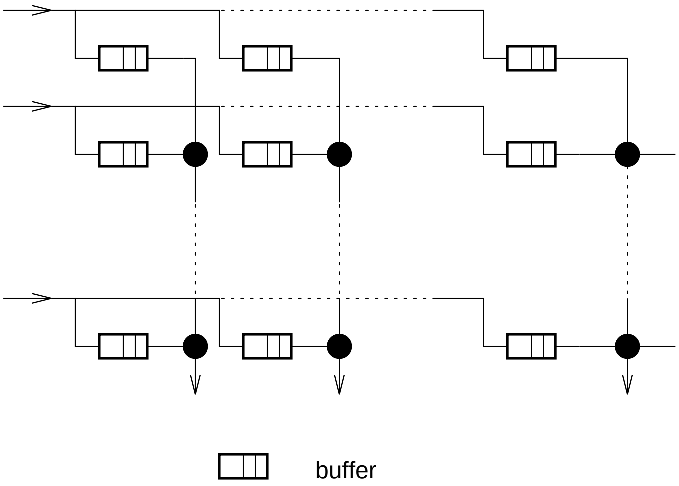


Figure 98: A matrix switching element with buffers at the matrix cross-points

passes through the switch, its time slot is interchanged with that of the user on the other side of the switch. The size of this kind of switching element is limited by the memory access speed of the hardware. This switching element demonstrates an example of **time domain multiplexing (TDM)**.

Although **single stage** switching elements and switches are possible using either space division or time division switches, many useful switching fabrics can be realised with just these two switching elements used in **multi stage** switches. They are usually written as an enumeration of the order in which the stages are connected, with a **S** representing the matrix (space division switch), and a **T** representing a time division switch. Common fabrics are TST or STS.

In a **central memory switching element** (Figure 100), the input controllers and output controllers are connected by central memory, which can be programmed to provide input and output buffering. This requires a highly parallel machine architecture for the purposes of memory access.

The interconnection network could also be arranged as a **Bus-Type switching element**, using a TDM bus (Figure 101). This is similar to a computer bus system. However, the bus would have to be at least as fast as the sum of the speeds of its inputs to prevent messages being dropped.

A **Ring-Type switching element** (Figure 102) passes round a time-slot that can be used by the input and output controllers connected to the ring, while having some of the constraints of the Bus-Type of switching element, the slot can be used several times in one cycle of the token.

When a large number of (often identical) switching elements are connected together in a network, we have a **multistage interconnection network (MIN)**. The most interesting MIN for our context is the **Delta Banyan** network. This is a **regular** network of 2×2 switching elements. An order N Delta Banyan switching fabric has the following useful properties:

1. They consist of $N^2/2$ identical switching elements.
2. They have the **self-routing** property, i.e. regardless of which input the message arrives at, it will always be routed to the same input. Such routing requires $\log_2 N$ bits of information.
3. They consist of $\log_2(2N)$ stages, with each stage having N switching elements.
4. As they are regular networks, and have a simple interconnection pattern, they are suited to VLSI techniques.

An example of a 4-stage Delta-Banyan network is given in Figure 103.

However, MINs such as this are susceptible to blocking. This can be reduced by using a preceding **distribution network** (Figure 104), but then this may cause sequencing problems without a complementary resequencing function at the outputs.

An alternative solution is to use a **sorting network** preceding the inputs, which orders the messages in a monotonous sequence, dependent on its destination output line. To prevent blocking, a **trap network** is placed after the sorting network, which feeds back messages with the same destination (Figure 105). Any fed back messages receive priority through the sorting network, in order to preserve sequencing. An example of such a network is the **Batcher Banyan network**.

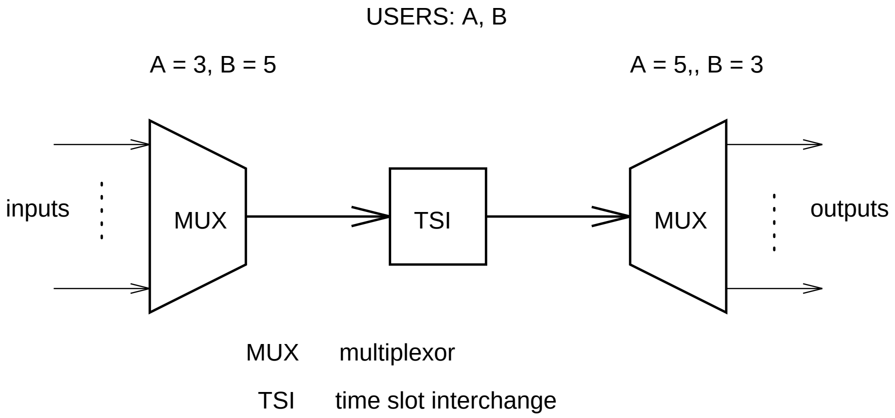


Figure 99: A time slot interchange switching element

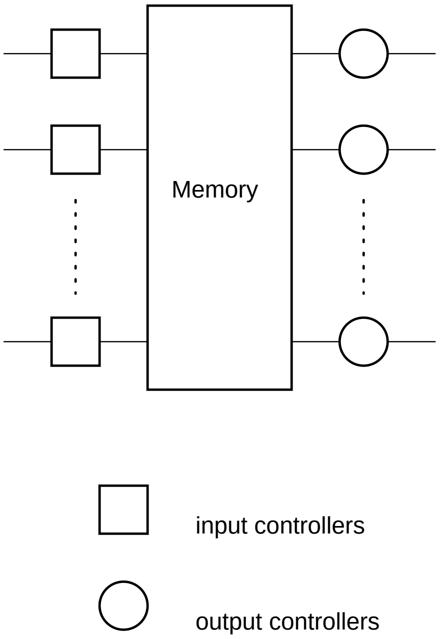


Figure 100: A central memory switching element

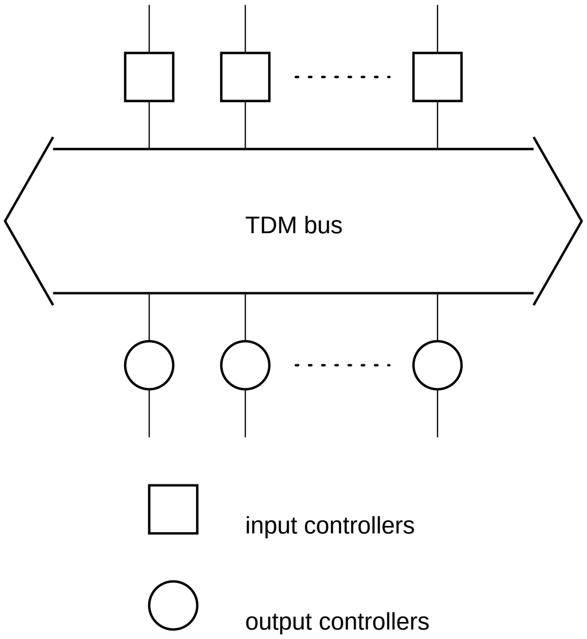


Figure 101: A TDM bus switching element

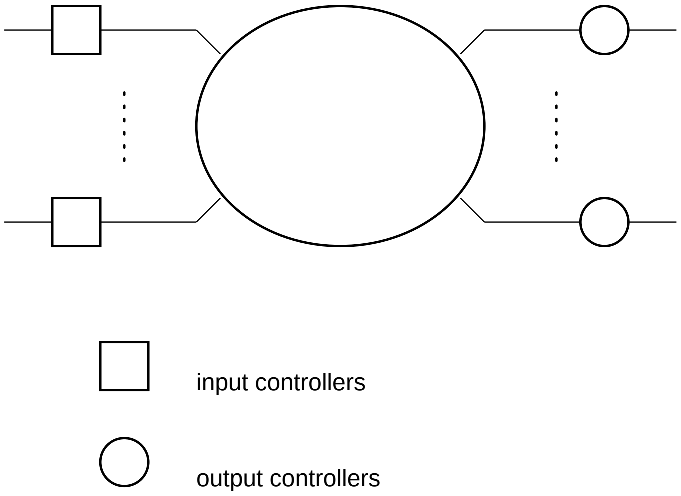


Figure 102: A Ring-Type switching element

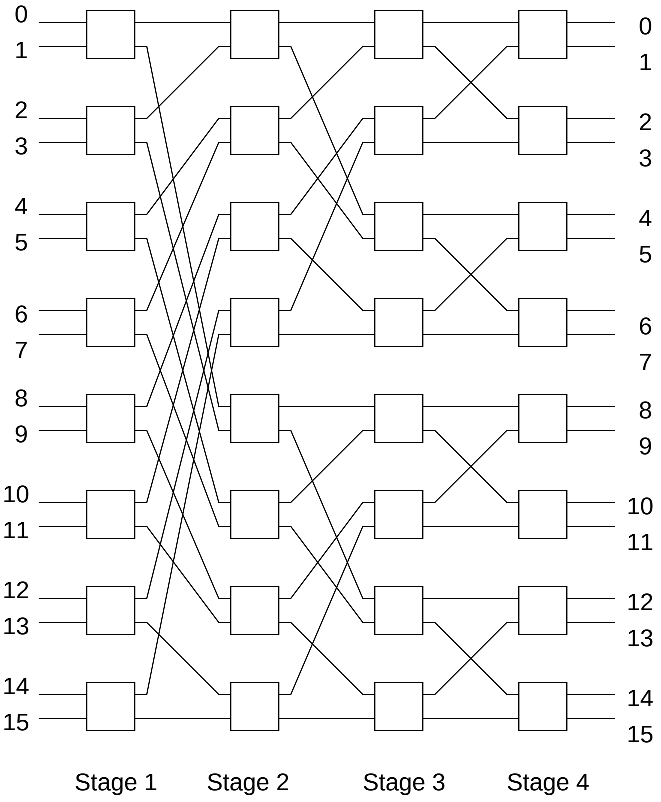


Figure 103: A Delta-2 4-Stage network

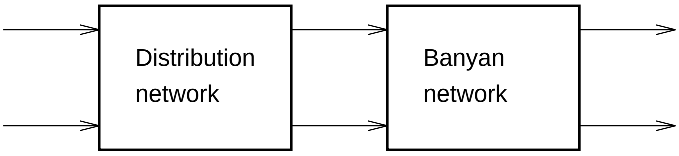


Figure 104: Structure of a distribution Banyan network

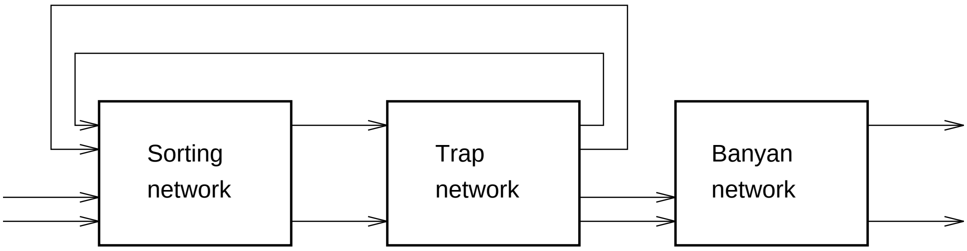


Figure 105: Structure of a sort-trap Banyan network

6.3 Switching services in networks

One of the most common uses of switching is in the **private automatic branch exchange (PABX)**. A PABX is a local switch that performs a similar function to a switch in a PSTN, however it is does not use the PSTN, saving on costs and making communication in an office environment more efficient. In a company or organisation, much of the communication is voice. This means that, until recently, much of the PABX equipment was based on carrying analogue, limited bandwidth signals. If data switching is required for computer equipment, most institutions also employ some sort of LAN arrangement for electronic messaging and other such data services. However, as people realise the need for integration of services such as data, voice and video, increasing number of PABXs are being produced that employ digital switching techniques, capable of switching data, as well as voice and video that has undergone analogue to digital conversion (ADC). Such PABXs are also know as **private digital exchanges (PDXs)**. PDXs can also offer additional services that may not have been possible in the analogue PABXs, such as call forwarding, call redirection, voice store-and-forward, etc.

Although the integration of voice and data using only one network via PDXs seems a natural step to take, it requires careful planning, and the management or corporate structure often assigns communications needs and data needs to different groups within in a company, so maintaining the separation between these services. Also, in an environment where there may be much data traffic with a very dynamic and unpredictable traffic profile, it may be unwise to switch voice traffic with data in case voice suffers at the cost of high bursts of data (e.g. UCL.CS!)

Another reason is that many data LANs are bus based – Ethernet remains the most widely used LAN environment – while the PDX based approach is naturally a star topology. With a LAN such as Ethernet, there is inherently a CL mode of operation while PDXs are circuit switches. It is likely that if PDXs are incorporated in to such a LAN environment they will be used to interconnect LAN segments. (However, there is now 10Mb/s and 100Mb/s Ethernet LAN technology available – **10Base-T** and **100Base-4T** – based on the use of UTP connections to form a star-hub arrangement.)

Circuit switching will continue in widespread use all over world – the world's telephone networks are circuit switched. Data requirements are often of a bursty nature and this is often because the exchange of data, unlike most phone calls, can often have very different data rates depending on direction. For instance, a user querying a database will probably make, for instance, simple menu choices which will result in the database providing much data in answer to his/her queries. This kind of data transfer is best suited to packet data, and can be facilitated by a **packet switching exchange (PSE)**. PSEs switch data by considering the header data provided as part of each packet.

PSEs can also provide another service, that of **virtual circuits (VC)**, each VC being referenced by a number called the **virtual circuit identifier (VCI)**. When a user requires this service, a **call request** containing the address of the destination and a VCI is sent to the PSE. The destination, on accepting the call, assigns another VCI to the call request and forwards it. VCIs only have local significance. When the call request reaches the destination, a VC is said to exist between the two end-points. The VC maybe set up through a route that has many PSEs, although this is transparent to the end-points. Such a packet switched network is the **X.25** network.

For the future, many people are looking to B-ISDN to provide integrated services - voice, video and data on one network using ATM. ATM switching can introduces new problems, not least the speed at which the switching must be achieved to maintain sufficient QoS. It seems likely that ATM switching will use the self-routing capabilities of Delta Banyan type switches. The main restriction in ATM switching is the **electronic** equipment that is required to process the **optical** signal. The optical signal must be converted to an electronic signal, processed, and then converted back to an optical signal from the electronic signal. Although CMOS, BiCMOS and ECL techniques can achieve Mb/s rates this may not be sufficient for the future Gb/s rates. GaAs technology may improve the situation but it seems that there will not be any significant progress until **optical switching** can be used. This may make it possible to fully exploit **wave division multiplexing (WDM)** or better coherent optical transmission techniques in modulation and switching.

7 High Speed Networks

Use of LANs is now common place in commercial, research, government and educational establishments. The evolution of hardware over the past decade has resulted in inexpensive solutions to interfacing computer systems to networks such as Ethernet and Token Ring. Such systems can typically offer around 10Mb/s on the shared media. Further, N-ISDN is available widely in Britain and in many places in Europe and it is now possible to buy hardware that will allow access to N-ISDN from a desktop machine.

The technology explosion has two important consequences:

- **Increase in the processing power available on the desktop.** It is possible to go to a high street shop and buy a PC with hardware that can offer facilities for real-time conferencing across a LAN, or perform simulations in real-time.
- **A diversity of software wanting to make use of the increased processing power and available network capacity.** As the hardware has become more capable so the applications designers have began to try and offer increasingly sophisticated services. Whereas LANS were initially used to provide access to, say, mailboxes or maybe a shared printer, we now find many applications requiring the use of the network to provide (often real-time) distributed services, for instance a distributed database or a network file service. Additionally, applications that make higher demands from the network are also becoming very popular – for example **multimedia conferencing** and **computer supported cooperative work (CSCW)**.

As well as network hungry applications resulting in increased network load, we find that there are now few computer installations that are **not** networked – virtually gone is the day of a stand alone PC used by a single person. So not only has the demand for network capacity increased per user but the number of users all contending for the same piece of LAN has increased.

To try and rationalise on the use of the available network capacity within an office or a site wide LAN, sites will often split the network up into **segments**. These segments are connected to a **backbone** network by use of **bridges**. The use of segments and bridges helps to localise traffic, so there is effectively less contention for the media compared to the case where there is only one segment. Further organisational changes like replication of data can also help to improve services for the user, but such schemes bring their own problems! While segmenting the network will help to create more localised traffic flows, the backbone may still need to contend with much inter-segment traffic.

Sometimes a user would simply like to have more network capacity available to him/her than the 10Mbp/s or so that, say, an Ethernet can offer, and if segments and bridges are used, the backbone needs to offer a greater capacity to cope with the many segments attached to it.

In this section we discuss some of the mechanisms of various high speed technologies that can bring 100Mbp/s and more to the LAN environment. Some of the technologies discussed here can be used to interconnect LANs over large public areas such as a town or city, and so are also referred to as **metropolitan area networks (MANs)**.

7.1 Medium access control at high data rates

Probably the most common form of LAN **medium access control (MAC)** is **carrier sense multiple access/collision detect (CSMA/CD)**, which is based on the use of a shared-bus network. We consider in this section why it is not feasible simply to run this kind of network 'faster' to get a higher throughput. Consider the following definitions for the propagation delay,

T_p, for a MAC layer frame being sent, and the time taken to actually transmit that MAC layer frame, T_t:

$$T_p = \frac{D}{V} \quad (85)$$

$$T_t = \frac{B_f}{R} \quad (86)$$

$$a = \frac{T_p}{T_t} \quad (87)$$

$$T_c = 2T_p \quad (88)$$

where:

D	separation of transmitter and receiver [m]
V	the velocity of propagation [m/s]
B_f	the number of bits in the frame
R	the link bit rate [b/s]
T_c	maximum time taken to detect a collision [s]

CSMA/CD networks work well provided that a is small, i.e. the propagation delay is small compared to the time required to transmit the frame. This is necessary in order to detect collisions during transmission. The maximum time that a node on a CSMA/CD network must wait to detect a collision is $T_c = 2T_p$ and so (from equations 85, 87 and 88) we see that $2a$ is the fraction of our frame that will be transmitted before a collision is detected. Hence, if T_p is much less than T_t then there is not much time wasted when a collision is detected. If we consider a 2.5Km bus and take $V = 2 \times 10^8$ m/s, then $T_c = 25\mu$s. So, in considering the detection of collisions, let us take a mean frame size of, say, 10,000 bits. Table 19 summarises information concerning collision detection as the bit rate of the link increases.

Bit rate [**Mb/s**]	T_t [μs]	$2a$	**Bits to detect a collision**
10	1000	0.05	500
100	100	0.5	5000
200	50	1.0	10,000

Table 19: Collision detection data for a 2.5Km CSMA/CD bus (10,000 bit frame)

Table 19 shows us that it is possible for the whole frame to have been transmitted before we can expect to detect a collision. This is obviously unacceptable. A solution to this problem might be to insist on an increasingly larger minimum frame size for higher transmission rates and longer cable lengths, but this is wasteful of network capacity. It is for these reasons that high speed networks – both LANs and MANs – must use other MAC methods.

7.2 Fast Ethernet

Ethernet switching uses the same CSMA/CD protocol that is used in popular LAN topologies, but connections consist of unshielded twisted pair (UTP) cable connected to hubs that act as fast switches, allowing each attached station to send and receive as required (Figure 106). There are actually four pairs of twisted pair wires used and so **Fast Ethernet** is sometimes also called **100Base-4T**. Some of the characteristics of 100Base-4T are given in Table 20.

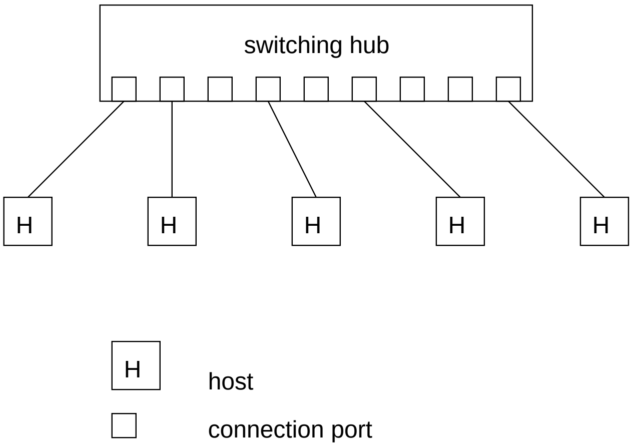


Figure 106: Ethernet switching hub schematic

The CSMA/CD procedure is implemented very simply. Each connected station uses 3 of its 4 set of wire pairs for transmission (to the hub) and another (different) set of 3 from 4 for receiving (from the hub). Each of the three wire pairs must achieve a data rate of 33.33Mb/s. This high data rate is achieved using each pair with a signalling rate of 25MHz (the maximum for UTP category 3), so before transmission, each 8 bits is converted to a 6 bit **ternary** code ($8/6 \times 25 = 33.33$) – **8B6T coding**. The three line levels in the ternary code are indicated by using the symbols + (for +V volts), **0** (for 0 volts) and **-** (for -V volts). For example, the 8 bit value 0x7f is given the 6 bit ternary symbol $+--+00$. There are similar translations for all other 8 bits values. The ternary pattern is chosen so such that **DC balance** is achieved on transmission, and also such that the patterns used all have two or more transitions.

When the hub receives a tranmission from a station, it simply retransmits the whole frame received on all the other connected ports. A collision is detected if a station finds that it is trying to use its send wire pairs at the same time as it is trying to use its receive wire pairs are active. A station can 'detect the carrier' when it sees that its transmission wires are not in use. Intelligence can be built into the hub so that it can learn the addresses of stations connected to its ports and when non-broadcast frames arrive, it can effectively switch the frame to its intended destination port only. In this way, it is possible for more than two stations to transmit (non-broadcast messages) at the same time.

The rest of the MAC protocol remains, essentially, unchanged. There is, however, one addition to the basic operation: an **end-of-sequence (EOS)** frame delimiter follows the normal CRC at the end of the MAC frame. This allows additional error checking and easy detection of the end of the

Characteristic	**Information**
Network capacity	100Mb/s
Network configuration	(half-duplex) point-to-point links
Maximum length	100m station-switch
Maximum distance between stations	200m station-switch-station
Maximum number of stations	switch dependent
MAC method	hub-star switched access
Physical media	4-pair category 3 UTP (category 5, 2-pair STP)
Channel coding	8B/6T coding
Line coding	Manchester coding

Table 20: Some characteristics of Fast Ethernet (100Base-4T)

frame across the wire pairs being used.

There are extensions of 100Base-4T planned known as **100BaseX**. The X denotes that there may be many different types of physical media, for instance optical fibre, and work is in progress within the standards bodies.

7.3 100BaseVG-AnyLAN

100BaseVG-AnyLAN is a hub-star topology. It is intended for use by small local workgroups. Also, it is intended to work over existing 10BaseT twisted pair voice grade (VG – category 3) cabling now being used for 10Mb/s CSMD/CD or Ethernet networks, so one of its main strengths is that it is relatively cheap to install. Some of the characteristics of 100BaseVG are given in Table 21.

While it offers an interface that is identical to the existing LAN MAC interface (Ethernet or CSMA/CD, token ring, token bus), the operation of the access protocol is different. The network topology is based around the use of special **repeater** stations. (These do not perform quite the same function as MAC repeaters, as will be explained later.)

100BaseVG uses a **round-robin** policy for allowing connected stations to transmit on the network. If we consider the simple set-up depicted in Figure 107. Stations connection to the repeater on **downlink ports**. These ports will simply be polled in and allowed access according to the numerical order of the ports. There is a priority mechanism and any stations making priority requests take precedence. If more than one attached station makes a priority request, then these are also scheduled in a round-robin fashion. To prevent priority requests from blocking normal request for too long, there is a timeout associated with each port (set at around 200ms to 300ms), after which the normal priority message takes precedence.

It is also possible to cascade the repeaters in a tree fashion by use of the **uplink port**. Such an arrangement is depicted in Figure 108. In this case, repeater A is the **root repeater** and takes control of the whole network. It has knowledge of which of its ports are connected to hosts and which to downstream repeaters. The round-robin scheduling is now applied, in order across the whole cascade arrangement, i.e. the order in which the ports would be served in Figure 108 is A1, B1, D2, D4, B7, A4, A5, A7, C2, C5. The repeaters takes care of any high priority requests by sending PREEMPT control signals to any lower level repeaters. The lower level repeaters then suspend their activity until the high priority request has been processed. They then continue their round-robin schedule from where they left off before the PREEMPT signal was received.

Round-robin access assures a high degree of fairness with respect to network access. With such fairness, it is easy to evaluate the maximum access delay of this network and so it is very much suited for use by workgroups that make use of multi-media or real-time applications with stringent

Characteristic	**Information**
Network capacity	100Mb/s
Network configuration	(half-duplex) point-to-point links
Maximum length	100m station-switch
Maximum distance between stations	200m station-switch-station
Maximum number of stations	switch dependent
MAC method	hub-star/tree switched access
Physical media	4-pair category 3/4/5 UTP, 2-pair STP, 2-pair optical fibre
Channel coding	5B/6B coding
Line coding	Manchester (UTP/STP) (?), NRZI (optical) (?)

Table 21: Some characteristics of 100Base-VG

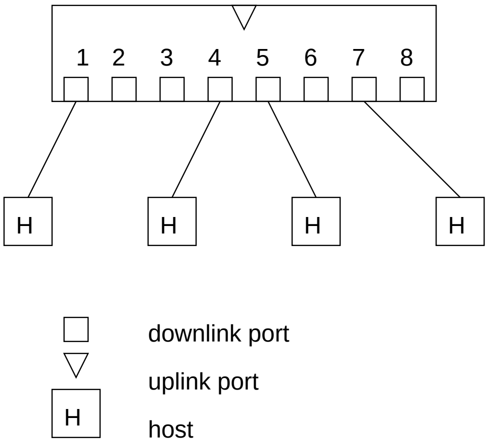


Figure 107: An example 100BaseVG hub set-up

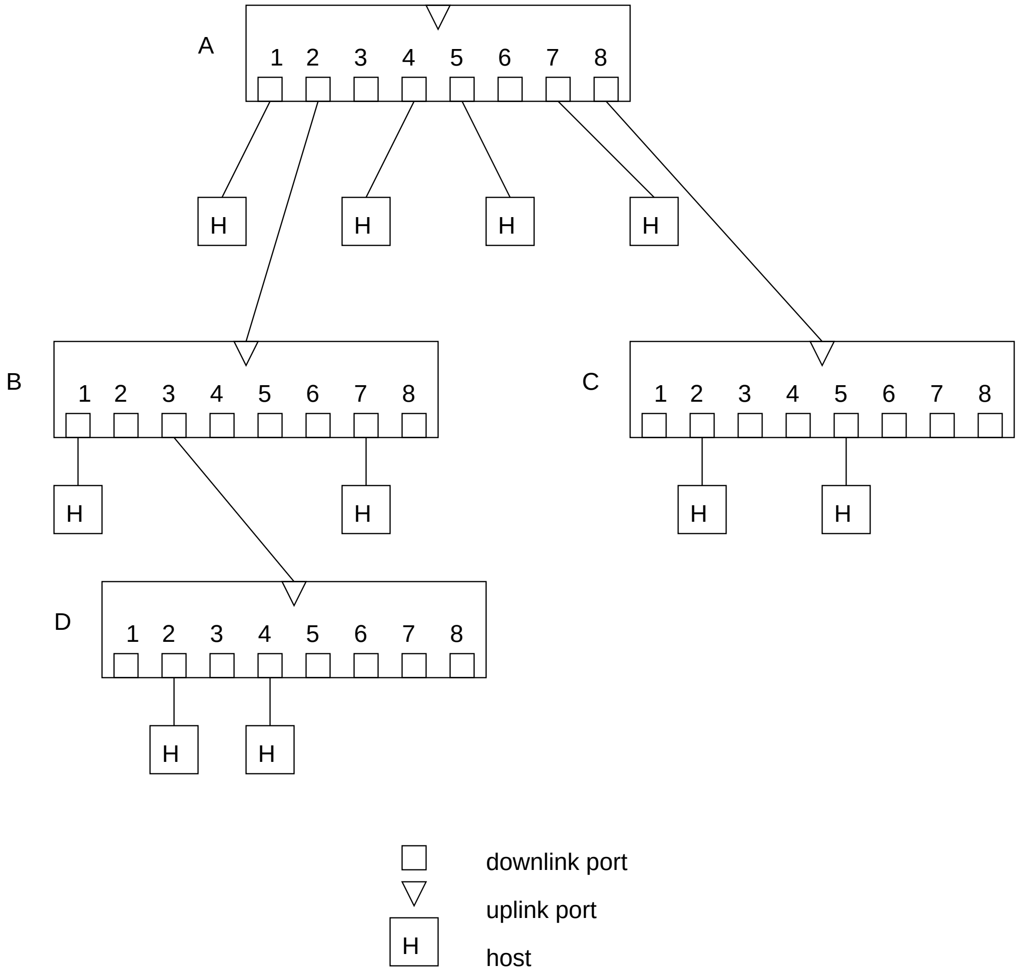


Figure 108: An example 100BaseVG cascade set-up

timing or synchronisation requirements.

The access protocol itself is fairly simple. A station that wishes to transmit sends a REQUEST signal to its repeater. The repeater will only pick up this signal when the port to which the sending station is attached. The root repeater eventually receives this REQUEST signal and sends an INCOMING signal to all other stations and repeaters to notify them to expect some data. After this its sends a CLEAR signal to the requesting station indicating that the station can now send data. The data is sent to all lower level repeaters so that they can detect the end of the frame. After the frame has been sent, the root repeater sends an IDLE signal to indicate that the network is free. High priority REQUESTs are treated using the PREEMPT signal as described above. Bridging, including to other 'traditional' MAC based LANs (Ethernet, etc), is possible by connecting a bridge to any downlink port, provided that the bridge is aware of the 100BaseVG access protocol.

7.4 High Performance Parallel Interface (HIPPI)

The **high performance parallel interface (HIPPI)** is a star-hub switch based technology for providing very high speed connectivity for short distances. The original HIPPI protocol specified the use of 50 copper twisted pairs to provide a uni-directional point-to-point link for a maximum length of 25 metres. HIPPI uses connections set up at very high speed through local switches. There is also work in progress to specify the use of HIPPI over **fibre channel** (an optical fibre based technology). Some of the characteristics of HIPPI are given in Table 22, with information about the fibre channel flavour of HIPPI given in brackets. The normal operating bit rate of HIPPI is 800Mb/s.

Characteristic	**Information**
Network capacity	800Mb/s (100, 200, 400, 800, 1600 Mb/s)
Network configuration	uni-directional point-to-point links
Maximum length	25m (10km) station-switch
Maximum distance between stations	50m (20Km) station-switch-station
Maximum number of stations	switch dependent
MAC method	hub-star switched access
Physical media	50/100 twisted-pair bundles (fibre channel)

Table 22: Some characteristics of HIPPI

A protocol suite is defined for the various functions to be performed in a HIPPI network (Figure 109).

The **HIPPI physical layer (HIPPI-PH)** is responsible for the electrical, mechanical and signalling aspects of HIPPI. The data is transferred in parallel normally as 32 bit words, but the cable can be doubled up (often called double wide HIPPI) to transfer 64 bit words and so offer 1600Mb/s. As HIPPI has unidirectional links, upto four separate wires are required to support duplex 1600Mb/s data exchange. HIPPI uses a block parity based error control scheme at transmission time.

The **HIPPI switching control (HIPPI-SC)** function is responsible for setting up connections between HIPPI switches. The HIPPI network can be composed of many HIPPI switches (Figure 110). The boxes marked as **HIPPI device** would be high performance machines (such as supercomputers), interworking units (for instance to B-ISDN) or intelligent peripheral devices. The switch control function uses information in the I-field of a HIPPI frame (see below). A HIPPI source requests a connection giving the relevant information in the I-field and the switch will try and make the connection to the destination, then data transfer can proceed, after which the connection is released. The control bits in the I-field and their affect on connection establishment is explained below. It is possible to get switches that can set up connections in less than one microsecond!

HIPPI specifies a framing protocol for transmitting information (Figure 111). The unit of transmission for data is a burst of 256 words (32 bits or 64 bits). The I-field contains information that is used for switching by the hub(s) (Figure 112). The various bits of the I-field are:

- **Locally defined (L)**: if this bit is set to one, it signifies that the rest of the I-field has locally (privately) defined format and not the standard format.
- **Width (W)**: if this bit is set it indicates that a double wide HIPPI connection (64 bit words) should be attempted. If the source or the destination can not support this, the connection request is rejected.
- **Direction (D)**: if this bit is set it indicates that the source and destination address bits should be swapped round. This provides a method for devices to simply route return messages, but is not useful when a LAN type arrangement is in operation such as that depicted in Figure 110.
- **Path selection (PS)**: determines how the I-field address information is treated. It can be treated as a 24 bit block which contains a series of port numbers that effectively define s route through a set of switches, or it can be used as two 12 bit addresses, as depicted in Figure 112. The former mode is called **source route** mode and the latter **logical address** mode.
- **Camp on (C)**: is set, this bit instructs the switch not to reject the connection if the destination is busy, but to wait until the destination is free and then re-attempt the connection.
- **Source/Destination address**: this information in logical address mode is often encoded as a 6 bit switch number and a 6 bit port number.

The **HIPPI link encapsulation (HIPPI-LE)** offers an interface that supports IEEE 802.2 Logical link PDUs, i.e. it looks like a CSMA/CD or token-bus/ring LAN but 80/160 times as fast! This provides a way for other standard technologies to be used over HIPPI, e.g. the Internet Protocol suite, IP, UDP and TCP.

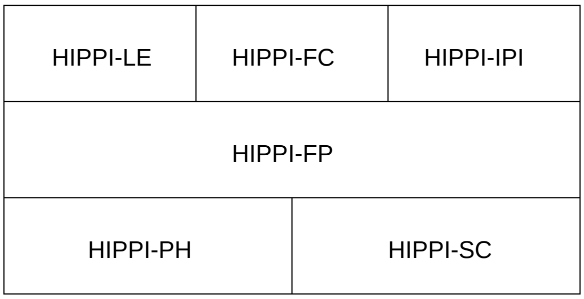
The **HIPPI fibre channel (HIPPI-FC)** interface is still under development, but when completed, hopes to offer a increased capabilities, as shown by the bracketed information in Table 22.

The **HIPPI intelligent peripheral interface (HIPPI-IPI)** will allow high speed computer controlled peripherals (such as optical disc drives) and the HIPPI-FC developments intends to add support for the **small computer systems interface (SCSI)** command set.

7.5 ATM LANs

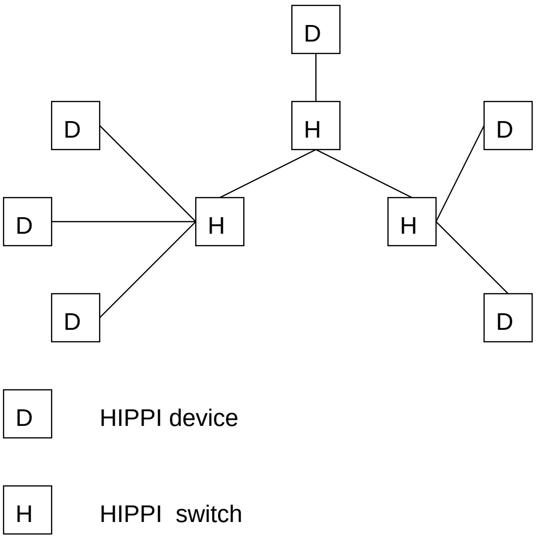
The concept of **ATM LANs** (or **local ATM (LATM)**) is very simple: use ATM switches to build a LAN and so allow users to have access extremely high network capacity. Such an ATM LAN may be as depicted in Figure 113. In fact, it is expected that this will be the most common usage of ATM at least in the near future. While there are no readily available B-ISDN services and much work is progressing on high throughput applications, there is a need for high speed reliable network environments for workgroups. Using ATM LANs also offer the advantage that the applications developed to use them may be used over DQDB MANs and any B-ISDN infrastructure as it appears.

For such local use, the ATM forum have defined several low cost interfaces that can run at lower rates (e.g. 25Mb/s, 34Mb/s, 55Mb/s) and use existing wiring.



HIPPI	high performance parallel intreface
HIPPI-LE	HIPPI LAN encapsulation (IEEE 802.2)
HIPPI-FC	HIPPI fibre channel
HIPPI-IPI	HIPPI intelligent peripheral interface
HIPPI-FP	HIPPI framing protocol
HIPPI-PH	HIPPI physical layer
HIPPI-SC	HIPPI switch control

Figure 109: HIPPI protocol suite



HIPPI high performance parallel interface

Figure 110: A LAN constructed using HIPPI switches

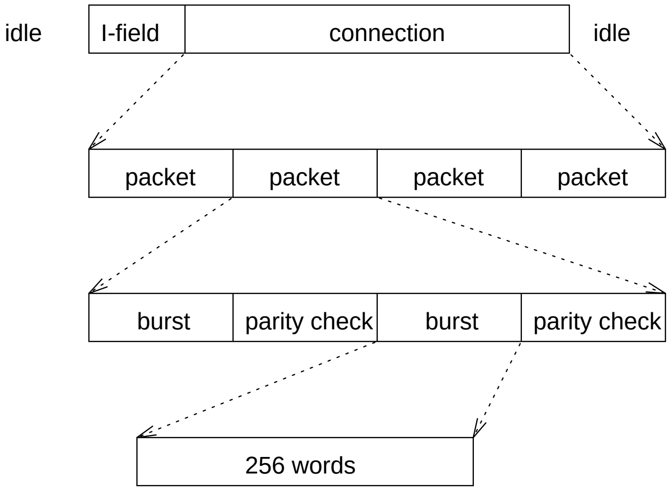


Figure 111: HIPPI data framing hierarchy

BITS	1	2	1	1	2	1	12	12
	L		W	D	PS	C	source address	destination address

L	1 indicates locally defined format
W	1 indicates 64 bit connection
D	1 indicates swap source and destination addresses
PS	path selection
C	camp on

Figure 112: HIPPI data frame I-field

This concept can be extended further by saying that it is possible to bring high speed links to the desktop, e.g. 155Mb/s, and that this cab be exploited by peripherals directly. In a similar fashion to the HIPPI-IPI interface, people are considering peripheral devices that connect to a **desktop area network (DAN)** that is effectively and extension of the desktop machine's back-plane bus. All the peripheral components, for example video cameras, disc drives, printers, etc., would be able to communicate with the processor or memory directly using ATM. This would mean that the peripherals could be located remotely or locally and this could be made transparent to the application, but also that any peripheral could easily be connected to the desktop machine via a high speed interface. Currently, most workstations and personal computers use parallel buses. However, in the future, we may well see such machines using **switched back plane** busses that use network protocols.

7.6 Fibre Distributed Data Interface (FDDI)

The **fibre distributed data interface (FDDI)** is a token based technology using fibre optical links. The FDDI specification defines the use of OSI layers 1 and 2 (Physical layer and Data Link layer). Some characteristics of FDDI are given in Table 23. FDDI's operation is similar to that of **token ring**.

Dual counter-rotating rings are used to improve reliability. The rings are labeled the **primary**

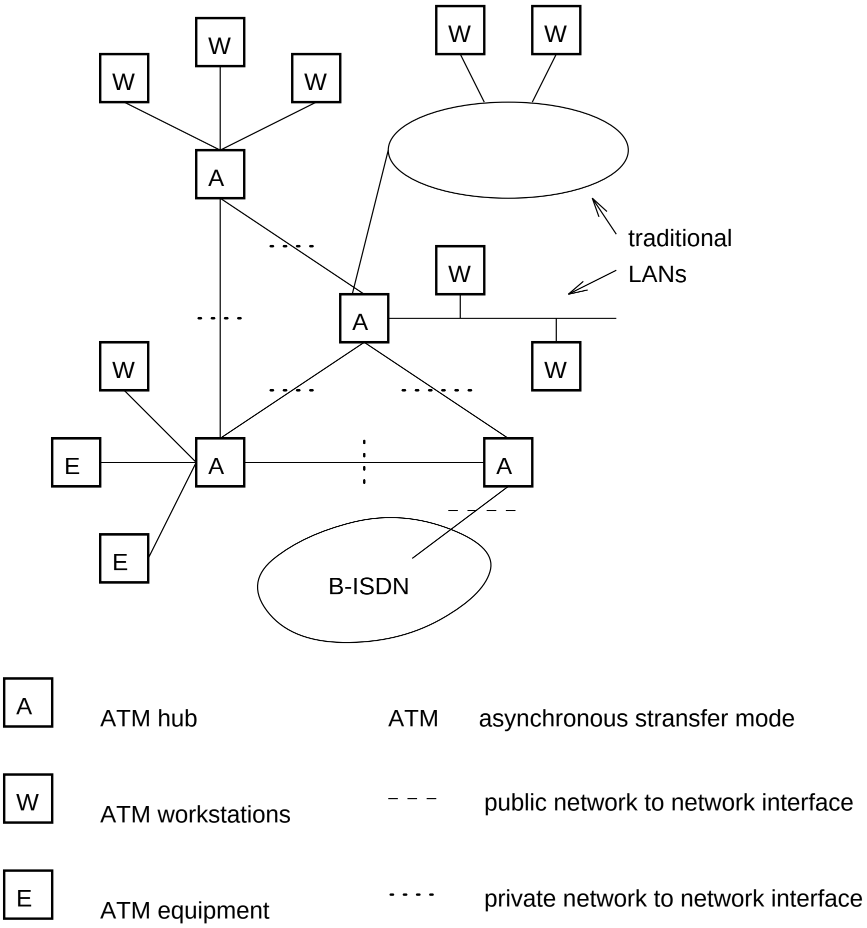

W
W
W
W
W
A
traditional
LANs
W
A
W
W
E
A
A
E
B-ISDN
A
ATM hub
ATM asynchronous stransfer mode
W
ATM workstations
– – – public network to network interface
E
ATM equipment
- - - - private network to network interface

Figure 113: An example ATM LAN configuration

ring and the **secondary ring**. Stations attached to the FDDI may be connected to both rings – **dual attach stations (DASs)** – or only to the primary ring – **single attach station (SASs)**. Although, the stations are logically attached in a ring, the physical connection is more conveniently realised in a hub-star fashion by using wiring concentrators.

The FDDI may be used as a LAN, but is more often used as a backbone and so most of the attached stations will be bridges that are dual attached.

In a token ring, the single active ring monitor encodes the clock signal into the token that it generates, using Manchester encoding. However, for a 100Mb/s rate, Manchester encoding requires 200Mbaud signalling rate. So in an FDDI, NRZI coding is used (signal transition when a 1 is transmitted and no transition when a 0 is transmitted) and all stations have their own **local clock** which is used in data transmission. When receiving frames, stations synchronise using the incoming signal. The physical interface is depicted in Figure 114.

The **4B/5B encoder** takes each group of 4 bits and replaces them with a 5-bit symbol (Table 24), and the **4B/5B decoder** performs the reverse operation. The use of 4B/5B coding and NRZI ensures a signal transition every 2 bits. The **latency buffer** is where 2 5-bit symbols are used to give correct symbol boundary alingment.

The FDDI frame formats are shown in Figure 115. The various fields are as follows (the various control symbols are given in Table 25):

Characteristic	**Information**
Network capacity	100Mb/s
Network configuration	dual counter-rotating rings
Maximum length	100Km
Maximum distance between stations	2Km
Maximum number of stations	1000
MAC method	circulating token
Physical media	mono-mode or multi-mode fibre
Optical wavelength	1300nm
Bit error rate	10^{-9}
Channel coding	4B/5B code
Line coding	NRZI

Table 23: Some characteristics of FDDI

4-bit data	**5-bit symbol**
0000	11110
0001	01001
0010	10100
0011	10101
0100	01010
0101	01011
0110	01110
0111	01111
1000	10010
1001	10011
1010	10110
1011	10111
1100	11010
1101	11011
1110	11100
1111	11101

Table 24: FDDI 4B/5B data symbols

- **Preamble (PA)**: 16 (or more) IDLE symbols. Causes line signal changes every bit to ensure receiver clock synchronisation at the beginning of a frame.

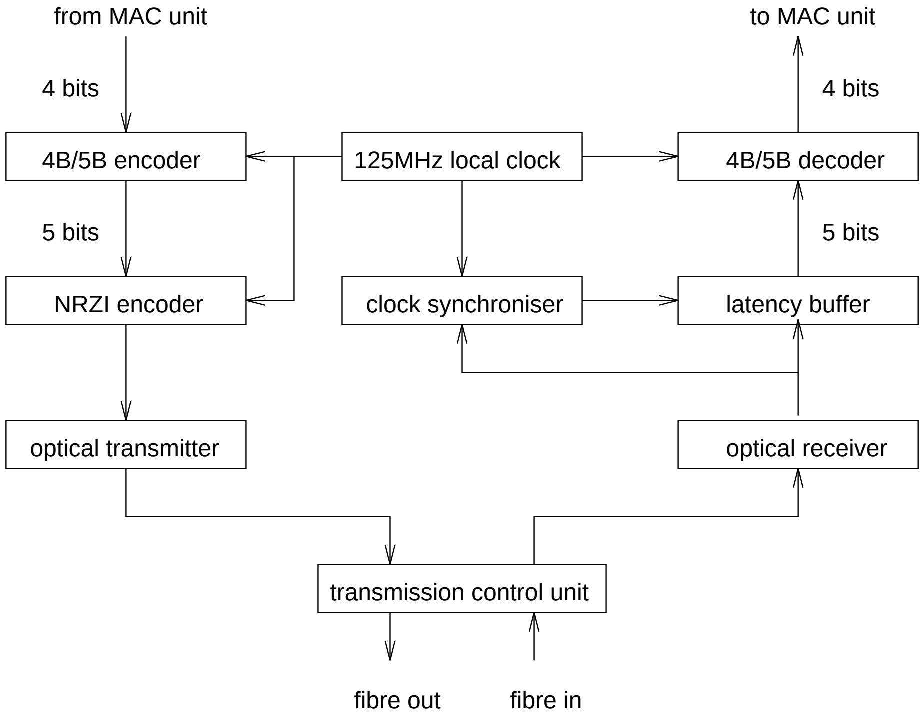
- **Start delimiter (SD)**: the 2 symbols J and K are used to show the start of the frame and also to allow interpretation of correct symbol boundaries.

- **Frame control (FC)**: 2 symbols indicating whether or not this is an information frame or a MAC frame (e.g. the token), with some additional control information for the station identified by the DA.

- **Destination address (DA)**: 4 or 12 symbols identifying the destination station. 12 symbols are used for a full 48-bit MAC address, 4 symbols for a 16-bit local addressing mechanism. If the first bit of the (decoded) address is a 1 then this identifies a group address.

- **Source address (SA)**: 4 or 12 symbols identifying the source station.

- **Information**: this is usually set to about 9000 symbols (4500 decoded octets) in length and is determined by the maximum length of time that a station can hold the token.



MAC medium access control NRZI non-return to zero inverted

Figure 114: Schematic of FDDI physical interface

- **Frame check sequence (FCS)**: 8 symbols containing a 32-bit CRC. The FCS covers the fields FC, DA, SA, information and FCS.
- **End delimiter (ED)**: 1 or 2 T control symbols.
- **Frame status (FS)**: 3 symbols which are a combination of R and S symbols indicating if the frame has been seen by the destination station and if it has been copied by the destination station.

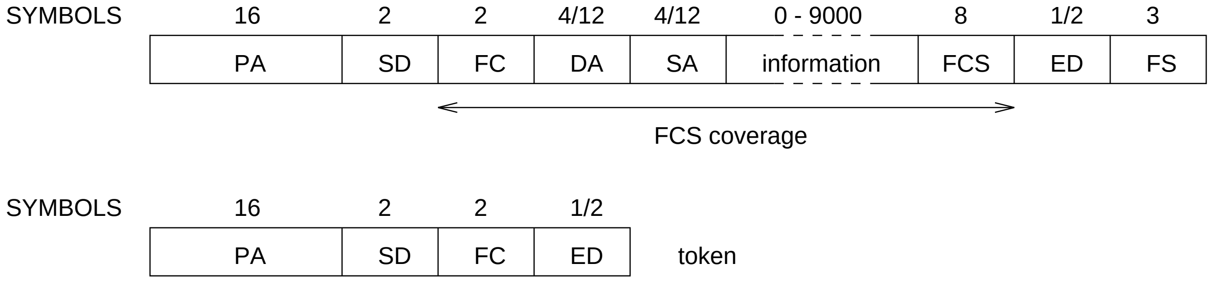
The operation of FDDI is much the same as token ring. A station must be in possession of a token before it can transmit an information frame. Once it has seen the frame go around the ring it can then regenerate the token allowing someone else to transmit. However, the potentially large size of the FDDI ring means that it has a higher latency than token ring and so more than one frame may be circulating around the ring at a given time. The physical interface will then repeat the PA, SD, FC and DA field before it knows if this is its own frame which it should remove from the ring. If this occurs, the station stops sending any more of the frame and instead sends out IDLE symbols until it receives an SD indicating another frame. This will lead to many frame fragments around the ring which should be removed by receiving stations.

All stations also keep a note of the **token rotation time (TRT)** which is the time elapsed since the station last saw the token. As the load on the FDDI network this time will increase. The TRT can be compared to a preset value called the **target token rotation time (TTRT)** to allow a priority operation scheme: only priority frames can be transmitted if the TRT is greater than the TTRT.

A timer called the **token hold timer (THT)** also controls the normal transmission of data. When a station receives the token it transfers the TRT to the THT which starts to count down.

Control symbol	**5-bit symbol**
IDLE	11111
J	11000
K	10001
T	01101
R	00111
S	11001
QUIET	00000
HALT	00100

Table 25: FDDI control code symbols



PA	preamble	SA	source address
SD	start delimiter	FCS	frame check sequence
FC	frame control	ED	end delimiter
DA	destination address	FS	frame status

Figure 115: FDDI frame formats

The station can then continue to transmit frames as long as the THT remains greater than the TTRT. In fact the THT determines the maximum number of octets/symbols that can be sent in one FDDI frame, as the the THT determines the maximum time that a station can remain transmitting data.

FDDI is a connectionless system and is designed for data purposes only. **FDDI-II** is an evolution of FDDI which can be seen as a superset of FDDI. It was developed a standards project that proposed adding isochronous transmission services to FDDI so that it was possible to support real-time applications such as multimedia. The proposal includes a mechanism for dividing the the 100Mb/s into separate channels, each of which would be allocated for a certain use, e.g. video, bursty data, etc. This mode of operation is called **hybrid mode** as opposed to the **basic mode** operation that is effectively FDDI.

The use of the transmission media is now structured as **cycles**, each cycle allowing the transport of traffic as various **wideband channels (WBCs)**. This is depicted in Figure 116. The cycle is 12,500 bits long and has a preamble, header and payload. The payload is split into 12 **dedicated packet groups (DPGs)** and each groups is identified by a single octet at the start of a 128byte block in the payload. The 128byte block is arranged as 16 × 8 byte channels – WBCs. There may be many cycles at one time on the FDDI-II ring.

Cycles are generated by the **cycle master**, who is chosen from all the stations on the ring during an initialistion phase. The cycle master controls utilisation of the channels. Stations negotiate use of a channel with the cycle master using the management subsystem of FDDI as a messaging system. This is akin to setting up a 'connection' before data transfer can take place, however a logical connection does not exist. After this negotiation has been finished, the use of the channels is controlled using the **cycle header** as depicted in Figure 117:

- **Start delimiter (SD)**: indicates start of cycle (same as FDDI).
- **Synchronisation control (C1)**: allows synchronisation of stations. A non-synchronised station sets the symbol R. The cycle master sets the symbol S when all stations have synchronised.
- **Sequence control (C2)**: used in conjunction with C1 when chosing a new cycle master. Set to R during the chosing phase, and then set to S by the cycle master.
- **Cycle sequence (CS)**: takes a value between 0 and 255. The values 0 to 63 are reserved for use during the chosing of the cycle master. Stations write a **monitor value** into this field, and the station which gives the highest value becomes cycle master. Once the cycle master has been chosen, this filed increments monotonically from 64 to 255, wrapping back to 64, acting a cycle identifier.
- **Programming template (P0 – P15)**: this set of 16 symbols indicates the use of each WBC – Pn indicates the use of WBCn. The symbol R in Pn indicates that WBCn is being used for asynchronous data and a symbol S indicates isochronous use.
- **Isochronous maintenance channel**: use not yet defined.

The channels can be allocated for use in such a way that it is possible to have multiples of 8Kb/s channels.

However, because existing FDDI hardware would be unable to have plug-in upgrades and with the advent of ATM LANs and DQDB, this option of FDDI may not be so heavily deployed.

7.7 Distributed Queue Dual Bus (DQDB)

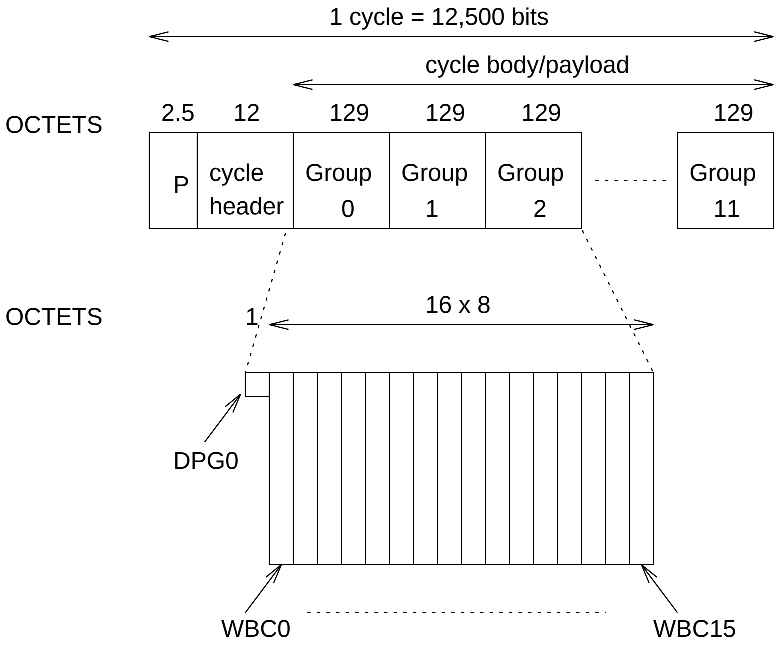
The **distributed queue dual bus (DQDB)** network uses a different kind of MAC method based on the use of a distributed queuing algorithm called **queued-packet distributed-switch (QPSX)** and a slotted ring arrangement. It uses two unconnected unidirectional buses, which are normally implemented as a series of point-to-point segments. DQDB also expects the use of optical fibre links. Some of the characteristics of DQDB are given in Table 26.

The DQDB LAN or MAN transports data in fixed size **cells**, which happen to look very much like ATM cells (Figures 118 and 119). However, as DQDB offers a MAC service, the MAC frame may need to be segmented into several cells before transmission (Figure 120). The cell structure for DQDB is almost identical to that of an ATM cell. This similarity between the cell structure has been made so that the DQDB will be compatible with B-ISDN.

A DQDB cell differs from an ATM cell in its header and its payload. In the header of the DQDB cell there is no virtual path identifier (VPI) and the virtual channel identifier (VCI) has an additional 4 bits. The first 8 bits of the DQDB header form an access control field (ACF) (in ATM this is the first 8 bits of the VPI). Also, the next 4 bits (the final 4 bits of the VPI in ATM) form the first 4 bits of the VCI.

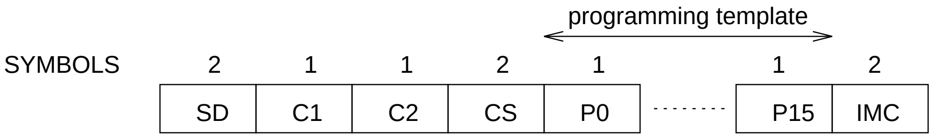
Also, the cell payload is structured:

- **Segment type (ST)**: identifies the cell as one of the following:



P	preamble
DPG	dedicated packet group
WBC	wideband channel

Figure 116: Cycle structure for FDDI-II



SD	start delimiter	CS	cycle sequence
C1	syncronisation control	P	programming symbol
C2	sequence control	IMC	isochronous maintenance channel

Figure 117: FDDI-II cycle header

  - **single segment**: only this segment (no MAC fragmentation was required).
  - **first segment**: the first cell of a segmented MAC frame;
  - **intermediate segment**: the intermediate cells in a fragmented MAC frame.
  - **last segment**: the final cell of a segmented MAC frame¿

- **Message identifier (MID)**: the MID is the same for all DQDB cells from the same MAC frame. This allows the identification of intermediate segments.
- **Information**: (part of) the MAC frame contents.
- **Length (LEN)**: the length of the information field.
- **CRC**: covering everything the whole cell payload.

The cell header contains the following information:

- **Access control field (ACF)**: this contains the BUSY and REQUEST bits that are used in the operation of the QPSX mechanism. The BUSY bit indicates the the slot is in use. The REQUEST bit is set in a slot by a node that is waiting to transmit.
- **Virtual channel identifier (VCI)**: This is not used in a DQDB MAN as there are no logical connections which require multiplexing – the ST and MID fields in the payload are used instead. In a DQDB LAN with a private UNI, there is the possibility for applications to make use of this field.
- **Payload type (PT)**: same as ATM.
- **Cell loss priority (CLP)**: same as ATM.
- **Header error control (HEC)**: CRC for the header.

A DQDB network is comprised of two buses that transmit cells in opposite directions and each node is connected to both buses. The connections are normally point-to-point, but are often depicted in the tapped-bus type configuration as shown in Figure 121. Both these buses always have a constant number of slots circulating on them. A slot on one bus is copied to the other.

The way that the DQDB network operates is by setting up a notion of a distributed queue at each node. This is done by using a counter – REQUEST counter (RC) – that records how many nodes

Characteristic	**Information**
Network capacity	150Mb/s (600Mb/s planned)
Network configuration	two uni-directional buses
Maximum length	160Km
Maximum distance between stations	–
Maximum number of stations	512
MAC method	QPSX with slotted access
Physical media	mono-mode or multi-mode fibre
Optical wavelength	1300nm
Bit error rate	10^{-9}
Channel coding	8B/9B coding
Line coding	NRZI

Table 26: Some characteristics of DQDB

are waiting to transmit ahead of this node in the queue. Consider Figure 122 which shows a node on the DQDB network. Only the inputs from the two buses to the node are shown, for simplicity. By *ahead* in the queue we mean any nodes upstream (relative to the flow of cells/slots) from this node. The count of waiting nodes is incremented by counting any set BUSY bits in passing slots on Bus A. The count is decremented by seeing any set REQUEST bits on Bus B.

For this node, we will consider what happens when it wishes to transmit on Bus A. Firstly, it waits for a passing slot on Bus B with an available REQUEST bit, which it then sets to indicate it is waiting to transmit. It then transfers the RC value to a down counter (DC). DC now contains the number of nodes ahead of this node in the queue. DC is decremented by slots passing by with BUSY bits set, indicating slots/cells used by other nodes that are ahead in the queue. When DC becomes 0, then this node can transmit on Bus A. Although we have only considered Bus A, an identical procedure is followed if the node wishes to transmit on Bus B.

DQDB also has 4 priority transmission levels, which are also based around a distributed queue system using counters, i.e. each node effectively has 5 queues, with a RC and DC for each queue. The priority levels are indicated by 4 further REQUEST bits in the ACF in the cell header, labeled R1, R2, R3 and R4. The highest priority is R4. To cater for the priority levels, the behaviour of RC and DC must be modified as follows:

- RC must count all request bits of its priority level or higher.
- DC decrements if an empty cell passes on Bus A, as this will be used by a waiting node with a higher priority further downstream.
- DC increments if it sees a a REQUEST bit for a higher priority queue on Bus B.

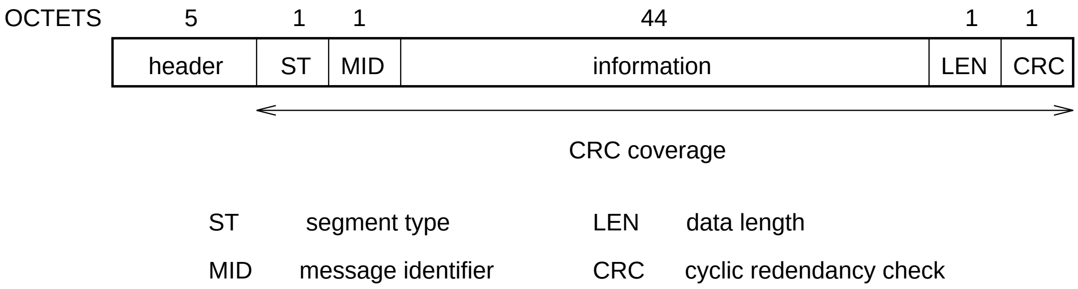
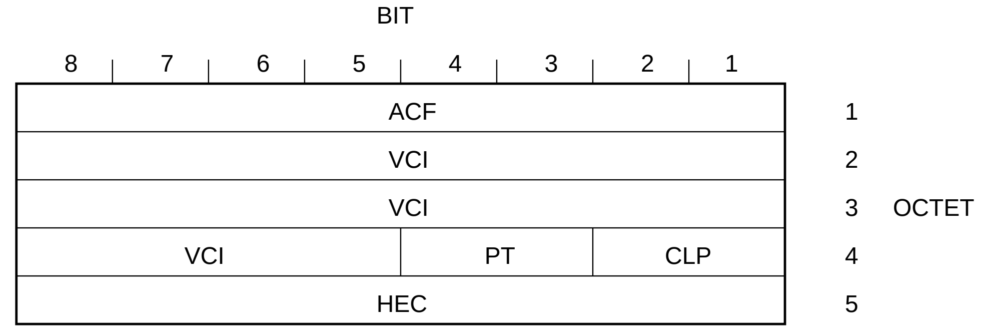
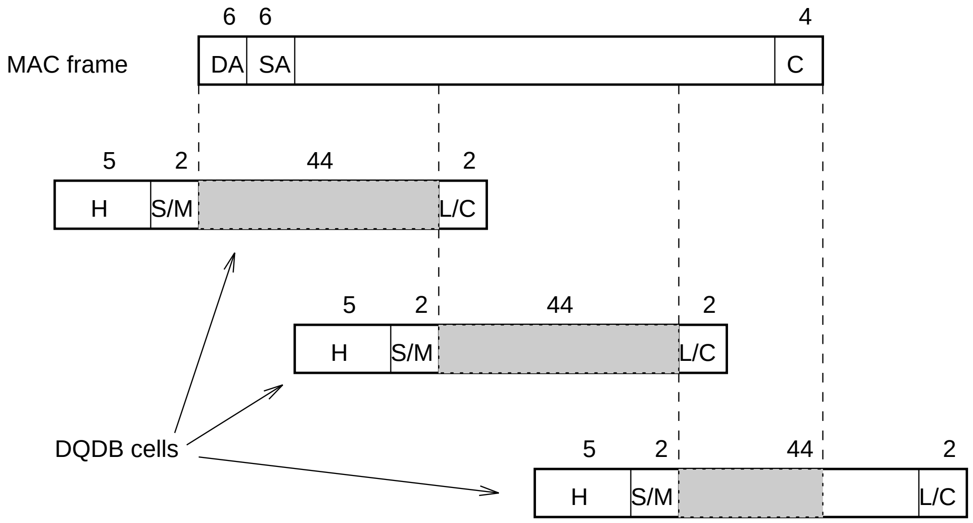


Figure 118: A DQDB cell



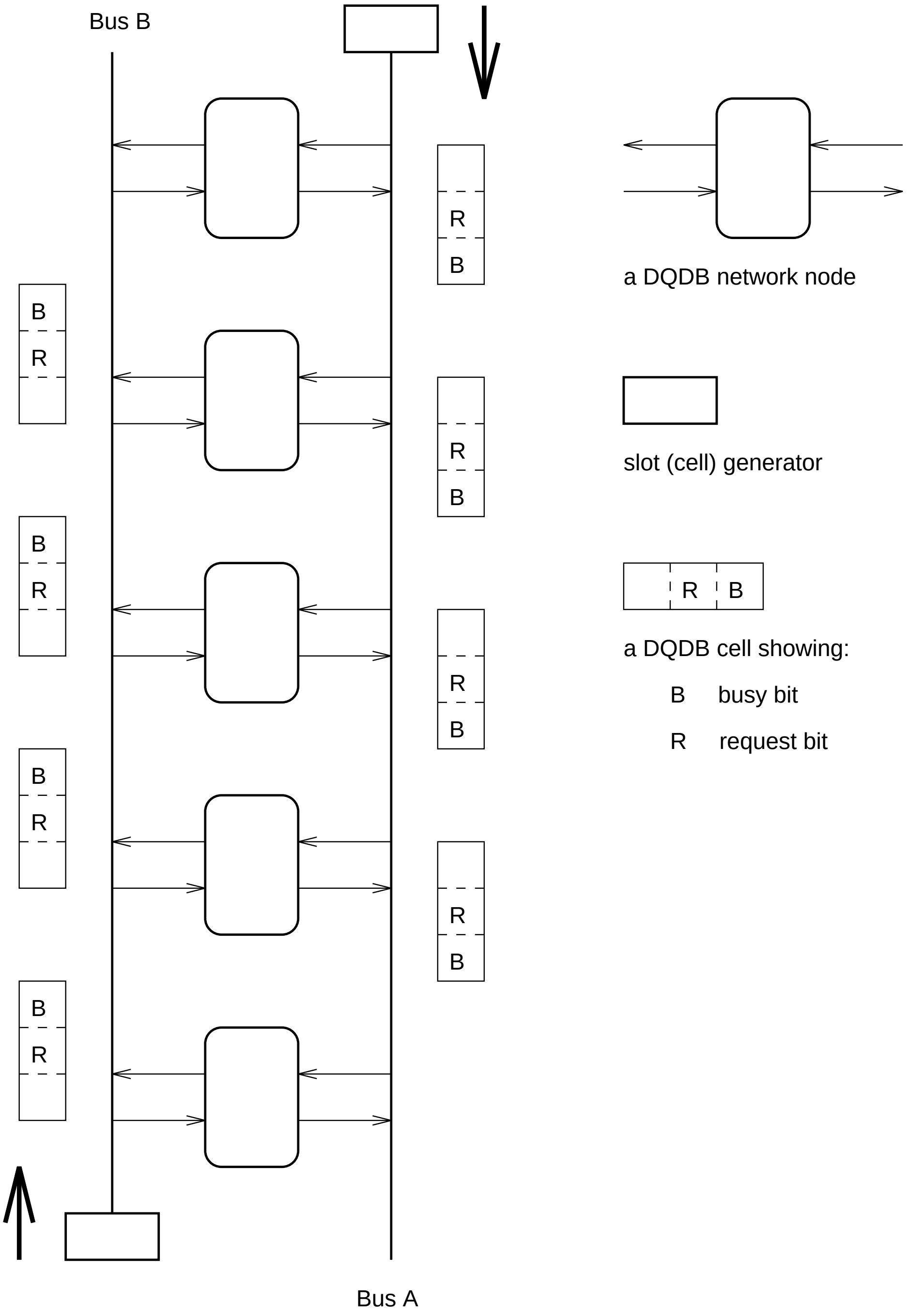
VPI	virtual path identifier	PT	payload type
VCI	virtual channel identifer	CLP	cell loss priority
HEC	header error control	ACF	access control field

Figure 119: A DQDB cell header



MAC	medium access control	H	header
DQDB	distributed queue dual bus	DA	destination address
S/M	segemt type and message identifier	SA	source address
L/C	length and cyclic redundancy check	C	cyclic redundancy check

Figure 120: MAC frame segmentation in DQDB



DQDB distributed queue dual bus

Figure 121: The operation of a DQDB network

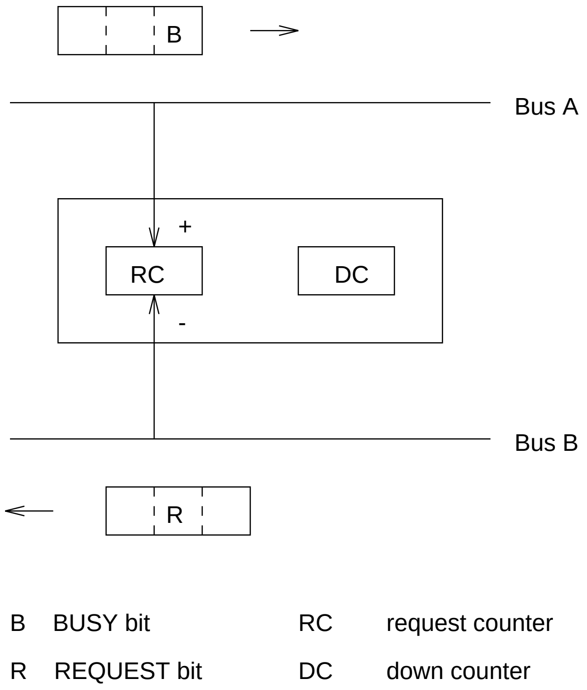


Figure 122: DQDB node collecting queue information about Bus A

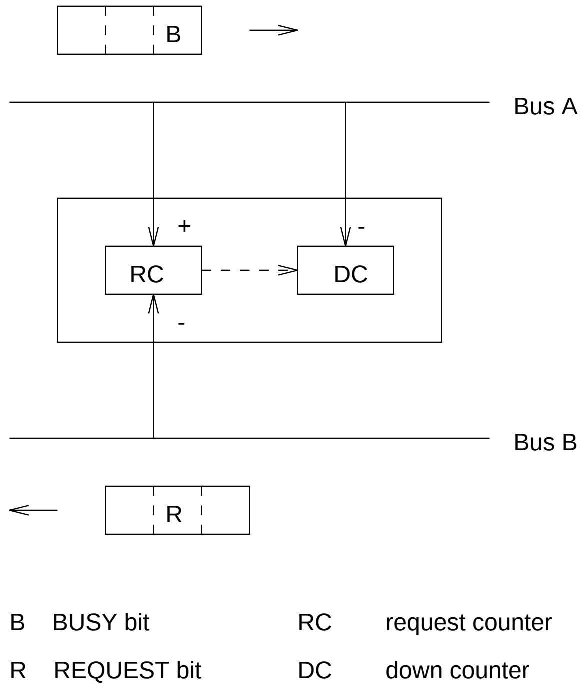


Figure 123: DQDB node waiting to send on Bus A