Module – 2

Chapter 3: Multicarrier Modulation:

- OFDM basics
- OFDM in LTE
- Timing and Frequency Synchronization
- Peak to Average Ratio(PAR)
- Single Carrier-Frequency Division Equalization(SC-FDE)

Chapter 4: OFDMA and SC-FDMA:

- OFDM with FDMA, TDMA, CDMA, OFDMA, SC-FDMA
- OFDMA and SC-FDMA in LTE

Chapter 5: Multiple Antenna Transmission and Reception:

- Spatial Diversity overview
- Receive Diversity
- Transmit Diversity
- Interference cancellation and signal Enhancement
- Spatial Multiplexing, Choice between Diversity
- Interference Suppression and Spatial Multiplexing

Chapter 3: Multicarrier Modulation:

3.1 Introduction:

Name the technologies adopted in multicarrier modulation technique

- Multicarrier modulation used in many of the most successful modern wireless systems, including
  - *Digital Subscriber Lines (DSL).*
  - *Wireless LANs (802.11a/g/n).*
  - *Digital Video Broadcasting.*
  - *Beyond 3C cellular technologies such as WiMAX and LTE*.
- The common feature of multicarrier modulation techniques is the use of *multiple parallel subcarriers*, invariably generated by the Discrete Fourier Transform (DFT).
- The most common type of multicarrier modulation is Orthogonal Frequency Division Multiplexing (OFDM). Other examples Discrete Multi-Tone (DMT)

3.2 The Multicarrier Concept:

Explain the concept of multicarrier modulation.

What are the purpose of using multicarrier modulation?

- The main purpose of using multicarrier modulation to achieve high data rates and mitigate Inter Symbol Interference (ISI) in broadband channels.
- In order to have a channel that does not have ISI the symbol time T_S, has to be much larger than the channel delay spread τ and transmission bandwidth less than coherence bandwidth (C_B)
- **Concept:**
  (Which are the channel condition are necessary to make ISI free channel?)
  1. *To achieve $T_S \gg \tau$, In multicarrier modulation divides the high-rate transmit bit stream into L lower-rate sub-streams, where L is chosen so that each of the subcarriers has effective symbol time T_S L $\gg \tau$ and is hence effectively ISI-free. These individual sub-streams can then be sent over L parallel subcarriers, maintaining the total desired data rate.*
  2. *The data rate on each of the subcarriers is much less than the total flat a rate, and so the corresponding subcarrier bandwidth is much less than the total system bandwidth. The number of sub-streams is chosen to ensure that each subcarrier has a bandwidth less than the coherence bandwidth (C_B) of the channel.*

3.2.1 An Elegant Approach to Inter Symbol Interference:

- Multicarrier modulation divides the wideband incoming data stream into L narrow band sub-streams.
- Each of which is then transmitted over a different orthogonal frequency subcarrier. The number of sub-streams L is chosen to make the symbol time T_s each sub-stream much greater than the delay spread of τ the channel or, equivalently, to make the sub-stream bandwidth less than the channel coherence bandwidth. This ensures that the sub-streams will not experience significant ISI.
- A simple illustration of multicarrier transmitter and receiver is given in Fig 3.1 & 3.2
- *Multicarrier transmitter& Receiver*: In Fig 3.1 a high-rate data signal of rate 'R' bps and with a pass bandwidth 'B' is broken into 'L' parallel sub-streams each with data rate 'R/L' and pass band bandwidth 'B/L'.
- After passing through the channel H(f), the received signal would appear as shown in Figure 3.3, no subcarrier overlap since the subcarrier bandwidth very much smaller than the coherence bandwidth C_B, i,e. $B/L << C_B$, then it can be ensured that each sub-carrier experiences approximately fiat fading.

- The mutually orthogonal signals can then be individually detected, as shown in Figure 3.2.

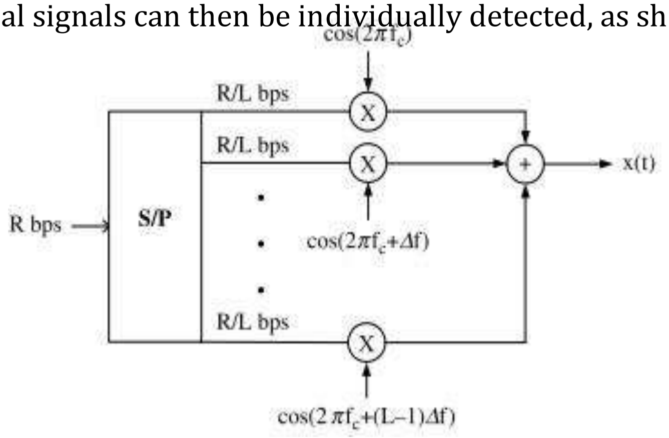


Figure 3.1: A basic multicarrier transmitter*: a high-rate stream of R bps is broken into L parallel sub-streams each with rate R/L and then multiplied by a different carrier frequency*

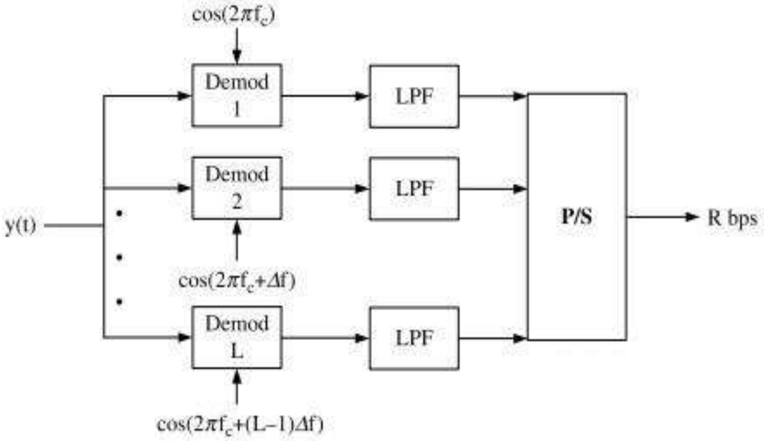


Figure 3.2: A basic multicarrier receiver: *each subcarrier is decoded separately, requiring L independent receivers.*

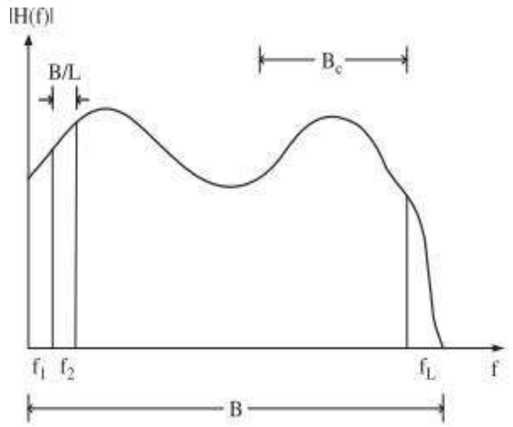


Figure 3.3 The transmitted multicarrier signal experiences approximately flat fading on each sub-carrier since B/L ≪ C_B, even though the overall channel experiences frequency selective fading, that is, B > C_B

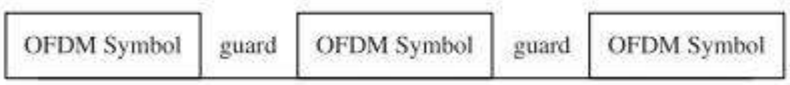
OFDM Basics

- OFDM employs the Fast Fourier Transform (FFT) to achieve ‘L’ RF radios path in both the transmitter and receiver. IFFT are able to create a multitude of orthogonal subcarriers using just a single radio.

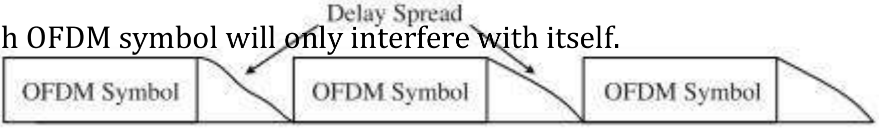
3.2.1 Block Transmission with Guard Intervals:

With respect to OFDM, explain block transmission with guard intervals.

- Grouping ‘L’ data symbols into a block known as an OFDM symbol with a duration of T seconds. Where T = LTs.
- Guard time ‘ T_g ‘ introduce in between OFDM symbol to keep independent of the others after going through a wireless channel as shown below:



- Receiving a series of OFDM symbols, as long as the guard time T_g is larger than the delay spread of the channel τ, each OFDM symbol will only interfere with itself.



- OFDM transmissions allow ISI within an OFDM symbol. But by including a sufficiently large guard band, it is possible to guarantee that there is no interference between subsequent OFDM symbols.

*3.2.2 The Cyclic Prefix (CP)****

What is cyclic prefix? Why is it used? Explain in detail?

- The cyclic prefix acts as a buffer region or guard interval to protect the OFDM signals from ISI.
- The CP is obtained by taking the last v samples from the length N block of OFDM symbols, and it is appended at the start of the symbol block. As a result, the transmitted OFDM symbol block is of length N + v *as shown in* fig 3.4 . For each OFDM symbol to be independent and to avoid any ISI and ICI, the length v of the CP should be at least equal to the channel order.

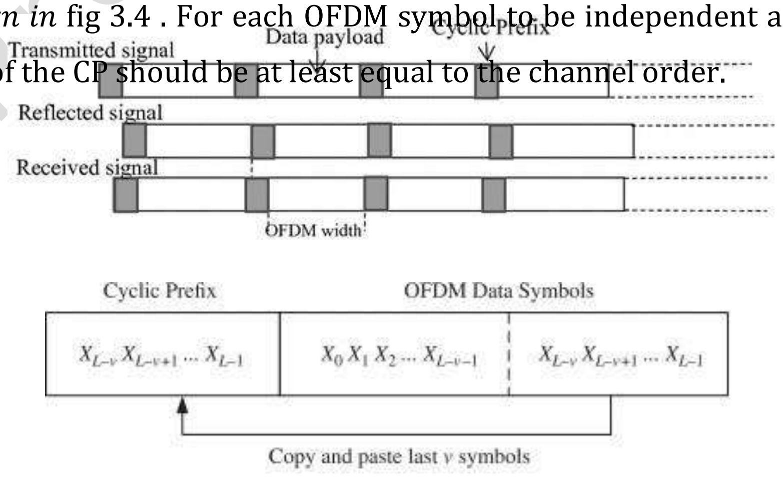

Transmitted signal
Data payload
Cyclic Prefix
Reflected signal
Received signal
OFDM width
Cyclic Prefix
OFDM Data Symbols
$X_{L-v} X_{L-v+1} \dots X_{L-1}$
$X_0 X_1 X_2 \dots X_{L-v-1}$
$X_{L-v} X_{L-v+1} \dots X_{L-1}$
Copy and paste last v symbols

Figure 3.4 The OFDM Cyclic prefix

- The cyclic prefix performs two main functions.
  1. *It provides a guard interval to eliminate ISI from the previous symbol.*
  2. *It repeats the end of the symbol so the linear convolution of a frequency-selective multipath channel can be modeled as circular convolution, which in turn may transform to the frequency domain via a DFT. This approach accommodates simple frequency domain processing, such as channel estimation and equalization.*

- FFT/IFFT algorithms are used to realize OFDM in practice with reduced computational complexity.
- The IFFT operation at the transmitter allows all the subcarriers to be created in the digital domain, and thus requires only a single radio to be used.
- In order for the IFFT/FFT to create an ISI-free channel, the channel must appear to provide a circular convolution.
- If a cyclic prefix is added to the transmitted signal, as shown in Figure 3.4, then this creates a signal that appears to be $x[n]_L$, and so $y[n] = x[n] \circledast h[n]$.
- If the maximum channel delay spread has a duration of v +1 samples, then by adding a guard band of at least v samples between OFDM symbols, each OFDM symbol is made independent of those coming before and after it, and so just a single OFDM symbol can be considered.
- Representing such an OFDM symbol in the time domain as a length L vector gives

$$\boldsymbol{X} = [x_{1,2,}x_{3,} \ldots \ldots \ldots . . x_{L,}] \qquad (3.1)$$

- After applying a cyclic prefix of length v, the actual transmitted signal is

$$\mathbf{x}_{cp} = [\underbrace{x_{L-v}\ x_{L-v+1}\ \cdots\ x_{L-1}}_{\text{Cyclic prefix}}\ \underbrace{x_0\ x_1\ \cdots\ x_{L-1}}_{\text{Original data}}].$$

- The output of the channel is by definitionY_{cp}, = $h * X_{cp}$, where h is a length v + 1 vector describing the impulse response of the channel during the OFDM symbols.
- The output Y_{cp} has samples = Length of OFDM symbol + Length of the channel response - 1

  = (L + v) + (v + 1) - 1

  = L + 2v samples.
- The first v samples of Y_{cp}, contain interference from the preceding OFDM symbol, and so are discarded. The last v samples disperse into the subsequent OFDM symbol, and so also are discarded. This leaves exactly L samples for the desired output $'y\,'$, which is precisely what is required to recover the L data symbols embedded in X.
- These L samples of y will be equivalent to y = h$\circledast$ x.

- The circular convolution operation y[n] = x[n]⊛ h[n] as shown below figure 3.5.

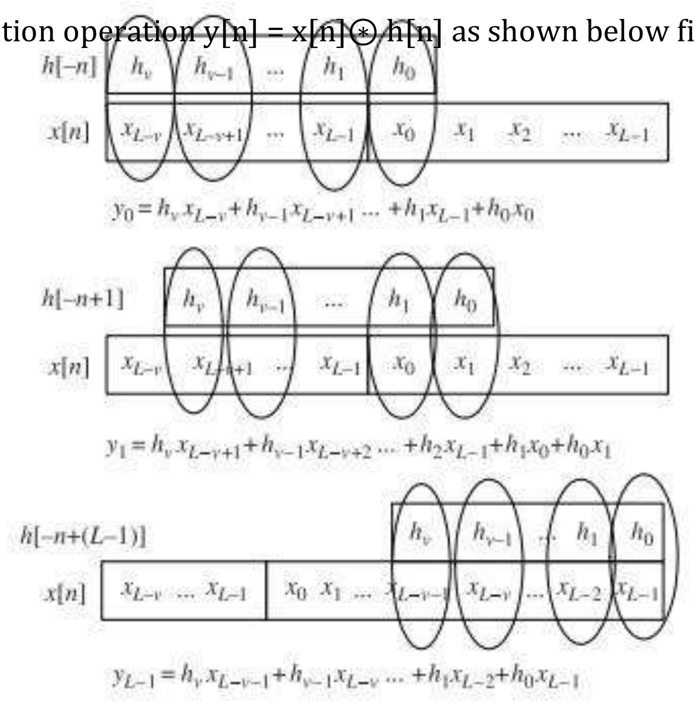

h[−n]
h_v
h_{v-1}
...
h_1
h_0
x[n]
x_{L-v}
x_{L-v+1}
x_{L-1}
x_0
x_1
x_2
x_{L-1}
$y_0 = h_v x_{L-v} + h_{v-1} x_{L-v+1} \cdots + h_1 x_{L-1} + h_0 x_0$
h[−n+1]
$y_1 = h_v x_{L-v+1} + h_{v-1} x_{L-v+2} \cdots + h_2 x_{L-1} + h_1 x_0 + h_0 x_1$
h[−n+(L−1)]
x_{L-v-1}
x_{L-v}
x_{L-2}
$y_{L-1} = h_v x_{L-v-1} + h_{v-1} x_{L-v} \cdots + h_1 x_{L-2} + h_0 x_{L-1}$

Figure 3.5 The OFDM cyclic prefix creates a circular convolution at the receiver (signal y) even though the actual channel causes a linear convolution.

- Due to the cyclic prefix y_0 depends on x_0 and the circularly wrapped values $x_{L-v} \ldots \ldots x_{L-1}$,That is:

$$\begin{aligned} y_0 &= h_0 x_0 + h_1 x_{L-1} + \cdots + h_v x_{L-v} \\ y_1 &= h_0 x_1 + h_1 x_0 + \cdots + h_v x_{L-v+1} \\ &\vdots \\ y_{L-1} &= h_0 x_{L-1} + h_1 x_{L-2} + \cdots + h_v x_{L-v-1} \end{aligned}$$

- Channel output y to be decomposed into a simple multiplication of the channel frequency response H = DFT {h} and the channel frequency domain input X = DFT{x}.
- The drawback of cyclic prefix need more bandwidth and power penalty.
- Since v redundant symbols are sent, the required Bandwidth of OFDM in increase from $B\ to \frac{L+v}{L} B$ and power penalty of $10 log_{10} \frac{L+v}{L}$ dB
- In summary, the use of cyclic prefix entails data rate and power losses that are both

$$Rate\ Loss\ = Power\ Loss = \frac{L+v}{L}$$

- The "wasted" power has increased importance in an interference-limited wireless system, since it causes interference to neighboring users.

3.2.3 Frequency Equalization:

What is frequency equalization in OFDM?

- Equalization is the process of adjusting the balance between frequency components within a received OFDM signal.
- Frequency domain equalizers (FEQs) have been applied extensively in multicarrier systems to enhance transmission rate by reducing transmit redundancy in the form of guard interval.
- Received symbols to be estimated, the complex channel gains for each subcarrier must be known, which corresponds to knowing the amplitude and phase of the subcarrier.
- After the FFT is performed, the data symbols are estimated using a one-tap frequency domain equalizer, or FEQ, as

$$\overline{X_l} = Y_l / H_l$$

Where H_l is the complex response of the channel at the frequency $f_c + (l-1)\,\Delta f$, and therefore it both corrects the phase and equalizes the amplitude before the decision device.

*3.2.5 An OFDM Block Diagram****

With block diagram explain, Multicarrier transmitter and receiver.

- The key steps in an OFDM communication system are briefly shown in Figure 3.6.

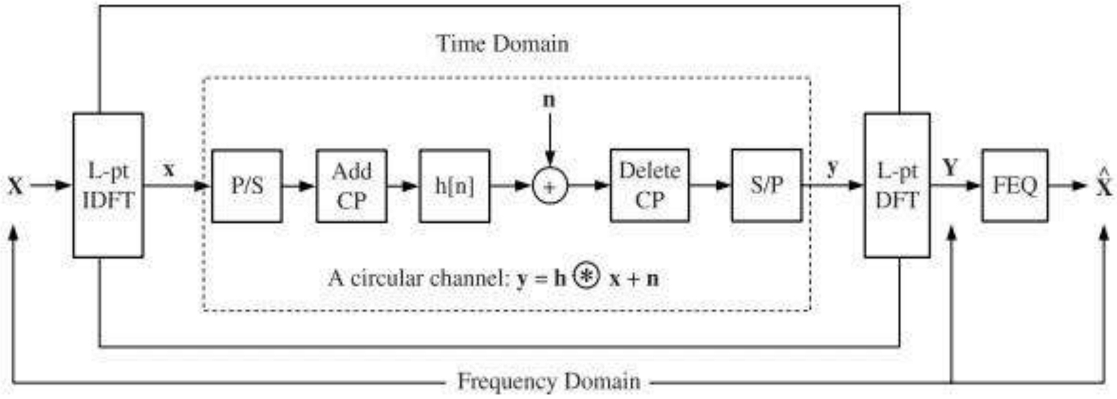


Figure 3.6: An OFDM system in vector notation. In OFDM, the encoding and decoding is done in the frequency domain, where X, Y. and $\hat{X}$contain the L transmitted, received, and estimated data symbols.

- *Transmitter operations:*
  - *Step 1*: In OFDM, break a wideband signal of bandwidth B into L narrowband subcarriers each of bandwidth B/L and each subcarrier experiences flat fading, or ISI-free communication, as long as a cyclic prefix that exceeds the delay spread is used. The L subcarriers for a given OFDM symbol are represented by a vector X, which contains the L current symbols.
  - *Step 2: L* independent narrow band subcarriers are created digitally using an IFFT operation.

- *Step 3:* IFFT/FFT decompose the ISI channel into orthogonal subcarriers, a cyclic prefix of length v must be appended after the IFFT operation. The resulting $L + v$ symbols are then sent in serial through the wideband channel.

- *Receiver operations:*
  - At the receiver, the cyclic prefix is discarded, and the L received symbols are demodulated using an FFT operation, which results in L data symbols, each of the form $Y_l = H_l X_l + N_l$ for subcarrier l.
  - Each subcarrier can then be equalized via an FEQ by simply dividing by the complex channel gain H[i] for that subcarrier. This results in $\tilde{X}_l = X_l + \frac{N_l}{H_l}$

3.3 OFDM in LTE:

With a neat diagram explain OFDM baseband to pass band transmitter in LTE

- LTE systems used as an example to brief time and frequency domain interpretations of OFDM.
- Figure 3.7 shows view of a pass band OFDM modulation engine. The inputs to this figure are L independent QAM symbols (the vector X), and these L symbols are treated as separate subcarriers.

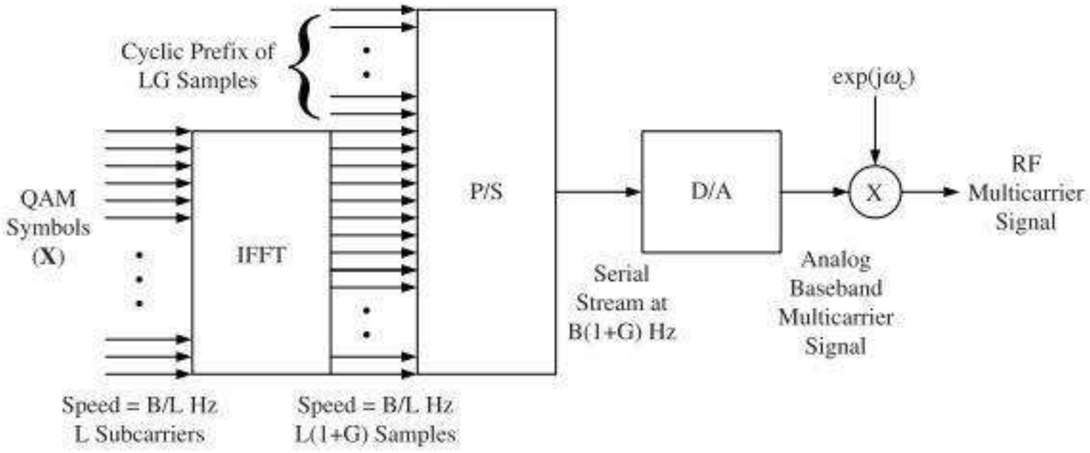
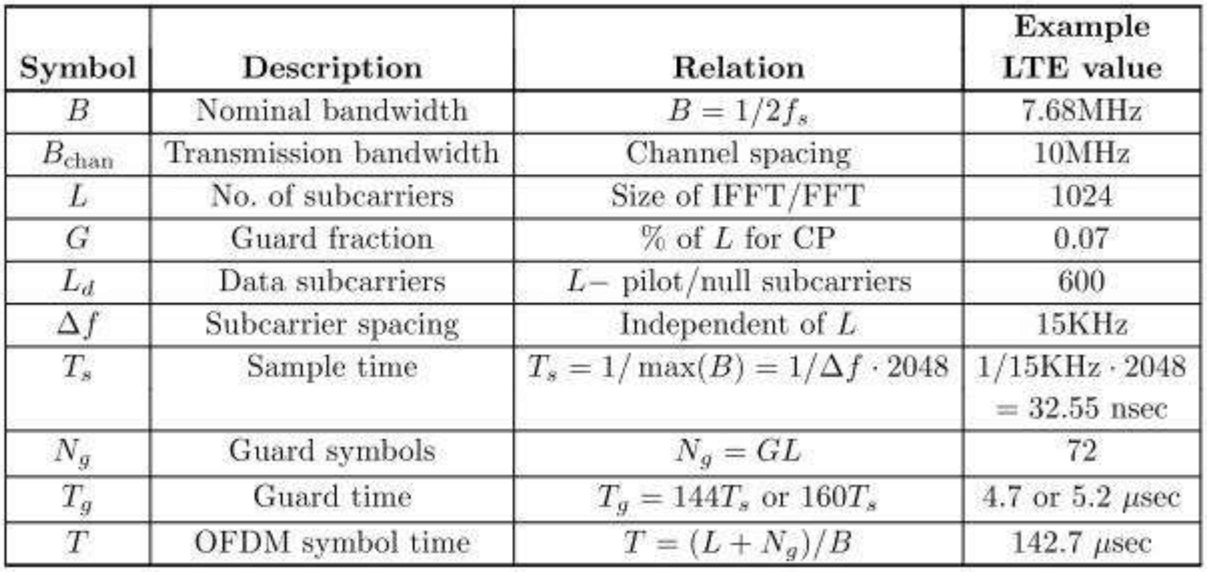


Figure 3.7: A close-up of the OFDM baseband to pass band transmitter.

- These L data-bearing symbols can be created from a bit stream by a symbol mapper and serial-to-parallel convertor (S/P).
- The L-point IFFT then creates a time domain L-vector x that is cyclic extended to length L(1 + G), where G is the fractional overhead. In LTE G 0.07 for the normal cyclic prefix and G = 0.25 for the extended cyclic prefix.
- This longer vector is then parallel-to-serial (P/S) converted into a wideband digital signal that can be amplitude modulated with a single radio at a carrier frequency of $f_c = \omega_c/2\pi$.
- The key OFDM parameters are summarized in table below

Table 3.1 Summary of Key OFDM Parameters in LTE and Example Values for 10MHz



Symbol	Description	Relation	Example LTE value
B	Nominal bandwidth	$B = 1/2f_s$	7.68MHz
B_{chan}	Transmission bandwidth	Channel spacing	10MHz
L	No. of subcarriers	Size of IFFT/FFT	1024
G	Guard fraction	% of L for CP	0.07
L_d	Data subcarriers	$L-$ pilot/null subcarriers	600
Δf	Subcarrier spacing	Independent of L	15KHz
T_s	Sample time	$T_s = 1/\max(B) = 1/\Delta f \cdot 2048$	$1/15\text{KHz} \cdot 2048$ $= 32.55$ nsec
N_g	Guard symbols	$N_g = GL$	72
T_g	Guard time	$T_g = 144T_s$ or $160T_s$	4.7 or 5.2 μsec
T	OFDM symbol time	$T = (L + N_g)/B$	142.7 μsec

For example, if 16QAM modulation was used *(M* = 16) with the normal cyclic prefix, the raw (neglecting coding) data rate of this LTE system would be:

$$\begin{aligned} R &= \frac{B}{L}\frac{L_d \log_2(M)}{1+G} \\ &= \frac{10^7\,\text{MHz}}{1024}\frac{600\log_2(16)}{1.07} = 21.9\,\text{Mbps}. \end{aligned}$$

3.4 Timing and Frequency Synchronization:

Define timing and frequency synchronization and explain in brief

- Synchronization of an OFDM signal is required to find the symbol timing and carrier frequency offset (CFO).
- In order to demodulate an OFDM signal, there are two important synchronization tasks that need to be performed by the receiver
- First, the timing offset of the symbol and the optimal timing instants need to be determined. This is referred to as timing synchronization.
- Second, the receiver must align its carrier frequency as closely as possible with the transmitted carrier frequency. This is referred to as frequency synchronization.
- Figure 3.8 shows a representation of an OFDM symbol in time (top) and frequency (bottom).
- In the time domain, the IFFT effectively modulates each data symbol onto a unique carrier frequency.
- In Figure 3.8 only two of the carriers are shown: the actual transmitted signal is the superposition of all the individual carriers.

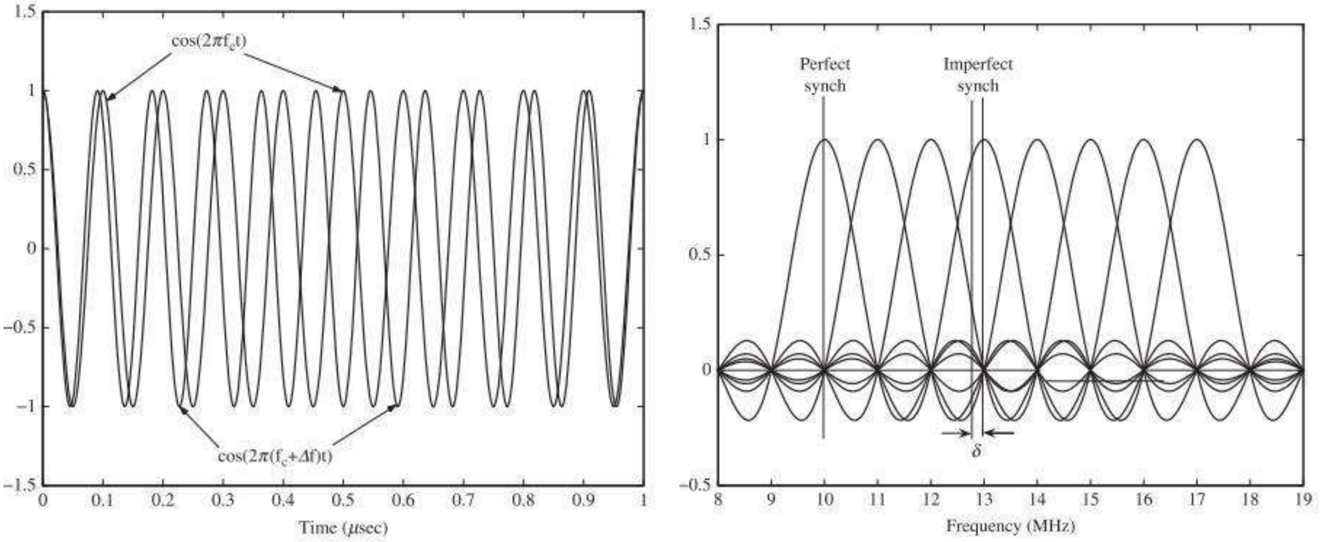


Figure 3.8 OFDM synchronization in time (top) and frequency (bottom). Here, two subcarriers in the time domain and eight subcarriers in the frequency domain are shown, where fc=10MHz and the subcarrier spacing Δf = 1Hz.

- In above figure time window size is T = 1 μsec and it has frequency response of each subcarrier becomes a "sine function with zero crossings every 1/T = 1MHz. This frequency response is shown for L = 8 subcarriers in the right part of Figure 3.8.
- *The challenge of timing and frequency synchronization*: If the timing window is slid to the left or right, a unique phase change will be introduced to each of the sub-carriers. It result carrier frequency is misaligned by some amountδ, then some of the desired energy is lost, and it is referred to as Inter-Carrier *I*nterference (ICI).
- The following two subsections will provide solution good timing and frequency synchronization algorithms for LTE systems. Synchronization is one of the most challenging problems in OFDM implementation.

3.4.1 Timing Synchronization:

- The effect of timing errors in symbol synchronization is relaxed in OFDM due to the presence of a cyclic prefix.
- If the cyclic prefix length N$_g$is equivalent to the length of the channel impulse response ν, successive OFDM symbols can be decoded ISI free.
- The tolerable a timing offset of τ seconds without any degradation in performance as long as 0 $\ll \tau \ll$ ($T_g - T_m$), where $T_g$the guard is time (cyclic prefix duration) and $T_m$is the maximum channel delay spread.
- As long as τ remains constant, it includes a fixed phase offset and it can be corrected by the FEQ without loss or performance.

- This acceptable range of τ is referred to as the timing synchronization margin, and is shown in Figure 3.9.

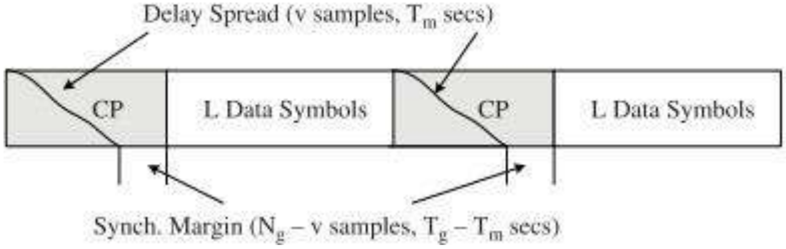

Delay Spread (v samples, T_m secs)
CP
L Data Symbols
CP
L Data Symbols
Synch. Margin (N_g – v samples, $T_g - T_m$ secs)

Figure 3.9 Timing synchronization margin.

(Write the SNR expression for timing synchronization and list the important performance observation of the expression.)

- If the timing offset τ is not within this window $0 \ll \tau \ll (T_g - \ )$, ISI occurs. The desired energy is lost while interference from the preceding symbol is included in the receive window. For both of these scenarios, the SNR loss can be approximated by

$$\Delta SNR(\tau) \approx -2\left(\frac{\tau}{LT_s}\right)^2$$

- Important observations from this expression are
  - SNR decreases quadratically with the timing offset.
  - Longer OFDM symbols are increasingly immune from timing offset, that is, more subcarriers helps.
  - Since in general $\ll L\ T_s$, timing synchronization errors are not that critical as long as the induced phase change is corrected.

- *Conclusion*: To minimize SNR loss due to imperfect timing synchronization, the timing errors should be kept small compared to the guard interval, and a small margin in the cyclic prefix length is helpful.

3.4.2 Frequency Synchronization:

Write the equation for SNR loss due to frequency offset and write your remarks.

- OFDM achieves a high degree of bandwidth efficiency compared to other wideband systems.
- In OFDM, the subcarrier packing is extremely tight compared to conventional modulation techniques, which require a guard band on the order of 50% or more.
- Frequency offsets is very sensitive in OFDM due to the fact that the subcarriers overlap, rather than having each subcarrier truly spectrally isolated.
- The zero crossings of the frequency domain sine pulses all line up as seen in Figure 3.8, as long as

the frequency offset δ= 0, there is no interference between the subcarriers.

- In practice, of course, the frequency offset is not always zero. The major causes for this are
  - *Mismatched oscillators at the transmitter and receiver*
  - *Doppler frequency shifts due to mobility*
  - *Precise crystal oscillators are expensive, tolerating some degree of frequency offset is essential in a consumer OFDM system like LTE*

  - Hence the received samples of the FFT will contain interference from the adjacent subcarriers, called inter-carrier interference (ICI) and it effect on OFDM performance.
  - The matched filter receiver corresponding to subcarrier l can be simply expressed for the case of rectangular windows (neglecting the carrier frequency) as

$$x_l(t) = X_l e^{j\frac{2\pi lt}{LT_s}}, \tag{3.17}$$

where $1/LT_s = \Delta f$, and again LT_s is the duration of the data portion of the OFDM symbol, that is, $T = T_g + LT_s$. An interfering subcarrier m can be written as

$$x_{l+m}(t) = X_m e^{j\frac{2\pi(l+m)t}{LT_s}}. \tag{3.18}$$

If the signal is demodulated with a fractional frequency offset of δ, $|\delta| \leq \frac{1}{2}$

$$\hat{x}_{l+m}(t) = X_m e^{j\frac{2\pi(l+m+\delta)t}{LT_s}}. \tag{3.19}$$

The ICI between subcarriers l and $l+m$ using a matched filter (i.e., the FFT) is simply the inner product between them:

$$I_m = \int_0^{LT_s} x_l(t)\hat{x}_{l+m}(t)dt = \frac{LT_s X_m \left(1 - e^{-j2\pi(\delta+m)}\right)}{j2\pi(m+\delta)}. \tag{3.20}$$

It can be seen that in the above expression, $\delta = 0 \Rightarrow I_m = 0$, and $m = 0 \Rightarrow I_m = 0$, as expected. The total average ICI energy per symbol on subcarrier l is then

$$ICI_l = E\left[\sum_{m \neq l} |I_m|^2\right] \approx C_0(LT_s\delta)^2\mathcal{E}_x, \tag{3.21}$$

where C_0 is a constant that depends on various assumptions and $\mathcal{E}_x$ is the average symbol energy [23, 33]. The approximation sign is due to the fact that this expression assumes that there are an infinite number of interfering subcarriers. Since the interference falls off quickly with m, this assumption is very accurate for subcarriers near the middle of the band, and is pessimistic by a factor of 2 at either end of the band.

The SNR loss induced by frequency offset is given by

$$\begin{aligned} \Delta SNR &= \frac{\mathcal{E}_x/N_o}{\mathcal{E}_x/\left(N_o + C_0(LT_s\delta)^2\mathcal{E}_x\right)} && (3.22) \\ &= 1 + C_0(LT_s\delta)^2 SNR && (3.23) \end{aligned}$$

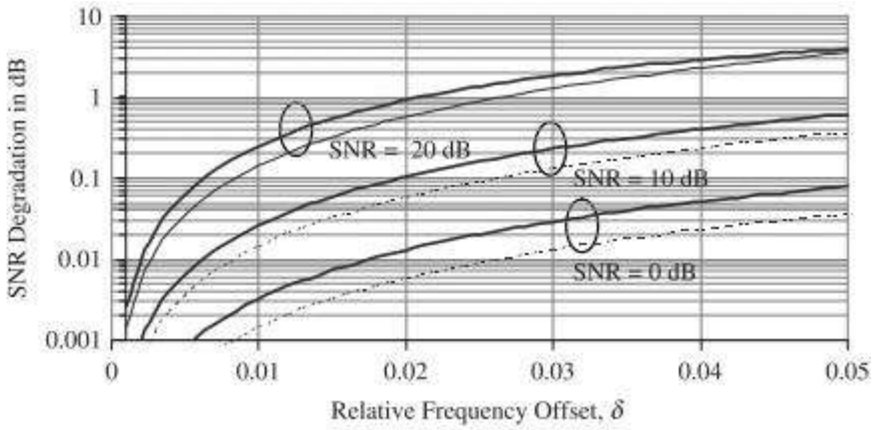


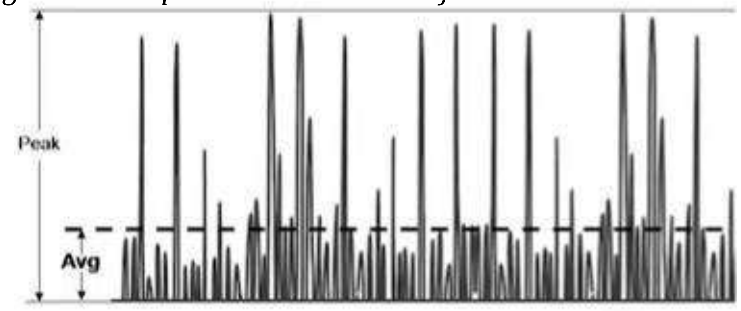
Figure 3.10: SNR loss as a function of the frequency offset 8, relative to the subcarrier spacing. The solid lines are for a fading channel and the dotted lines are for an AWGN channel.

- Important observations from the ICI expression (3.23) and Figure 3.10 are that:
  - *SNR decreases quadratically with the frequency offset.*
  - *SNR decreases quadratically with the number of subcarriers.*
  - *The loss in SNR is also proportional to the SNR itself.*
- In order to keep the loss negligible, say less than 0.1 dB, the relative frequency offset needs to be about 1-2% of the subcarrier spacing, or even lower to preserve high SNRs.
- Therefore, this is a case where reducing the CP overhead by increasing the number of subcarriers causes an offsetting penalty, introducing a tradeoff.
- In order to further reduce the ICI for a given choice of L, non-rectangular windows can also be used.

3.5 The Peak-to-Average Power Ratio (PAPR) ***

Define PAPR. Explain the PAPR problems? How PAR can be reduced using clipping?

- Definition: *The PAPR is the ratio the maximum power of a sample in a given OFDM transmit symbol to the average power of that OFDM symbol. In simple terms, PAPR is the ratio of peak power to the average power of a signal. It is expressed in the units of dB*.



PAPR of 10 dB means that for transmitting an average power of 0.2 W, the transmitter should be able to handle power peaks of 2.0 W (10 time higher). The result is a very low efficiency or in other words: **high battery power consumption.**

- PAPR occurs when in a multicarrier system the different sub-carriers are out of phase with each other.
- OFDM signals have a higher peak-to-average ratio (PAPR). This high PAR is one of the most important implementation challenges that face OFDM because it reduces the efficiency and hence increases the cost of the RF power amplifier, which is one of the most expensive components in the LTE transmitter.
- Alternatively, the same power amplifier (PA) can be used but the input power to the PA must be reduced: this is known as input backoff (IBO) and results in a lower average SNR at the receiver, and hence a reduced transmit range.

3.5.1 The PAR Problem:

- When a high-peak signal is transmitted through a nonlinear device such as a high-power amplifier (HPA) or digital-to-analog converter (DAC), it generates out-of-band energy and in- band distortion. These degradations may affect the system performance severely.
- The nonlinear behavior of HPA can be characterized by amplitude modulation/amplitude modulation (AM/AM) and amplitude modulation/phase modulation (AM/PM) responses.
- Figure 3.11 shows a typical AM/AM response for an HPA, with the associated input and output backoff regions. IBO and OBO, respectively.

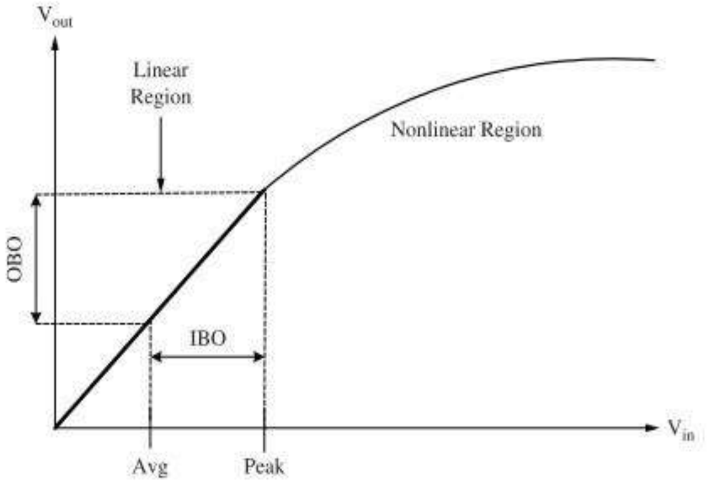


Figure 3.11: A typical power amplifier response.

- Operation in the linear region is required in order to avoid distortion, so the peak value must be constrained to be in this region, which means that on average, the power amplifier is underutilized by a "backoff" amount.
- To avoid the undesirable nonlinear effects just mentioned, a waveform with high-peak power must be transmitted in the linear region of the HPA by decreasing the average power of the input signal. This is called input backoff (IBO) and results in a proportional output backoff (OBO).

- High backoff reduces the power efficiency of the HPA, and may limit the battery life for mobile applications.
- In addition to inefficiency in terms of power, the coverage range is reduced and the cost of the HPA is higher than would be mandated by the average power requirements.
- The input backoff is defined as

$$IBO = 10log_{10} \frac{P_{inSat}}{P_{in}}$$

Where P_{inSat} is the saturation power and P_{in} is the average input power.

- The amount of backoff is usually greater than or equal to the PAR of the signal.
- The power efficiency of an HPA can be increased by reducing the PAR of the transmitted signal. It would be desirable to have the average and peak values be as close together as possible in order to maximize the efficiency of the power amplifier.
- In addition to the large burden placed on the HPA, a high PAR requires high resolution for both the transmitter's DAC and the receiver's ADC, since the dynamic range of the signal is proportional to the PAR.
- High-resolution D/A & A/D conversion places an additional complexity, cost, and power burden on the system.

3.5.2 Quantifying the PAR:

Explain quantifying of PAR.

- The OFDM carries L narrowband signals. In particular, each of the L output samples from an L-point IFFT operation involves the sum of L complex numbers, the resulting output values $\{x_1, x_2, \ldots, x_L\}$ can be accurately modelled and the amplitude of the output signal is

$$|x[n]| = \sqrt{(\Re\{x[n]\})^2 + (\Im\{x[n]\})^2}$$

Where $\Re$ and $\Im$ give the real and imaginary parts. Since x[n] is complex Gaussian, the output power is

$$|x[n]|^2 = (\Re\{x[n]\})^2 + (\Im\{x[n]\})^2$$

- Which is exponentially distributed with mean $2\sigma^2$. The important thing to note is that the output amplitude and hence power are random, so the PAR is not a deterministic quantity either.
- The PAR of the transmitted analog signal can be defined as

$$PAR \triangleq \frac{\max_t |x(t)|^2}{E[|x(t)|^2]},$$

- The discrete-time PAR can be defined for the IFFT output as

$$PAR \triangleq \frac{\max_{l \in (0, L+N_g)} |x_l|^2}{E[|x_l|^2]} = \frac{\mathcal{E}_{\max}}{\mathcal{E}_x}.$$

- The maximum possible value of the PAR is L or 10 $\log_{10}$ L dB, which would occur if all the subcarriers add up constructively at a single point.

3.5.3 Clipping and Other PAR Reduction Techniques:

- ***Clipping Techniques:***
  - In this technique "clip" off the highest peaks, at the cost of some hopefully minimal distortion of the signal. Second and conversely, it can be seen that even for a conservative choice of IBO, say 10 dB, there is still a distinct possibility that a given OFDM symbol will have a PAR that exceeds the IBO and causes clipping.
  - Clipping, sometimes called "soft limiting," truncates the amplitude of signals that exceed the clipping level as

$$\tilde{x}[n] = \begin{cases} Ae^{j\angle x[n]}, & \text{if } |x[n]| > A \\ x[n], & \text{if } |x[n]| \le A, \end{cases}$$

Where x (n) is the original signal and $\tilde{x}(n)$ is the output after clipping, and A is the clipping level, that is, the maximum output envelope value. The clipping ratio can be used as a metric and is defined as

$$\gamma \triangleq \frac{A}{\sqrt{E\{|x[n]|^2\}}} = \frac{A}{\sqrt{\mathcal{E}_x}}$$

- *Conclusion*:
  - Clipping reduces the PAR at the expense of distorting the desired signal.
  - The two primary drawbacks from clipping are
    1. *Spectral regrowth (frequency domain leakage), which causes unacceptable interference to users in neighboring RF channels,*
    2. *Distortion of the desired signal.*
- ***Spectral Regrowth:***
  - It is frequency domain leakage noise due to clipping. The clipping noise can be expressed in the frequency domain through the use of the DFT.
  - The resulting clipped frequency domain signal $\tilde{X}$ is

$$\tilde{X} = X_k + C_k \qquad\qquad K = 0 \ \ldots\ldots\ldots\ldots\ldots\ldots L-1$$

Where C_k represents the clipped off signal in the frequency domain.

- In Figure 3.14, the power spectral density of the original (X), clipped ($\tilde{X}$), and clipped-off (C) signals are plotted for different clipping ratios 7 of 3, 5, and 7 dB.

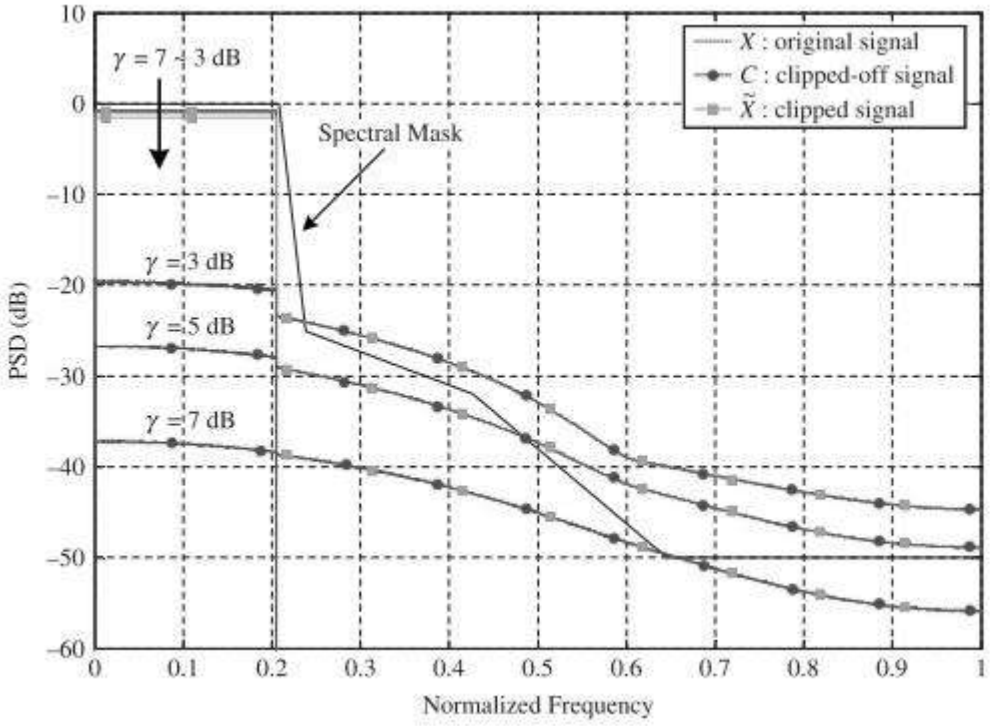


Figure 3.14 Power spectral density (PSD) of the unclipped (original) and clipped (nonlinearly distorted) OFDM signals with 2048 block size and 64 QAM when clipping ratio (γ) is 3, 5, and 7 dB in soft limiter

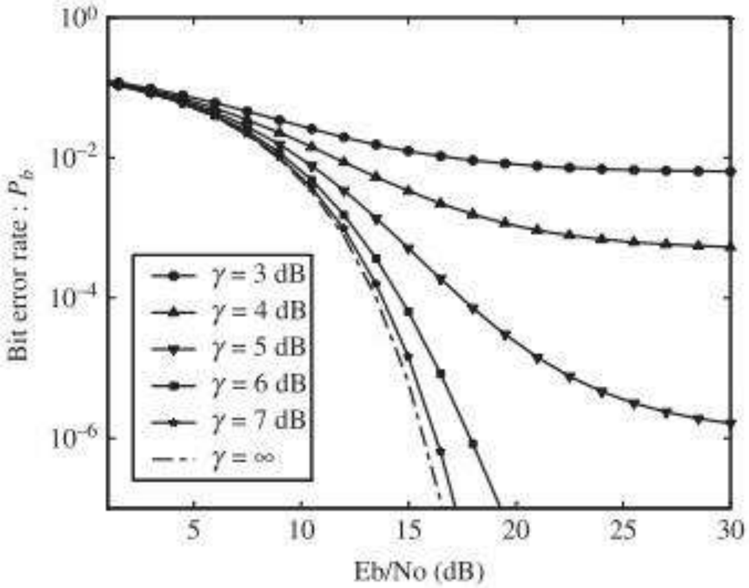
- The following deleterious effects are observed.
  1. The clipped-off signal C_k is strikingly increased as the clipping ratio is lowered from 7 dB to 3 dB.
  2. This increase shows the correlation between X_k, and C_k inside the desired band at low clipping ratios, and causes the in-band signal to be attenuated as the clipping ratio is lowered.
  3. It can be seen that the out-of-band interference caused by the clipped signal X is determined by the shape of clipped-off signal C_k.

3.5.4 LTE's Approach to PAR in the Uplink:

Explain LTE's approach to PAR with graph

- PAR is less important because the base stations are fewer in number and generally higher in cost, and so are not especially sensitive to the exact PAR.
- If the PAR is still considered to be too high, a number of techniques can be utilized to bring it down, all with some complexity and performance tradeoffs. Typically, the high PAR is basically tolerated and sufficient input power backoff is undertaken in order to keep the in-band distortion and spectral regrowth at an acceptable level.

- Figure below Bit error rate probability for a clipped OFDM signal in AWGN with different clipping ratios.



3.6 Single-Carrier Frequency Domain Equalization (SC-FDE)

- SC-FDE maintains OFDM's three most important benefits:

  (1) Low complexity even for severe multipath channels

  (2) Excellent BER performance, close to theoretical bounds.

  (3) Decoupling of ISI from other types of interference, notably spatial interference, which is very useful when using multiple antenna transmission.

- By utilizing single-carrier transmission, the peak-to-average ratio is also reduced significantly (by several dB) relative to multicarrier modulation.

3.6.1 SC-FDE System Description

Explain SC-FDE with block diagram.

- The block diagrams for OFDM and SC-FDE are compared in Figure 3.17
- IFFT is moved to the end of the receive chain rather than operating at the transmitter, to create a multicarrier waveform as in OFDM.
- An SC-FDE system still utilizes a cyclic prefix at least as long as the channel delay spread, but now the transmitted signal is simply a sequence of QAM symbols, which have low PAR, on the order of 4-5 dB depending on the constellation size.
- Considering that an unmodulated sine wave has a PAR of 3 dB, it is clear that the PAR cannot be lowered much below that of an SC-FDE system.

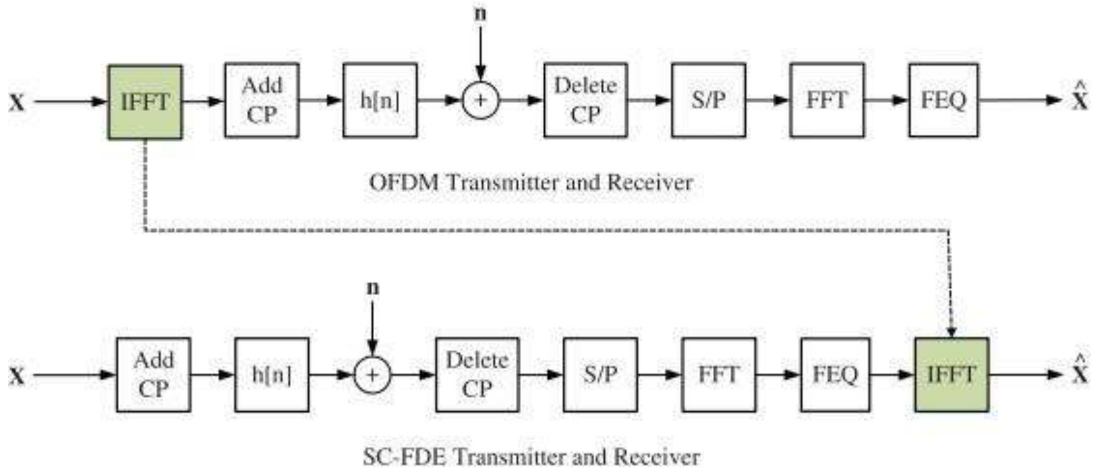


Figure 3.17 Comparison between an OFDM system and an SC-FDE system. The principle difference is that the IFFT formerly in the transmitter is in the SC-FDE receiver

- As in an OFDM system, an FFT is applied, but in an SC-FDE system this operation moves the received signal into the frequency domain.
- Because of the application of the cyclic prefix, the received signal appears to be circularly convolved, that is, y[n] = x[n]⊛ h[n] + w[n], where w[n] is noise. Therefore,

$$FFT\{\,y(n)\} \triangleq Y[m] = H[m]X[m] + \ W[m]$$

- After the FFT, a simple 1-tap FEQ can be applied that inverts each virtual subcarrier, so that

$$\tilde{X}[m] = \frac{Y\,[m]}{H\,[m]}$$

- Use IFFT operation to obtain resulting signal back into the time domain using ie x[n], which are estimates of the desired data symbols. Naturally, in practice H[m] must be estimated at the receiver using pilot signals or other standard methods.

3.6.2 SC-FDE Performance vs. OFDM:

Differentiate the performances of SC-FDE and OFDM.

SLNo.	OFDM	SC-FDE
1.	*OFDM provides high performance*	*Relatively less performance*
2.	*The high Peak-to-Average Power Ratio (PAPR) associated with OFDM*	*The low Peak-to-Average Power Ratio (PAPR) associated with SC-FDE*
3.	*SNR ratio of each data symbol is doesn't change by multiplying constant factor at receiver.*	*SNR ratio of each data symbol is change by multiplying constant factor receiver.*
4.	*OFDM has a nominally less dispersive spectrum.*	*SC-FDE has a nominally more dispersive spectrum.*

5.	*OFDM's sharper spectrum results in less CCI and/or less restrictive RF roll-off requirements.*	*Due to dispersive spectrum results more co-channel interference and/or more restrictive RF roll-off requirements.*
6.	*In OFDM, short-scale variations in SNR would generally be addressed by coding and interleaving.*	*In SC-FDE, however, the FEQ does not operate on data symbols themselves but rather on the frequency domain dual of the data symbols*
7.	*The noise amplification is isolated, hence it does not affects all the symbols prior to decoding and detection*.	*The noise amplification is not isolated to a single symbol in SC-FDE, but instead affects all the symbols prior to decoding and detection.*
8.	*On the whole, OFDM continues to be much more popular than SC-FDE*	*SC-FDE less popular than OFDM.*

3.6.3 Design Considerations for SC-FDE and OFDM:

Differentiate between design considerations of SC-FDE and OFDM

SLNo.	OFDM	SC-FD
1	*In General OFDM is more complex*	*Relatively less complex*
2	*It has a lower-complexity receiver*	*It has higher-complexity receiver*
3	*It has medium-complexity transmitter*	*It has lower-complexity transmitter*
4	*LTE downlink could utilize OFDM*	*LTE uplink could utilize SC-FDE*
5	*The Base Station as transmitter would perform 3 IFFT/FFT operations*	*It perform only a single FFT operation at receiver*
6	*The PAR is high in OFDM and high cost and more power requirement.*	*It benefits of reduced PAR and the reduced cost and power savings.*
7	*The channel estimation and synchronization are accomplished via a preamble of known data symbols, and then pilot tones.*	*It include preamble is in the time domain so it is not as straightforward to estimate the frequency domain values.*
8	*The preamble can be inserted at known positions in all subsequent OFDM symbols*	*It is not possible to insert pilot tones on a per frame basis. Hence it uses DFT and IFFT at the transmitter*
9	*OFDM has a nominally less dispersive spectrum.*	*SC-FDE has a nominally more dispersive spectrum.*
10	*OFDM's sharper spectrum results in less co-channel interference and/or less restrictive RF roll-off requirements.*	*Due to dispersive spectrum results more co-channel interference and/or more restrictive RF roll-off requirements.*
11	*The combination of OFDM with MIMO is a natural and best combination for performance improvement in fading channel.*	*The combination of SC-FDE with MIMO is not as natural because detection cannot be done in the frequency domain. Not possible to use maximum likelihood detection for MIMO with SC-FDE*
12	*On the whole, OFDM continues to be much more popular than SC-FDE*	*SC-FDE less popular than OFDM.*

Module – 2

Chapter 4: Frequency Domain Multiple Access: OFDMA and SC-FDMA

4.1 Introduction: Following multiple access strategy used in cellular communication.

- *First Generation (IG, example AMPS :* ***FDMA***
- *Second Generation (2G) example GSM or IS-54:* ***TDMA, CDMA****.*
- *Third Generation (3G), example UMTS:* ***WCDMA.***
- *Fourth Generation (4G), example LTE:* ***OFDMA for down link, SC-FDMA for uplink****.*

4.2 Multiple Access for OFDM Systems

- OFDM has wide acceptance in wireless communications as an appropriate broadband modulation scheme.
- OFDM divides a wideband frequency-selective channel into narrowband flat fading sub-channels.
- In multi-user systems, these sub-channels can be allocated among different users to provide multiple access schemes
- The use of adaptive techniques in these sub-channels can further increase the spectral efficiency of the wireless system.
- Therefore, a main advantages of OFDM is the flexibility in combining adaptive modulation and multiple access techniques

4.1.1 Multiple Access Overview

List the different types of multiple accesses methods

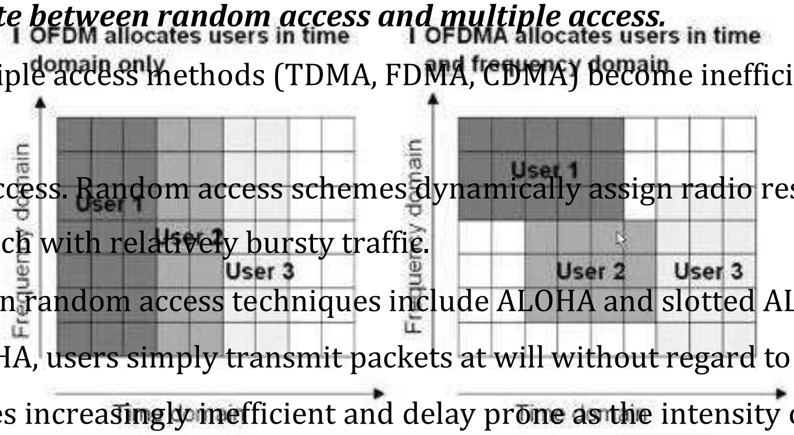
- Multiple-access strategies typically attempt to provide non-interfering, communication channels for each active base station-subscriber link.
- The most common ways to divide the available channel among the multiple users is through
  1. *Frequency Division Multiple Access (FDMA)*: Each user receives a unique carrier frequency and bandwidth.
  2. *Time Division Multiple Access (TDMA):* Each user is given a unique time slot, either on demand or in a fixed rotation.
  3. *Orthogonal Code Division Multiple Access (CDMA):* Systems allow each user to share both the bandwidth and time slots with many other users.
- TDMA, FDMA, and orthogonal CDMA all have the almost same theoretical capacity in an additive noise channel.
- *Limitation of above multiple access*:
  - FDMA, TDMA, CDMA are bandwidth or interference limited system*.*
  - Orthogonally is not possible in dense wireless systems.

- The above techniques only guarantee orthogonality between users in the same cell.
- Different multiple access techniques have different delay characteristics and so may be appropriate for different types of data.

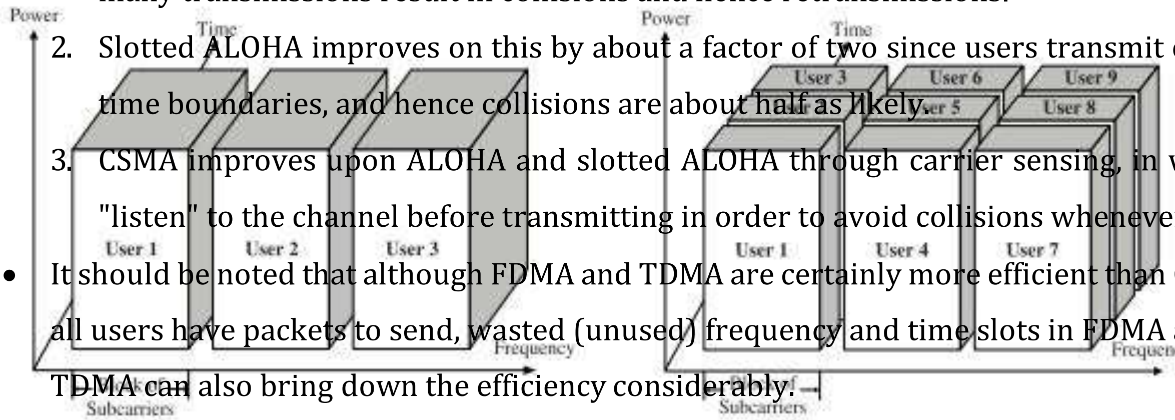
- *Conclusion*: The above limitation of conventional multiple access can be mitigated by principle merits of OFDMA.

4.1.2 Random Access vs. Multiple Access

Differentiate between random access and multiple access.


I OFDM allocates users in time domain only
I OFDMA allocates users in time and frequency domain
Frequency domain
User 1
User 2
User 3
Time domain
Frequency domain
User 1
User 2
User 3
Time domain

- Fixed multiple access methods (TDMA, FDMA, CDMA) become inefficient when the traffic is bursty.
- Random Access. Random access schemes dynamically assign radio resources to a large set of users, each with relatively bursty traffic.
- Well-known random access techniques include ALOHA and slotted ALOHA and CSMA.
  1. In ALOHA, users simply transmit packets at will without regard to other users. This scheme becomes increasingly inefficient and delay prone as the intensity of the traffic increases, as many transmissions result in collisions and hence retransmissions.
  2. Slotted ALOHA improves on this by about a factor of two since users transmit on specified time boundaries, and hence collisions are about half as likely.
  3. CSMA improves upon ALOHA and slotted ALOHA through carrier sensing, in which users "listen" to the channel before transmitting in order to avoid collisions whenever possible.



- It should be noted that although FDMA and TDMA are certainly more efficient than CSMA when all users have packets to send, wasted (unused) frequency and time slots in FDMA and TDMA can also bring down the efficiency considerably.
- In fact, around half the bandwidth is typically wasted in TDMA and FDMA voice systems.
- CDMA system has proven so successful for voice.
- The efficiency of a connection-oriented MAC can approach 90%, compared to at best 50% or less in most CSMA wireless systems such as 802.11.
- *Conclusion:* The need for extremely high spectral efficiency and low delay in LTE make impossible to use of CSMA, and the burden of resource assignment is placed on the base stations.

Explain three multiple access techniques for OFDM systems.

There are three fundamental multi-carrier based multiple access techniques for OFDM systems:

1. ***OFDM-FDMA***
2. ***OFDM-TDMA***
3. ***OFDM-CDMA***

4.1.3 Frequency Division Multiple Access (OFDM-FDMA)

- Frequency Division Multiple Access (FDMA) can be readily implemented in OFDM systems by assigning different users their own sets of subcarriers.
- Available sub-carriers are distributed among all the users for transmission at any time instant
- Each user is allocated a pre-determined band of subcarriers. Allows adaptive techniques per sub-carrier, based on sub-channel condition.

Figure 4.1 FDMA (left) and a combination of FDMA with TDMA (right).

- The simplest method is a static allocation of subcarriers to each user, as shown on the left of Figure 4.1. For example, in a 64-subcarrier OFDM system, user 1 could take subcarriers 1-16, with users 2, 3, and 4 using subcarriers 17-32, 33-48, and 49-64, respectively.
- The allocations are enforced with a multiplexer for the various users before the IFFT operation.
- OFDMA in LTE, however, has explicit time-sharing and procedures to allow for the dynamic allocation of subcarriers.
- In LTE use dynamic subcarrier allocation based upon channel state conditions. For example, due to frequency selective fading, user 1 may have relatively good channels on subcarriers 33-48, while user 3 might have good channels on subcarriers 1-16. Obviously, it would be mutually beneficial for these users to swap the static allocations.

4.1.4 Time Division Multiple Access (OFDM-TDMA)

- A particular user is given all the sub-carrier of the system for any particular symbol duration.

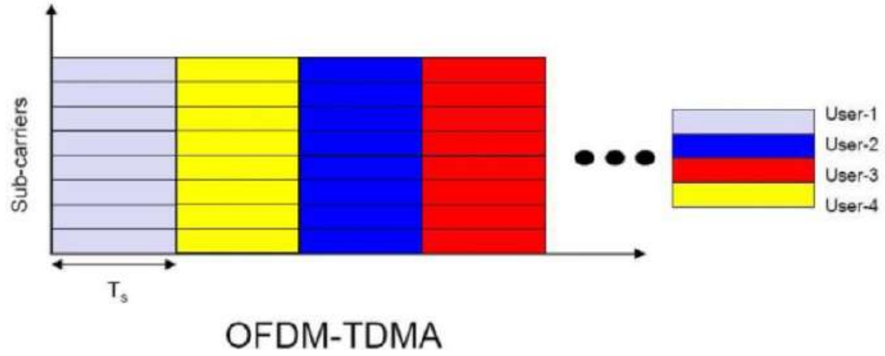


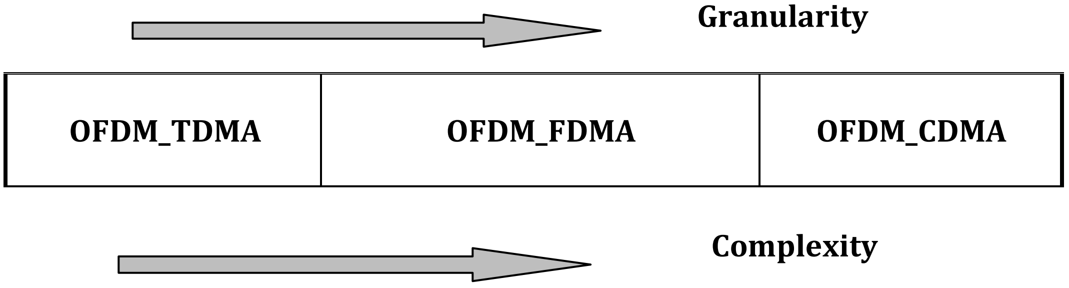
Figure 4.2: Combination of FDMA with TDMA (right).

- Each user is assigned a time slot during which all the sub-carriers can be used for the particular user
- Adaptive loading can be performed on all the subcarriers, depending on channel conditions.
- The number of symbols per frame can be varied based on each user’s requirement.
- Power consumption reduction (less activity). Degrading performance should be taken into account in delay constrained systems.
- A packet-based system like LTE can employ more sophisticated scheduling algorithms based on queue-lengths, channel conditions, and delay constraints to achieve much better performance than static TDMA.

4.1.5 Code Division Multiple Access (OFDM-CDMA or MC-CDMA)

- User data is spread over several sub-carriers and/or OFDM symbols using spreading codes, and combined with signals from other users.
- Hybrid access scheme that combines benefits:
  1. OFDM: *Provides a simple method to overcome the ISI effect of the multi-path frequency selective channel*
  2. CDMA: *Provides frequency diversity and multi-user access scheme*
- Several users transmit over the same sub-carriers.
- In wireless broadband networks the data rates already are very large, so spreading the spectrum further is not viable.
- OFDM and CDMA are not fundamentally incompatible; they can be combined to create a Multicarrier CDMA (MC-CDMA) waveform. MC-CDMA is not part of the LTE standard.

- ***Comparison between OFDM_TDMA, OFDM_FDMA, OFDM_CDMA***



Multiple types	Advantages	Disadvantages
OFDM_TDMA	• Simple implementation • Flexibility	• Frequency-reuse factor ≥ 3
OFDM_FDMA	• Power savings Simple resource allocation • Easiest to implement	• Relatively high latency • Frequency-reuse factor ≥ 3 • Lowest flexibility
OFDM_CDMA	• Spectral efficiency • Frequency diversity • MAI and ICI interference resistance • Frequency-reuse factor = 1 • Highest flexibility	• Requirement of power control • Implementation complexity

4.2 Orthogonal Frequency Division Multiple Access (OFDMA)

- OFDMA systems allocate subscribers time-frequency slices (in LTE, "resource grids").
- A resource block (RB) is the smallest unit of resources that can be allocated to a user.
- It consisting of M subcarriers over some number of consecutive OFDM symbols in time.
- The M subcarriers can either be
  1. *Spread out over the band*: It often called a "distributed," "comb," or "diversity" allocation. The distributed allocation achieves frequency diversity over the entire band, and would typically rely on interleaving and coding to correct errors caused by poor subcarriers. In a highly mobile system, then a distributed allocation would typically be preferred in order to maximize diversity.
  2. *Bunched together in M contiguous subcarriers*: Which is often called a "band AMC," "localized," or "grouped" cluster. The band AMC mode, instead attempts to use subcarriers where the SINR is roughly equal and to choose the best coding and modulation scheme for that SINR. If accurate SINR information can be obtained at the receiver about each band's SINR, then band AMC outperforms distributed subcarrier allocation.

- Table 4.1 summarizes the notation used in the explanation

Table 4.1 OFDMA Notation

K	*Number of active users*
L	*Total number of subcarriers*
M_k, M	*Number of subcarriers per active user k*
$h_{k,l}$	*Envelope of channel gain for user k in subcarrier l*
$P_{k,l}$	*Transmit power allocated for user k in subcarrier l*
σ^2	*AWGN power spectrum density*
P_{tot}	*Total transmit power available at the base station*
B	*Total transmission bandwidth*

4.2.1 OFDMA: How It Works: What is OFDMA? Explain the working of OFDMA with block diagrams of uplink and downlink transmitters and receivers.

- The block diagram for a downlink OFDMA system is shown in Figures 4.3 and 4.4.

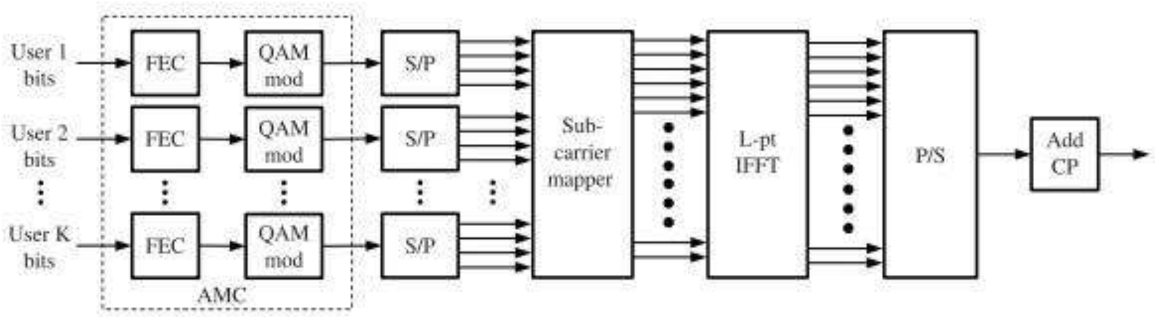


Figure 4.3 OFDMA downlink transmitter.

- The basic flow is very similar to an OFDM system except for now K users share the L subcarriers, with each user being allocated M_k subcarriers.
- In theory it is possible
- to have users share subcarriers, this never occurs in practice, so $\sum_k M_k = L$ and each subcarrier only has one user assigned to it.

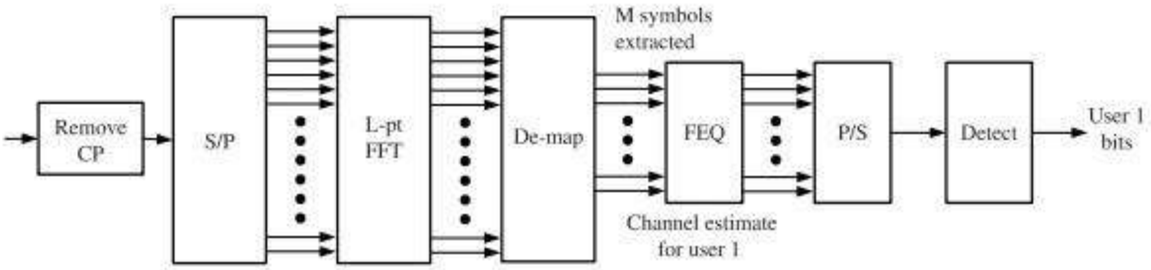


Figure 4.4 OFDMA downlink receiver for user 1. Each of the K active users—who by design have orthogonal subcarrier assignments—have a different receiver that only detects the Mk subcarriers intended for it

- At each receiver, the user cares only about its own M_k subcarriers, but still has to apply an L point FFT to the received digital waveform in order to extract the desired subset of subcarriers.

- Receiver has to know which time-frequency resources it has been allocated in order to extract the correct subcarriers: the control signaling that achieves.
- OFDMA downlink receiver must mostly demodulate the entire waveform, which wastes power, but digital separation of users is simple to enforce at the receiver and the amount of residual inter user interference is very low compared to either CDMA or FDMA.
- OFDMA uplink block diagrams in Figures 4.5 and 4.6 to clearly show the differences and numerous similarities between OFDMA and SC-FDMA.

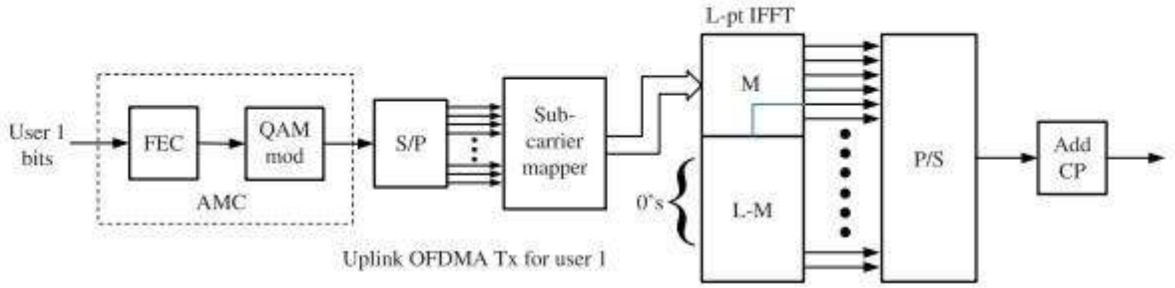


Figure 4.5 OFDMA uplink transmitter for user 1, where user 1 is allocated subcarriers 1, 2..............M of L total subcarriers.

- The transmitter modulates user $k's$ bits over just the M_k subcarriers of interest: in this case, we have chosen $M_k = M$ for all users, and shown user 1 occupying subcarriers 1,2, • • • , M of the L total subcarriers. All the users' signals collide at the receiver's antenna, and are collectively demodulated using the receiver's FFT.
- Assuming each subcarrier has only a single user on it, the demodulated subcarriers can be de-mapped to the detectors for each of the K served users.

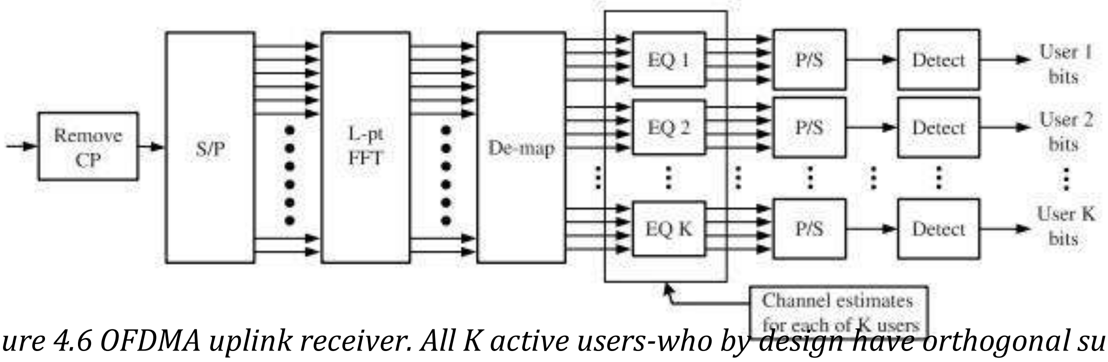


Figure 4.6 OFDMA uplink receiver. All K active users-who by design have orthogonal subcarrier assignments—are aggregated at the receiver and demultiplexed after the FFT.

- It should be noted that uplink OFDMA is considerably more challenging than downlink OFDMA since the uplink is naturally asynchronous, that is the users' signals arrive at the receiver offset slightly in time (and frequency) from each other.

- This is not the case in the downlink since the transmitter is common for all users. These time and frequency offsets can result in considerable self-interference if they become large.
- Particularly in the distributed subcarrier mode, sufficiently large frequency offsets can severely degrade the orthogonality across all subcarriers.
- The timing offsets also must typically be small, within a fraction of a cyclic prefix.
- In LTE the uplink multi access scheme uses only the localized subcarrier mode due to the SC-FDMA nature of the uplink.
- In this case, the lack of perfect frequency and time synchronization between the multiple users leads to some ICI but this is limited only to the subcarriers at the edge of the transmission band of each user.
- Frequency and timing synchronization for the uplink is achieved relative to the downlink synchronization, which is done using the synchronization channels.
- A higher level view of OFDMA can be seen in Figure 4.7. Here, a base station is transmitting a band AMC-type OFDMA waveform to four different devices simultaneously.

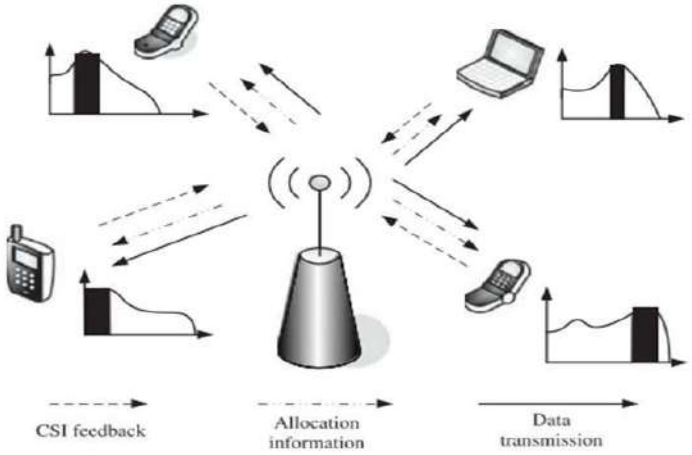


Figure 4.7 In OFDMA, the base station allocates each user a fraction of the subcarriers, preferably, in a range where they have a strong channel.

- The three arrows for each user indicate the signaling that must happen in order for band AMC-type OFDMA to work.
  - First, the mobiles measure and feedback the quality of their channel, or channel state information (CSI) to the base station.
  - Usually, the CSI feedback would be a measurement corresponding to SINR. The base station would then allocate subcarriers to the four users and send that subcarrier allocation information to the four users in an overhead message.
  - Finally, the actual data is transmitted over the subcarriers assigned to each user.
- Here, it can be seen that the base station was successful in assigning each user a portion of the spectrum where it had a relatively strong signal.

4.2.2 OFDMA Advantages and Disadvantages

List the advantages and disadvantages of OFDMA.

- *Advantages of OFDMA*:
  1. OFDMA is a flexible multiple access technique that can accommodate many users with widely varying applications, data rates, and QoS requirements.
  2. OFDMA provide robust multipath suppression, relatively low complexity, and the creation of frequency diversity.
  3. Multiple access is performed in the digital domain, dynamic, flexible, and efficient bandwidth allocation is possible.
  4. Lower data rates (such as voice) and bursty data are handled much more efficiently in OFDMA than in single-user OFDM (i.e., OFDM-TDMA) or with CSMA.
  5. It makes receiver simple as it eliminates intra-cell interference which avoids multi- user detection of CDMA type. Here only FFT processing is needed.
  6. Fading environment leads to better BER performance.
- *Disadvantages of OFDMA*
  1. Since the switching between users would have to be very rapid, more frequency overhead signaling would be required, reducing the overall system throughput.
  2. The permutation and depermuation rules of subcarriers for allocation and deallocation to sub channels are complex. This makes transmitter and receiver algorithms complex for data processing/extraction unlike OFDM.
  3. OFDMA has higher PAPR (Peak to Average Power Ratio). Hence large amplitude variations lead to increase of in-band noise.
  4. OFDMA requires very tight time/frequency/channel equalizations between users. This is achieved with the help of preamble, pilot signals and other signal processing techniques.
  5. Co-channel interference is more complex compare to CDMA technique.

4.3 Single-Carrier Frequency Division Multiple Access (SC-FDMA)

- SC-FDMA is employed in the LTE uplink.
- Conceptually, this system evolves naturally from SC-FDE modulation approach.
- SC-FDE is a single-carrier modulation technique, it is not possible for an uplink user to use only part of the spectrum.SC-FDMA can reasonably be called "FFT (or DFT) pre-coded OFDMA.
- SC-FDMA more closely resembles OFDMA because it still requires an IFFT operation at the transmitter in order to separate the users.
- The goal of SC-FDMA are

  Take the low peak-to-average ratio (PAR) properties of SC-FDE.

  Achieve an OFDMA-type system that allows partial usage of the frequency band.

4.3.1 SC-FDMA Advantages and Disadvantages

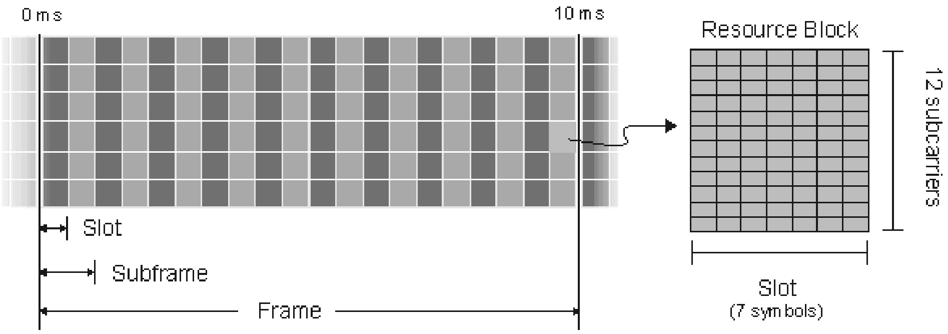
- ***SC-FDMA Advantages***
    - PAR of SC-FDMA is significantly lower than OFDMA.
    - Low cost and power constraints experienced by mobile handsets.
    - Only part of the frequency spectrum is used by any one user at a time, like in OFDM
- ***SC-FDMA Advantages***
    - SC-FDMA can experience more spectral leakage than OFDMA.
    - Achieve frequency diversity differently, leading to slight differences in performance.
    - SC-FDMA has a more complexity both the transmitter and receiver compare to OFDM.
    - Need additional FFT of size M_k has to be performed for each user at the transmitter and receiver.

4.5 OFDMA and SC-FDMA in LTE

- Any OFDMA-based standard specify following things in order for the system to work.
    1. *It must specify the "quanta," or units, of time-frequency resource (RB) that can be assigned.*
    2. *It must specify messaging protocols that allow the MS to request resources when necessary, and to know what resources they have been assigned, both for transmission and reception.*
    3. *Ranging procedures must be specified so that simultaneous uplink transmissions from several different mobile units can be reliably decoded at the base station.*

4.5.1 The LTE Time-Frequency Grid

- In LTE, mobile units are allocated groups of subcarriers over time and frequency known as a resource block (RB).
- The size of the resource block is chosen to balance a tradeoff between granularity and overhead.
- For example
    - *If assign any subcarrier to any user in any time slot increase very large amount of overhead to specify the current allocation to all the mobile units.*
    - *Much lower overhead would be achieved by an OFDM-TDMA type system, but not efficient in many respects including total throughput, delay, and the required peak power.*
- The Structure of LTE Time –Frequency Grid(LTE FDD frame of 1.4 MHz channel) as shown below



- A typical resource block consists of 12 subcarriers over 7 OFDM symbols, also referred to as a timeslot.
- A timeslot in LTE spans 0.5 msec and two consecutive timeslots create a subframe.
- Resources are allocated to users in units of resource blocks over a subframe, that is, 12 subcarriers over 2 x 7 = 14 OFDM symbols for a total of 168 "resource elements," which in practice are QAM symbols.
- Not all the 168 resource elements can be used for data since some are used for various layer 1 and layer 2 control messages.
- The subcarriers of a resource block can be allocated in one of two ways.

  1. ***Distributed subcarrier allocation****:*
     - *It takes advantage of frequency diversity by spreading the resource block hop across the entire channel bandwidth.*
     - *This can be accomplished by using a "comb" pattern at any given point of time for a given user, so that its subcarriers occur at even intervals across the entire frequency bandwidth.*
     - *This approach is typically used in the downlink (OFDMA) when distributed subcarrier allocation is used.*
     - Frequency diversity can be achieved by hopping a contiguous block of subcarriers in time. Frequency diversity is achieved as long as sufficient interleaving is employed: this is certainly the case in LTE systems, which are heavy on interleaving.
     - This approach is used in the uplink, since SC-FDMA transmitters in general operate on contiguous sets of subcarriers.
  2. ***Adjacent subcarrier allocation***:
     - *This approach relies on a channel-aware allocation of resources, so that each user can be allocated a resource block where they have a strong channel.*
     - *Since a block of 12 subcarriers is typically smaller than the coherence bandwidth of the channel, frequency diversity is not achieved, which is helpful as long as the scheduler is able to assign "good" blocks to each user.*

4.5.2 Allocation Notification and Uplink Feedback

- In LTE uplink, notification and feedback signaling between BS and MS done on a logical control channel. Specifically PDCCH (physical downlink control channel).
- These signaling carries to use in downlink reception and uplink transmission for MS.
- The BS must broadcast information to the pool of active users in its cell.
- The PDCCH specifies the following:

  1. *Downlink resource block allocation*
  2. *Uplink resource block allocation*

3. *QAM constellation to use per resource block*
4. *Type and rate of coding to use per resource block*

- Once a user is able to decode the PDCCH, it knows precisely where to receive (downlink) or to transmit (uplink), and how.
- The PDCCH is sent over the first 2-3 OFDM symbols of each subframe across all the subcarriers.
- PDCCH uses about 14-21% of the total downlink capacity is used by the PDCCH. Additional downlink capacity is also used by other control channels and the pilot symbols.
- To aid the base station in uplink scheduling, LTE units utilize buffer status reporting (BSR), wherein each user can notify the BS about its queue length, and channel quality information (CQI) feedback.
- Once the BS is well informed about the channels to/from the users and their respective queue lengths, it can more appropriately determine the optimum allocation among the various users.
- In the downlink, the BS has inherent knowledge of the amount of buffered data for each user, while in the uplink it can estimate the channel from each user.
- Hence, BSR feedback is only used for uplink scheduling while CQI feedback is only used for downlink scheduling and AMC-mode selection.
- The CQI reporting can be either periodic or aperiodic, wideband or sub-band, and multiple CQI feedback modes are defined for different scenarios.

4.5.3 Power Control

- Power control is a kind of a solution equalize SINR values over the cell and hence control the Inter
- Cell Interference (ICI) in both uplink and down link.
- It manages self-interference and is related to imperfect time-frequency-power synchronization between the different uplink users.
- If power control is not used, the different signals may be received with very different powers, which causes a dynamic range problem when the signal is A/D converted.
- If power control is not used, the strong users will dominate the A/D dynamic range and the weak users will experience severe quantization noise, making digital reconstruction of those signals difficult or impossible.
- In short, some uplink power control is needed in OFDMA (or SC-FDMA) systems.
- In LTE, closed-loop power control is possible in the uplink where the BS can explicitly indicate the maximum transmit power density (power per resource block) that can be used by each user.
- PDCCH carries power control information, when the uplink allocation for each user is specified.
- The uplink loop power control algorithm in LTE is flexible in terms of the amount of channel inversion it performs.

- If no power control can be used—all users transmit at full power—which results in high average spectral efficiency but low battery efficiency and poor fairness, as cell edge users are disadvantaged.
- These two extremes can be balanced by fractional power control. Fractional power control is the open-loop power control scheme in LTE.
- In the downlink, no closed-loop power control is specified in the standard; however, LTE systems can specify a relative power offset between different users.
- This is done using a higher layer message and thus can only be performed at much longer timescales compared to uplink power control.
- By allocating different power offsets among the different users according to their location, the system can try to improve the fairness in terms of the data rate of a user who is at the cell edge relative to that of a user closer to the BS.

Module-2

Chapter 5: Multiple Antenna Transmission and Reception

5.1 Introduction:

- The basic concept of diversity: transmit the signal via several independent diversity branches to get independent signal replicas via
  - *Time diversity*
  - *Frequency diversity*
  - *Space diversity*
  - *Polarization* diversity.
- High probability: all signals not fade simultaneously and the deepest fades can be avoided. It provides protection against fading.

Name three categories of multiple antenna techniques.

- Multiple antenna techniques can be grouped into roughly three different categories:
  1. ***Diversity***: *It allows a number of different versions of the signal to be transmitted and/or received, and provides considerable resilience against fading.*
  2. ***Interference suppression***: *It uses the spatial dimensions to reject interference from other users, either through the physical antenna gain pattern or through other forms of array processing such as linear precoding, post coding, or interference cancellation.*
  3. ***Spatial multiplexing***: *It allows two or more independent streams of data to be sent simultaneously in the same BW, and hence is useful primarily for increasing the data rate.*
- All three of these different approaches are often collectively referred to as multiple input-multiple output (MIMO) communication.

5.1 Spatial Diversity Overview

What is spatial diversity? Explain the two types of gains achieved in spatial diversity.

- Spatial diversity is exploited through two or more antennas, which are separated by enough distance so that the fading is approximately decorrelated between them.
- The primary advantage of spatial diversity is that no additional bandwidth or power is needed.
- It is limited by need additional antenna, RF transmit and/or receive chain, and DSP signal processing required to modulate or demodulate multiple spatial streams.
- When multiple antennas are used, there are two forms of gain available, which we will refer to as *Array Gain* and *Diversity Gain*.

5.1.1 Array Gain

- Array gain means a power gain of transmitted signals that is achieved by using multiple-antennas at transmitter and/or receiver, with respect to single-input single-output case.
- It can be simply called power gain. The array gain is almost exactly proportional to the length of the array.
- It provide performance enhancement by coherently combining the energy of each of the antennas to gain an advantage versus the noise signal on each antenna.
- *Array gain for correlated channel*s: Channels are completely correlated (as might happen in a line-of-sight system with closely spaced antennas) the received SNR increases linearly with the number of receive antennas N_r. For a $N_t \times N_r$ system, the array gain is , which can be seen for a $1 \times N_r$, as follows.
    - For each antenna $i \epsilon (1, N_r.)$ receives a signal that can be characterized as:

$$y_i = h_i x + \quad = hx + n_i \qquad (5.1)$$

Where $h_i = h$, for all the antennas since they are perfectly correlated.

- Hence, the SNR on a single antenna is

$$\gamma_i = \overline{\frac{|h|^2}{\sigma^2}} \qquad (5.2)$$

Where σ^2 is noise power of each correlated path

- If all the receive antenna paths are added, the resulting signal is

$$y = \sum_{i=1}^{Nr} y_i = N_r hx + \sum_{i=1}^{Nr} n_i \qquad (5.3)$$

- *Array gain for uncorrelated channels*: Assuming that the noise on each branch is uncorrelated, is

$$\gamma\Sigma = \frac{|N_r h|^2}{N_r \sigma^2} \qquad (5.4)$$

- *Conclusion:* The received SNR also increases linearly with the number of receive antennas even if those antennas are correlated. However, because the channels are all correlated in this case (in fact, identical), there is no diversity gain.

5.1.2 Diversity Gain

- Diversity gain is the increase in SNR ratio due to some diversity scheme, or how much the transmission power can be reduced when a diversity scheme is introduced, without a performance loss. It is usually expressed in decibels, and sometimes as a power ratio.
- The main objective of spatial diversity has been to improve the communication reliability by decreasing the sensitivity to fading.
- The physical layer reliability is typically measured by the outage probability or average bit error rate. The bit error probability (BEP) can be written for virtually any modulation scheme as:

$$P_b \approx c_1 e^{-c_2 \gamma} \qquad (5.5)$$

Where c_1 and c_2 are constants that depend on the modulation type

γ = received SNR.

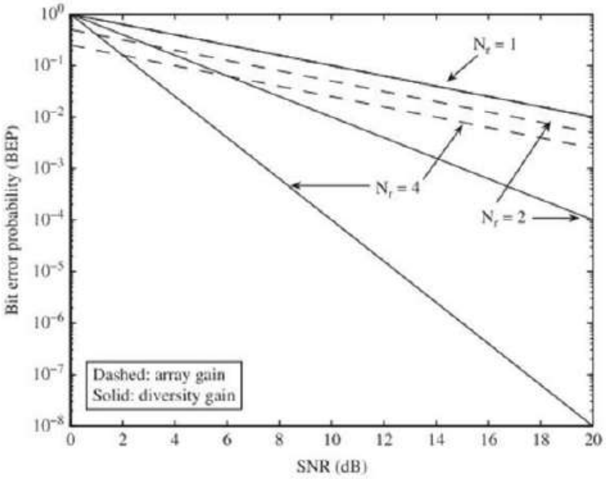
- With reference to equation (5.5) the error probability is exponentially decreasing with SNR, the few instances in a fading channel when the received SNR is low dominate the BEP, since even modestly higher SNR values have dramatically reduced BEP.
- *Fading channel without diversity:* The SNR becomes a random variable in fading channel and so the BEP is also a random variable., the average BEP decreases very slowly, and can be written as

$$\bar{P_b} \approx c_3 \gamma^{-1} \qquad (5.6)$$

- *Conclusion*: This simple inverse relationship between SNR and BEP is much, much weaker than a decaying exponential. Results in terrible reliability for unmitigated fading channels.
- *Fading channel diversity:* The probability of all the uncorrelated channels having low SNR is very small, the diversity order has a dramatic effect on the system reliability. With diversity, the average BEP improves to:
- $\bar{P_b} \approx c_3 \gamma^{-N_d}$ (5.7)

Where N_d *is diversity order* $= N_t \times N_r$. Which is an enormous improvement. For example, if the BEP without any diversity was about 1 in 10 which is awful. The BEP with two antennas at both the transmitter and receiver would be closer to 1 in 10,000. Diversity gain is very powerful.

Figure 5.1 Relative bit error probability (BEP) curves for M = 1, Nr = (1, 2, 4). The BEP (0 dB) is normalized to 1 for each technique. Statistical diversity has a very large impact on BEP, whereas the array gain only results in a fixed shift of the curve.



5.1.3 Increasing the Data Rate with Spatial Diversity

How to increase the data rate with spatial diversity? Explain

- The Shannon capacity formula gives the maximum achievable data rate of a single communication link in additive white Gaussian noise (AWGN) as:

$$= Blog_2\,(1+\gamma) \qquad (5.9)$$

Where C is the "capacity," or maximum error-free data rate, B is the bandwidth of the channel, and γ is again the SNR (or SINR).

- By using advance coding, and with sufficient diversity, it may be possible to approach the Shannon limit in some wireless channels.
- Antenna diversity increases the SNR linearly, diversity techniques increase the capacity only logarithmically with respect to the number of antennas. In other words, the data rate benefit rapidly diminishes as antennas are added.
- However, when the SNR is low, the capacity increase is close to linear with SNR, since $\log(1+x) \approx x$ for small x. Hence, in low SNR channels, diversity techniques increase the capacity about linearly, but the overall throughput is generally still poor due to the low SNR.
- *Conclusion*: (1). More substantial data rate increase at higher SNRs, the multi-antenna channel can instead be used to send multiple independent streams. Spatial multiplexing has the ability to achieve a linear increase in the data rate with the number of antennas at moderate to high SINRs through the use of sophisticated signal processing algorithms.

  (2). In a system with N_t transmit and N_r receive antennas, often known as $N_t \times N_r$ spatial multiplexing system, the peak data rate is proportional to min (N_t, N_r).

5.1.4 Increased Coverage or Reduced Transmit Power

Explain, how spatial diversity increase the coverage with reduced transmit power.

- The benefits of diversity is increase the coverage area with reduced transmit power.
- *Increase in coverage area due to spatial diversity*:
  - Assume that there are N_r receive antennas and just one transmit antenna.
  - Due to gain, the average SNR is approximately $N_r \times \gamma$. Where γ is the average SNR per branch.
  - Consider simplified path loss model $Pr = Pt \times Po\ d^{-\alpha}$.
  - It can be found that the increase in coverage range is $Nr^{\,1/\alpha}$.
  - The coverage area improvement is $Nr^{\,2/\alpha}$, without even considering the diversity gain.
  - *Conclusion:* The system reliability greatly enhanced even with this range extension.
- *Reduced Transmit Power:*
  - The required transmit power can be reduced by $10log_{10}\,N_r\,dB$ while maintaining a diversity gain of $N_t \times N_r$.

5.2 Receive Diversity

What is receive diversity?

- The most prevalent form of spatial diversity is receive diversity.
- This type of diversity is nearly ubiquitous $N_r = 2$. It is most common receiver configuration on cellular BSs and wireless LAN access points. It is mandatory for LTE BSs and handsets.
- Receive diversity on its own places no particular requirements on the transmitter, but requires a receiver that processes the N_r received streams and combines them in some fashion.

Explain the two combining algorithms of receive diversity.

- Two majorly used combining algorithms are
  1. *Selection combining (SC)*
  2. *Maximal Ratio Combining (MRC)*

Compare selection combing v/s maximal ration combing (MRC)

5.2.1 Selection Combining:

- *Principle of SC algorithm*: It simply estimates the instantaneous strengths of each of the N_r streams, and selects the highest one (see fig 5.2)

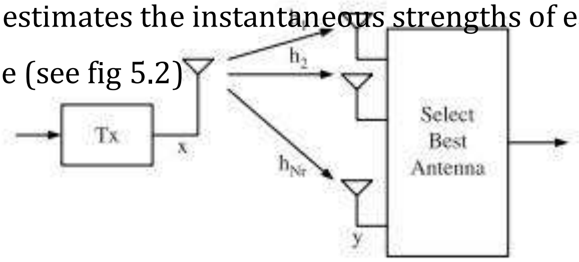


Figure 5.2 Receive diversity: selection combining

- *Advantages:* It is the simplest type of combiner. Its simplicity and reduced hardware and power requirements make it attractive for narrowband channels.
- *Limitation:* It ignores the useful energy on the other streams, it is clearly suboptimal. Not suitable for wideband channel.
- The diversity gain from employing selection combining can be confirmed by considering the outage probability.
- *Outage probability* ($_t$) : It is defined as the probability that the received SNR drops below some required threshold, $P_{out} = P[\gamma < \gamma_0] = P$.
- Assuming N_r uncorrelated receptions of the signal,

$$\begin{aligned} P_{out} &= P[\gamma_1 < \gamma_o,\ \gamma_2 < \gamma_o,\ \ldots,\ \gamma_M < \gamma_o], \\ &= P[\gamma_1 < \gamma_o]P[\gamma_2 < \gamma_o]\ldots P[\gamma_M < \gamma_o], \\ &= p^{N_r}. \end{aligned}$$

- For a Rayleigh fading channel:

$$p = 1 - e^{-\gamma_o/\bar{\gamma}},$$

where $\bar{\gamma}$ is the average received SNR at that location

.

- Thus, selection combining decreases the outage probability to:

$$P_{out} = (1 - e^{-\gamma_o/\bar{\gamma}})^{N_r}.$$

- The average received SNR for N_r branch SC can be derived in Rayleigh fading to be:

$$\begin{aligned}\gamma_{sc} &= \bar{\gamma}\sum_{i=1}^{N_r}\frac{1}{i}, \\ &= \bar{\gamma}(1+\frac{1}{2}+\frac{1}{3}+\ \ldots\ +\frac{1}{N_r}).\end{aligned}$$

- *Conclusion:* Each added (uncorrelated) antenna does increase the average SNR. The average BEP can be derived by averaging (integrating) the appropriate BEP expression in AWGN against the exponential distribution. Plots of the BEP with different amounts of selection diversity are shown in Figure 5.3.

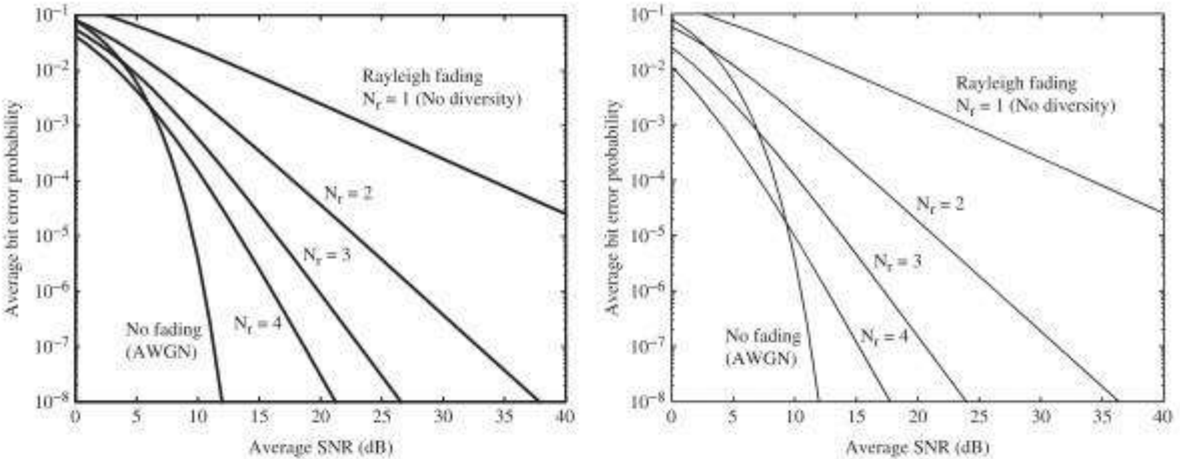
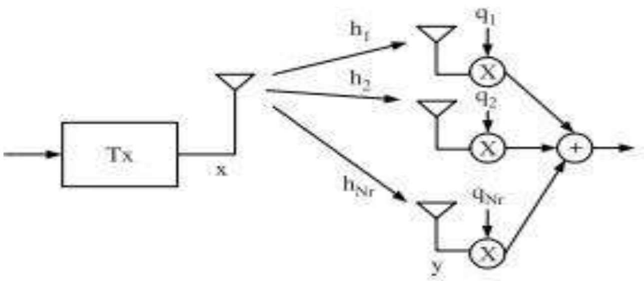


Figure 5.3 . Average bit error probability for selection combining (left) and maximal ratio combining (right) using coherent BPSK. MRC typically achieves a few dB better SNR than SC due to its array gain.

- The performance improvement with increasing N_r diminishes, the improvement from the first few antennas is substantial.
- For example, at a target BEP of 10^{-4} about 15 dB of improvement is achieved by adding a single receive antenna, and the improvement increases to 20 dB with an additional antenna.

5.2.2 Maximal Ratio Combining

- Maximal ratio combining (MRC) combines the information from all the received branches in order to maximize the ratio of signal-to-noise power (see the figure below)



- The combined signal can then be written as:

$$y(t) = x(t)\sum_{i=1}^{N_r}|q_i||h_i|\exp\{j(\phi_i+\theta_i)\}.$$

Where $q_i = |q_i|e^{j\emptyset}$ *is the complex weighting factor of each branch.*,

$h_i = |h_i|e^{j\emptyset}$ *channel response of each branch with input signal* $x(t)$.

- Let the phase of the combining coefficient $\emptyset_i = \theta_i$ for all the branches, then the signal-to-noise ratio of y(t) can be written as:

$$\gamma_{\mathrm{mrc}} = \frac{\mathcal{E}_x\left(\sum_{i=1}^{N_r}|q_i||h_i|\right)^2}{\sigma^2\sum_{i=1}^{N_r}|q_i|^2},$$

Where ε_x is the transmit signal energy. Maximizing this expression by taking the derivative with respect to q_i gives the maximizing combining values. In other words, branches with better signal energy should be enhanced, whereas branches with lower SNRs should be given relatively less weight.

- The resulting signal-to-noise ratio can be found to be:

$$\gamma_{\mathrm{mrc}} = \frac{\mathcal{E}_x\sum_{i=1}^{N_r}|h_i|^2}{\sigma^2} = \sum_{i=1}^{N_r}\gamma_i.$$

- *Conclusion:* The total SNR is achieved by simply adding up the branch SNRs when the appropriate weighting coefficients are used.
- *Limitation of MRC:* It may not be optimal in many cases since it ignores interference power.
- *Alternate:* Equal gain combining (EGC), which only corrects the phase and achieves a post-combining SNR of:

$$\gamma_{\mathrm{EGC}} = \mathcal{E}_x\frac{\sum_{i=1}^{N_r}|h_i|^2}{N_r\sigma^2} = \frac{1}{N_r}\sum_{i=1}^{N_r}\gamma_i.$$

5.3 Transmit Diversity (TD)

What is transmit diversity? Why it not suitable for uplink?

- This method is utilized in the downlink of LTE using 2 or 4 transmit antenna at the eNodeB.
- The receiver (UE) may have 1 or more receive antenna. Here, similar modulation symbols are transmitted to improve the signal quality (SINR).
- It is a type of spatial diversity where N number of transmit antenna and one receive antenna.
- Transmit diversity is particularly useful in the downlink since the base station can usually accommodate more antennas than the mobile station.
- Advantage:
  - *This method does not require any feedback information from the receiver.*
  - *It is effective when the receiver is in a low SINR radio environment.*
  - *It improves the SINR and thus reduces required retransmissions attempts.*
  - *It allows the transmitter to utilize aggressive coding & modulation scheme.*
  - *The specific method LTE uses for transmit diversity is SFBC (Space Frequency Block Coding), providing both spatial and frequency diversity.*
  - *SFBC improves cell coverage and/or improves cell-edge throughput.*
  - *Multiple antennas are already present at the BS for uplink receive diversity, the incremental cost of using them for transmit diversity is small.*
- Limitation of TD*:*
  - *Need additional DSP is required both at both the transmitter and receiver in order to achieve diversity while removing or at least attenuating the spatial interference.*

    Name the two classes of transmit diversity.

- Multiple antenna transmit schemes are often categorized into two classes: *Open-loop and closed-loop time diversity*.
- *Open-loop transmit diversity*: It refers to systems that do not require knowledge of the channel at the transmitter as shown in Figure 5.4.

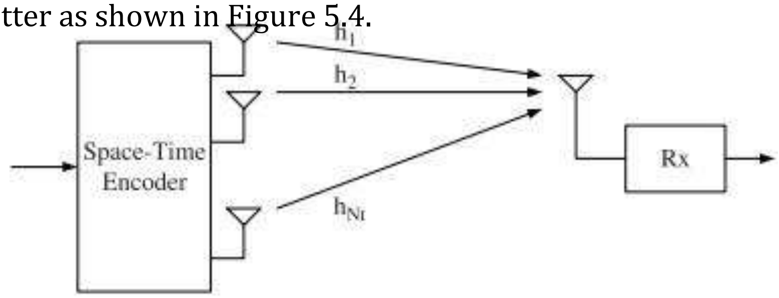


Figure 5.4 Open-loop transmit diversity (no feedback).

- *Closed-loop transmit diversity*: It require channel knowledge at the transmitter, thus more commonly a TDD feedback channel from the receiver to the transmitter.

5.3.1 Open-Loop Transmit Diversity:

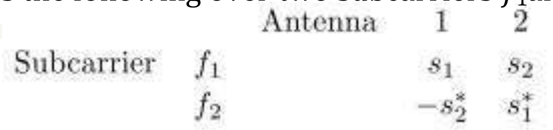
Explain open loop transmit diversity. Mention the advantages and limits of transmit diversity.

Following are the open loop Transmit Diversity approaches

1. *2 x 1 Space-Frequency Block Coding*
2. *With more antennas :* $2 \times 2\ SFBCs, 4 \times 2\ SFBCs, 4 \times 2\ in\ LTE$
3. *Transmit Diversity vs. Receive Diversity*

1). ***2 x 1 Space-Frequency Block Coding***

- It is most popular open-loop transmit diversity scheme is space frequency block coding (SFBC).
- *Principle of this approach*: Where a particular code known to the receiver is applied at the Tx.
- This simple code has become the most popular means of achieving transmit diversity due to its ease of implementation and conceived for a narrowband fading channel.
- STBCs can easily be adapted to a wideband fading channel using OFDM by utilizing adjacent subcarriers rather than consecutive symbols.
- Mathematically and conceptually, there is no difference between SFBCs and the more common STBCs: STBCs use consecutive symbols in time. SFBCs are preferred to STBCs because they experience less delay and are less likely to suffer from channel variations. STBCs would require two OFDM symbols to be encoded (and decoded) over, which significantly increases delay while also increasing the likelihood of channel variation over the code block.
- The simplest SFBC corresponds to two transmit antennas and a single receive antenna. If two symbols to be transmitted are S_1 and S_2,
- The Alamouti code sends the following over two subcarriers $f_1$and f_2



		Antenna	1	2
Subcarrier	f_1		s_1	s_2
	f_2		$-s_2^*$	s_1^*

$$\begin{array}{c|cc} & \text{Antenna 1} & \text{Antenna 2} \\ \hline \text{Frequency 1} & S_1 & -S_2^* \\ \text{Frequency 2} & S_2 & S_1^* \end{array}$$

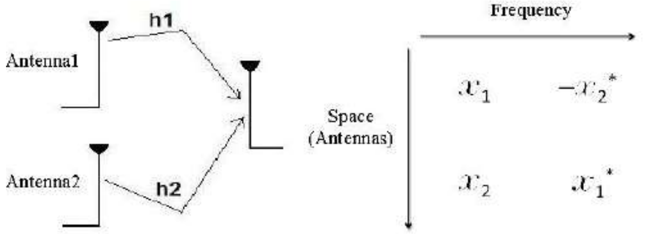


Figure 1 - Description of Alamouti SFBC 2x1 scheme

In above, one transmit antenna transmits modulation symbols S1 and S2 and other transmit antenna transmits phase shifted versions of these modulation symbols (S2* and -S1*).

- Thus, utilizing two subcarriers to transmit two modulation symbols doesn't double the data rate but it certainly improves the signal quality (SINR) of the transmitted signal and thus increasing the achievable data rates. The 2 x 1 Alamouti SFBC is referred to as a rate 1 code.

- Since modulation symbol S1 is transmitted from one antenna on frequency f1, and its phase shifted version (-S1*) is transmitted from another antenna (space diversity) on another frequency f2 (frequency diversity), this method is known as *Space Frequency Block Coding (SFBC) in LTE.*
- The received signal $r(f)$ from subcarriers $f_1$andf_2 can be written as:

$$\begin{aligned} r(f_1) &= h_1 s_1 + h_2 s_2 + n(f_1), \\ r(f_2) &= -h_1 s_2^* + h_2 s_1^* + n(f_2), \end{aligned}$$

- Where h1 (f1) is the complex channel gain from transmit antenna 1 to the receive antenna and h2 (f2) is from transmit antenna 2. n (f1, f2) sample of white Gaussian noise.

- The following diversity combining scheme can then be used, assuming the channel is known at the receiver:

$$\begin{aligned} y_1 &= h_1^* r(f_1) + h_2 r^*(f_2), \\ y_2 &= h_2^* r(f_1) - h_1 r^*(f_2). \end{aligned}$$

Hence, for example, it can be seen that:

$$\begin{aligned} y_1 &= h_1^*(h_1 s_1 + h_2 s_2 + n(f_1)) + h_2(-h_1^* s_2 + h_2^* s_1 + n^*(f_2)), \\ &= (|h_1|^2 + |h_2|^2) s_1 + h_1^* n(f_1) + h_2 n^*(f_2), \end{aligned}$$

and proceeding similarly that:

$$y_2 = (|h_1|^2 + |h_2|^2) s_2 + h_2^* n(f_1) - h_1 n^*(f_2).$$

- Hence, this very simple decoder that just linearly combines the two received samples r(f1) and r*(f2) is able to eliminate all the spatial interference. The resulting SNR can be computed as:

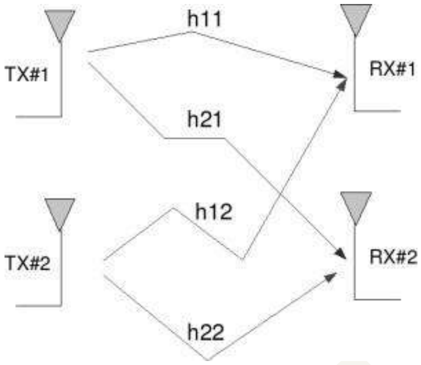
$$\begin{aligned} \gamma_\Sigma &= \frac{(|h_1|^2 + |h_2|^2)^2}{|h_1|^2\sigma^2 + |h_2|^2\sigma^2} \frac{\mathcal{E}_x}{2}, \\ &= \frac{(|h_1|^2 + |h_2|^2)}{\sigma^2} \frac{\mathcal{E}_x}{2}, \\ &= \frac{\sum_{i=1}^{2} |h_i|^2}{\sigma^2} \frac{\mathcal{E}_x}{2}. \end{aligned}$$

- *Conclusion*: The 2 x 1 Alamouti code achieves the same diversity order and data rate as a 1 x 2 receive diversity system with MRC, but with a 3-dB penalty due to the redundant transmission that is required to remove the spatial interference at the receiver. An equivalent statement is that the Alamouti code sacrifices the array gain of MRC, while achieving the same diversity gain.

5.3.2 Open-Loop Transmit Diversity with More Antennas

- In this approach achieve the gains of both MRC and the SFBC simultaneously
- The overview two other popular open-loop transmit diversity approaches are as follows

- **2 x 2 SFBC:**
  - The 2 x 2 SFBC uses the same transmit encoding scheme as for 2 x 1 transmit diversity.



- The channel description can be represented as a 2 x 2 matrix rather than a 2 x 1 vector.

$$\mathbf{H} = \begin{bmatrix} h_{11} & h_{12} \\ h_{21} & h_{22} \end{bmatrix}.$$

The resulting signals at subcarriers f_1 and f_2 on antennas 1 and 2 can be represented as:

$$\begin{aligned} r_1(f_1) &= h_{11}s_1 + h_{21}s_2 + n_1(f_1), \\ r_1(f_2) &= -h_{11}s_2^* + h_{21}s_1^* + n_1(f_2), \\ r_2(f_1) &= h_{12}s_1 + h_{22}s_2 + n_2(f_1), \\ r_2(f_2) &= -h_{12}s_2^* + h_{22}s_1^* + n_2(f_2). \end{aligned} \tag{5.23}$$

- Using the following combining scheme:

$$\begin{aligned} y_1 &= h_{11}^* r_1(f_1) + h_{21} r_1^*(f_2) + h_{12}^* r_2(f_1) + h_{22} r_2^*(f_2), \\ y_2 &= h_{21}^* r_1(f_1) - h_{11} r_1^*(f_2) + h_{22}^* r_2(f_1) - h_{21} r_2^*(f_2), \end{aligned}$$

yields the following decision statistics:

$$\begin{aligned} y_1 &= (|h_{11}|^2 + |h_{12}|^2 + |h_{21}|^2 + |h_{22}|^2)s_1 + 4 \text{ noise terms}, \\ y_2 &= (|h_{11}|^2 + |h_{12}|^2 + |h_{21}|^2 + |h_{22}|^2)s_2 + 4 \text{ noise terms}, \end{aligned}$$

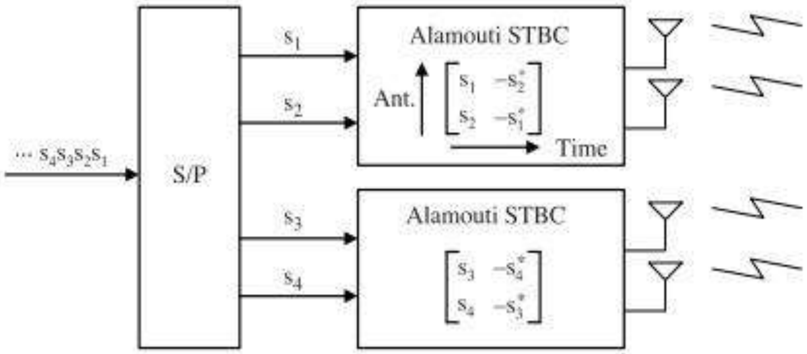
and results in the following SNR:

$$\gamma_\Sigma = \frac{\left(\sum_j \sum_i |h_{ij}|^2\right)^2}{\sigma^2 \sum_j \sum_i |h_{ij}|^2} \frac{\mathcal{E}_x}{2} = \frac{\sum_{j=1}^{2} \sum_{i=1}^{2} |h_{ij}|^2}{\sigma^2} \frac{\mathcal{E}_x}{2}.$$

- This is like MRC with four receive antennas, where again there is a 3-dB penalty due to transmitting each symbol twice.

- Conclusion*:* An orthogonal, full-rate, full-diversity SFBC over a $N_t \times N_r$ channel will provide
  - *Diversity gain equivalent to that of an MRC system with* $N_t \times N_r$ *antennas.*
  - *Power penalty of a 10* $\log_{10} N_t$ *dB transmit power.*

- **4 x 2 Stacked STBCs**
  - In LTG, it will be common to have four transmit antennas at the base station.
  - Here, two data streams can be sent using a double space-time transmit diversity (DSTTD) scheme that essentially consists of operating two 2 x 1 Alamouti code systems in parallel.
  - DSTTD, also called "stacked STBCs," combines transmit diversity and maximum ratio combining techniques along with a form of spatial multiplexing as shown in Figure 5.5.



··· s4s3s2s1
S/P
s1
s2
s3
s4
Alamouti STBC
Ant.
$\begin{bmatrix} s_1 & -s_2^* \\ s_2 & -s_1^* \end{bmatrix}$
Time
Alamouti STBC
$\begin{bmatrix} s_3 & -s_4^* \\ s_4 & -s_3^* \end{bmatrix}$

Figure 5.5 4 × 2 stacked STBC transmitter.

- The received signals at subcarriers $f_1$and $f_2$on antenna 1 and 2 can be represented with the equivalent channel model as

$$\left[\begin{array}{c} r_1(f_1) \\ r_1^*(f_2) \\ \hline r_2(f_1) \\ r_2^*(f_2) \end{array}\right] = \left[\begin{array}{cc|cc} h_{11} & h_{12} & h_{13} & h_{14} \\ h_{12}^* & -h_{11}^* & h_{14}^* & -h_{13}^* \\ \hline h_{21} & h_{22} & h_{23} & h_{24} \\ h_{22}^* & -h_{21}^* & h_{24}^* & -h_{23}^* \end{array}\right] \left[\begin{array}{c} s_1 \\ s_2 \\ \hline s_3 \\ s_4 \end{array}\right] + \left[\begin{array}{c} n_1(f_1) \\ n_1^*(f_2) \\ \hline n_2(f_1) \\ n_2^*(f_2) \end{array}\right].$$

Then, the equivalent matrix channel model of DSTTD can be represented as:

$$\begin{bmatrix} \mathbf{r}_1 \\ \mathbf{r}_2 \end{bmatrix} = \begin{bmatrix} \mathbf{H}_{11} & \mathbf{H}_{12} \\ \mathbf{H}_{21} & \mathbf{H}_{22} \end{bmatrix} \begin{bmatrix} \mathbf{s}_1 \\ \mathbf{s}_2 \end{bmatrix} + \begin{bmatrix} \mathbf{n}_1 \\ \mathbf{n}_2 \end{bmatrix}.$$

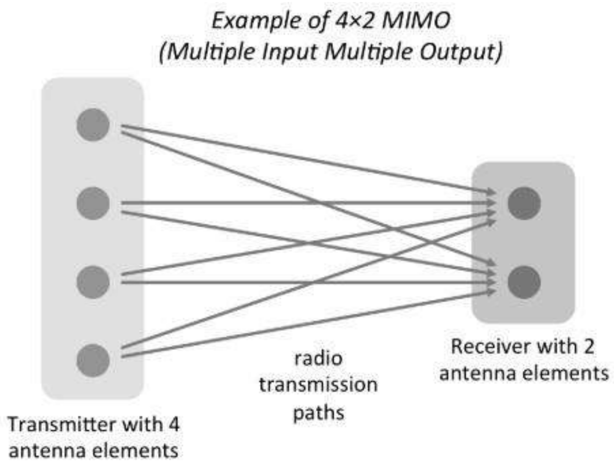
- Thus, DSTTD can achieve a diversity order of $N_d = 2N_r$.
- If the same linear combining scheme is used as in the 2 x 2 STBC case, then the following decision statistics can be obtained:

$$\begin{aligned} y_1 &= (|h_{11}|^2 + |h_{12}|^2 + |h_{21}|^2 + |h_{22}|^2)s_1 + I_3 + I_4 + 4 \text{ noise terms}, \\ y_2 &= (|h_{11}|^2 + |h_{12}|^2 + |h_{21}|^2 + |h_{22}|^2)s_2 + I_3 + I_4 + 4 \text{ noise terms}, \\ y_3 &= (|h_{13}|^2 + |h_{14}|^2 + |h_{23}|^2 + |h_{24}|^2)s_3 + I_1 + I_2 + 4 \text{ noise terms}, \\ y_4 &= (|h_{13}|^2 + |h_{14}|^2 + |h_{23}|^2 + |h_{24}|^2)s_4 + I_1 + I_2 + 4 \text{ noise terms}, \end{aligned}$$

- Where I_i is the interference from the i^{th} transmit antenna due to transmitting two simultaneous data stream.

- **4 x 2 in LTE**

  - In LTE, when four transmit antennas are available, a combination of SFBC and frequency switched transmit diversity (FSTD) is used. See figure below


Example of 4×2 MIMO
(Multiple Input Multiple Output)
Transmitter with 4 antenna elements
radio transmission paths
Receiver with 2 antenna elements

  - This combination of SFBC and FSTD is a rate 1 diversity scheme, i.e., four modulation symbols are sent over four OFDM symbols using the following space-frequency encoder, where the columns correspond to the subcarrier index and the rows to the transmit antenna:

$$\frac{1}{\sqrt{2}}\begin{bmatrix} s_1 & s_2 & 0 & 0 \\ 0 & 0 & s_3 & s_4 \\ -s_2^* & s_1^* & 0 & 0 \\ 0 & 0 & -s_4^* & s_3^* \end{bmatrix}$$

  - The first and second symbols s1 and s2 are sent over antenna ports 0 and 2 on the first two OFDM subcarriers in the block, s3 and s4 symbols are sent using antenna port 1 and 3.
  - It can be detected using a simple linear ML receiver.

5.3.3 Transmit Diversity vs. Receive Diversity

Differentiate between transmit diversity and receive diversity

- Both transmit and receive diversity are capable of providing an enhanced diversity that increases the robustness of communication over wireless fading channels.
- The manner in which this improvement is achieved is quite different.

1. ***Receive Diversity***: In Receiver diversity, received SNR continuously grows as antennas are added, and the growth is linear, that is:

$$\gamma_{\text{mrc}} = \frac{\mathcal{E}_x}{\sigma^2}\sum_{i=1}^{N_r}|h_i|^2 = \sum_{i=1}^{N_r}\gamma_i.$$

The expected value or average combined SNR can thus be found as:

$$\bar{\gamma}_{\text{mrc}} = N_r\bar{\gamma},$$

where $\bar{\gamma}$ is the average SNR on each branch.

2. ***Transmit Diversity:*** Due to the transmit power penalty inherent to transmit diversity techniques, the received SNR does not always grow as transmit antennas are added. the received combined SNR in an orthogonal STBC scheme is generally of the form:

$$\gamma_\Sigma = \frac{\mathcal{E}_x}{N_t\sigma^2}\sum_{i=1}^{N_t}|h_i|^2.$$

As the number of transmit antennas grows large, this expression becomes

$$\gamma_\Sigma = \frac{\mathcal{E}_x}{\sigma^2}\frac{|h_1|^2+|h_2|^2+\ldots+|h_{N_t}|^2}{N_t} \rightarrow \frac{\mathcal{E}_x}{\sigma^2}E\left[|h_1|^2\right],$$

Transmit diversity it eliminates the effects of fading but does not actually increase the average amount of useful received signal-to-noise ratio.

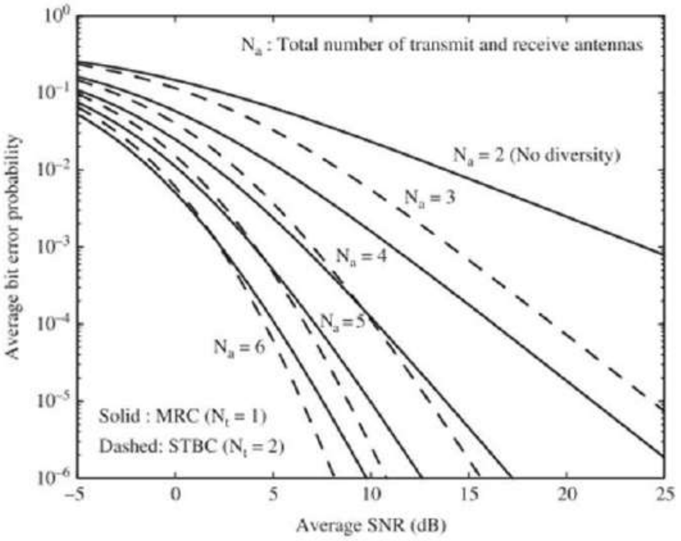


Figure 5.6 Comparison of the SFBC with MRC for coherent BPSK in a Rayleigh fading channel.

5.3.4 Closed-Loop Transmit Diversity

What is closed loop transmit diversity? Explain its types

- If feedback is added to the system, then the transmitter may be able to have knowledge of the channel between it and the receiver.
- There is a substantial gain in many cases from possessing Channel State Information (CSI) at the transmitter, particularly in the spatial multiplexing setup.
- The basic configuration for closed-loop transmit diversity is shown in Figure 5.7.

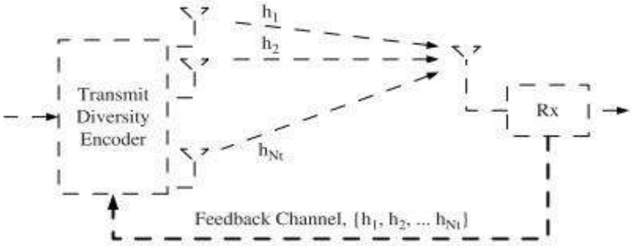

h$_1$
h$_2$
Transmit Diversity Encoder
Rx
h$_{N_t}$
Feedback Channel, {h$_1$, h$_2$, ... h$_{Nt}$}

Figure 5.7 Closed-loop transmit diversity.

- Two important types of closed-loop transmit diversity are as follows
  1. *Transmit selection diversity*
  2. *Linear diversity precoding*

1. ***Transmit selection diversity:*** It is the simplest form of transmit diversity. Only a subset $N^* < N_t$ of the available N_t antennas is used at a given time. The selected subset typically corresponds to the best channels between the transmitter and receiver.

- *Some advantages of transmit antenna selection are*
  1. Hardware cost and complexity are reduced,
  2. Spatial interference is reduced since fewer transmit signals are sent.
  3. It does not incur the power penalty relative to receive selection diversity
  4. The diversity order is still $N_t \times N_r$ even though only N* of the N_t antennas are used.
- *The main drawback: T*he gain from selecting the best antenna averaged over all the coherence bands is likely to be small in a wide band channel.

2. ***Linear diversity precoding:*** Precoding is a technique which exploits transmit diversity by weighting the information stream, i.e. the transmitter sends the coded information to the receiver to achieve pre-knowledge of the channel. This technique will reduce the corrupted effect of the communication channel. Linear precoding is a general technique for improving the data rate or the link reliability by exploiting the CSI at the transmitter.

5.4 Interference Cancellation Suppression and Signal Enhancement

- The available antenna elements at either the transmitter or receiver can be used to suppress undesired signals and/or enhance the power of the desired signal.

***Explain the three approaches for interference suppression and signal enhancement**.*

- Depending on the amount of information available about the interfering channels, following are three approaches for interference suppression and signal enhancement
  1. *DOA-Based Beam steering*
  2. *Linear Interference Suppression: Complete Knowledge of Interference Channels*
  3. *Linear Interference Suppression: Statistical Knowledge of Interference Channels*

1. DOA-Based Beam steering:

- Electromagnetic waves can be physically steered to create beam patterns at either the transmitter or the receiver.
- At the transmitter, this causes energy to be sent predominantly in a desired direction.
- The more antennas are used, the more control over the beam pattern.
- The most common and simple form of this is static pattern-gain beam steering, which is known as "sectoring" concept in cellular system.

- The beam steering approaches performs to produce beam patterns can be finely and, dynamically adjusted to attenuate undesired signals while amplifying desired signals.
- The various signals can be characterized in terms of the direction of arrival (DOA) or angle of arrival (AOA) of each received signal.
- Each DOA can be estimated using signal processing techniques such as the MUSIC, ESPRIT, and MLE algorithms.
- From the acquired DOAs, a beam former extracts a weighting vector for the antenna elements and uses it to transmit or receive the desired signal of a specific user while suppressing the undesired interference signals.

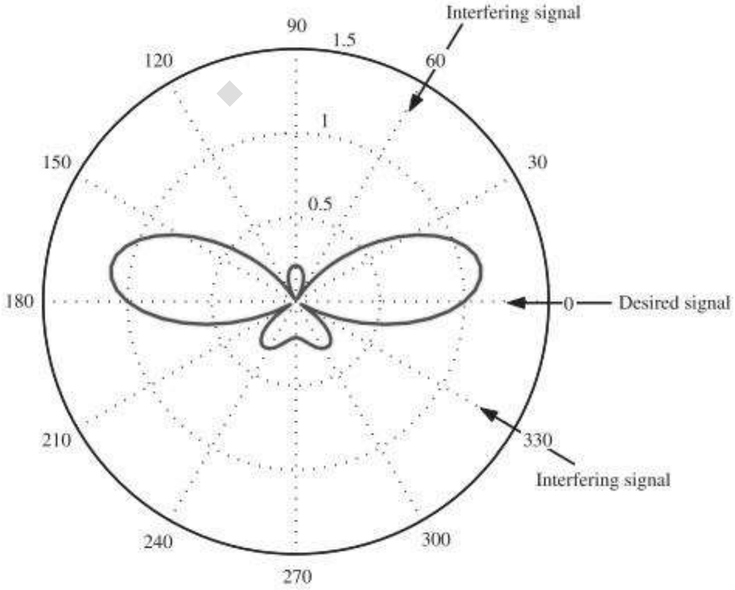


Figure 5.8 Null-steering beam pattern for the DOA-based beamforming using three-element ULA with $\lambda/2$ spacing at transmit antennas. The AOAs of the desired user and two interferers are 0, $\pi/3$, and $-\pi/6$, respectively.

5.4.2 Linear Interference Suppression: Complete Knowledge of Interference Channels

- Consider a single transmitter with N_t antennas trying to communicate to a receiver with $N_r >$ N_t Antenna, in the presence of L_i, interfering transmitters, thus the interfering sources.

$$L = \sum_{i=1}^{L_I} N_{t,i}$$

- For total of two transmitted streams, to a two-antenna receiver, as shown in Figure 5.9. The received signal model is therefore:

$$= \boldsymbol{H}\,\mathrm{x} + \mathrm{n}$$

Where **H** is a 2 x 2 matrix of both the desired and interfering channels. If we assume the receiver knows not only its own channel vector but the interfering channel as well, then detection of its desired signal x_1 is straightforward.

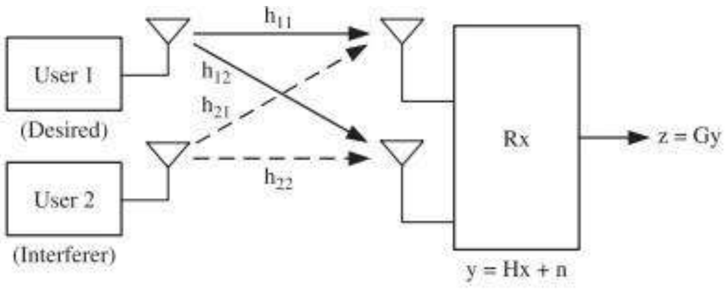


Figure 5.9 Simple two-user interference cancellation example for Sections 5.4.2 and 5.4.3.

- For example, a zero-forcing receiver G = H^{-1} would do the trick and produce as long as H is well-conditioned.

$$\mathbf{z} = \mathbf{x} + \mathbf{H}^{-1}\mathbf{n}.$$

- Drawback: It is an over design concept.

5.4.3 Linear Interference Suppression: Statistical Knowledge of Interference Channels

- To suppress multiple interferers, need to have only statistical knowledge of the interference.
- Consider a general setup where again the desired transmitter has N_t antennas for transmission and the desired receiver N_r antennas for reception in a flat fading channel.
- Then L_i distinct co-channel interferers each equipped with $N_{t,\,i}$ antenna elements. Other channels are suppressed with only statistical knowledge of the interference level rather than the instantaneous channel matrix. Then beamforming vector w_i. Then N_r dimensional received signal vector at the receiver is given by

$$\mathbf{y} = \mathbf{H}\mathbf{w}_t x + \mathbf{H}_I \mathbf{x}_I + \mathbf{n}$$

Where x is the desired symbol with energy.

5.5 Spatial Multiplexing***

What is spatial multiplexing?

- *Concept:* Several different data bits are transmitted via several independent (spatial) channels.
- Spatial multiplexing refers to breaking the incoming high rate data stream into M parallel data streams, as shown in Figure 5.11 for M = N_t and $N_t \leq N_r$.

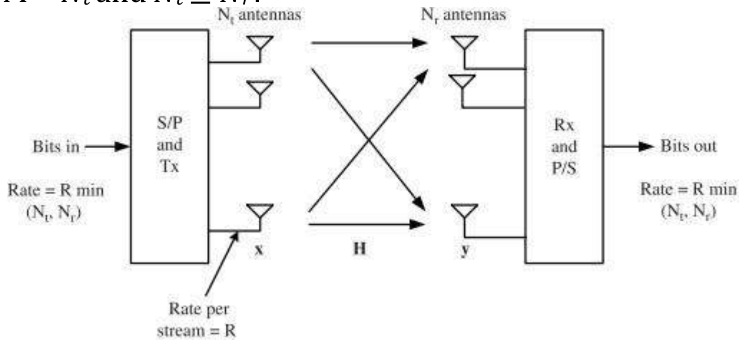


Figure 5.11 A spatial multiplexing MIMO system transmits multiple sub-streams to increase the data rate.

- Here spectral efficiency is increased by a factor of M. It implies that adding antenna elements can greatly increase the data rate without any increase in bandwidth.
- ***Characteristics of spatial multiplexing***
  - No bandwidth expansion.
  - Space–time equalization needed in the receiver.
    - Conventionally: number of Rx antenna ≥ number of Tx antenna.
  - The data streams can be separated by the equalizer, if fading processes of the spatial channels are (nearly) independent.
  - Actual MIMO channel with capacity linearly increasing the number of antenna or more precisely independent spatial channels.
  - Alternative to spatial diversity: multiplexing–diversity trade–off.

5.5.1 An Introduction to Spatial Multiplexing

Explain the operation principle with 2×2 spatial multiplexing system

- ***Principles of Operation****: If the transmitter and receiver both have multiple antennas, then we can set up multiple parallel data streams between them, so as to increase the data rate. In a system with N_t transmit and N_r receive antennas, often known as an $N_t \times N_r$ spatial multiplexing system, the peak data rate is proportional to $min(N_t, N_r)$*

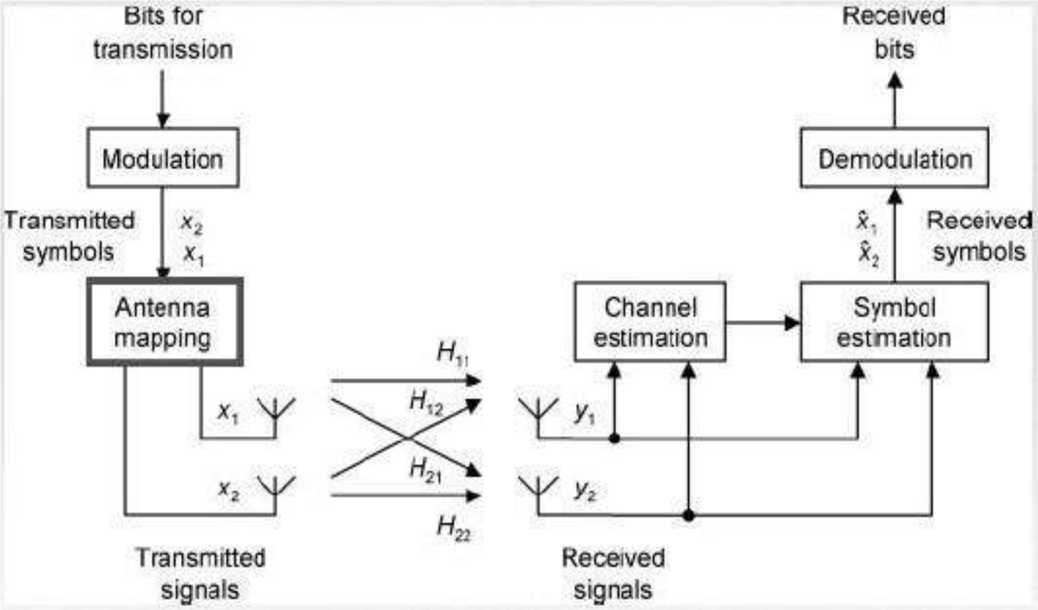


Figure 5.12 Basic principles of a 2x2 spatial multiplexing system

- A basic spatial multiplexing system, in which the transmitter and receiver both have two antennas as shown in figure 5.12
  - *In the transmitter, the antenna mapper takes symbols from the modulator two at a time, and sends one symbol to each antenna*
  - *The antennas transmit the two symbols simultaneously, so as to double the transmitted data rate.*

- *The symbols travel to the receive antennas by way of four separate radio paths, so the received signals can be written as follows*

  $$y1 = H_{11}\, x_1 + H_{12}\, x_2 + n_1$$

  $$y2 = H_{21}\, x_1 + H_{22}\, x_2 + n_2$$

- *x_1 and x_2 are the signals sent from the two transmit antennas and y1 and y2 are the signals that arrive at the two receive antennas and n1 and n2 represent the received noise and interference.*
- *H_{ij} expresses the way in which the transmitted symbols are attenuated and phase-shifted, as they travel to receive antenna i from transmit antenna j*

- In general ,the standard mathematical model for spatial multiplexing is

  $$y = \boldsymbol{H}\, \mathrm{x} + \mathrm{n}$$

Where y is the size of the received vector $N_r \times 1$. The channel matrix H is $N_t \times N_r$ the transmit vector x is $N_t \times 1$, and the noise n is $N_r \times 1$.

- The channel matrix in particular is of the form:

$$\mathbf{H} = \begin{bmatrix} h_{11} & h_{12} & \cdots & h_{1N_t} \\ h_{21} & h_{22} & \cdots & h_{2N_t} \\ \vdots & \vdots & \ddots & \vdots \\ h_{N_r1} & h_{N_r2} & \cdots & h_{N_rN_t} \end{bmatrix},$$

- The entries in the channel matrix and the noise vector are complex Gaussian. In other words, the spatial channels all experience uncorrelated Rayleigh fading and Gaussian noise.
- This model enables a rich framework for mathematical analysis for MIMO systems based on random matrix theory, information theory, and linear algebra.
- The key points we would like to summarize regarding this single-user MIMO system model are
  1. *The capacity, or maximum data rate, grows as $(N\ , N_r\ )\ log(1 + SNR)$ when the SNR is large. When the SNR is high, spatial multiplexing is optimal.*
  2. *When the SNR is low, the capacity is much smaller than at high SNR, it still grows approximately linearly with $(N\ , N_r\ )$ since capacity is linear with SNR in the low-SNR regime.*
  3. *Both of these cases are superior in terms of capacity to space-time coding, where the data rate grows at best logarithmically with N_r.*
  4. *The average SNR of all N_t streams can be maintained without increasing the total transmit power relative to a SISO system.*

  (What is MIMO? Name two types of MIMO)

- Spatial multiplexing can be performed with or without channel knowledge at the transmitter. Accordingly there two classes of MIMO

1. ***Open loop MIMO****: Spatial multiplexing without channel feedback. The principal open-loop techniques; will always assume that the channel is known at the receiver through pilot symbols or other channel estimation techniques. The open-loop techniques for spatial multiplexing attempt to suppress the interference that results from all N_t streams being received by each of the N_r antennas.*

2. ***Closed loop MIMO****: Spatial multiplexing with channel feedback. The potential gain from transmitter channel knowledge is quite significant in spatial multiplexing systems. For example using singular value decomposition (SVD) that shows the potential gain of closed- loop spatial multiplexing methods.*

5.6 How to Choose Between Diversity, Interference Suppression, and Spatial Multiplexing ***

- In MIMO, diversity techniques provides diversity gain and aimed at improving the reliability.
- In MIMO, spatial-multiplexing techniques provides degrees of freedom or multiplexing gain and aimed at improving the data rate of the system
- Diversity provides robustness to fades and interference suppression (IS) provides robustness to interference.
- In particular, diversity increases and steadies Signal (s), while interference suppression reduces I.
- On the other hand, spatial multiplexing creates more parallel streams but does not necessarily increase the per-stream SINR.
- Interference suppression is often considered impractical in a cellular system, and of questionable utility.
- Diversity-Multiplexing Tradeoff (DMT) The DMT stipulates that both diversity gain and multiplexing gain can be achieved in a multiple antenna channel but that there is a fundamental tradeoff between how much of each gain can be achieved.
- Conclusion is that all the spatial degrees of freedom should be used for multiplexing and none for spatial diversity. In short, there is no tradeoff! This is well-captured in Figure 5.17. We see that for all but the highest SNR values, transmit diversity indeed outperforms spatial multiplexing.
- In fact, spatial multiplexing even does worse than no transmit diversity, because so many errors are made on the weakest streams.
- Modern wireless systems (like LTE) have many forms of diversity, most notably time and frequency diversity, which are exploited using coding, interleaving, retransmissions (ARQ), OFDMA, and adaptive modulation.

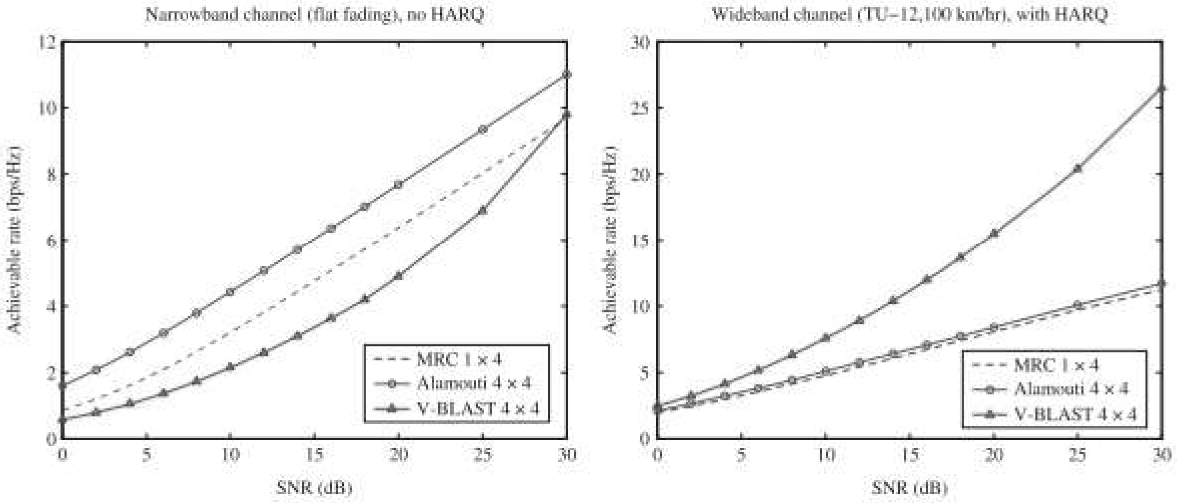


Figure 5.17 The Diversity-Multiplexing Tradeoff, for a narrowband system with no other forms of diversity (left) and for a wideband system with ARQ (right).

Why OFDM is not suitable for LTE uplink. Which is suitable modulation scheme for uplink.

OFDMA is a variant of OFDM which has the advantage that **receiver complexity** is at reasonable level, it can handle scalable bandwidth requirements and supports various modulation schemes from BPSK, QPSK, 16QAM and 64 QAM. This allows adaptive modulation on a per user base.

In uplink variant of OFDMA called SC-FDMA (Single carrier Freq division multiple access) is used. It has advantage against OFDMA to have a lower PAPR (Peak to average Power ratio) since it is a sum of many narrowband signals, which leads to lower power consumption and less expensive Power amplifier in the user terminal.

SC-FDMA enables an orthogonal UL (within the cell) which reduces the UL intra cell interference compared to the non-orthogonal UL in WCDMA.

While LTE’s downlink uses OFDM, the uplink uses a different modulation scheme known as single-carrier frequency-division multiplexing (SC-FDMA). OFDM signals have a high peak to average power ratio (PAPR), requiring a linear power amplifier with overall low efficiency. This is a poor quality for battery-operated handsets. While complex, SC-FDMA has a lower PAPR and is better suited to portable implementation.

Module – 4

Uplink Channel Transport Processing:

- Uplink Channel Transport Processing Overview
- Uplink shared channels
- Uplink Control Information
- Uplink Reference signals
- Random Access Channels
- H-ARQ on uplink

Physical Layer Procedures:

- Hybrid – ARQ procedures
- Channel Quality Indicator CQI feedback
- Pre-coder for closed loop MIMO Operations
- Uplink channel sounding
- Buffer status Reporting in uplink
- Scheduling and Resource Allocation
- Cell Search
- Random Access Procedures
- Power Control in uplink

Module 4

Chapter 8. Uplink Channel Transport Processing

8.1 Background:

- Low complexity and high power efficiency are among the major factors for the transmitter design in the uplink.
- To achieve above requirement LTE uplink is based on SC-FDMA.
- SC-FDMA based uplink, each UE can only be allocated contiguous resource blocks.
- The uplink only supports a limited number of MIMO modes compared to the downlink.
- Similarities between the downlink and uplink transport channel processing are
    - The same channel coding processing is applied on both downlink and uplink shared channels
    - The time-frequency structure of the uplink resource blocks is similar to that of the downlink.

8.2 Uplink Transport Channel Processing Overview

Construct the overview of uplink transport channel processing.
Which are design factors are consider during uplink transport channel processing.

The transport channel processing in the uplink include two distinct steps

1. Channel coding
2. Modulation, as shown in Figure 8.1.

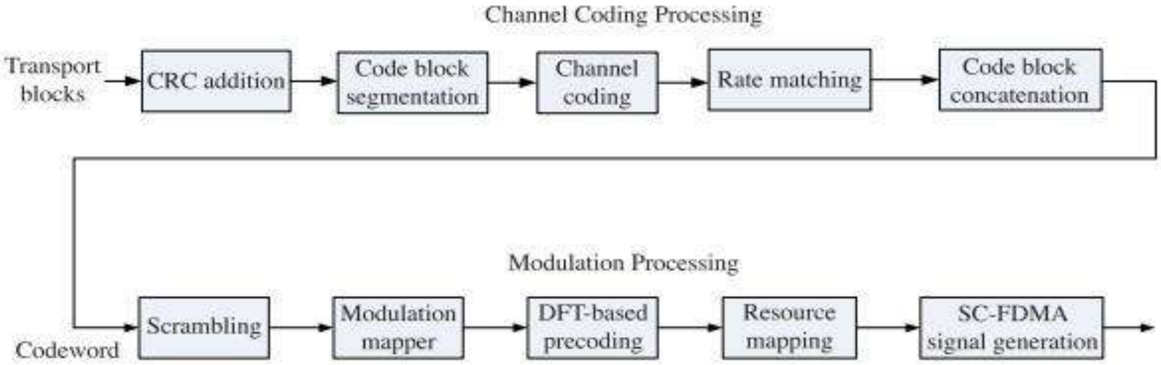
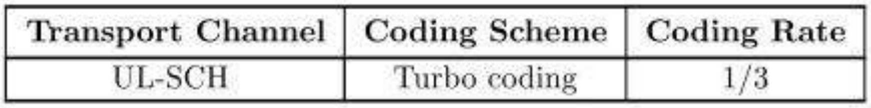


Figure 8.1 Overview of uplink transport channel processing

1. *Channel Coding Processing:*

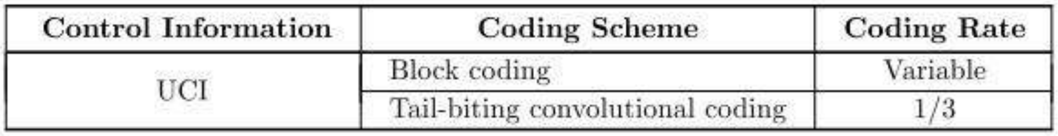
- The channel coding processing in the uplink includes
  - *CRC addition and Code block segmentation*
  - *Channel coding and Rate matching and*
  - *Code block concatenation* (See figure 8.1)
- The usage of the channel coding scheme and coding rate for the uplink shared channel and control information is specified in Table 8.1 and Table 8.2, respectively.

Table 8.1 Usage of Channel Coding Scheme and Coding Rate for Uplink Transport Channels



Transport Channel	Coding Scheme	Coding Rate
UL-SCH	Turbo coding	1/3

Table 8.2 Usage of Channel Coding Scheme and Coding Rate for Uplink Control Information



Control Information	Coding Scheme	Coding Rate
UCI	Block coding	Variable
	Tail-biting convolutional coding	1/3

- The turbo encoder used for uplink shared channels.
- Control information, the channel coding scheme depends on the type of control information and also on the type of the physical channel that carries the control information.
- The control information in the uplink can be mapped either to the Physical Uplink Shared Channel (PUSCH) or the Physical Uplink Control Channel (PUCCH).

2. *Modulation Processing:*

- Modulation processing in the uplink includes scrambling and modulation mapping.
- In uplink a UE specific scrambling is applied in order to randomize the interference.
- Since spatial multiplexing is not supported in the uplink there is no layer mapping or MIMO precoding.
- The main difference from the downlink, the generation of the SC-FDMA baseband signal is illustrated in Figure 8.2

  With a neat block diagram briefly explain generation of SC-FDMA baseband signal

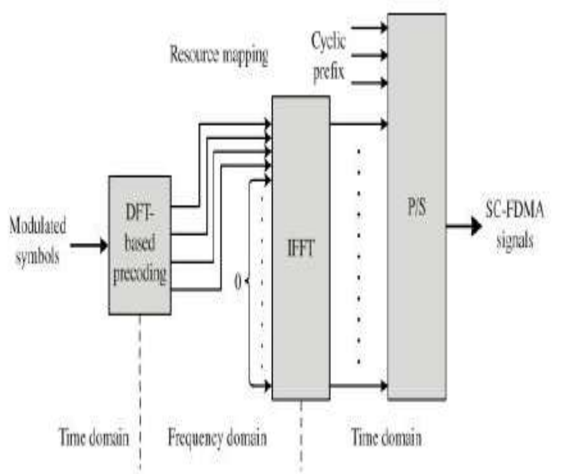
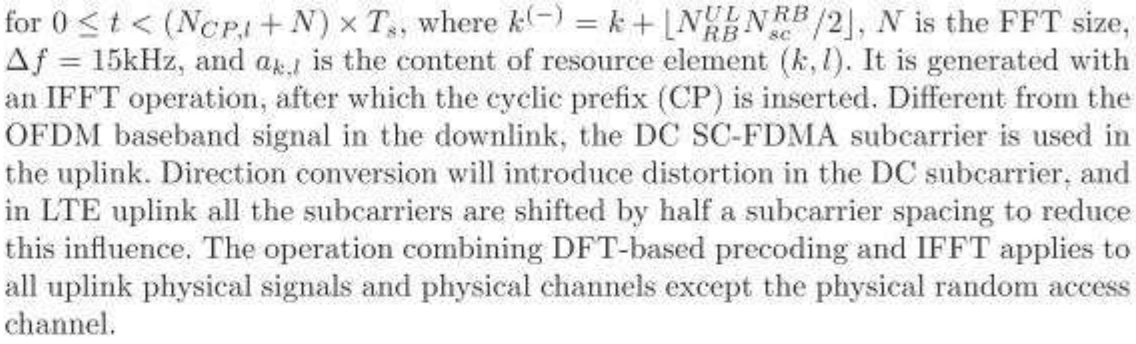


Figure 8.2 Generation of SC-FDMA baseband signals, where P/S denotes the parallel-to-serial converter.

- *Generation Of SC_FDMA baseband signal:*
  a. First, the DFT-based precoding is applied to the block of complex-valued modulation symbols, which transforms the time-domain signal into the frequency domain.
  b. In LTE, the DFT size is constrained to a tradeoff between the complexity of the implementation and the flexibility on the assigned bandwidth and also depends on the number of resource blocks assigned to the UE.
  c. The output of the DFT-based precoder is mapped to the resource blocks that have been allocated for the transmission of the transport block.
  d. In LTE, only localized resource allocation is supported in the uplink, that is, contiguous resource blocks are assigned to each UE.
  e. The baseband signal $S_l(t)$in SC-FDMA symbol l in an uplink slot is defined by:

$$s_l(t) = \sum_{k=-\lfloor N_{RB}^{UL} N_{sc}^{RB}/2 \rfloor}^{\lceil N_{RB}^{UL} N_{sc}^{RB}/2 \rceil - 1} a_{k^{(-)},l} \cdot e^{j2\pi(k+1/2)\Delta f(t-N_{CP,l}T_s)} \qquad (8.1)$$



for $0 \le t < (N_{CP,l} + N) \times T_s$, where $k^{(-)} = k + \lfloor N_{RB}^{UL} N_{sc}^{RB}/2 \rfloor$, N is the FFT size, $\Delta f = 15\text{kHz}$, and $a_{k,l}$ is the content of resource element (k, l). It is generated with an IFFT operation, after which the cyclic prefix (CP) is inserted. Different from the OFDM baseband signal in the downlink, the DC SC-FDMA subcarrier is used in the uplink. Direction conversion will introduce distortion in the DC subcarrier, and in LTE uplink all the subcarriers are shifted by half a subcarrier spacing to reduce this influence. The operation combining DFT-based precoding and IFFT applies to all uplink physical signals and physical channels except the physical random access channel.

8.2 Uplink Shared Channels:

- The description of transport channel processing for Uplink Shared Channels (UL-SCH) below. The UL-SCH is the only transport channel that carries traffic data.

  Explain channel mapping around the uplink shared channel.

- The channel mapping around the UL-SCH is shown in Figure 8.3.

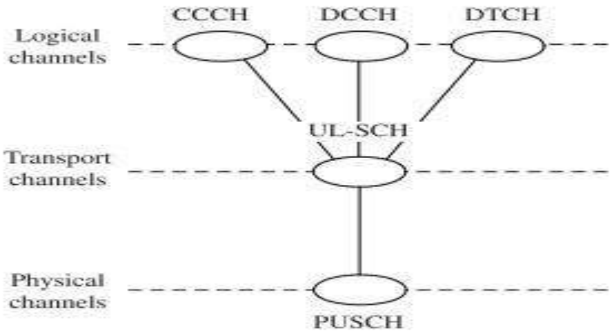


Figure 8.3 Channel mapping around the uplink shared channel.

- Uplink Shared Channels
- (UL-SCH) processing includes

  1. *Channel Encoding and Modulation*
  2. *Frequency hopping*
  3. *Multi-antenna transmission*

1. Channel Encoding and Modulation for UL-SCH:

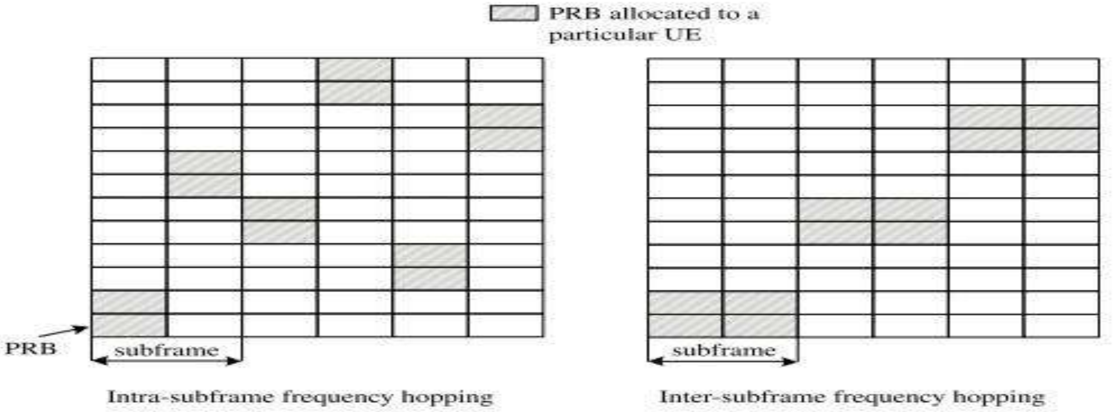
Brief the encoding and modulation process involved in uplink shared channels

- The channel coding scheme: *UL-SCH uses 1/3 turbo encoder is used to encode the transport block. Effective code rates other than 1/3 are achieved by either puncturing or repetition of the encoded bits, depending on the transport block size, the modulation scheme, and the assigned radio resource. The encoded symbols are scrambled prior to modulation. UE-specific scrambling is applied in the uplink.*
- Modulation for UL-SCH: *The UL-SCH supports QPSK, 16QAM, and 64QAM modulation schemes. The QPSK and 16QAM modulation schemes are mandatory and support for the 64QAM modulation is optional and depends on the UE capability.*

2. ***Frequency Hopping: Illustrate the frequency hopping on PUSCH channels.***

   Explain frequency hopping requirement in LTE uplink channels

- LTE supports frequency hopping on PUSCH, which provides additional frequency diversity gain in the uplink.
- Frequency hopping can also provide interference averaging when the system is not 100% loaded. In LTE both intra-subframe and inter-subframe frequency hopping are supported, as illustrated in Figure 8.4.



▨ PRB allocated to a particular UE
PRB
subframe
Intra-subframe frequency hopping
subframe
Inter-subframe frequency hopping

Figure 8.4 Illustrations of frequency hopping on PUSCH.

- *Intra-subframe hopping*: The UE hops to another frequency allocation from one slot to another within the same subframe.
- *Inter-subframe hopping*: The frequency resource allocation changes from one subframe to another.
- Higher layers determine if the hopping is "inter-subframe" or "intra- and inter-subframe."
- Intra-subframe hopping provides higher frequency diversity gain since this gain can be extracted over a single H-ARQ transmission, which always spans only one subframe.
- In the case of "inter-subframe" hopping, multiple H-ARQ transmissions are needed in order to extract the frequency diversity gain.
- Single bit Frequency Hopping (FH) field in the corresponding PDCCH with DCI format 0 is set to 1, the UE shall perform PUSCH frequency hopping.
  - ***No frequency hopping:*** If uplink frequency hopping is disabled (FH = 0), the set of physical resource blocks to be used for transmission is given by $n_{PRB} = n_{VRB}$,
    Where n_{VRB} is the virtual resource block index obtain from uplink scheduling grant.
  - ***Frequency hopping:*** set FH = 1, there are two frequency hopping types.
    1. *Type I hopping:* It uses an explicit offset in the second slot, determined by parameters in DCI format 0.
    2. *In Type 2 hopping*: The set of physical resource blocks to be used for transmission is given by the scheduling grant together with a predefined hopping pattern.

The UE first determines the allocated resource blocks after applying all the frequency hopping rules, and then the data is mapped onto these resources.

3. Multi-antenna Transmission

- LTE only supports a limited number of multi-antenna transmission schemes in the uplink due reduce UE complexity and cost. It has two sections
  1. ***Transmit antenna selection: Explain different types of transmit antenna selection*** When UE has two or more transmit antennas transmit antenna selection can be applied, which is able to provide spatial diversity gain. It ca be operated in three modes. They are
     a. *No antenna selection:* If transmit antenna selection is disabled or not supported by the UE, the UE shall transmit from antenna port 0.
     b. *Closed-loop (CL) antenna selection:* This mode is enabled by higher layers, the UE shall perform transmit antenna selection in response to commands received via DCI format 0 from the eNode-B. The DCI format 0 is scrambled with the antenna selection mask, which enables the UE to determine which antenna port to select.
     c. *Open-loop (OL) antenna selection:* This mode is enabled by higher layers and the transmit antenna to be selected by the UE is not specified. The UE can determine the optimum antenna based on H-ARQ ACK/NAK feedbacks. The UE can transmit from antenna 0 for some time instance and then switch to antenna 1 for a next time instance. During both of these time instances, the UE also monitors the H-ARQ ACK/NAK ratio. If the ACK/NAK ratio in the time instance when antenna 0 was used is less than the ACK/NAK ratio for the time instance when antenna 1 was used, then clearly antenna 1 is a better choice and vice versa
  2. ***Multiuser MIMO (MU-MIMO) in uplink****: Write a note on MU-MIMO in uplink*

     It is also referred to as "virtual" MIMO trans- mission. Two UEs transmit simultaneously on the same radio resource, forming a virtual MIMO channel, and the eNode-B separates the data streams for each UE, for example, using multiuser detection. This transmission mode provides a spatial multiplexing gain to increase the uplink spectrum efficiency, even with single-antenna UEs. CQI calculation and the scheduling process will change due to the interaction between data streams for different UEs. As the eNode-B can estimate the channel information from the uplink reference signal, it is capable of performing CQI calculation and scheduling without further feedback from UEs, which makes it easier to implement MU-MIMO in the uplink than in the downlink.

8.3 Uplink Control Information (UCI)

Explain the physical layer control information

- UCI is to assist physical layer procedures by providing the following types of physical layer control information:
  - *a.* Downlink CQI: *It is used to assist the adaptive modulation and coding and the channel-dependent scheduling of the downlink transmission. The CQI indicates highest modulation and coding rate that can be supported in the downlink.*
  - b. H-ARQ acknowledgment (H-ARQ-ACK): *It is associated with the downlink H-ARQ process*.
  - *c.* Scheduling Request (SR): *It is used to request radio resources for the uplink transmission.*
  - *d.* Precoding Matrix Indicator (PMI): *It is for downlink MIMO transmission.*
  - e. Rank Indication (RI): *RI indicates the maximum number of layers that can be used for spatial multiplexing in the downlink, while PMI indicates the preferred precoding matrix.The channel mapping for control information:* It is shown in figure
- 8.5, which has three different physical control channels, there is only one physical control channel defined for the UCI—the PUCCH. The UCI can also be mapped onto PUSCH when the UE has been assigned uplink radio resources.

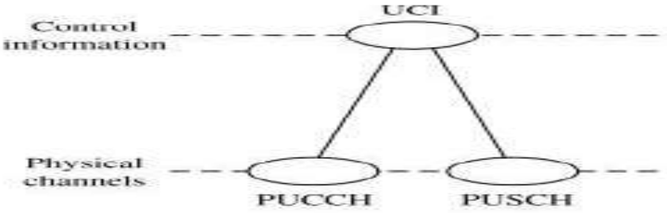


Figure 8.5 Channel mapping for control information in the uplink.

8.3.1 Channel Coding for Uplink Control Information

Briefly explain a) UCI on PUCCH b) UCI on PUSCH with UL-SCH data c)UCI on PUSH without UL-SCH data

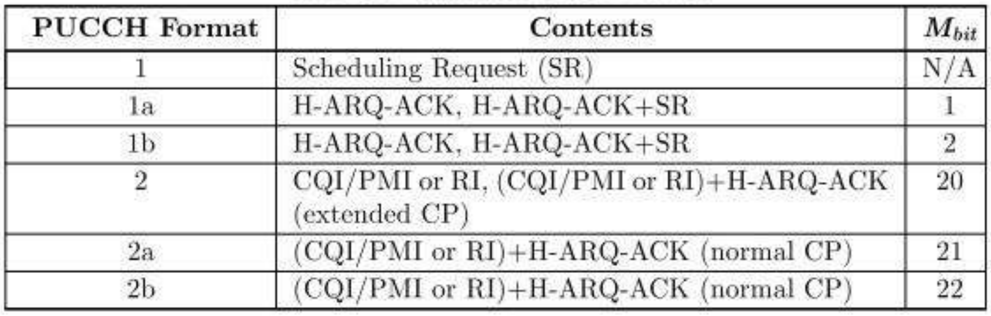
- The channel coding for UCI therefore depends on whether it is carried on the PUCCH or PUSCH.
- Different types of control information are encoded differently, which allows individual adjustments of transmission energy using different coding rates.
- ***UCI on PUCCH:*** When the UCI is transmitted on the PUCCH, three channel coding scenarios are considered:
  1. *Encoding CQI/PMI but not H-ARQ-ACK:* The CQI/PMI is encoded using (20, N_{CQI}) code, with codewords being a linear combination of the 13 basis sequences. Denote a_i, $i = 1, \ldots . N_{CQI}$ as the input channel quality bits, and the encoding is performed as:

$$b_i = \sum_{n=0}^{N_{CQI}} (a_n \cdot M_{i,n}) \bmod 2, \ i = 0, 1, \ldots, 19. \qquad (8.2)$$

2. *Encoding H-ARQ-ACK and SR:* The H-ARQ-ACK bits and SR indication are received from higher layers. Each positive acknowledgement (ACK) is encoded as a binary '1' while each negative acknowledgment (NAK) is encoded as a binary '0'.
3. *Encoding both CQI/PMI and H-ARQ-ACK.:* When CQI/PMI and H-ARQ-ACK are transmitted in the same subframe, the following coding scheme is used:
   - *With the normal CP*: The CQI/PMI is encoded using the (20, N_{CQI}) code and then the H-ARQ-ACIC bits are added at the end of the resulting codeword.
   - *With the extended CP*: The CQI/PMI and H-ARQ-ACK are jointly encoded using the same (20, N_{CQI}) code as that for encoding CQI/PMI alone, with N as the sum of CQI/PMI bits and H-ARQ-ACK bits.

- Based on different types of control information carried on the PUCCH, there are six different PUCCH formats defined in LTE, as shown in Table 8.4. The parameter M_{bit} is the number of encoded bits for each PUCCH format.

Table 8.4 Supported PUCCH Formats



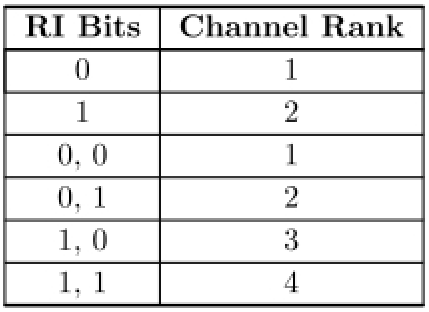
PUCCH Format	Contents	M_{bit}
1	Scheduling Request (SR)	N/A
1a	H-ARQ-ACK, H-ARQ-ACK+SR	1
1b	H-ARQ-ACK, H-ARQ-ACK+SR	2
2	CQI/PMI or RI, (CQI/PMI or RI)+H-ARQ-ACK (extended CP)	20
2a	(CQI/PMI or RI)+H-ARQ-ACK (normal CP)	21
2b	(CQI/PMI or RI)+H-ARQ-ACK (normal CP)	22

- ***UCI on PUSCH with UL-SCH Data:***
  - The UCI can be multiplexed with the UL-SCH data on the PUSCH channel and there is no need to send SR.
  - In this case, the channel coding for H-ARQ-ACK, RI, and CQI/PMI is done independently.
  - Different coding rates can be achieved by allocating different numbers of coded symbols, depending on the amount of allocated radio resource.
  - ***Coding for H-ARQ-ACK:*** For the FDD mode, there is one or two H-ARQ-ACK bits. For the TDD mode, two ACK/NAK feedback modes are supported with different information bits:
    - *ACK/NAK bundling*: It consists of one or two bits of information.
  - *Both FDD and TDD ACK/NAK multiplexing with NBARQ < 2:* The output sequence from the channel encoder is obtained by concatenating multiple encoded H-ARQ-ACK blocks.

- ○ *TDD with ACK/NAK bundling*: The output sequence from the channel encoder is obtained by scrambling the concatenation of multiple encoded H-ARQ-ACK blocks with a specified scrambling sequence.
- ○ *TDD with ACK/NAK multiplexing with NHARQ > 2*: The H-ARQ-ACK bits are encoded using a linear combination of a set of basis sequences.

- ▪ ***Coding for RI***: The mapping between the RI bits and the channel rank is shown in Table 8.5. N$_{RI}$ denote as the number of RI bits encoded into a $N_{RI}Q_m$ codeword, and then multiple concatenated RI blocks are concatenated to form a bit sequence.

Table 8.5 RI Mapping



RI Bits	Channel Rank
0	1
1	2
0, 0	1
0, 1	2
1, 0	3
1, 1	4

- ▪ ***Coding for CQI/PMI***: The coding scheme for CQI/PMI depends on the total number of CQI and PMI bits. After channel encoding, the CQI encoded sequence is multiplexed with the UL- SCH data, the output of which is interleaved with the RI and H-ARQ-ACK encoded sequence as depicted in Figure 8.6. The multiplexing ensures that control and data information bits are mapped to different modulation symbols.

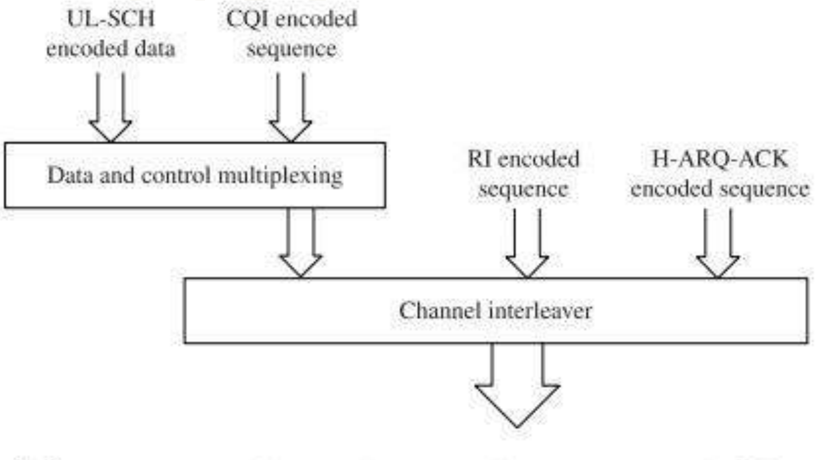


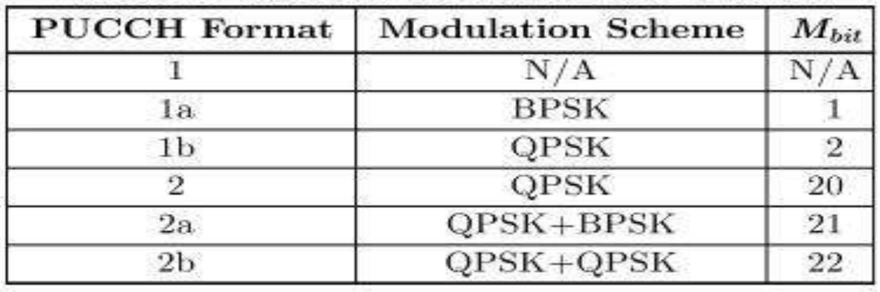
Figure 8.6 Multiplexing of data and control information on the PUSCH channel.

- ▪ ***UCI on PUSCH without UL-SCH Data:*** For this case, the channel coding for CQI, RI, and H- ARQ-ACK information is performed in the same manner as if the UCI is transmitted with UL- SCH data, and then the coded sequences are interleaved. The same interleaves as in Figure 8.6 is applied without the UL-SCH data.

8.3.2 Modulation of PUCCH

- When the UCI is transmitted on the PUSCH, the modulation scheme is determined by the scheduler in the MAC layer.
- The modulation scheme and the number of bits per subframe for different PUCCH formats are specified in Table 8.6.

Table 8.6 Modulation for Different PUCCH Formats

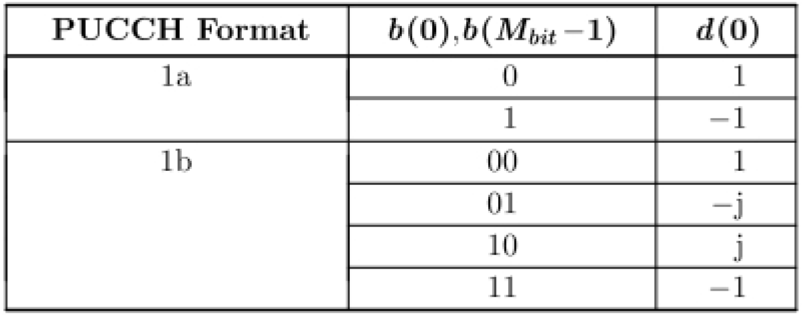


PUCCH Format	Modulation Scheme	M_{bit}
1	N/A	N/A
1a	BPSK	1
1b	QPSK	2
2	QPSK	20
2a	QPSK+BPSK	21
2b	QPSK+QPSK	22

- All PUCCH formats use a cyclic shift of a based sequence to transmit in each SC-FDMA symbol, so UCI from multiple UEs can be transmitted on the same radio resource through code division multiplexing (CDM). Two classes of PUCCH formats. They are

1. ***PUCCH Formats 1, la, and lb:***
   - These format are used to transmit H-ARQ-ACK and/or SR, without CQI bits.
   - When both ACK/NAK and SR are transmitted in the same subframe, a UE shall transmit the ACK/NAK on its assigned ACK/NAK PUCCH resource for a negative SR transmission and transmit the ACK/NAK on its assigned SR PUCCH resource for a positive SR transmission. As shown in Table 8.4, one or two explicit bits are transmitted, respectively, the modulation for which is described in Table 8.7.

Table 8.7 Modulation Symbol $d(0)$ for PUCCH Formats 1*a* and 1*b*

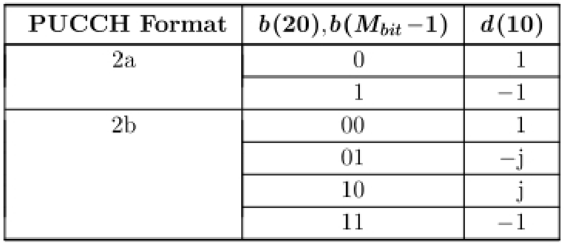


PUCCH Format	$b(0), b(M_{bit}-1)$	$d(0)$
1a	0	1
	1	−1
1b	00	1
	01	−j
	10	j
	11	−1

2. ***PUCCH Formats 2, 2a, and 2b:***
   - Format 2, 2a, or 2b is used when $M_{bits} \geq 20$, as mentioned in table 8.4
   - The block of the first 20 bits, b(0), , b(19), shall be scrambled with a UE-specific scrambling sequence, producing a block of scrambled bits b(0),……… b(19). Then the scrambled bits will be QPSK modulated, resulting in a block of complex-valued Table 8.8 Modulation symbol d(10) for PUCCH formats 2a and 2b complex-valued symbols is multiplied with a length-12

cyclically shifted version of a Zadoff-Chu sequence. This allows FUCCHs from multiple UEs to be transmitted on the same resource block with CDM.

Table 8.8 Modulation symbol $d(10)$ for PUCCH formats 2a and 2b



PUCCH Format	$b(20), b(M_{bit}-1)$	$d(10)$
2a	0	1
	1	−1
2b	00	1
	01	−j
	10	j
	11	−1

- ○ The modulation of these H-ARQ-ACK bits are described in Table 8.8. The resulting modulated symbol d(10) will be used in the generation of the reference signal for PUCCH format 2a and 2b, from which the eNode-B can decode the ACK/NAK information.

8.3.3 Resource Mapping

Explain resource mapping

- ○ PUCCH is time-division multiplexed with the PUSCH from the same UE. This is done in order to retain the single-carrier property of SC-FDMA.
- ○ PUCCH can be FDM with the PUSCH from other UEs in the same subframe.
- ○ For frame structure type 2 (the TDD mode), the PUCCH is not transmitted in the UpPTS field, which is only for the transmission of uplink sounding reference signals or random access.
- ○ The PUCCH uses one resource block in each of the two slots in a subframe.
- ○ The physical resource blocks to be used for PUCCH transmission in slot n_s are given by:

$$n_{PRB} = \begin{cases} \lfloor \frac{m}{2} \rfloor & \text{if } (m + n_s \bmod 2) \bmod 2 = 0 \\ N_{RB}^{UL} - 1 - \lfloor \frac{m}{2} \rfloor & \text{if } (m + n_s \bmod 2) \bmod 2 = 1 \end{cases} \tag{8.3}$$

Where the parameter m depends on the PUCCH format.

- ○ The mapping of PUCCH to physical resource blocks in one subframe is shown in Figure 8.7 for different values of m.

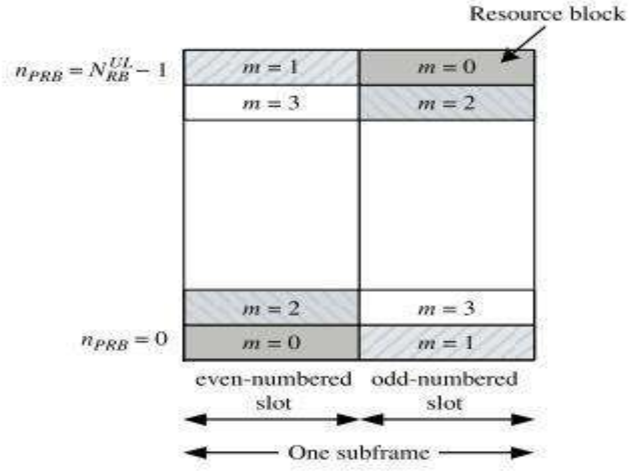


Figure 8.7 Mapping to physical resource blocks for PUCCH.

- PUCCH is transmitted at the bandwidth edge, which is to provide the contiguous bandwidth in the middle for data transmission as only localized resource allocation is allowed in the uplink.
- The frequency hopping between different slots provides frequency diversity.
- The PUCCH symbols are mapped to resource elements not used for RS transmission.

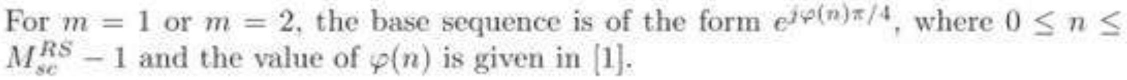
8.4 Uplink Reference Signals

Discuss about 2 types of reference signal defined in the uplink

- In LTE there are two types of reference signals defined in the uplink:
  1. *Demodulation reference signals*: These reference signals are used for coherent demodulation of data and control information at the eNode-B. As PUCCH cannot be transmitted simultaneously with PUSCH, there are demodulation reference signals defined for each of them, that is, there are demodulation reference signals for PUSCH and demodulation reference signals for PUCCH.
  2. *Sounding reference signals:* There are wideband reference signals for the eNode-B to measure uplink CQI for uplink resource allocation. They are not associated with the transmission of PUSCH or PUCCH.

8.4.1 Reference Signal Sequence:

What is reference signal sequence

- Both the demodulation reference signal and the sounding reference signal are defined by a cyclic shift of the same base sequence.
- The generation of the base sequence depends on the reference signal sequence length, which which is $M_{sc}^{RS} = mN_{sc}^{RB}$ with $1 \leq m \leq N_{RB}^{max,UL}$, Where m in is the size of the resource blocks assigned to the UE.
  - If $m \geq 3$ (the UE is assigned three resource blocks or more), the base sequence is based on prime-length Zadoff-Chu sequences that are cyclically extended to the desired length.
  - 
    For $m = 1$ or $m = 2$, the base sequence is of the form $e^{j\varphi(n)\pi/4}$, where $0 \leq n \leq M_{sc}^{RS} - 1$ and the value of $\varphi(n)$ is given in [1].
  - The reference signal in the uplink is always UE-specific.

8.4.2 Resource Mapping of Demodulation Reference Signals

Explain about resource mapping of demodulation reference signal

- For PUSCH, the demodulation reference signal sequence is mapped to resource elements (k,l) with l = 3 for normal CP and = 2 for extended CP, with increasing order first in k and then in the slot number.
- An example of demodulation reference signal mapping for PUSCH is shown in Figure 8.8, with the normal CP.
- PUCCH supports six different formats, and the resource mapping to SC-FDMA symbols for

different formats is listed in Table 8.9.

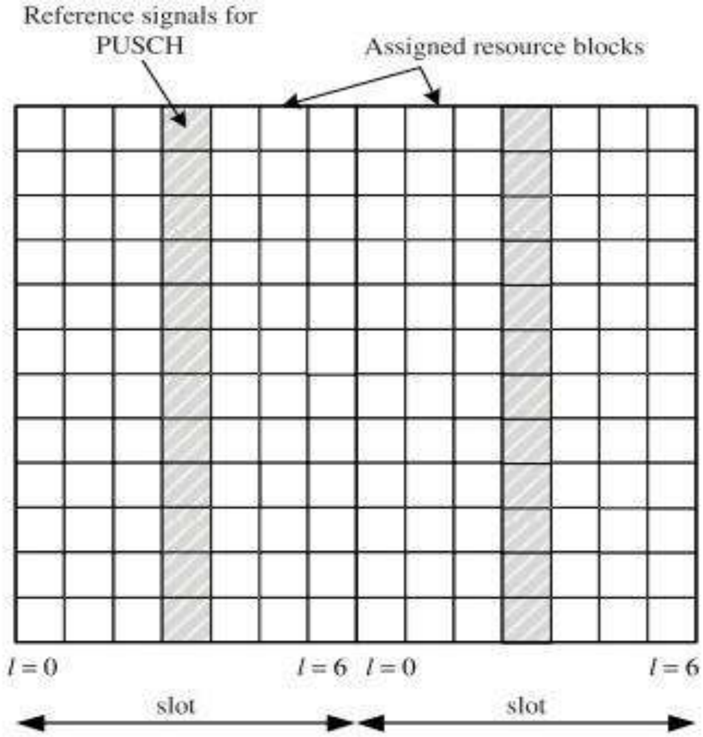
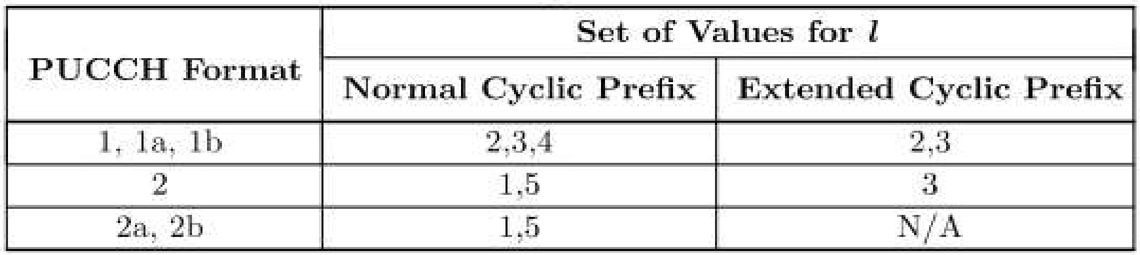


Figure 8.8 Resource mapping of demodulation reference signals for PUSCH with the normal CP.

Table 8.9 Demodulation Reference Signal Location for Different PUCCH Formats



PUCCH Format	Set of Values for l	
	Normal Cyclic Prefix	Extended Cyclic Prefix
1, 1a, 1b	2,3,4	2,3
2	1,5	3
2a, 2b	1,5	N/A

- The number of PUCCH demodulation reference symbols are different for different formats, which is related to the number of control symbols for each format.
- There are 10 CQI/PMI modulated symbols for PUCCH format 2/2a/2b, and there are 2 reference symbols in each slot as shown in Table 8.9, so there are a total of 14 symbols that fill the whole subframe, which is of 14 SC-FDMA symbols.
- PUCCH format 1/1a/1b has fewer information bits than PUCCH format 2/2a/2b, there aremore reference symbols for format 1/1a/1b than there are for format 2/2a/2b, which can be used to improve the channel estimation performance.

8.4.3 Resource Mapping of Sounding Reference Signals

Explain about resource mapping of sounding reference signal

- In FDD mode, it is transmitted in the last SC-FDMA symbol in the specified subframe.
- In the TDD mode, the sounding reference signal is transmitted only in configured uplink subframes or the UpPTS field in the special subframe.

- The subframes in which the sounding reference signals are transmitted are indicated by the broadcast signaling, and there are 15 different configurations

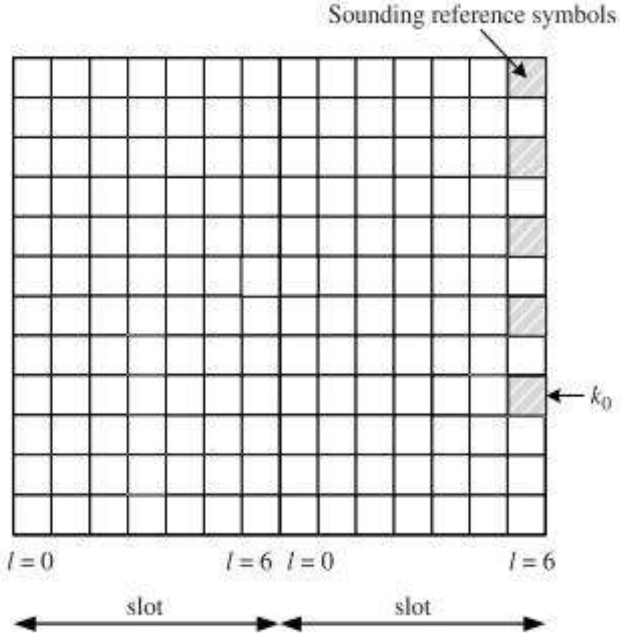


Figure 8.10 An example of resource mapping of sounding reference signals, with the normal CP.

The bandwidth of sounding reference signals is configured by higher layers and also depends on the system bandwidth. An example of resource mapping of sounding reference signals is shown in Figure 8.10.

8.5 Random Access Channels (RCH)

- The uplink random access procedure is used during initial access or to re-establish uplink synchronization.
- As shown in Figure 8.11, the random access preamble consists of a CP of length T_{cp} and a sequence part of length T_{SEQ}.

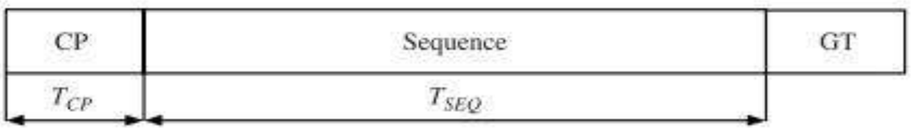
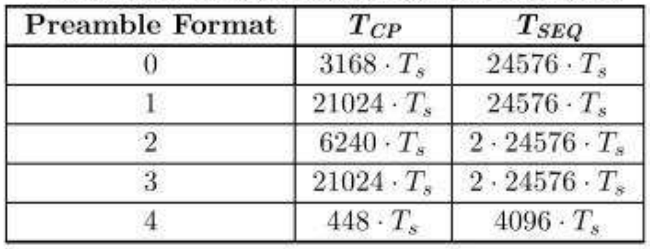


Figure 8.11 The random access preamble format.

- Guard Time (GT) is also needed to account for the round trip propagation delay between the UE and the eNode-B.
- The values of T_{cp} and T_{SEQ} depend on the cell size and base station implementation. There are five different preamble formats defined in LTE, specified in Table 8.10

Table 8.10 Random Access Preamble Parameters

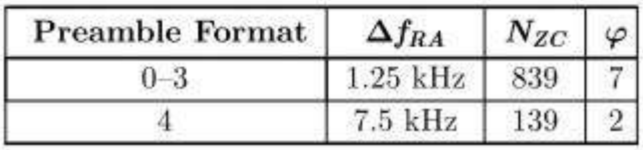


Preamble Format	T_{CP}	T_{SEQ}
0	$3168 \cdot T_s$	$24576 \cdot T_s$
1	$21024 \cdot T_s$	$24576 \cdot T_s$
2	$6240 \cdot T_s$	$2 \cdot 24576 \cdot T_s$
3	$21024 \cdot T_s$	$2 \cdot 24576 \cdot T_s$
4	$448 \cdot T_s$	$4096 \cdot T_s$

- *Format 0*: It is for normal cells.
- *Format 1:* It is also known as the extended format, is used for large cells.
- *Format 2 and format 3:* These are use repeated preamble sequences to compensate for increased path loss, and are used for small cells and large cells, respectively.
- *Format 4:* It is defined for frame structure type 2 only.
- The network configures the set of preamble sequences that the UE is allowed to use.
- In each cell, there are 64 available preambles, which are generated from one or several root Zadoff-Chu sequences.
- There is no intra-cell interference from multiple random access attempts using different preambles in the same cell due to Zadoff-Chu sequences.
- The Physical Random Access Channel (PRACH) resources within a radio frame are indicated by a PRACH configuration index, which is given by higher layers.
- In the frequency domain, the random access burst occupies a bandwidth corresponding to six consecutive resource blocks (72 subcarriers) in a subframe or a set of consecutive subframes.
- The PRACH uses a different subcarrier spacing (Δf_{RA}) than other physical channels, which is

listed in Table 8.11 together with the preamble sequence length N_{zc}.

Table 8.11 Parameters for Random Access Preamble



Preamble Format	Δf_{RA}	N_{ZC}	φ
0–3	1.25 kHz	839	7
4	7.5 kHz	139	2

- The continuous-time random access signal is defined by:

The image part with relationship ID rId37 was not found in the file.

8.6 H-ARQ in the Uplink

Brief out H-ARQ in the UPLINK

Explain FDD and TDD mode of H-ARQ in LT uplink

- The H-ARQ retransmission protocol is also used in the LTE uplink, so the eNode-B has the capability to request retransmissions of incorrectly received data packets.
- Uplink H-ARQ process, the corresponding ACK/NAK information is carried on the PHICH.
- LTE uplink applies the synchronous H-ARQ protocol, that is, the retransmissions are scheduled on a periodic interval
- Synchronous retransmission is preferred in the uplink because it does not require to explicitly signal the H-ARQ process number so there is less protocol overhead.
- The number of H-ARQ processes and the time interval between the transmission and retransmission depend on the duplexing mode and the H-ARQ operation type.
- There are two types of H-ARQ operation in the uplink:
  1. *The non-subframe bundling operation (normal H-ARQ operation)*
  2. *Subframe bundling operation (also called TTI[3] bundling)*

8.6.1 The FDD Mode: For the FDD mode

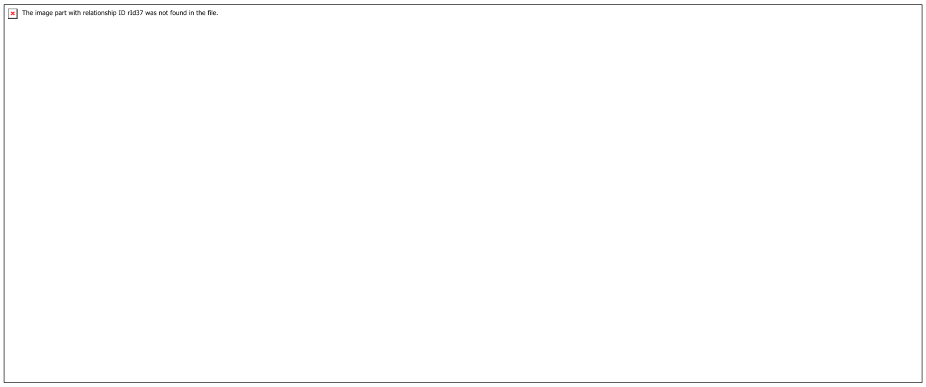
- There are eight parallel H-ARQ processes in the uplink for the non-subframe bundling operation, and four H-ARQ processes for the subframe bundling operation.
- For the FDD mode with the normal H-ARQ operation, upon detection of an NAK in subframe n, the UE retransmits the corresponding PUSCH in subframe n + 4
- For the FDD mode with the subframe bundling operation, upon detection of an NAK in subframe n - 5, the UE retransmits the corresponding first PUSCH transmission in the bundle in subframe n + 4.

8.6.2 The TDD Mode: For the TDD mode,

- The number of H-ARQ processes is determined by the DL/UL configuration, listed in Table 8.12.
- For TDD UL/DL configurations 1-6 and the normal H-ARQ operation, upon detection of an NAK in subframe n, the UE retransmits in subframe n + k, with k given in Table 8.13.

☒ The image part with relationship ID rId37 was not found in the file.

- *For TDD UL/DL configuration 0 and the normal H-ARQ operation*: Upon detection of an NAK in subframe n, the UE will retransmit in subframe n + 7 or n + k with k given in Table 8.13, which depends on the UL index field in DCI and the value of n.


The image part with relationship ID rId37 was not found in the file.

- *For TDD UL/DL configurations 1 and 6 with subframe bundling*: Upon detection of an NAK in subframe n. — I with 1 given in Table 8.14, the UE retransmits the corresponding first PUSCH transmission in the bundle in subframe $n + k$, with k given in Table 8.13.

The image part with relationship ID rId37 was not found in the file.

- *For TDD UL/DL configuration 0 and the subframe bundling operation*: Upon detection of an NAK in subframe $n - l$ with$l$ given in Table 8.14, the UE retransmits in subframe n + 7 or $n + k$ with k given in Table 8.13, depending on the UL index field in DCI and the value of n.

Chapter 9

Physical Layer Procedures and Scheduling

1. Introduction:

- The physical layer procedures that provide crucial services to higher layers.
- The physical layer procedures or functions in LTE includes
  - *Hybrid-ARQ (H-ARQ) and Channel Quality Indicator (CQI)*
  - *Dynamic channel-dependent scheduling and MIMO transmission.*
  - *Precoding for MIMO closed-loop operations.*
  - *The Rank Indicator (RI) and Precoder Matrix Indicator (PMI) feedback.*
  - *Cell search and Random accesses procedures*.
  - *Power control in uplink*.

9.2 Hybrid-ARQ (H-ARQ) Feedback

- In LTE, the H-ARQ protocol is applied to improve the transmission reliability over the wireless channel.
- HARQ = ARQ+FEC (Forward Error Correction)/Soft Combining.
- Soft Combining is an error correction technique in which the bad packets are not discarded but stored in a buffer. The basic idea is that 2 or more packets received with insufficient information can be combined together in such a way that total signal can be decoded.
- A mechanism H-ARQ is implemented to correct the error packets in the PHY layer.
- Principle of H-ARQ: *Works at PHY layer but controlled by MAC layer. If the received data has an error then the receiver buffers the data and requests a re-transmission from the sender. When the receiver receives the re-transmitted data, it then combines it with buffered data prior to channel decoding and error detection. This helps the performance of the re-transmissions.*
- Two methods of H-ARQ protocol uses in LTE
  1. *Asynchronous adaptive H-ARQ protocol*: Used by down link transmission
     - The retransmissions can take place whenever in time, due to scheduling purposes.
     - Need an appropriate signaling to make the transmitter aware of which HARQ process we are considering.
     - The TTI and resource allocation for the retransmission is dynamically determined by the scheduler.
     - Asynchronous HARQ increases flexibility because re-transmissions doesn't have to be scheduled during every sub-frame.
  2. *Synchronous adaptive H-ARQ protocol:* Used by uplink transmission
     - Re-transmissions are scheduled at fixed time intervals.
     - Generates lower over-head as it doesn't need to include HARQ process Id in the outgoing data

       3. Always works in cycle even if no resources are allocated during a specific sub frame; which means that the 1st process will repeat itself after every 8 ms.
- The H-ARQ feedback is different for FDD and TDD mode.

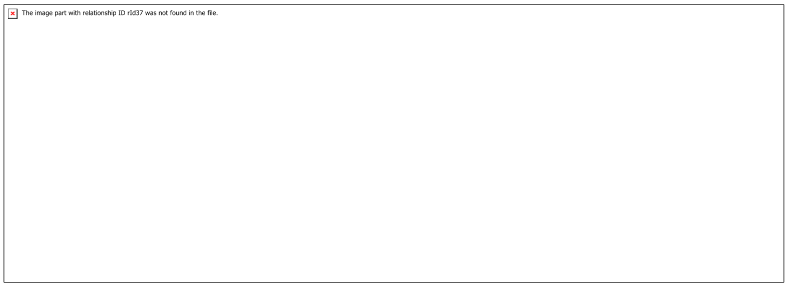
9.2.1 H-ARQ Feedback for Downlink (DL) Transmission:

What is meant by H-ARQ feedback for downlink transmission and uplink transmission

Procedures are

- UEs need to feedback the associated ACK/ NAK information on PUCCH or PUSCH.
- One ACK/NAK bit is transmitted in case of single-codeword downlink transmission, while two ACK/NAK bits are transmitted in case of two-codeword downlink transmission.
- H-ARQ retransmission happens when the channel is static or experiences little or no variation between subsequent H-ARQ transmissions.

- *H-ARQ feedback for FDD mode*: Procedure are
  1. UE transmits H-ARQ-ACK in subframe n for a PDSCH transmission in subframe $n-4$.
  2. The reason for the 4 subframe delay in the transmission of an ACK/NACK message is due to the processing delay of about 3 ms at the receiver.
  3. Both H-ARQ-ACK and Scheduling Request (SR) are transmitted in the same subframe.
  4. UE shall transmit the H-ARQ-ACK on its associated H-ARQ-ACK PUCCH resource for a negative SR transmission.
  5. UE transmit the H-ARQ-ACK on its assigned SR PUCCH resource for a positive SR transmission.
- H-ARQ feedback for TDD mode: Procedure are
  1. In this mode, the time association between the data transmission and the ACK/NACK cannot be maintained due to the variable numbers of DL and UL subframes being present in a frame.
  2. The UL and DL delay between data and ACK is dependent on the TDD configuration chosen. Hence, a fixed delay between a transmission and the HARQ ACK/NACK is not possible in TDD-LTE.
  3. For TDD, two ACK/NAK feedback modes are supported by higher layer configuration:
     - *ACK/NAK bundling using PUCCH format 1a or 1b, which is the default mode and consists of one or two bits of information*
     - *ACK/NAK multiplexing using PUCCH format 1b, which consists of between one and four bits of information*
- In TDD, the delay between the transmission and the HARQ ACK/NACK depends on both the TDD configuration and the subframe in which the data was transmitted.
- For example, in configuration 1 which is shown below in Figure 9.1, there are some DL subframes for which the nearest UL subframe (greater than a separation of 4 or more subframes) is 7 subframes away. In the Figure 8.12 shown below, this can be seen clearly for the DL data transmission case.



☒ The image part with relationship ID rId37 was not found in the file.

Figure 9.1: HARQ ACK/NACK Timing for configuration 1

- The feedback of H-ARQ-ACK in the UL subframe n corresponds to the detection of the PDSCH transmission within subframe(s) n — k, where the parameter k is different for different UL/DL configurations and different subframe and $k \in K$ with K specified in Table 9.1. For ACK/NAK bundling.
- Multiple acknowledgments are combined by a logical AND operation, and then the bundled 1 or 2 ACK/NAK bits are transmitted using PUCCH boost 1a and PUCCH format 1b, respectively.

☒ The image part with relationship ID rId37 was not found in the file.

9.2.2 H-ARQ Indicator for Uplink (UL) Transmission: Procedure are

- Only a single-bit H-ARQ Indicator (HI) needs to be sent to each scheduled UE, which is carried on the PHICH physical channel.
- For the FDD mode, an ACK/NAK received on the PHICH assigned to a UE in subframe n is associated with the PUSCH transmission in subframe n - 4.
- For the TDD mode, different from the feedback for downlink transmission, there is no problem to transmit multiple acknowledgments on PHICH.
- For UL/DL configurations 1-6, an ACK /NAK received on the PHICH in subframe n is associated with the PUSCH transmission in the subframe n - k as indicated in Table 9.2. For TDD with UL/DL configuration 0:
  - *If there is PUSCH transmission in subframe 4 or 9, an ACK/NAK received on the PHICH in subframe n is associated with the PUSCH transmission in the subframe n - 6.*
  - *Otherwise, an ACK/NAK received on the PHICH in subframe 11 is associated with the PUSCH transmission in the subframe n - k with k indicated in Table 9.2.*

☒ The image part with relationship ID rId37 was not found in the file.

9.3 Channel Quality Indicator (CQI) Feedback: It includes

1. *Introduction*
2. *CQI estimation*
3. *CQI feedback modes*

Explain channel quality indicator feedback

9.3.1 **Introduction***: CQI is an indicator carrying the information on how good/bad the communication channel quality is. The CQI basically includes CQI, PMI (Precoding Matrix Indicators), RI (Rank Indicator) components. The requirement for each of these components depend on transmission mode. All transmission modes need UE to provide CQI feedback.*

- *CQI reporting contains information sent from a UE to the eNode-B to indicate a suitable downlink transmission data rate, i*.e., a Modulation and Coding Scheme (MCS) value.
- CQI is a 4-bit integer and is based on the observed signal to-interference-plus-noise ratio (SNIR) at the UE.
- The CQI estimation process takes into account the UE capability such as the number of antennas and the type of receiver used for detection.
- The CQI reported value are used by the eNode-B for downlink scheduling and link adaptation, which are important features of LTE.
- LTE supports wideband and subband CQI reporting.
  1. A wideband CQI reporting:
     - *The wideband report provides one CQI value for the entire downlink system bandwidth.*
     - *It is a value of single 4-bit integer that represent an effective SINR as observed by the UE.*
     - *It is most efficient in terms of uplink bandwidth consumption since it requires only a single 4-bit feedback.*
     - *With wideband CQI, the variation in the SINR across the channel due to frequency selective nature of the channel is masked out.*
     - *It is not suitable for frequency selective scheduling.*
     - *It is the preferred mode to use for high speeds where the channel changes rapidly since frequent subband reporting would exhaust a large portion of the uplink bandwidth.*
     - *Wideband CQI is also the preferred mode for services such as VoIP where a large number of simultaneous UEs are supported and latency is more critical than the overall throughput since VoIP is typically a low data rate application with very strict latency requirement.*
  2. A subband CQI reporting:
     - *To support frequency selective scheduling, each UE needs to report the CQI with a fine frequency granularity, which is possible with subband CQI reporting.*

- *A subband CQI report consists of a vector of CQI value, where each CQI value is representative tithe SINR observed by the UE over a sub-band.*
- *A subband is a collection of n adjacent Physical Resource Blocks (PRBs) where the value tint can be 2, 3, 4, 6, or 8 depending on the channel bandwidth and the CQI feedback mode.*
- *It requires more uplink bandwidth but is more efficient since it allows for a frequency selective scheduling, which maximizes the multiuser diversity gain.*

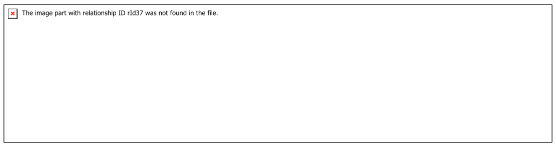
- Note:

  (1). *One of the critical aspects of designing the CQI feedback mechanism for LTE is the optimization between the downlink system performance and the uplink bandwidth consumed by the feedback mechanism*.

  (2). The LTE standard does not specify how to select between wideband and subband CQI reporting depending on the UE speed or the QoS requirements of the application. It is left up to the equipment manufacturer to develop their proprietary algorithms in order to accomplish this.

9.3.2 A Primer on CQI Estimation ***Brief out primer on CQI estimation***

- Downlink cell-specific *Reference Signals (RS)* are used by each UE to estimate the MIMO LTE channel from the eNode-B.
- The estimated MIMO channel along with the known reference signal is then used to calculate the other-cell interference level.
- The important thing is to understand that reference signals are sent in both UL and DL while the CQI is sent in UL only (either on PUCCH or PUSCH) but it reports the DL signal strength based on the channel estimation.
- The UE uses the estimated channel and interference plus noise variance to compute the SINR on the physical resource element (PRE) carrying the reference signal.
- The UE computes SINR samples over multiple OFDM symbols and subcarriers, which are then used to calculate an effective SINR. The effective SINR is given as:



The image part with relationship ID rId37 was not found in the file.

Where N is the number of samples. α_1 *and* α_2 adapt to different modulation and coding schemes.

- Exponential Effective SINR Mapping (EESM) and the mutual information-based methods are preferred since they have been shown to give a more accurate estimate of the channel quality.

- In the case of wideband CQI feedback, the UE measures the SINR from the reference signal over all the PRBs, and then computes its CQI based on the effective SINR across the entire channel bandwidth.
- In subband CQI the UE measures the SINR over the PRBs contained in the given subband, and then computes the CQI.
- If a UE reports a CQI value for a particular subband, it is called *subband feedback*.
- If a UE reports a single CQI value for whole system bandwidth, it is called *wideband feedback*.
- Based on the estimated effective SINR, the UE picks the CQI index that indicate the highest MCS level (modulation and code rate) that can be supported with a 10% BLER on the first H- ARQ transmission.
- The CQI feedback is used by the eNode-B to select an optimum PDSCH transport block with a combination of modulation scheme and transport block size corresponding to the CQI index that could be received with target block error probability after the first H-ARQ transmission.
- The target block error probability is left open as an implementation choice, typical values are in the range of 10-25%.
- The supported CQI indices and their interpretations are given in Table 9.3. In total, there are 16 CQI values, which require a 4-bit CQI feedback. In Table 9.3, the efficiency for a given CQI index is calculated as: efficiency = Q,,, × code rate,

  Where Q,,, is the number of bits in the modulation constellation. Taking CQI index 4 as an example, as Q,,, = 2 for QPSK. We have efficiency = $2 \times \frac{308}{1024} \approx 0.6016\ bits/symbol$.

Table 9.3 4-Bit CQI Table

☒ The image part with relationship ID rId37 was not found in the file.

9.3.3 CQI Feedback Modes *How reporting of CQI,PMI and RI in the time domain are categorized*

- There are two reporting CQI feedback modes in the time domain
  1. Periodic reporting*: The UE reports CQI, PMI, and RI with reporting periods configured by the higher layer. PUCCH is used for this report.*
  2. Aperiodic reporting: *It can be used to provide large and more detail reporting in a single reporting instance via PUSCH. Report Timing is triggered by DCI*
- Note: *In cases where both periodic reporting on the PUCCH and the aperiodic reporting PUSCH happen to be on the same subframe, the UE will only transmit the aperiodic report over the PUSCH and ignore the periodic PUCCH report.*
- Both periodic and aperiodic reporting modes support wideband and subband CQI reporting.

In LTE there are two distinct reporting mechanisms for subband CQI feedback when the aperiodic reporting mode is used:

(*Differentiate between higher layer configured subband and UE selected subband report)*

1. *Higher Layer Configured Subband Report*: In this case, the UE reports the subband CQI for each band in a single feedback report.
2. The size of a band is specified by a higher layer message and is contingent on the system bandwidth.
3. *UE Selected Subband Report*: In this case, the UE reports the subband CQI for the 'M' bands with the highest CQI values. The CQI for the rest of the bands is not reported.

- The average per sector downlink throughput for various wideband and subband CQI feedback modes as shown Figure 9.2. These result are typical of a 10MHz FDD system in a multicell.

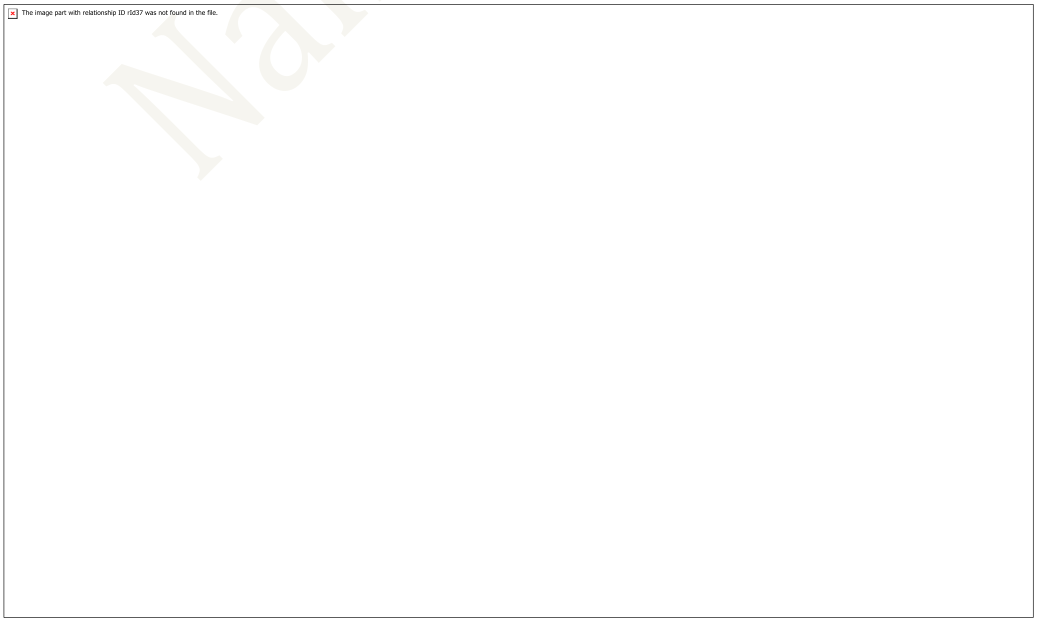
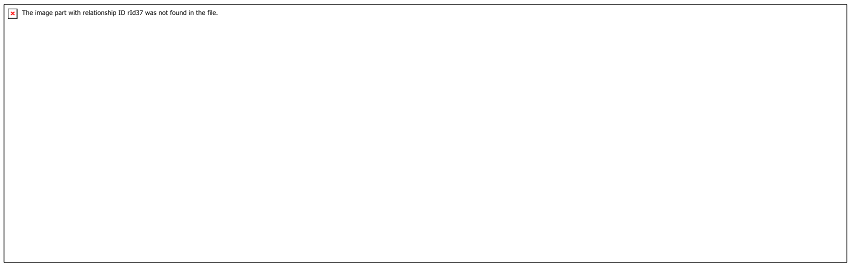

The image part with relationship ID rId37 was not found in the file.

Figure 9.1: The average downlink throughput per sector for various CQI feedback

- Each reporting class supports a number of different reporting modes. Where each reporting mode is characterized by a specific CQI feedback type and a PMI feedback type. Which are listed in Table 9.4 and Table 9.5 for periodic reporting and aperioclic reporting respectively.


The image part with relationship ID rId37 was not found in the file.

The image part with relationship ID rId37 was not found in the file.

- There are seven transmission modes in the downlink and each of them supports a specific subset of the reporting mode the details of which are shown in Table 9.6.

The image part with relationship ID rId37 was not found in the file.

- **Periodic CQI Reporting:**

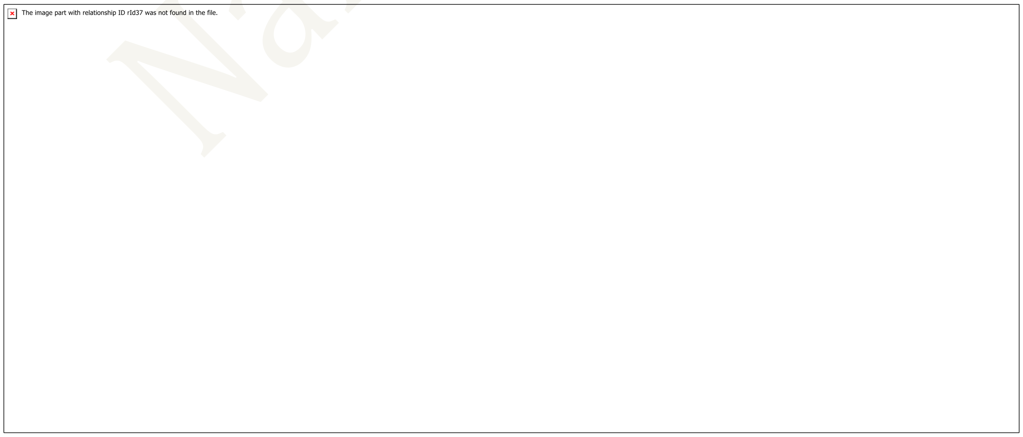
  UE is semi-statically configured by higher layers. UE periodically feedback CQI on the PUCCH in one of the reporting mode given in Table 9.4. Note that

- Mode 1-0 and 2-0 do not report PMI and they are used for OL MIMO modes and single-antenna port transmission.
- Mode 1-1 and mode 2-1 report a single PMI for CL NIIMO modes, i.e., only the wideband PMI is reported.
- The periodic CQI feedback is useful for scheduling and adaptive modulation and coding, and can also be used to check or change semi-static parameters such as the MIMO mode or transmission mode. Considering the reporting for CQI/PMI and RI, there are four different reporting types supported for each of these reporting modes as given in Table 9.7:
  1. Type 1: report supports CQI feedback for the UE selected subbands.
  2. Type 2: report supports wideband CQI and Phil feedback.
  3. Type 3: report supports RI feedback.

4. Type 4: report supports wideband CQI.

The image part with relationship ID rId37 was not found in the file.

- **Aperiodic CQI Reporting**
  - The UE shall perform aperiodic CQI, PMI and RI reporting using the PUSCH channel in subframe n + k. The value of it is specified as follows:
    a. For FDD, k = 4.
    b. For TDD UL/DL configuration 1-6, k, is given in Table 9.10.
    c. For TDD UL/DL configuration 0:


The image part with relationship ID rId37 was not found in the file.

  - As shown in Table 9.5, there are three different aperioclic CQI feedback types:
    1. Wide-band feedback
    2. Higher layer-configured subband feedback,
    3. UE-selected subband feedback.
    4. Five reporting modes.
    5. Modes 2-0 arid 3-0 are for single-antenna-port transmission and OL MIMO modes, while Mode 3-1 with single PMI and Modes 1-2 and 2-2 with multiple P511 are for

CL MEMO modes.

- Aperiodic CQI reporting supporting both Wideband and subband feedback.
    i. *Wideband feedback*: a UE selects a preferred precoding matrix for each subband, assuming transmission only in that subband. Then each UE reports one wideband CQI value for each codeword, assuming the use of the selected precoding matrix in each subband, and it also reports the selected PMI for each subband.
    ii. *Higher Layer-Configured Subband Feedback:* There are two different reporting modes with higher layer-configured subband feedback: Mode 3-0 (without PMI) and Mode3-1 (with single PMI). The supported subband size k is the same as that for the periodic reporting, as in Table 9.8. As a separate CQI is reported for each subband, this reporting type provides the finest frequency granularity but also has the highest overhead.

Explain a) UE selected Subband CQ b) Wideband CQI c) A periodic CQI reporting

Summarize UE-Selected Subband feedback

9.4 Precoder for Closed-Loop MIMO Operations

9.4.1 Introduction:

Explain 4 different MIMO modes

- MIMO transmission is a key technique in LTE and can provide a significant *throughput and gains*, especially with the spatial multiplexing mode.
- The amount of feedback required to provide the full CSI (*Channel State Information*) to the eNode-B is large, particularly in multicarrier systems.
- In order to mitigate the feedback issue, limited *feedback mechanisms* are used in LTE.
- The UE chooses the optimum rank and precoder for downlink transmission based on a predefined set of precoders, also known as a "*Codebook"*.
- Instead of indicating the full precoding matrix, the UE only needs to indicate the *index of the precoding matrix* from the codebook.
- RI is reported by the UE to indicate the number of layers, i.e., the number of data streams used in spatial multiplexing.
- For CL MEMO modes, i.e., the transmission modes 4, 5, and 6, the preferred precoding matrix in the predefined codebook needs to be reported, which is provided by the PMI.

9.4.2 Precoder Estimation for Multicarrier Systems ***What is precode estimation for multicarrier system***

- The precoder estimation at the UE can be done based on a few different metrics. The most common metric is the capacity-based one.
- The precoder is chosen to maximize the MIMO capacity of the effective channel, which includes the radio channel and the precoder.

- CL MIMO system, the interference is dynamic in nature, as the precoders used at interfering cells change from one TTI to the next. Thus, choosing a precoder based on the instantaneous interference seen by a UE can lead to suboptimal performance.
- It is better to choose the precoder based on long-term characteristic of the interference such as the interference variance at each receive antenna.
- For the I^{th}subcarrier, the achievable rate for Minimum mean square error (MMSE) receiver is

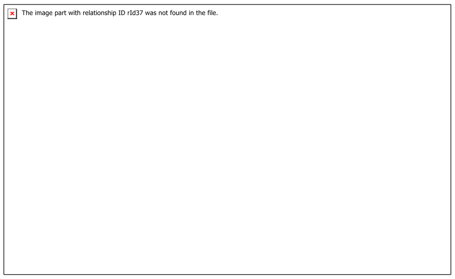
The image part with relationship ID rId37 was not found in the file.

- The precoder is chosen to maximize R_{sum} for a given subband (subband PMI) or the entire bandwidth (wideband PMI).

9.4.3 Precoding Matrix Index (PMI) and Rank Indication (RI) Feedback

What is precoding matrix index and rank indication feedback

- The RI report is determined from the supported set of RI values for the corresponding eNode-B and the UE antenna configuration.
- The value of RI can be 1 or 2 for two-antenna ports and from 1 to 4 for four-antenna ports.
- The mapping between RI bits and the channel rank is shown in Table 9.13.
- UEs need to report RI for both CL and OL MIMO modes.


The image part with relationship ID rId37 was not found in the file.

- For the CL spatial multiplexing, the RI report, together with the PMI, informs the eNode-B to select the suitable precoder.
- For OL MIMO, the RI report supports selection between transmit diversity (RI = 1) and OL

spatial multiplexing (RI > 1).

- Only wideband RI reporting is supported, i.e., only a single RI is reported for the whole bandwidth, as subband RI reporting provides little performance gain.
- In addition, as the channel rank normally changes slowly, the reporting period for RI is longer than CQI in periodic reporting.
- PMI reports the channel-dependent preceding matrix for CL NEMO mode.

9.5 Uplink Channel Sounding *What is uplink channel sounding*

- Channel sounding is mainly used for uplink channel quality measurement at the eNode-B.
- The Sounding Reference Symbol (SRS) is transmitted by the UE in the uplink for the eNode-B to estimate the channel state information, which includes the MIMO channel of the desired signal. SINR, noise. Interference level, etc.
- The SRS can also be used for uplink timing estimation and uplink power control.
- The SRS transmission is always in the last SC-FDMA symbol in the configured subframe, on which PUSCH data transmission is not allowed.
- The eNode-B can either request an individual SRS transmission from a SE or configure a UE to periodically transmit SRS.
- The periodicity may take any value of 2, 5, 10, 20, 40, 80, 160, and 320 ms.
- The UE-specific SRS parameters include
  - The starting physical resource block assignment
  - Duration of SOS transmission
  - SRS periodicity and SRS subframe offset
  - SRS bandwidth
  - Frequency hopping bandwidth and cyclic shift.
- The above parameters are semi-statically configured by higher layers.
- A UE shall not transmit SRS in the following scenarios:
  - *If SRS and PUCCH format 2/2a/2b transmissions happen to coincide in the same subframe*
  - *Whenever SRS and ACK/NAK and/or positive SR transmissions happen to coincide in the same subframe unless the parameter Simultaneous-AN-and-SRS is TRUE*

9.6 Buffer Status Reporting in Uplink

Explain buffer status reporting in uplink

- A Buffer Status Report (BSR) is sent from the UE to the serving eNode-B to provide information about the amount of pending data in the uplink buffer of the UE.
- The buffer status along with other information, such as priorities allocated to different logical channels is useful for the uplink scheduling process to determine which UEs or logical channels

should be granted radio resources at a given time.

- A BSR is triggered if any of the following events occurs:
  - a. *Regular BSR:* Uplink data for a logical channel becomes available for transmission, and either the data belongs to a logical channel with higher priority than the priorities of the data available for transmission for any of the logical channels.
  - b. *Padding BSR*: Uplink resources are allocated and the number of padding bits is equal to or larger than the size of the BSR MAC control element.
  - c. A serving cell change occurs, in which case the BSR is referred to as *"regular BSR."*
  - d. The retransmission BSR timer expires and the UE has data available for transmission, in which case the BSR is referred to as *"regular BSR."*
  - e. *Periodic BSR*: BSR timer expires, in which case the BSR is referred to as "periodic BSR."
- The buffer status is reported on a per radio bearer' (logical channel) group basis.
- There are two BSR formats used in the LTE uplink: short BSR that reports only one radio bearer group, and long BSR that reports multiple radio bearer groups.
- For regular and periodic BSR, if more than one radio bearer group has data available for transmission in the TTI where the BSR is transmitted,
- long BSR is reported: otherwise. Short BSR is reported. For padding BSR:
  - a. When the number of padding bits is equal to or larger than the size of the short BSR plus its subheader but smaller than the size of the long BSR plus its subheader truncated BSR with the highest priority logical channel is reported if more than one logical channel group has buffered data: otherwise, short BSR is reported.
  - b. If the number of padding bits is equal to or larger than the size of the long BSR plus its subheader, long BSR is reported.
- When the BSR procedure determines that at least one BSR has been triggered, and then if the UE has been allocated uplink resources, a buffer status report is transmitted.
- If a regular BSR has been triggered and the UE has no allocated uplink resource, a scheduling request for a BSR transmission is triggered.
- A MAC PDU shall contain at most one MAC BSR control element, even when multiple events trigger.
- In this case, the regular BSR and the periodic BSR shall have precedence over the padding BSR. All triggered BSRs shall be cancelled in the following two scenarios:

1. *The uplink grant can accommodate all pending data available for transmission but is not sufficient to additionally accommodate the BSR MAC control element*
2. *A BSR is include in a MAC PDU for transmission.*

9.7 Scheduling and Resource Allocation *What is meant by scheduling and resource allocation*

- The main aim of scheduling and resource allocation is to efficiently allocate the available radio resources to UEs to optimize a certain performance metric with QoS requirement constraints.
- Scheduling algorithms for LTE can be divided into two categories:

1. ***Channel-dependent scheduling***:
   - Dynamic channel-dependent scheduling is one of the key feature to provide high spectrum efficiency in LTE.
   - The allocation of RBs to a UE is based on the channel condition, e.g., proportional fairness scheduler, max Carrier Interfernce scheduler.
   - To better exploit the channel selectivity, the packet scheduler is located in the eNode-B, which allocates physical layer resource for both the DL-SCH and UL-SCH transport channels every TTI.
   - Scheduling depends heavily on the channel information available at the eNode-B, which is provided by the uplink CQI reporting for the downlink channel and by channel sounding in the uplink channel.
   - The scheduler should also take account of the traffic volume and the QoS requirement of each UE and associated radio bearers.
   - Due to the use of OFDMA/SC-FDMA, LTE is able to exploit the channel variation in both the time and frequency domain, which is a major advantage compare to HSPA (3G).
2. ***Channel dependent scheduling***:
   - The objective of channel-dependent scheduling is to exploit multiuser diversity to improve the spectrum efficiency.
   - Here allocate resource blocks to a UE is random and not based on channel condition, e.g., round-robin scheduler.
   - It should also consider such issues as fairness and QoS requirements.
   - In addition, scheduling is tightly integrated with link adaptation and the H-ARQ process.
   - The scheduling algorithm is not standardized and it is vendor specific.

- In a multicarrier system such as LTE, channel-dependent scheduling can be further divided into two categories:
  1. ***Frequency diverse scheduling***: The UE selection is based on wideband CQI. However, the PRB allocation in the frequency domain is random. It can exploit time selectivity and frequency diversity of the channel.
  2. ***Frequency selective scheduling:*** The UE selection is based on both wideband and subband CQI, and the PRB allocation is based on the subband CQI. This can exploit both time and frequency selectivity of the channel.

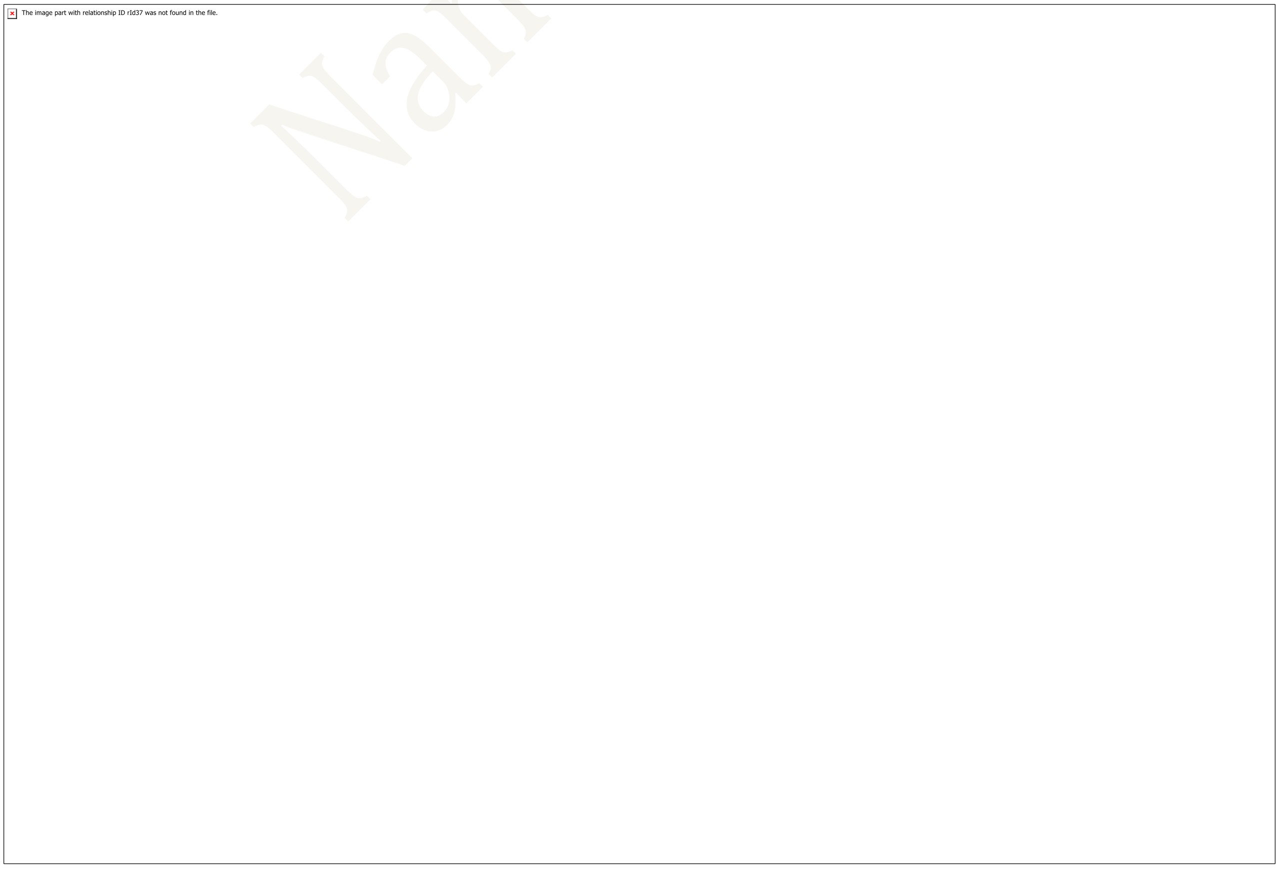
9.7.1 Signaling for Scheduling in Downlink and Uplink

Explain signalling for scheduling in downlink and uplink

Brief out changes in the signalling structure in semi-persistent scheduling

- eNode-B scheduler dynamically controls time-frequency resources are allocated to a certain UE in both downlink and uplink.
- The resource assignments, including the assigned time/frequency resources and respective transmission formats, are conveyed through downlink control signaling.
- The minimum size of radio resource that can be allocated to a UE corresponds to two resource blocks, which is 1 ms duration in the time domain and 180 kHz in the frequency domain.
- Both localized and distributed resource allocations are supported in the downlink, while in the uplink UEs are always assigned contiguous resources.
- In addition, there is a strict constraint on the UE transmit power in the uplink.

- ***Signaling for Downlink Scheduling:*** *Producers are*
  - The channel state information (CSI) at the eNode-B for the downlink scheduling is obtained through CQI reporting from UEs.
  - Based on the CQI, eNode-B dynamically allocates resources to UEs at each TTI.
  - A UE always monitors the PDCCH for possible allocations and decode the PDCCH with CRC.
  - The UE shall decode the PDCCH and any corresponding PDSCH according to the respective combinations defined in Table 9.16.


The image part with relationship ID rId37 was not found in the file.

- ***Signaling for Uplink Scheduling***
  - In the uplink, the CSI is estimated at the eNode-B with the help of sounding reference signals.
  - A UE always monitors the PDCCH in order to find possible allocation for uplink transmission.
  - Only contiguous resource blocks can be allocated to a UE due to the SCFDMA nature of the UL transmission.
  - Frequency hopping can be applied to provide additional diversity.
  - The UE obtains the uplink resource allocation as well as frequency hopping information from the uplink scheduling grant received four subframes earlier.
  - To determine the modulation order, redundancy version, and transport block size for the PUSCH, the UE shall first read the 5-bit "modulation and coding scheme and redundancy version" field (I_{MCS}) in the DCI.
  - The mapping between I_{MCS}, modulation order, and I_{TBS} for the PUSCH is shown in Table 9.18.

☒ The image part with relationship ID rId37 was not found in the file.

9.7.2 Multiuser MIMO Signaling

Write a note on multiuser MIMO signalling

- If MU-MINIO is used in the uplink, then it is transparent to the UE with the exception that two UEs should transmit orthogonal reference signals in order for the eNocle-B to separate them.
- The uplink resource allocation is indicated on PDCCH using DCI format 0, which contains a 3-bit field to indicate the cyclic shift in the reference signal to be used by each UE.
- When MU-MIMO is used in the downlink, two rank-1 UEs are multiplexed on the same physical resource.
- In this case the power for each UE is reduced by 3 dB. This is indicated by the power offset field in DCI format ID which is used for MU-MIMO scheduling.

9.8 Cell Search*** *Explain the cell search process in LTE.*

- Cell Search means the collective term representing the combined procedure of Measurement, Evaluation and Detection process.
- This is very tightly related to Cell Selection process because UE goes through this search process first before it goes through the cell selection.
- Also this process influence greatly on energy consumption of UE during the idle mode.
- When a UE powers on, it needs to acquire time and frequency synchronization with a cell and detect the physical-layer cell ID of that cell through the cell search procedure or synchronization procedure.
- During cell search, different types of information need to be identified by the UE, including *symbol and frame timing, frequency*, *cell identification*, *transmission bandwidth*, *antenna configuration, and the cyclic prefix length.*
- LTE uses a hierarchical cell search scheme similar to WCDMA, demonstrated in Figure 9.4.

☒ The image part with relationship ID rId37 was not found in the file.

- **Cell search procedure**: *A cell search is nothing but a procedure that a UE shall perform in order to get more details about the nearby eNodeB/cell. So this is the first step that a UE shall do, as soon as it is powered on. The cell search procedure, is the UE's way of acquiring cell specific information and this, the UE has to perform once or several times based on the condition of the network.*
- ***Cell Search step 1***: The UE detects the symbol timing and the cell ID index $N_{ID}^{(2)}$ from the primary synchronization signal. This is achieved through matched filtering between the received signal and the primary synchronization sequences. Three orthogonal sequences defined for the primary synchronization signal, the cell ID index $N_{ID}^{(2)}$, can be detected by identifying the received sequence. Frequency and Time synchronization can be performed based on the primary synchronization signal. OFDM symbol timing can be detected, but as there are two primary synchronization signals transmitted in each frame that are indistinguishable, frame timing cannot be detected.
- ***Step 2***: The UE detects the cell ID group index $N_{ID}^{(1)}$ and frame timing from the secondary synchronization signal. The index $N_{ID}^{(1)}$; is detected by identifying the shift in the m-sequence in the received signal. For detecting the frame timing, the pair of secondary synchronization signals in a radio frame has a different structure than primary synchronization signals.
- ***Step 3:*** After the cell search, the UE can detect the broadcast channel to obtain other physical layer information, e.g., system bandwidth, number of transmit antennas, and system frame number.
- ***Step 4:*** The system information is divided in to Master Information Block (MIB) transmitted on the PBCH and System Information Blocks (SIB) transmitted on the PDSCH. At this stage, the UE detects MIB from the PBCH. To maintain the uplink intra-cell orthogonality, uplink transmissions from different UEs should arrive at the eNode-B within a cyclic prefix. This is achieved through the timing advance procedure.
- ***Step 5:*** The timing advance is obtained from the uplink received timing and scot by the eNode-B to the UE. The UE advances or delays its timing of transmissions to compensate for propagation delay and thus time-aligns its transmissions with other UEs. The timing advance command is on a per-need basis with a granularity in the step size of $0.52\mu s$

9.9 Random Access Procedures (RACH): *What do you meant by random access procedure, Why it is needed?*

- In order to be synchronized with the network, RACH procedure is used by UE.
- Suppose a UE wants to access the network, so first it will try to attach or synchronize with the network. In LTE a separate channel PRACH (Physical Random Access Channel) is provided for initial access to the network.

- The UEs also obtain uplink timing information from the initial handshake.
- In LTE, there are two random access mechanisms:

  1. Non-synchronized random access: *Non-synchronized random access is used ohm the UE uplink has not been time synchronized, or when the UE uplink loses synchronization. Its main purpose is to obtain synchronization of the uplink, notify the eNode-B that the UE has data to transmit, or transmit a small amount of control information and data packets.*
  2. Synchronized random access: *Synchronized random access is used when uplink synchronization is present. Its main purpose is to request resources for uplink data transmission from the eNode-B scheduler.*

- ***Non-synchronized random access procedure:***
  - Prior to initiation of the non-synchronized random access procedure, each UE obtains the following information broadcast from eNode-B:
    - Random access channel parameters
    - Including POACH configuration
    - Frequency position and preamble forma
    - Parameters for determining the root sequences and their cyclic shifts in the preamble sequence set for the cell.
  - The non-synchronized random access procedure consists of following steps is depicted in the figure 9.5 and described as follows

The image part with relationship ID rId37 was not found in the file.

1. Multiple UEs transmit randomly selected random access code.
2. eNode - B conducts a multiuser detection process and allocates resources to the detected UEs.
3. Each UE transmits detailed information losing allocated resources.
4. The eNode-B transmits the contention-resolution message on the DL-SCH. When the previous steps are finished successfully. eNode-B and each UE initiate data communication.

Step 1: Random Access Preamble Transmission:

- The UE randomly selects a random access preamble transmitted by eNodeB, and transmits on the PRACH physical channel.
- Open-loop power control is used to determine the initial transmit power level.
- Multiple UEs may transmit their random access preambles simultaneously through the same channel, and the eNode-B monitors the random access channel and conducts multiuser detection identifying each RACH transmission.
- The RACH signals from the different UEs are based on the Zadoff-Chu sequence with different cyclic shift resulting in a zero cross-correlation between them.
- The eNode-B also calculates the timing correction for the uplink transmission for each UE.

Step 2: Random Access Response:

- eNode-B transmits the corresponding random access response on the DL-SCH, which contains the identity of the detected preamble, the timing correction for uplink transmission, a temporary identity for transmission in following steps, and an initial uplink resource grant.
- The random access response message can also include a backoff indicator to instruct the UE to back off for a period of time before retrying another random access attempt.
- The uplink scheduling grant for the following uplink transmission contains 20 bits, and the content is illustrated in Table 9.20.

☒ The image part with relationship ID rId37 was not found in the file.

- Once the random access preamble is transmitted, it will monitor the PDCCH for random access response identified by the Random Access Radio Network Temporary Identifier (RA- RNTI), as the time-frequency slot carrying the preamble is associated with an RA-RNTI.
- If the received random access response matches the transmitted preaanble, the UE may stop monitoring.

Step 3: Scheduled Transmission:

- After step 2, the UE is uplink synchronized and can transmit additional messages on scheduled UL-SCH.
- This step is to assist contention resolution.
- If the UEs that perform random access attempts in the same time-frequency resource use different preambles.
- Different UEs can be identified by the eNode-B and there is no collision. However. it is possible that multiple UEs select the same preamble, which causes a collision.
- To resolve the contention for access, the UE that detects a random access preamble transmits a message containing a terminal identity.
- If the UE is connected to a cell, Cell Radio Network Temporary Identifier (C-RNTI) will be tool, which is a unique UE ID at the cell level; otherwise, a core network identifier is used. In step 1 the H-ARQ protocol is supported to improve the transmission reliability.

Step 4: Contention Resolution:

- In this step, the eNode-B transmits the contention-resolution message on the DL-SCH, which contains the identity of the winning UE.
- The UE that observes a match between this identity and the identity transmitted in step 3 declares a success and completes its random access procedure.
- If this UE has not been assigned a C-RNTI, the temporary identity is then set as its C-RNTI.
- The H-ARQ protocol is supported in this step and the UE with successful access will transmit an H-ARQ acknowledgment.

9.10 Power Control in Uplink: *Brief out power control in uplink*

- As compared to Downlink, in case of Uplink in LTE, Power control is used. As the battery of the phone (UE) is power limited compared to base station power in the DL.
- Uplink power control is used mainly for the following two reasons.
  1. *Limit intracell and intercell interference*
  2. *Reduce UE power consumption*
- In LTE, the power control in the uplink is to control the interference caused by UEs to neighboring cells while maintaining the required SINR at the serving cell.
- The power control scheme for the PUSCH transmission in the uplink. Usually in Uplink. Power control is done in two ways. One is

1. ***Conventional Power Control (CPC)****:* Conventional power control in the uplink is to achieve the same SINR for different UEs at the base station, also known as *full compensation*. But it suffers low spectral efficiency as the common SINR is limited by the cell-edge UEs.

- ***Fractional Power control (FPC)***: It is an open-loop power control scheme, which allows for full or partial compensation of path loss and shadowing. FPC allows the UEs with higher path loss, i.e., cell-edge UEs, to operate with lower SINR requirements so that they generate less interference to other cells, while having a minor impact on the cell-interior UEs so that they are able to transmit at higher data rates. Besides open-loop power control, there is also a closed-loop power control component, which is to further adjust the UE transmission power to optimize the system performance.
- FPC scheme, based on which the UE adjusts the transmission power according to:

☒ The image part with relationship ID rId37 was not found in the file.

- Considering both open-loop and closed-loop components, the UE sets its total transmission power using the following formula:

☒ The image part with relationship ID rId37 was not found in the file.

The image part with relationship ID rId37 was not found in the file.

Module – 5

10. Radio Resource Management and Mobility Management:

- PDCP overview
- MAC/RLC overview
- RRC overview
- Mobility Management
- Inter-cell Interference Coordination

Module 5

Chapter 10. Radio Resource Management (RRM) and Mobility Management (MM)

10.1 Data flow services in LTE:

What is bearer service? Briefly explain relevance of bearer service

- LTE is an end-to-end packet-switched network, designed for high speed data services.
- LTE uses "bearer" concept to provide varying QoS requirements as per the applications.
- EPS bearer is defined between the PDN-GW and UE.
- Each EPS bearer maps to a specific set of QoS parameters like Data rate, latency, packet error rate etc.
- Applications with very different QoS requirements such as e-mail and voice can be put on separate bearers that will allow the system to simultaneously meet their QoS requirements.
- The end-to-end connectivity through the network is made via the bearer service, and the bearer service architecture is shown in Figure 10.1.

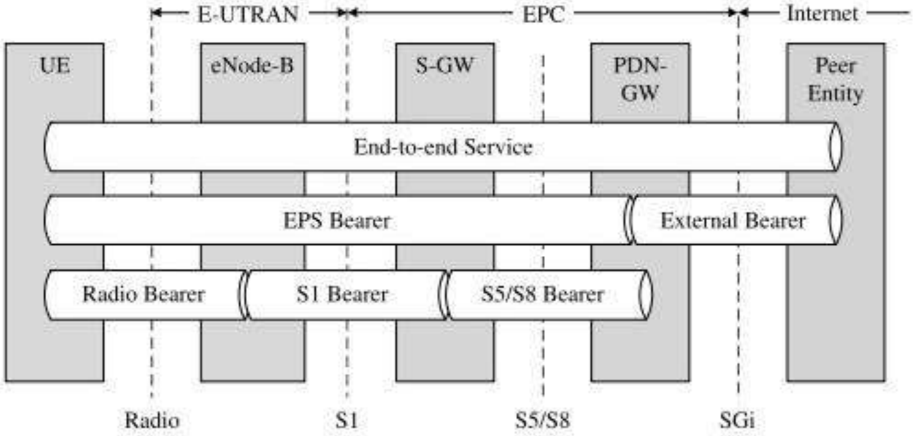
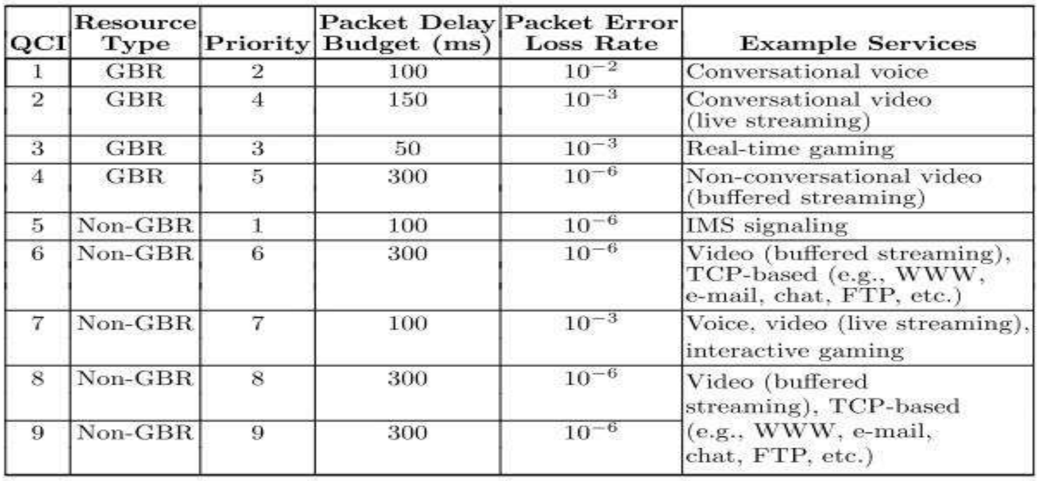


Figure 10.1 EPS bearer service architecture.

- An EPS bearer has to cross multiple interfaces, and across each interface it is mapped to a transport layer bearer.
- An S5/S8 bearer transports the packets of an EPS bearer between a Serving GW (S-GW) and a PDN-GW.
- S1 bearer transports the packets of an EPS bearer between an eNode-B and an S-GW.
- Over the radio interface the bearer is referred to as the radio bearer, which transfers data between a UE and the E-UTRAN.
- Signaling Radio Bearers (SRBs) carry the Radio Resource Control (RRC) signaling messages.
- Data Radio Bearers (DRBs) carry the user plane data.
- Radio bearers are mapped to logical channels through Layer 2 protocols.

  [Which are two classes of bearer service]
- Broadly, the bearers can divided into two classes:
  1. **Guaranteed Bit Rate (GBR) bearers**: *These bearers define and guarantee a minimum bit rate that will be available to the UE. Bit rates higher than the minimum bit rate can be allowed if resources are available. GBR bearers are typically used for applications such as yoke, streaming video, and real-time gaming.*
  2. **Non-GBR bearers:** *These bearers do not define or guarantee a minimum bit rate to the UE. The achieved bit rate depends on the system load, the number of UEs served by the eNode-B, and the scheduling algorithm. Non-GBR bearers are used for applications such as web browsing, e-mail, FTP, and P2P file sharing.*
- ***EPS Bearer Service Architecture***: ***Explain in brief EPS bearer service architecture?***
  - Each bearer is associated with a QoS Class
  - Identifier (QCI), which indicates the priority, packet delay budget, acceptable packet error loss rate and the GBR/non-GBR clas- sification.
  - The nine standardized QCI defined in the LTE are shown in Table 10.1.

Table 10.1 Standardized QoS Class Identifiers (QCIs) for LTE



QCI	Resource Type	Priority	Packet Delay Budget (ms)	Packet Error Loss Rate	Example Services
1	GBR	2	100	10^{-2}	Conversational voice
2	GBR	4	150	10^{-3}	Conversational video (live streaming)
3	GBR	3	50	10^{-3}	Real-time gaming
4	GBR	5	300	10^{-6}	Non-conversational video (buffered streaming)
5	Non-GBR	1	100	10^{-6}	IMS signaling
6	Non-GBR	6	300	10^{-6}	Video (buffered streaming), TCP-based (e.g., WWW, e-mail, chat, FTP, etc.)
7	Non-GBR	7	100	10^{-3}	Voice, video (live streaming), interactive gaming
8	Non-GBR	8	300	10^{-6}	Video (buffered streaming), TCP-based (e.g., WWW, e-mail, chat, FTP, etc.)
9	Non-GBR	9	300	10^{-6}	

- One EPS bearer is established when the UE connects to a PDN.
- **Default Bearer**: *EPS bearer is established throughout the lifetime of the PDN connection to provide the UE with always-on IP connectivity to that PDN.*
- **Dedicated bearer**: *Any additional EPS bearer established to the same PDN.*

10.1.1 LTE Protocol Architecture:

Explain the divisions between the UE and the core network in the protocol architecture of LTE?

With diagram, brief the user and control plane protocol stack of LTE

- The protocol architecture in LTE between the UE and the core network is divided:
  1. *User plane protocol stack*
  2. *Control plane protocol stack*

1. ***User plane protocol stack:*** *U*ser plane protocol stack as shown in Figure 10.2

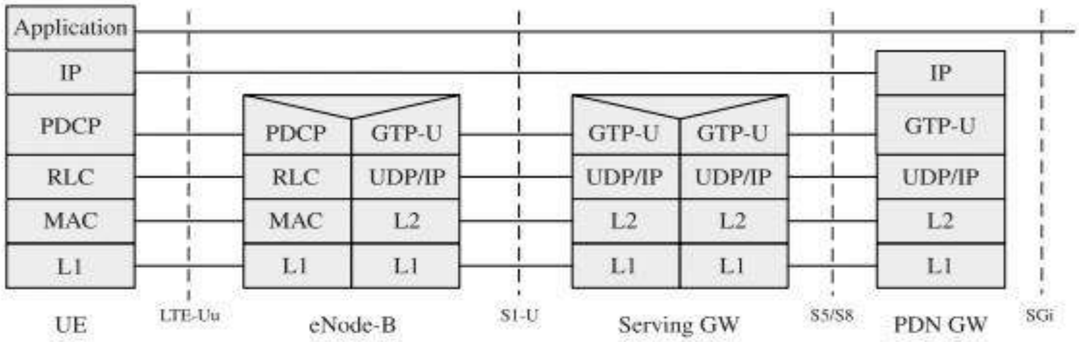


Figure 10.2 User plane protocol stack.

- The user plane is responsible for transporting IP packets carrying application-specific data from the PDN-GW to the UE.
- This is done by encapsulating the IP packets in an Evolved Packet Core (EPC)-specific protocol and tunneling them from the PDN-GW to the eNode-B using the GPRS Tunneling Protocol (GTP).
- From the eNode-B the packets are transported to the UE using the Packet Data Convergence Protocol (PDCP).

2. ***Control plane protocol stack:*** *Control* plane protocol stack as shown in Figure 10.3

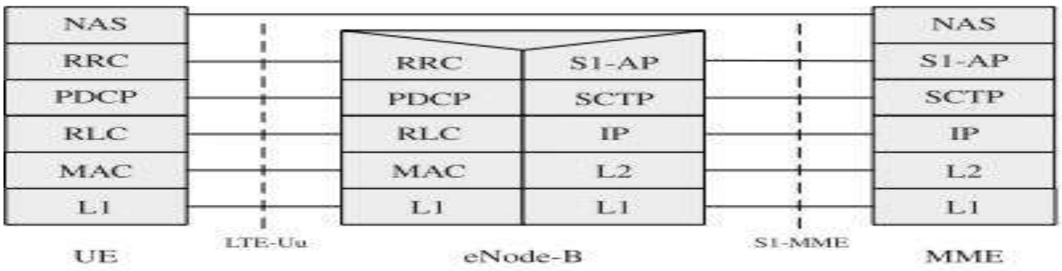


Figure 10.3 Control plane protocol stack.

- The control plane is used for transporting signaling between the Mobility Management Entity (MME) and the UE.
- The type of signaling handled over the control plane is typically related to bearer management, QoS management, and mobility management including functions such as handover and paging.
- In LTE, Layer 2 of the protocol stack is split into the following sublayers:
  - *a. Medium Access Control (MAC)*
  - *b. Radio Link Control (RLC) and*
  - *c. PDCP.*

10.1.2 The layer 2 structure for the downlink:

Explain Layer 2 structure for downlink with diagram?

It is depicted in Figure 10.4.

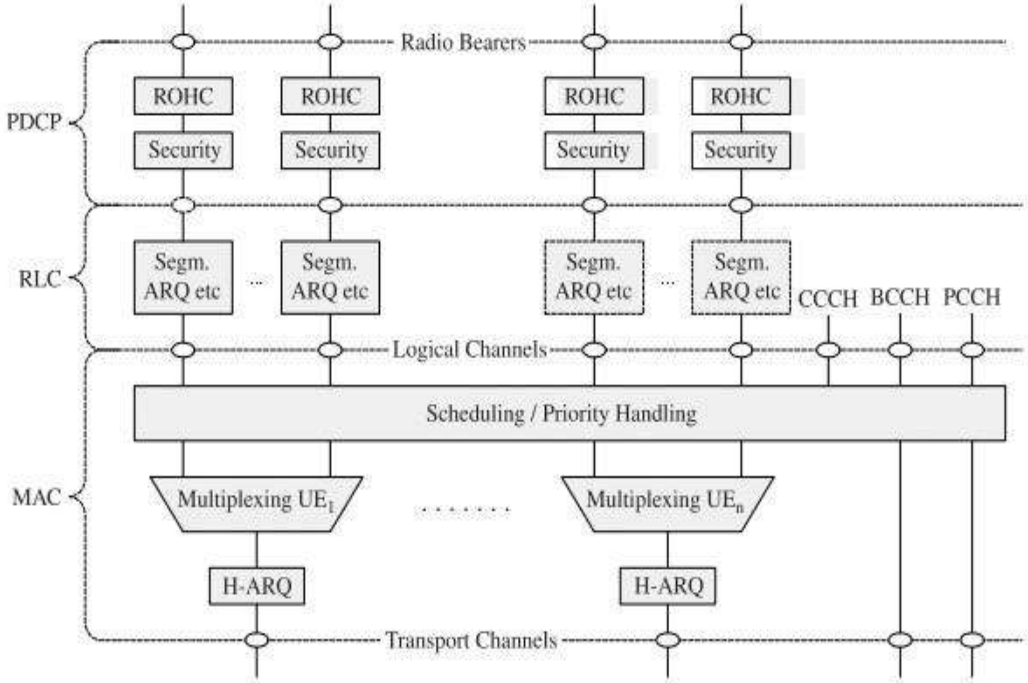


Figure 10.4 Layer 2 structure for downlink.

- audio bearers are mapped to logical channels through PDCP and RLC sublayers.
- The Service Access Point (SAP) between the physical layer and the MAC sublayer provides the transport channels that are used by the physical layer to provide services to the MAC.
- The SAP between the MAC sublayer and the RLC sublayer provides the logical channels that are used by the MAC layer to provide services to the RLC.
- One RLC and PDCP entity per radio bearer in the UE and eNode-B.
- MAC layer multiplexes the data and control information from all the radio bearers at the UE and eNode-B. i.e., MAC layer multiplexes several logical channels on the same transport channel.

10.2 Packet Data Convergence Protocol (PDCP) Overview:

Explain PDCP functions for the user plane and the control plane with a block diagram?

- A PDCP entity is associated either with the control plane or user plane depends on which radio bearer it is carrying data for.
- Each radio bearer is associated with one PDCP entity.
- Each PDCP entity is associated with one or two RLC entities depending on the radio bearer characteristic (uni-directional or bi-directional) and the RLC mode.
- PDCP is used only for radio bearers mapped on DCCH and DTCH types of logical channels.
- The main services and functions of the PDCP sublayer for the user plane and control plane as shown in Figure 10.5 are as follows.

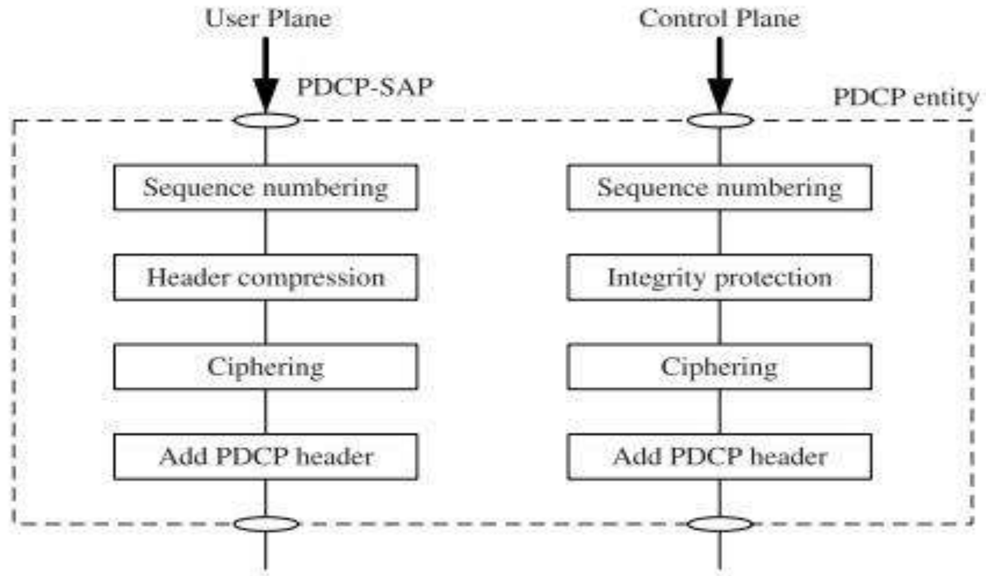


Figure 10.5 PDCP functions for the user plane and the control plane.

- ***Functions of PDCP sublayer for the user plane:***
  1. Header compression and decompression of IP data flows with the RObust Header Compression (ROHC) protocol.
  2. Ciphering and deciphering of user plane data.
  3. In-sequence delivery and reordering of upper-layer PDUs at handover
  4. Buffering and forwarding of upper-layer PDUs from the serving eNode-B to the target eNode-B during handover
  5. Timer-based discarding of SDUs in the uplink
- ***Functions of PDCP sublayer for the control plane:***
  1. Ciphering and deciphering of control plane data.
  2. Integrity protection and integrity verification of control plane data
  3. Transfer of control plane data

Explain the categories of PDCP PDU? Also explain PDCP data PDU formats for the user plane and the control plane?

- The PDCP PDUs can be divided into two categories:
  1. ***The PDCP data PDU:*** *It is used in both the control and user plane to transport higher layer packets. It is used to convey either user plane data containing a compressed /uncompressed IP packet or control plane data containing one RRC message and a Message Authentication Code for Integrity (MAC-I) field for integrity protection.*
  2. ***The PDCP control PDU:*** *It is used only within the user plane to convey a PDCP status report during handover and feedback information for header compression. It carries peer- to-peer signaling b/w the PDCP entities at two ends. It doesn't carry higher layer SDU.*
- The constructions of the PDCP data PDU formats from the PDCP SDU for the user plane and the control plane are shown in Figure 10.6.

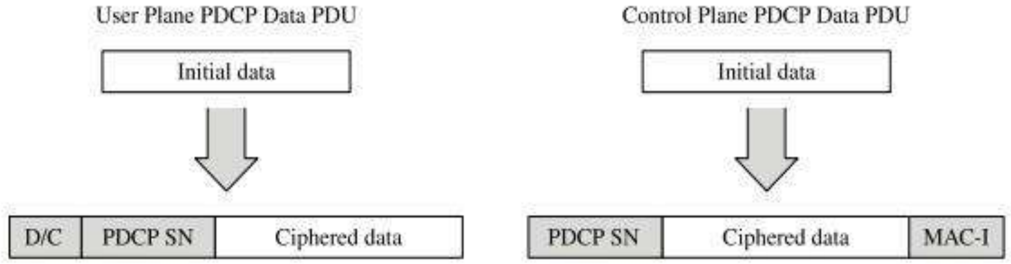
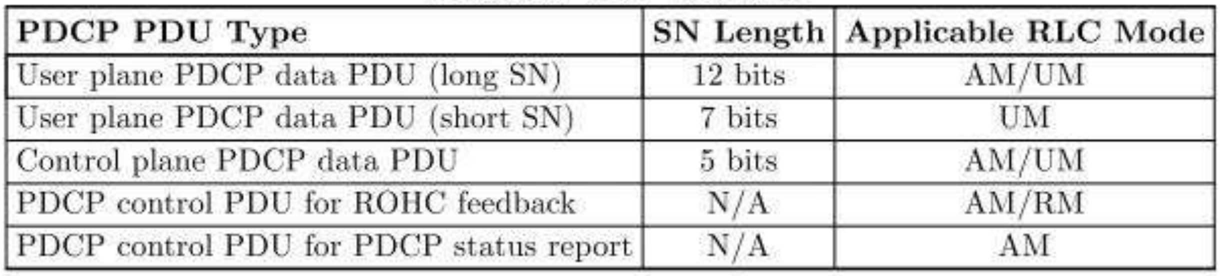


Figure 10.6 PDCP data PDU formats for the user plane and the control plane.

- The various types of PDCP PDU carried on the user and control plane are shown in Table 10.2. There are three different types of PDCP data PDUs, distinguished by the length of the Sequence Number (SN).

Table 10.2 PDCP Data Units



PDCP PDU Type	SN Length	Applicable RLC Mode
User plane PDCP data PDU (long SN)	12 bits	AM/UM
User plane PDCP data PDU (short SN)	7 bits	UM
Control plane PDCP data PDU	5 bits	AM/UM
PDCP control PDU for ROHC feedback	N/A	AM/RM
PDCP control PDU for PDCP status report	N/A	AM

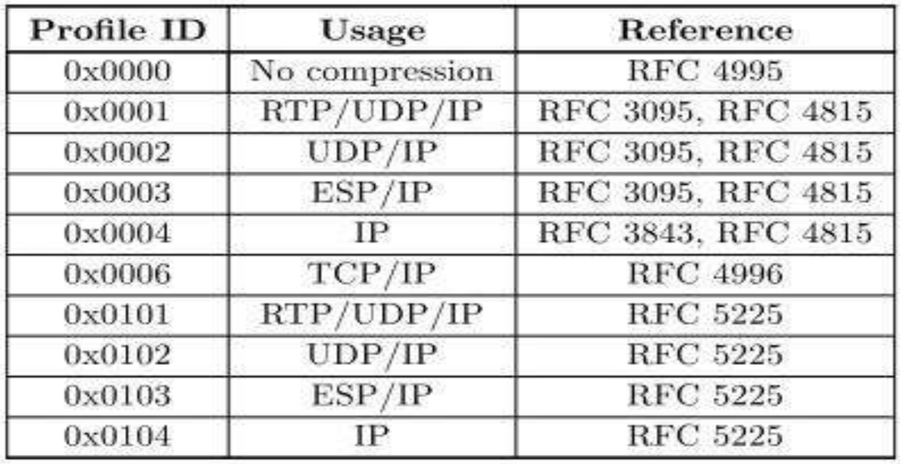
- The PDCP SN is used to provide robustness against packet loss and to guarantee sequential delivery at the receiver.
- The PDCP data PDU with the long SN is used for the Un-acknowledge Mode (UM) and Acknowledged Mode (AM) and the PDCP data PDU with the short SN is used for the Transparent Mode (TM).
- Besides the SN field and the ciphered data, the PDCP data PDU for the user plane contains a `D/C' field that is to distinguish data and control PDUs.
- This is required since the PDCP data PDU can carry both user plane and control plane data.
- PDCP performs ***Header compression, Integrity and Ciphering***

10.2.1 Header Compression:

Write a short note on Header Compression?

- ROHC framework defined by IETF provides Header Compression in LTE.
- The protocols at different layers network (IP), transport (TCP, UDP), Application (RTP) bring significant amount of header overhead.
- Efficient header compression scheme is needed for VoIP services.
- The ROHC framework contains "profiles" (Header Compression Algorithms).
- Each profile specifies particular network layer, transport layer or upper layer.
- The supported profiles in 3GPP Release 8 are listed in Table 10.3.

Table 10.3 Supported Header Compression Protocols and Profiles



Profile ID	Usage	Reference
0x0000	No compression	RFC 4995
0x0001	RTP/UDP/IP	RFC 3095, RFC 4815
0x0002	UDP/IP	RFC 3095, RFC 4815
0x0003	ESP/IP	RFC 3095, RFC 4815
0x0004	IP	RFC 3843, RFC 4815
0x0006	TCP/IP	RFC 4996
0x0101	RTP/UDP/IP	RFC 5225
0x0102	UDP/IP	RFC 5225
0x0103	ESP/IP	RFC 5225
0x0104	IP	RFC 5225

10.2.2 Integrity and Ciphering: *Write a short note on integrity and ciphering?*

- The security-related functions in PDCP include integrity protection and ciphering.
- A PDCP PDU counter, denoted by the parameter COUNT, is maintained and used as an input to the security algorithm.
- The format of COUNT is shown in Figure 10.7, which has a length of 32 bits and consists of two parts: the Hyper Frame Number (HFN) and the PDCP SN.

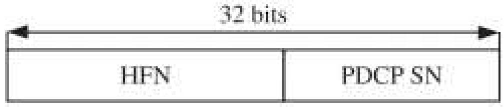


Figure 10.7 Format of COUNT.

- The SN is used for reordering and duplicate detection of RLC packets at the receive end.
- If the key does not match the MAC-I field, then the PDCP PDU does not pass the integrity check, and the PDCP PDU will be discarded.
- The ciphering function includes both ciphering and deciphering.
- It is performed on both control plane data and user plane data.

- For the control plane, the data unit that is ciphered is the data part of the PDCP PDU and the MAC-I; for the user plane, the data unit that is ciphered is the data part of the PDCP PDU.
- Neither integrity nor ciphering is applicable to PDCP control PDUs.
- The ciphering function is activated by upper layers, which also configures the ciphering algorithm and the ciphering key to be used.
- The ciphering is done by an XOR operation of the data unit with the ciphering stream.
- The ciphering stream is generated by the ciphering algorithm based on ciphering keys, the radio bearer identity, the value of COUNT, the direction of the transmission, and the length of the key stream.

10.3 MAC / RLC Overview: *Describe MAC/RLC Overview?*

- ***Functions of MAC Layer:***
  - *It performs multiplexing and demultiplexing of logical channels on to the transport channel.*
  - *At eNode-B, it performs multiplexing and prioritizing various UEs serving by the eNode-B.*
  - *At UE, it performs multiplexing and prioritizing various radio bearers associated with the UE.*
  - *It provides services to the RLC layer through logical channels.*
  - *It takes services from PHY layer through transport channels.*
- ***Functions of RLC Layer:***
  - *It performs segmentation and concatenation on PDCP PDUs based on size mentioned by the MAC layer.*
  - *Reorders RLC PDUs if they receive out of order due to H-ARQ process in the MAC layer.*
  - *RLC supports ARQ mechanism.*

What are the three different modes in which RLC entity can be operated? Explain in brief

10.3.1 Data Transfer Modes of RLC: RLC entity can be operated in three different modes:

1. *Transparent Mode (TM)*
2. *Unacknowledged Mode (UM)*
3. *Acknowledged Mode (AM)*

1. The Transparent Mode (TM): Following are feature of TM mode

- *The TM mode is the simplest mode*
- *The TM mode is not used for user plane data transmission*
- *RLC entity doesn't add any RLC header to the PDU.*
- *No data segmentation or concatenation.*
- *No retransmissions.*
- *Order of delivery is not guaranteed. Ex., RRC broadcast messages, paging messages uses TM.*

2. The Unacknowledged Mode (UM):

- *Order of delivery is guaranteed.*
- *DTCH logical channels operate in this mode.*
- *UM RLC entity performs data segmentation or concatenation RLC SDUs.*
- *No retransmissions of the lost PDU.*
- *Ex., delay-sensitive, error-tolerant real-time applications like VoIP*
- *Relevant RLC headers are included in the UM Data PDU.*
- *At the Rx, UM RLC entity performs duplicate detection and reordering*

3. The Acknowledged Mode (AM):

- *The AM mode is the most complex one, which requests retransmission of missing PDUs in addition to the UM mode functionalities.*
- *It is mainly used by error-sensitive and delay-tolerant applications.*
- *An AM RLC entity can be configured to deliver/receive RLC PDUs through DCCH and DTCH.*
- *An AM RLC entity delivers/receives the AM Data (AMD) PDU and the STATUS PDU indicating the ACK/NAK information of the RLC PDUs.*
- *When the AM RLC entity needs to retransmit a portion of an AMD PDU, which results from the ARQ process and segmentation, the transmitted PDU is called the AMD PDU segment.*
- *The operation of the AM RLC entity is similar to that of the UM RLC entity, except that it supports retransmission of RLC data PDUs.*
- *The receiving AM RLC entity can send a STATUS PDU to inform the transmitting RLC entity about the AMD PDUs that are received successfully and that are detected to be lost.*

10.3.2 Purpose of MAC and RLC Layers

What are the main services and functions of the RLC and MAC sublayer?

- The main services and functions of the RLC sublayer include
  - *Transfer/receive PDUs from upper layers.*
  - *Error detection using ARQ (only in AM mode)*
  - *Concatenation, segmentation and reassembly of RLC SDUs (only in UM and AM data transfer)*
  - *In-sequence delivery of upper-layer PDUs (only in UM and AM data transfer)*
  - *Duplication detection (only in UM and AM data transfer)*
  - *RLC SDU discard (only in UM and AM data transfer)*
  - *RLC re-establishment.*
- Services and functions of MAC sublayer:
  - *Mapping between logical channels and transport channels*
  - *Multiplexing/ Demultiplexing of MAC SDUs belonging to one or more logical channels from the same transport block*
  - *Scheduling for uplink and downlink transmission*

- *Error correction through H-ARQ*
- *Priority handling between logical channels of one UE or between UEs by means of dynamic scheduling.*
- *Transport format selection, i e, selection of MCS for link adaption.*
- *Padding if the MAC PDU is not fully filled with data.*

10.3.3 PDU Headers and Formats RLC PDU formats

- RLC PDUs can be categorized into RLC data PDUs and RLC control PDUs.
- The formats of different RLC Data PIDUs are shown in Figure 10.8

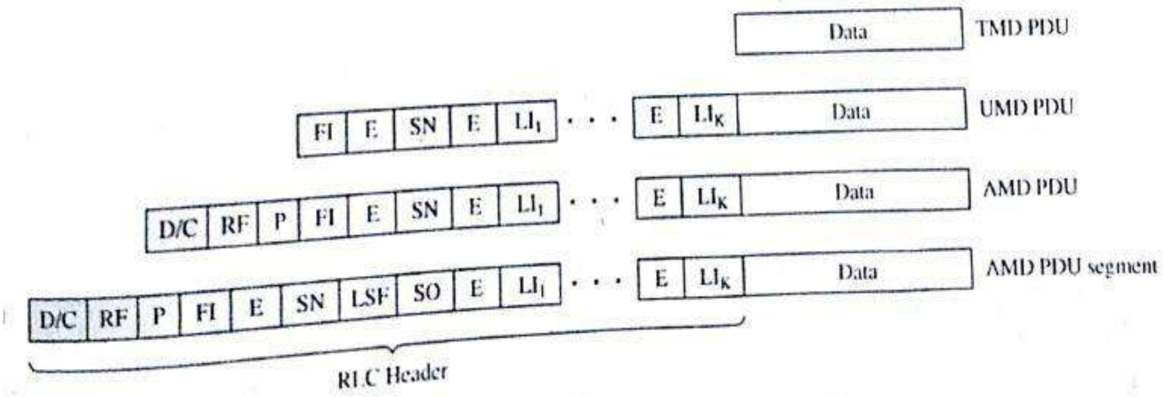


Figure 10.8 Formats of RLC Data PDUs.

- ***Framing Info (FI) field***: The FI field indicates whether a RLC SDU is segmented at the beginning and/or at the end of the Data field.
- ***Length Indicator (LI) field***: The LI field indicates the length in bytes of the corresponding Data field element present in the UMD or _AMD-PDU
- ***Extension bit (E) field:*** The E field indicates whether a Data field follows or a set of E field and LI field follows.
- ***SN: field***: The SN field indicates the sequence number of the corresponding UMD or AMD PDU. It consists of 10 bits for AMD PDU, AMD PDU segments and STATUS PDUs and 5 bits or 10 bits for UMD PDU. The PDU sequence number

- For AMD PDU and AMD PDU segments, additional fields are available:

  - **Data/Control (D/C) field:** The D/C field indicates whether the RLC PDU is an RLC Data PDU or an RLC Control PDU.
  - **Re-segmentation Flag (RF) field:** The RF field indicates whether the RLC PDU is an AMD PDU or an AMD PDU segment.
  - **Polling bit (P) field:** The P field indicates whether the transmitting side of an AM RLC entity requests a STATUS report from its peer AM RLC entity.

- Additionally, the RLC header of an AMD PDU segment contains special fields including:

- ***Segment Offset (SO) field***: The SO field indicates the position of the AMD PDU segment in bytes within the original AMD PDU.
- ***Last Segment Flag (LSF) field***: The LSF field indicates whether the last byte of the AMD PDU segment corresponds to the last byte of an AMD PDU.

A). The format of the STATUS PDU

Explain in brief the format of STATUS PDU?

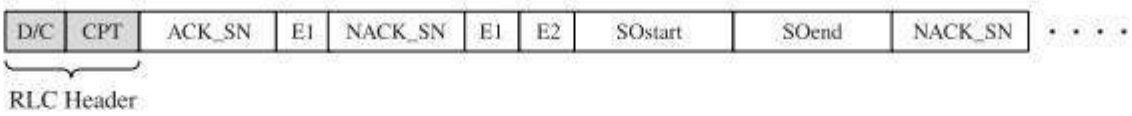


Figure 10.9 The format of STATUS PDU.

- ***Control PDU Type (CPT) field:*** The CPT field indicates the type of the RLC control PDU, and in Release 8 the STATUS PDU is the only defined control PDU.
- ***Acknowledgment SN (ACK_SN) field***: The ACKSN field indicates the SN of the next not received RLC Data PDU, which is not reported as missing in the STATUS PDU.
- ***Extension bit 1 (E1) field***: The El field indicates whether a set of NACK_SN, El, and E2 follows.
- ***Extension bit 2 (E2) field:*** The E2 field indicates whether a set of SOstart and SOend follows.
- ***Negative Acknowledgment SN (NACK_SN) field:*** The NACK_SN field indicates the SN of the AMD PDU (or portions of it) that has been detected as lost at the receiving side of the AM RLC entity.
- ***SO start (SOstart) field and SO end (SOend) field:*** These two fields together indicate the portion of the AMD PDU with SN = NACK_SN that has been detected as lost at the receiving side of the AM RLC entity

B). MAC PDU Formats:

Explain in brief the format of a typical MAC subheader?

MAC layer receives data from RLC as MAC SDUs and passes the MAC PDUs to PHY layer.

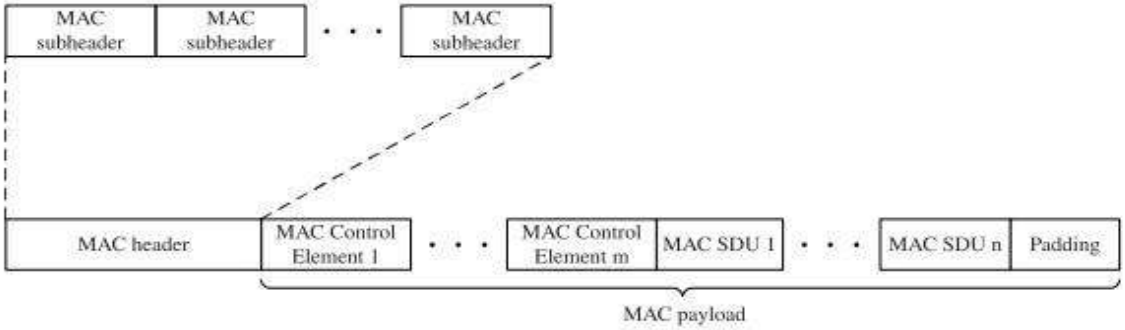


Figure 10.10 An example of MAC PDU consisting of MAC header, MAC control elements, MAC SDUs, and padding.

- The MAC PDU contains TWO fields: ***MAC PDU header & MAC payload***
  1. ***MAC Payload***: It contains zero or more MAC SDUs, zero or more MAC control elements, and optional padding.
  2. ***MAC PDU header***: It consists of one or more MAC PDU subheaders.
- The format of a typical MAC subheader is shown in Figure 10.11, which contains five different fields as explained in the following:

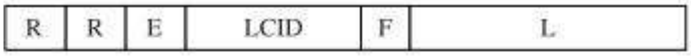


Figure 10.11 An example of the MAC subheader.

1. **R**: It is reserved field and set to '0' always.
2. **E**: It is an extension field to indicate the presence of more fields in the MAC header.
   - If E=1, set of R/R/E/LCID fields follows
   - If E=0, either MAC SDU, a MAC control element, or Padding follows.
3. **LCID**: Logical Channel ID: It indicates the logical channel instance of the corresponding MAC SDU or the type of the corresponding MAC control element, or padding.
4. **F**: It indicates the size of the Length field.
   - F=0, size of MAC SDU or MAC control element < 128 bytes.
5. **L**: It indicates the length of the corresponding MAC PDU or MAC control element in Bytes.

C). **The MAC PDU for random access response** It has a different format, as shown in Figure 10.12.

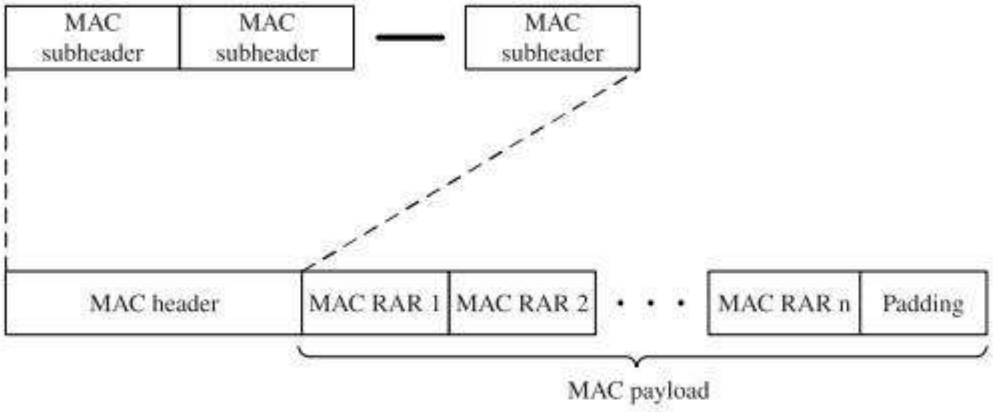


Figure 10.12 The MAC PDU for random access response.

- MAC header consists of one or more MAC PDU subheaders. Subheader contains payload information.
- Payload contains one or more MAC Random Access Responses (MAC RAR) and optional padding.

10.3.4 ARQ Procedures:

Explain in brief ARQ Procedures?

LTE applies two-layer retransmission scheme.

1. ***H-ARQ Protocol***: *It is a low latency and low overhead feedback protocol in MAC layer. It is responsible for handling transmission errors by retransmissions based on H-ARQ processes. H-ARQ is the use of conventional ARQ along with an Error Correction technique called "Soft Combining". 'Soft Combining' data packets that are not properly decoded are not discarded instead stored in a "buffer" and will be combined with next retransmission. Two or more packets received, each one with insufficient SNR to allow individual decoding can be combined in such a way that the total signal can be decoded.*

1. ***Selective Repeat ARQ protocol***: *It is a RLC layer protocol to correct residual H-ARQ errors due to the error in H-ARQ ACK feedback. The latency with RLC ARQ is high and it is used in only AM transfer mode. The ARQ NAK is received by STATUS PDU or H-ARQ delivery failure notification.*

10.4 RRC Overview: *RRC is responsible for:*

- *RRC connection Management*
- *Radio bearer control*
- *Mobility functions*
- *UE measurement reporting and control*
- *Broadcasting system information and paging.*

10.4.1 RRC States: *Explain Two RRC states in LTE?*

Two states of RRC as shown in figure

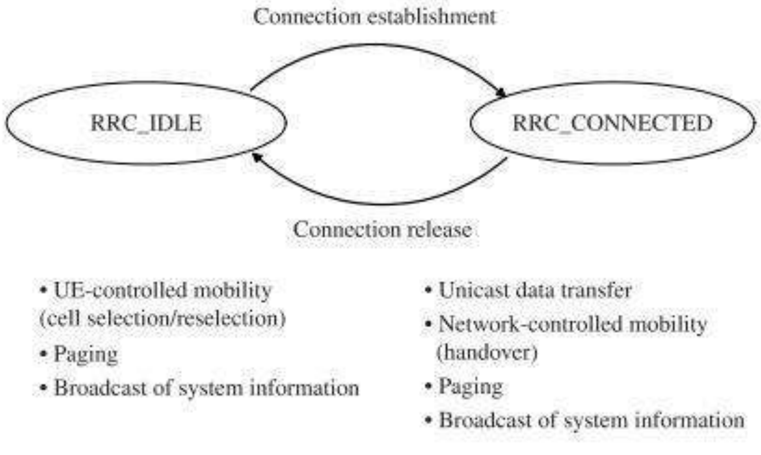


Figure 10.13 RRC states in LTE.

- ***RRC-IDLE STATE:***

1. *UE receives broadcasts of system information*
2. *Paging information.*
3. *Mobility control is handled by UE (performs neighboring cell measurements and cell selection/reselection.*

4. *UE monitors paging channel for incoming calls.*
5. *UE specifies the paging DRX cycle.*

- ***RRC states: RRC-CONNECTED:***

1. *UE transmits/receive data from the network (eNode-B).*
2. *UE monitors control channels (PDCCH) associated with the shared data channel for data.*
3. *UE reports channel quality information and feedback information to the network (eNode-B) to assist the data transmission.*
4. *UEs provide neighboring cell measurement information.*
5. *The n/w controls mobility/handover of the UE.*

10.4.2 RRC Functions:

What are the main functions of the RRC protocol?

- Signaling Radio Bearers SRBs are the radio bearers and used only for the transmission of RRC and NAS messages.
  - ***SRB 0****: It is FOR RRC messages using the CCCH logical channel.*
  - ***SRB 1****: It is for RRC messages and NAS messages.*
  - ***SIB 2****: It is for NAS messages using DCCH logical channel.*
- All SIBs other than Type 1 carries system information messages.
- Following are the main functions of the RRC protocol.

1. ***Broadcast of system information*****:**
   - It is divided in to Master Information Block (MIB) and a number of System Information Blocks (SIBs)
   - **MIB:** contains most essential and most frequently transmitted parameters which are needed to acquire other information from the cell.
   - **SIB Type 1**: contains parameters to determine the cell selection, information of time- domain scheduling of other SIBs.
   - **All SIBs** other than Type 1 carries system information messages.

2. ***RRC connection control*****:** Itincludes
   - RRC connection establishment, modification and release.
   - Paging and Initial security activation
   - Establishment of SRBs
   - Radio bearers carrying user data
   - Radio configuration control
   - QoS control
   - Recovery from radio link failure.

3. ***Measurement configuration and reporting***: It includes
    - Measurement establishment, modification and release.
    - Configuration and deactivation of measurement gaps
    - Measurement reporting for intra-freq. and inter-freq.
    - Inter RAT mobility
4. ***Other functions***: It include
    - Transfer of dedicated NAS information
    - Non-3GPP dedicated information
    - Transfer of UE radio access capability information.
    - Support for self-configuration and self-optimization

10.5 Mobility Management***

Write a short note on Mobility Management?

- LTE mobility management functions are classified into Two groups:
    1. ***Intra-LTE mobility***: Mobility is within the LTE system.
        - *It can happen either over S1 interface or X2 interface.*
        - *Mobility through X2 interface occurs when UE moves from one eNode-B to another eNode-B within the same RAN attached to same MME.*
    2. ***Mobility to other systems***: 3GPP systems (ex., UMTS), non 3GPP systems (inter-RAT mobility).
        - *Mobility through S1 interface occurs when the UE moves from one eNode-B to another that belongs to a different RAN attached to different MMEs.*
        - *In this case the PDCP context is not continued and the UE needs to re-establish its session once it moves to the target non-LTE system.*

S1 (S- One) Mobility :

Which are the three different phases involved in Mobility Management? Explain each in brief?

It contains 3 phases (as shown in figure 10.14)***

1. ***Preparation Phase***
2. ***Execution Phase***
3. ***Completion Phase***

1. **Preparation Phase:**
    - *Once a handover decision is made, and identifying target MME, eNode-B, the n/w needs to allocate resources on the target side for handover to happen.*
    - *MME sends a handover request to the target eNode-B and request for resource allocation to the UE.*

- *After resource allocation at the target eNode-B, it sends a handover request ACK to the MME.*
- *After receiving the handover request ACK by the MME, it sends a handover command to the UE via the source eNode-B*

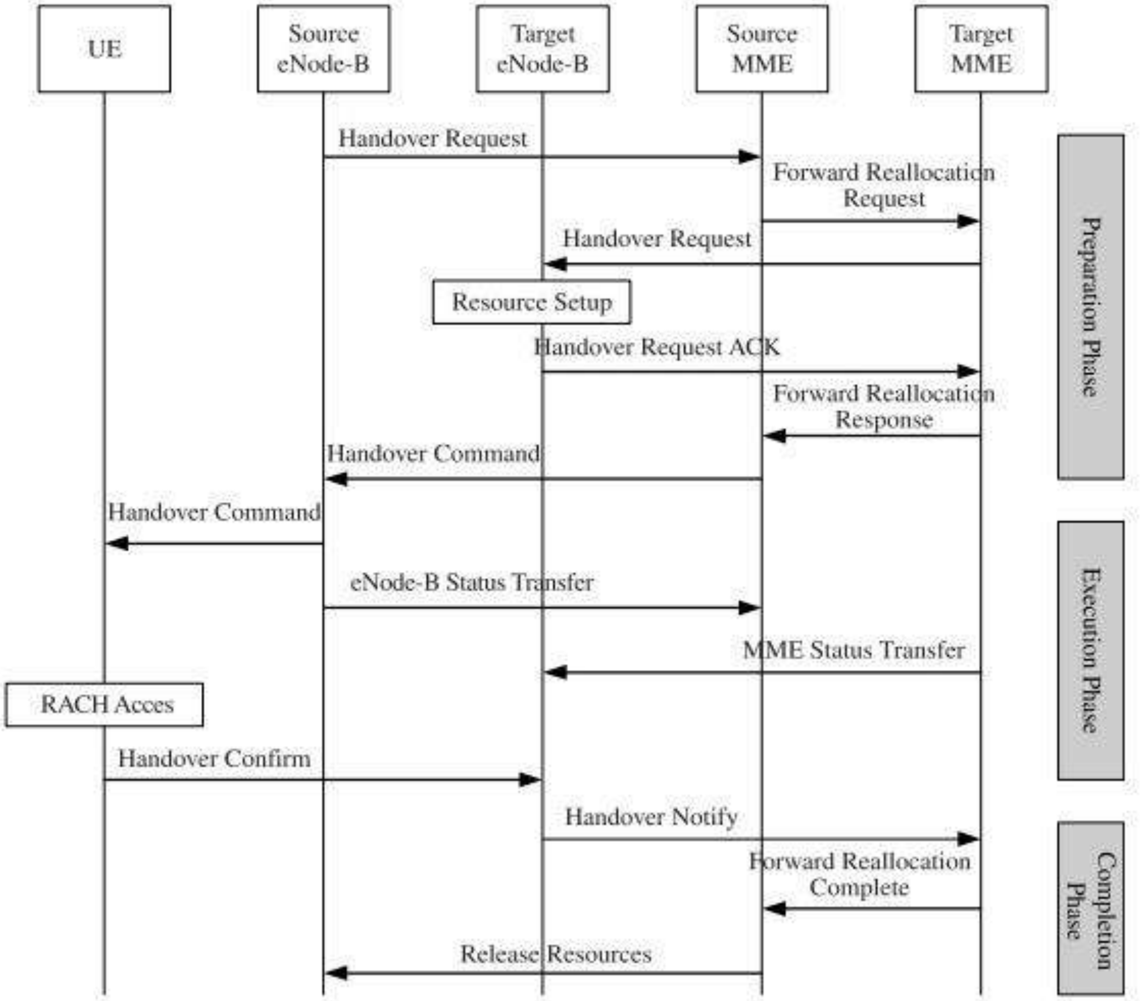


Figure 10.14 Mobility management over the S1 interface.

2. **Execution Phase:**
   - *Once UE receives the handover command, it access the target eNode-B using RACH.*
   - *Source eNode-B initiates the status transfer where the PDCP context of the UE is transferred to the target eNode-B.*
   - *Source eNode-B forwards the data stored in the PDCP buffer to the target eNode-B.*
   - *Now, UE is able to establish a RAB on the target eNode-B, it sends the handover confirm message to the target eNode-B.*
3. **Completion Phase***:*
   - *When the target eNode-B receives the handover confirm message, it sends a handover notify message to the MME.*

- *Now, MME informs the source eNode-B to release the resources.*

10.5.2 X2 Mobility:

Which are the phases involved mobility management over X2 interface? Explain each in brief?

- It is the default mode of operation in LTE.
- If X2 interface is not available between the source and target eNode-Bs, then S1 interface is triggered.
- Mobility over the X2 interface also consists of Three steps (shown in Figure 10.15):

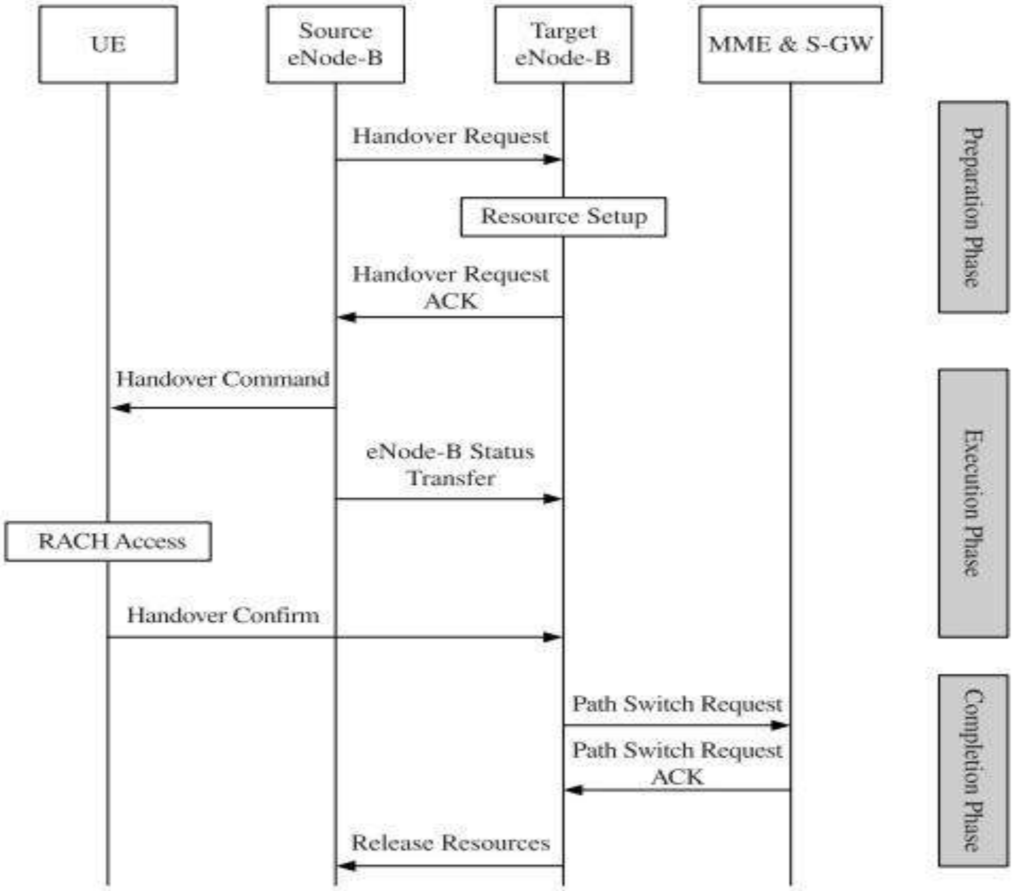


Figure 10.15 Mobility management over the X2 interface.

1. **Preparation Phase**:
   - *Once a handover decision is made by the source eNode-B, it sends a handover request message to the target eNode-B.*
   - *Now, the target eNode-B works with the MME & S-GW to set up the resources for the UE.*
   - *After handover process, the UE will have same RABs (same set of QoS) at the source and target eNode-Bs, it makes quick and seamless handover process.*
   - *Target eNode-B replies to source eNode-B with a handover request ACK.*
2. **Execution Phase:**
   - *Once handover request ACK is received by source eNode-B, it sends a handover command to the UE.*
   - *While, UE completes the various RAN-related handover procedures, the source eNode-B*

starts the status and data transfer to the target eNode-B.

3. **Completion Phase**:
   - *Once the UE completes the handover procedure, it sends a handoff completion message to the target eNode-B.*
   - *The target eNode-B sends a path switch request to the MME/S-GW and the S-GW switches the GTP tunnel from the source eNode-B to the target eNode-B.*
   - *Once the data path is switched in the user plane, the target eNode-B sends a message to the eNode-B to release the resources originally used by the UE.*

RAN Procedures for Mobility:

Explain RAN Procedures for Mobility?

- RAN related mobility management procedures decides when to handover should takes place, how a UE accesses the target eNode-B during handover.
- These procedures happen b/w UE and eNode-B or b/w UE and MME to enable UE to handover from one eNode-B to another.
- These procedures are classified into TWO cases:
  4. *Mobility in the RRC-IDLE state*
  5. *Mobility in the RRC-CONNECTED state*
- RRC states are designed such that no ping-pong between two eNode-Bs during state transition of UE, network sharing, country border, and home deployment (femtocells).

1. ***Mobility in the RRC-IDLE state***: In this state
   - UE decides when a handover is required.
   - UE chooses the cell/freq. based on the priority and radio link quality of the cell.
   - E-UTRAN allocates absolute priorities to the different frequency and conveyed by the system over BCH.
2. ***Mobility in the RRC-CONNECTED state:*** In this state
   - E-UTRAN determines the optimum cell and freq. for the target eNode-B to maintain the best radio link quality.
   - E-UTRAN initiates handover when one or more events that trigger a measurement report (A1-A5) and (B1-B2). Two handover types in this state
     - *Blind handover*: E-UTRAN initiates handover without any trigger events.
     - *Hard handover*: UE is connected to only one eNode-B at a time. Hard handover process is a “backward” handover.
     - *Backward handover*: Source eNode-B controls the handover and request the target eNode-B to prepare for the handover by allocating resources for the UE. Once the resources are allocated, the target eNode-B sends RRC message requesting the UE

to perform handover. UE uses the random access procedure in target eNode-B to establish a connection and execute handover.

- In LTE, radio link quality is the measure used in intra-freq. handover (i.e., UE selects the eNode-B with the best radio link quality).
- For a LTE cell, radio link quality is indicated by Reference Signal Received Power (RSRP).
- For a UMTS cell, radio link quality is indicated by Reference Signal Code Power (RSCP).
- For inter-frequency or inter-RAT handover, radio link quality is not the measure.
- Instead, UE capability, call type, QoS requirements etc., are included in deciding handover.
- Network triggers and controls handover procedure based on the measuring report from UE (contains radio link measurement for the neighboring eNode-B).
- The serving eNode-B provides the UE with list of neighboring cells and frequencies.

- ***For intra-LTE handover there are 5 Events (A1 to A5) that trigger measurement report:***
  - ***A1:*** *The serving cell radio link quality becomes greater than an absolute threshold.*
  - ***A2:*** *The serving cell radio link quality becomes less than an absolute threshold.*
  - ***A3:*** *The neighbor cell radio link quality becomes greater than an offset relative to the serving cell.*
  - ***A4:*** *The neighbor cell radio link quality becomes greater than an absolute threshold.*
  - ***A5:*** *The serving cell radio link quality becomes less than an absolute threshold and the neighbor cell radio link quality becomes greater than an absolute threshold.*
- ***For inter-RAT handover there are 2 events (B1 and B2) that trigger measurement report:***
  - ***B1:*** *Neighbor cell radio link quality on different RAT becomes greater than an absolute threshold.*
  - ***B2:*** *The serving cell radio link quality becomes less than an absolute threshold and the neighbor cell radio link quality on a different RAT becomes greater than another threshold.*
- ***Time to Trigger***: *Amount of time each of these events must satisfy before a measurement report is triggered. "Time to Trigger" parameter is used to prevent the UE from ping-pong between eNode-Bs. Measurement report is triggered when there is a significant change in the radio link quality.*

Paging: ***Write a short note on Paging?***

It is a connection control function of the RRC protocol. In RRC-IDLE state, UE monitors a paging channel for incoming calls. In RRC-IDLE and RRC-CONNECTED states. Paging message is used to inform the UEs about a system information change, Earthquake and Tsunami Warning System (ETWS) notification etc. Change of system information occurs at specific radio frames.

- ***Modification period***: system information is same within this period. UE will acquire system information at the beginning of modification period.
- ***Paging Frame (PF):*** It is a Radio frame in which the E-UTRAN can page the UE. One PF may contain one or multiple subframes. Each subframe is "Paging Occasion (PO)". UE uses DRX in

the idle mode and monitors one PO per (Discontinuous Reception) DRX cycle and switch off its Rx during other POs to save power. Paging information is carried by PDSCH (Physical down link shared channel) physical channel.

10.6 Inter-Cell Interference (ICI) Coordination:

Write a short note on Inter-Cell Interference Coordination

- LTE uses universal frequency reuse to meet the spectrum efficiency (i.e., same spectrum is used in each cell) which leads to higher ICI.
- ICI minimization can be achieved through BS coordination or networked MIMO.
- ICI minimization techniques used for both uplink and downlink.

10.6.1 Downlink ICI Minimization:

Explain the three basic approaches to mitigate ICI in the downlink

There are 3 approaches.

1. ***ICI randomization***: Scramble the codeword after channel coding with PR sequence. Scrambling is Cell specific.
2. ***ICI Cancellation***: UE decodes the interfering signals (from neighboring cells) and subtract them from the desired signal*. Disadvantage* is UE can't decode the PDCCH from neighboring cells and does not know the transmission format of neighbor cells. *Practical solution is to use* linear spatial interference cancellation with statistical knowledge of interference channels. ICI minimization in downlink is limited by the capability and no. of antennas at UEs.

   [Explain in brief ICI coordination/avoidance]
3. ***ICI Coordination/Avoidance:*** It is based on scheduler implementation at eNode-Bs. Restrictions to the downlink resource management in a coordinated between neighboring cells. The restrictions can be on time/frequency resources or Transmit power at eNode-B. It needs additional inter eNode-B communication and UE measurements and report. The *ICI Coordination/Avoidance technique can be static or semi static.*
   - ***Static ICI Coordination/Avoidance***: This is done during cell planning and doesn't need frequent reconfiguration. It doesn't consider cell load and user distribution.
   - ***Semi-Static ICI Coordination/Avoidance***: It needs reconfigurations on a time-scale (seconds or longer). The information (transmit power, traffic load) exchange between neighboring eNode-Bs is over X2 interface.

- LTE system defines eNode-B power restriction signaling in the downlink using Relative Narrowband Transmission Power (RNTP) indicator exchanged between eNode-Bs over X2 interface. Each bit of RNTP indicator corresponds to one PRB, it indicates maximum transmitter power on that PRB. Based on the RNTP indicators from neighboring cells, each eNode-B improve the performance of UEs in its own cell by scheduling and power allocation.

Figure 10.6 shows a simple example of power patterns in three neighboring cells

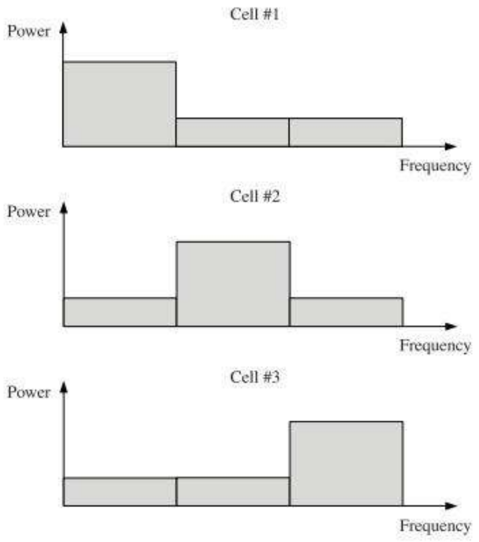


Figure 10.16 Possible downlink power levels of three neighboring cells. Edge users in each cell would be allocated to the higher power levels.

10.6.2 Uplink ICI minimization:

- ICI randomization: Scramble the encoded symbols prior to modulation.
- UE specific scrambling is used.
- *ICI cancellation*: It is more applicable in the uplink than downlink as the eNode-B has higher computational capability and more antennas.
- Uplink Power Control: LTE uses FPC to suppress ICI in the uplink.
- ICI Coordination/avoidance: Similar to downlink.

- ***Coordinated Multi-Point (CoMP) Reception***:
  - It is developed for uplink in LTE-Advanced.
  - There is a coordinated reception at multiple eNode-Bs of transmitted signals from geographically separated UEs in different cells.
  - As Uplink scheduling is performed at the eNode-B, coordinated inter-cell scheduling is applied to control ICI.
  - Uplink CoMP reception has limited impact on the radio-interface specifications.

Module – 1

Chapter 1: Evolution of Cellular Technologies

- Key Enabling Technologies and Features of LTE
- LTE Network Architecture

Chapter 2: Wireless Fundamentals

- Cellular concept Hardware Components
- Broadband wireless channel (BWC)
- Fading in Broadband wireless channel.
- Modeling of Broadband Fading Channel.
- Mitigation of Narrow band and Broadband Fading

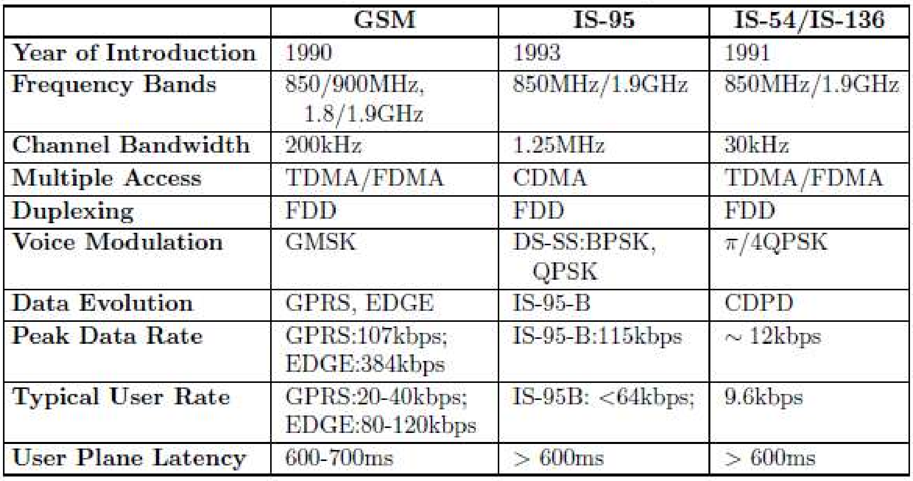
Chapter1: Evolution of Cellular Technologies

1.1 Evolution of Wireless Cellular Technologies:

- **First Generation (1G) Technology:**
  - 1G (or 1-G) refers to the first-generation
  - It is an analog based voice oriented telecommunications standards
  - AMPS (Advanced Mobile Phone system) were the popular 1G cellular system
  - Used analog FM modulation and FDD used to achieve Duplexing
  - Type of multiple access is FDMA and Channel B.W is 30Khz
  - Frequency band is 824-894 MHz.
  - Forward link and Reverse link separated by 45 MHz.
  - Operating Frequency: 150MHz / 900MHz
  - Examples for IG:
    - *Japan's Nippon Telephone and Telegraph Company (NTT) in 1979.*
    - *Nordic Mobile Telephone (NMT-400) system, deployed in Europe in 1981.*
    - *Advanced Mobile Phone Service (AMPS) in USA in 1983.*
  - Drawbacks of IG:
    - *Poor Voice Quality and Poor Battery Life*
    - *Large Phone Size and no Security*
    - *Limited Capacity and no roaming*
    - *Poor Handoff Reliability and no data services.*

- **2G and 2.5G Generation Technology:**
  - 2G is Digital based cellular system and launched in Finland in 1991.
  - 2G network use digital signaling.
  - Its data speed was up to 64Kbps.
  - Enables SMS, picture message and MMS (Multi Media Message).
  - Provides better quality and capacity.
  - Introduce two major multiplexing schemes called TDMA and CDMA.
  - Use digital modulation techniques to send digital control messages.
  - Use Digital encryption used for security and privacy.
  - Use of digital encoding and decoding schemes.
  - Use of error detection and correction codes for reliability.
  - Examples of 2G digital cellular systems include:
    - *Global System for Mobile Communications(GSM)*
    - *IS-95 CDMA, and IS-136 TDMA systems*
  - **2.5G:**Different technologies to increase the data services are over 2g networks:
    - *CDPD (Cellular Digital Packet Data)*
    - *HSCSD (High Speed Circuit Switched Data)*
    - *GPRS (General Packet Radio Service)*
    - *Packet data over CDMA and other technologies*
    - *E-Mails, Web browsing, Camera phones, Speed : 64-144 kbps*

Table 1.3 Major Second Generation Cellular Systems

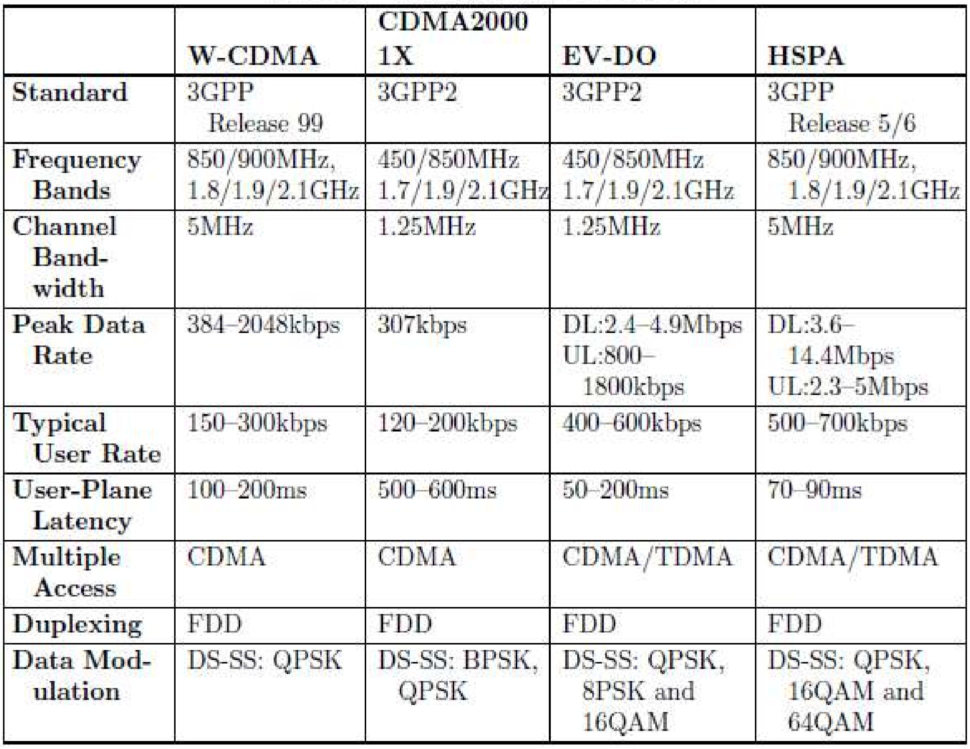


	GSM	IS-95	IS-54/IS-136
Year of Introduction	1990	1993	1991
Frequency Bands	850/900MHz, 1.8/1.9GHz	850MHz/1.9GHz	850MHz/1.9GHz
Channel Bandwidth	200kHz	1.25MHz	30kHz
Multiple Access	TDMA/FDMA	CDMA	TDMA/FDMA
Duplexing	FDD	FDD	FDD
Voice Modulation	GMSK	DS-SS:BPSK, QPSK	π/4QPSK
Data Evolution	GPRS, EDGE	IS-95-B	CDPD
Peak Data Rate	GPRS:107kbps; EDGE:384kbps	IS-95-B:115kbps	~ 12kbps
Typical User Rate	GPRS:20-40kbps; EDGE:80-120kbps	IS-95B: <64kbps;	9.6kbps
User Plane Latency	600-700ms	> 600ms	> 600ms

- Drawback of 2G:
  - *Limited data rates*
  - *Basically circuit switched system*
  - *Not supported for true mobility and less security.*

- **3G Generation technology**
  - 3G technology was introduced in year 2000s.
  - Data transmission speed increased from144Kbps to 2Mbps.
  - Increased bandwidth and data transfer rates.
  - Compatible with smart phones and Provides Web-based applications.
  - Frequency: 1.6 – 2.0 GHz and Bandwidth: 100MHz
  - Characteristic: Digital broadband, increased speed
  - Technology: CDMA-2000, UMTS, EDGE,HSPA
  - Advantages:
    - *Support high-speed data transfer from packet networks*
    - *Permit global roaming and Advanced digital services (i.e., Multimedia)*
    - *High speed web/ More security/ Video*
    - *Conferencing/ 3-D Gaming.*
    - *Large Capacities & Broadband Capabilities.*

Table 1.4 Summary of Major 3G Standards



	W-CDMA	CDMA2000 1X	EV-DO	HSPA
Standard	3GPP Release 99	3GPP2	3GPP2	3GPP Release 5/6
Frequency Bands	850/900MHz, 1.8/1.9/2.1GHz	450/850MHz 1.7/1.9/2.1GHz	450/850MHz 1.7/1.9/2.1GHz	850/900MHz, 1.8/1.9/2.1GHz
Channel Band-width	5MHz	1.25MHz	1.25MHz	5MHz
Peak Data Rate	384–2048kbps	307kbps	DL:2.4–4.9Mbps UL:800–1800kbps	DL:3.6–14.4Mbps UL:2.3–5Mbps
Typical User Rate	150–300kbps	120–200kbps	400–600kbps	500–700kbps
User-Plane Latency	100–200ms	500–600ms	50–200ms	70–90ms
Multiple Access	CDMA	CDMA	CDMA/TDMA	CDMA/TDMA
Duplexing	FDD	FDD	FDD	FDD
Data Mod-ulation	DS-SS: QPSK	DS-SS: BPSK, QPSK	DS-SS: QPSK, 8PSK and 16QAM	DS-SS: QPSK, 16QAM and 64QAM

- **Limitation of 3G:**
  - *Expensive fees for 3G Licenses Services*
  - *It was challenge to build the infrastructure for 3G*
  - *High Bandwidth Requirement*
  - *Expensive 3G Phones.*
  - *Large Cell Phones*

- **4G Generation technology**
  - It is an IP based packed switched network.
  - Speeds of 100 Mbps while moving and 1 Gbps while stationary.
  - High usability: anytime, anywhere, and with any technology.
  - Support for multimedia and integrated services at low transmission cost.
  - Smooth Handoff across heterogeneous networks.
  - Seamless connectivity and global roaming across multiple networks.
  - Interoperability with existing wireless standards.
  - Good QoS and high security.
  - It provides Dynamic bandwidth allocation, QoS and advanced Security
  - 4G can be described using MAGIC:
    - *Mobile Multimedia*
    - *Anytime Anywhere*
    - *Global Mobility Support*
    - *Integrated Wireless Solution*
    - *Customized Personal Services*
  - Example: LTE (Long Term Evolution**)**

- **5G Generation technology**
  - 5G was started from late2010s.
  - Complete wireless communication with almost no limitations.
  - It is highly supportable to WWWW (Wireless World Wide Web).
  - Aims at higher capacity than current 4G, allowing a higher density of mobile broadband users.
  - Supports
    - *Interactive multimedia*
    - *Voice streaming*
    - *Buckle up.. Internet*
    - *Enhanced security*

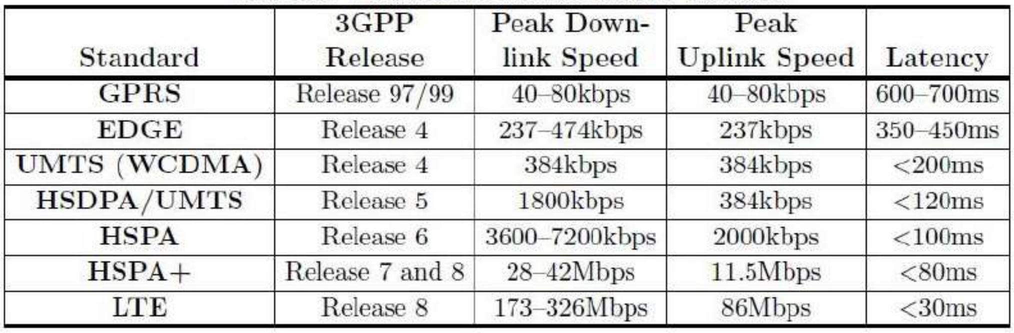
1.4 Key Enabling Technologies and Features of LTE***

1.4.1 LTE Background:

- Two groups within 3GPP (Third Generation Partnership Project) started work on developing a standard to support the expected heavy growth in IP data traffic.
  1. ***The Radio Access Network (RAN) group***: Initiated work on the Long Term Evolution (LTE) project. The LTE group developed a new radio access network called Enhanced UTRAN (E-UTRAN) as an evolution to the UMTS RAN
  2. ***Systems Aspects (SA) group***: Initiated work on the Systems Architecture Evolution (SAE) project. The SAE group developed a new all IP packet core network architecture called the Evolved Packet Core (EPC).
- Together, EUTRAN and EPC are formally called the Evolved Packet System (EPS).

- ***Demand Drivers for LTE:***
  - *Growth in high-bandwidth applications*
  - *Proliferation of smart mobile devices*
  - *Intense competition leading to flat revenues*
- ***Key Requirements of LTE Design:***
  - *Performance on Par with Wired Broadband*
  - *Flexible Spectrum Usage*
  - *Co-existence and Interworking with 3G Systems as well as Non-3GPP Systems*
  - *Reducing Cost per Megabyte*

Table 1.7 Performance Evolution of 3GPP Standards



Standard	3GPP Release	Peak Down-link Speed	Peak Uplink Speed	Latency
GPRS	Release 97/99	40–80kbps	40–80kbps	600–700ms
EDGE	Release 4	237–474kbps	237kbps	350–450ms
UMTS (WCDMA)	Release 4	384kbps	384kbps	<200ms
HSDPA/UMTS	Release 5	1800kbps	384kbps	<120ms
HSPA	Release 6	3600–7200kbps	2000kbps	<100ms
HSPA+	Release 7 and 8	28–42Mbps	11.5Mbps	<80ms
LTE	Release 8	173–326Mbps	86Mbps	<30ms

- **The key enabling technologies to achieve LTE features are**
  1. *Orthogonal Frequency Division Multiplexing (OFDM)*
  2. *SC-FDE and SC-FDMA*
  3. *Channel Dependent Multi-user Resource Scheduling*
  4. *Multi-antenna Techniques*
  5. *IP-Based Flat Network Architecture*

1.4.2 Orthogonal Frequency Division Multiplexing (OFDM) ***

- 3G systems such as UMTS and CDMA2000 are based on CDMA technology.
    - *Advantage*: CDMA Performs remarkably well for low data rate communications, where a large number of users can be multiplexed to achieve high system capacity.
    - *Limitation*: For high-speed applications, CDMA becomes untenable due to the large bandwidth needed to achieve useful amounts of spreading.
- OFDM has emerged as a technology of choice for achieving high data rates.
- It is the core technology used by a variety of systems including Wi-Fi and WiMAX.
- **The following advantages of OFDM led to its selection for LTE:**
    1. ***Elegant solution to multipath interference***: The critical challenge to high Bit-rate transmissions in a wireless channel is inter symbol interference (ISI) caused by multi path. At high data rates, the symbol time is shorter; hence, it only takes a small delay to cause ISI.OFDM is a multicarrier modulation technique that overcomes this challenge in an elegant manner. It increases the symbol duration of each stream such that the multipath delay spread is only a small fraction of the symbol duration. In OFDM, the subcarriers are orthogonal to one another over the symbol duration, thereby avoiding the need to have non-over lapping subcarrier channels to eliminate ISI.
    2. ***Reduced computational complexity***: OFDM can be easily implemented using Fast Fourier Transforms (FFT/IFFT), and the computational requirements grow only slightly faster than linearly with data rate or bandwidth. The computational complexity of OFDM = (BlogBTm), where B is the bandwidth and Tm is the delay spread. Reduced complexity is particularly attractive in the downlink as it simplifies receiver processing and thus reduces mobile device cost and power consumption.
    3. ***Graceful degradation of performance under excess delay***: The performance of an OFDM system degrades gracefully as the delay spread exceeds the designed value. OFDM is well suited for adaptive modulation and coding, which allows the system to make the best of the available channel conditions.
    4. ***Exploitation of frequency diversity***: OFDM facilitates coding and interleaving across subcarriers in the frequency domain, which can provide robustness against burst errors caused by portions of the transmitted spectrum undergoing deep fades. OFDM also allows for the channel bandwidth to be scalable without impacting the hardware design of the base station and the mobile station.

5. ***Enables efficient multi-access scheme***: OFDM can be used as a multi-access scheme by partitioning different subcarriers among multiple users. This scheme is referred to as OFDMA and is exploited in LTE.
6. ***Robust against narrowband interference***: OFDM is relatively robust against narrowband interference, since such interference affects only a fraction of the subcarriers.
7. ***Suitable for coherent demodulation***: It is relatively easy to do pilot-based channel estimation in OFDM systems, which renders them suitable for coherent demodulation schemes that are more power efficient.
8. ***Facilitates use of MIMO***: MIMO refers to a collection of signal processing techniques that use multiple antennas at both the transmitter and receiver to improve system performance. For MIMO techniques to be effective, it is required that the channel conditions are such that the multipath delays do not cause ISI interference OFDM, however, converts a frequency selective broad band channel into several narrowband flat fading channels where the MIMO models and techniques work well.
9. ***Efficient support of broadcast services***: It is possible to operate an OFDM network as a single frequency network (SFN). This allows broadcast signals from different cells to combine over the air to significantly enhance the received signal power, thereby enabling higher data rate broadcast transmissions for a given transmit power.

- ***Disadvantages of OFDM:***
  - *Peak-to-Average Ratio (PAR)*: OFDM has high PAR, which causes non-linearity and clipping distortion when passed through an RF amplifier. It increases the cost of the transmitter and is wasteful of power. OFDM is tolerated in the downlink as part of the design, for the uplink LTE selected a variation of OFDM that has a lower peak-to-average ratio. The modulation of choice for the uplink is called Single Carrier Frequency Division Multiple Access (SC-FDMA).

1.4.3 SC-FDE and SC-FDMA:

- ***Single-Carrier Frequency Domain Equalization (SC-FDE):***
  - It is a single-carrier (SC) modulation combined with frequency-domain equalization (FDE).
  - It is an alternative approach to inter symbol interference (ISI) mitigation.
  - It uses QAM rather than IFFT used OFDM to send data with a cyclic prefix added.

- SC-FDE retains all the advantages of OFDM such as multipath resistance and low complexity, while having a low peak-to-average ratio of 4-5dB.
- It keeps the MS cost down and the battery life up.
- LTE incorporated a SC-FDE as a power efficient transmission scheme for the uplink.

- ***Single-Carrier Frequency Division Multiple Access(* SC-FDMA)**
  - A multi-user version of SC-FDE, called SC-FDMA.
  - The uplink of LTE implements uses to SC-FDMA, which allows multiple users to use parts of the frequency spectrum.
  - SC-FDMA closely resembles OFDMA and also preserves the PAR properties.
  - The drawback of SC-FDE is increases the complexity of the transmitter and the receiver.

1.4.4 Channel Dependent Multi-user Resource Scheduling

- The OFDMA scheme used in LTE provides enormous flexibility in how channel resources are allocated.
- OFDMA allows for allocation in both time and frequency and it is possible to design algorithms to allocate resources in a flexible and dynamic manner to meet arbitrary throughput, delay, and other requirements.
- The standard supports dynamic, channel-dependent scheduling to enhance overall system capacity.

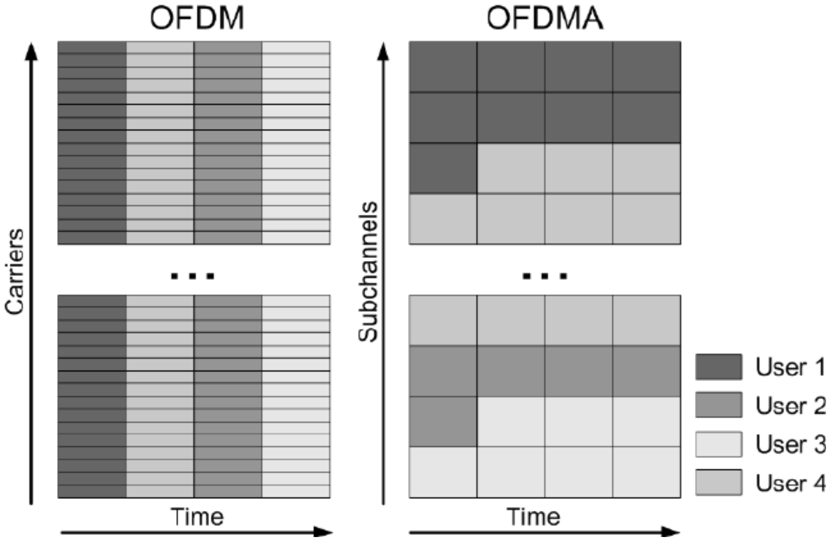


Figure 1. Resource mapping in OFDMA

- In OFDM, It is possible to allocate subcarriers among users in such a way that the overall capacity is increased. This technique, called frequency selective multiuser scheduling, calls for focusing transmission power in each user's best channel portion.

- In OFDMA, frequency selective scheduling can be combined with multi-user time domain scheduling.
- Capacity gains are also obtained by adapting the modulation and coding to the instantaneous signal-to-noise ratio conditions for each user subcarrier.
- For high-mobility users, OFDMA can be used to achieve frequency diversity. By coding and interleaving across subcarriers. Frequency diverse scheduling is best suited for control signaling and delay sensitive services.

1.4.5 Multi-antenna Techniques:

- The LTE standard provides extensive support for implementing advanced multi-antenna solutions to improve link robustness, system capacity, and spectral efficiency.
- Multi-antenna techniques supported in LTE include:

1. ***Transmit diversity***: Diversity means send copies of the same signal by using two or more communication channels with different characteristics. This is a technique to combat multipath fading in the wireless channel.

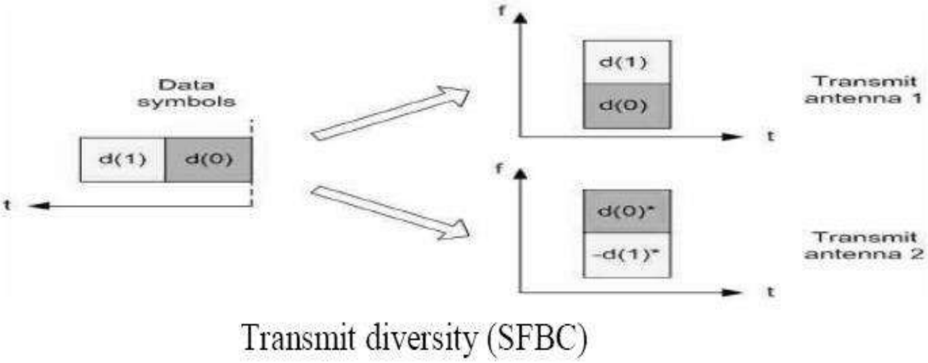


Figure 2. Resource mapping in OFDMA

LTE transmit diversity is based on space-frequency block coding (SFBC) techniques. Transmit diversity is primarily intended for common downlink channels that cannot make use of channel-dependent scheduling. It increases system capacity and cell range.

2. ***Beamforming***: It is a type of RF (radio frequency) management and signal processing technique in which an access point uses multiple antennas to send out the same signal. Multiple antennas in LTE may also be used beamforming technique to transmit the beam in the direction of the receiver and away from interference, thereby improving the received signal-to-interference ratio. It can provide significant improvements in coverage range, capacity, reliability, and battery life. It can also be useful in providing angular information for user tracking. LTE supports beamforming in the downlink.

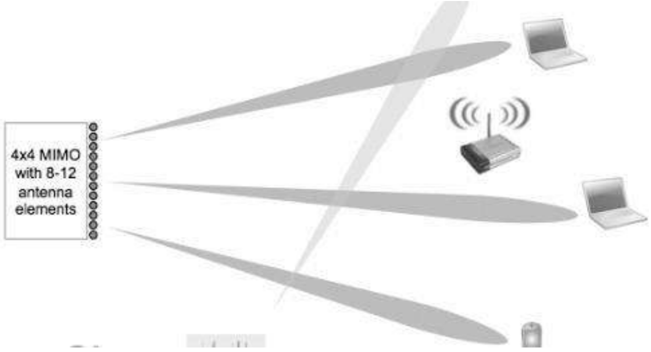


Figure 3. Beamforming with MIMO

3. ***Spatial multiplexing***: In spatial multiplexing, multiple independent streams can be transmitted in parallel over multiple antennas and can be separated at the receiver using multiple receive chains through appropriate signal processing. Spatial multiplexing provides data rate and capacity gains proportional to the number of antennas used. It works well under good SNR and light load conditions. LTE standard supports spatial multiplexing with up to four transmit antennas and four receiver antennas.

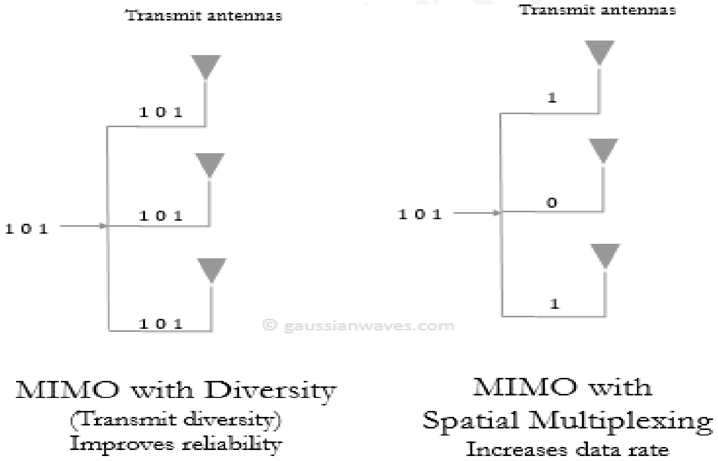


Figure4: Comparison of MIMO with Diversity and spatial multiplexing

4. ***Multi-user MIMO***: Since spatial multiplexing requires multiple transmit antennas, it is currently not supported in the uplink due to complexity and cost considerations. However, multi-user MIMO (MU-MIMO), which allows multiple users in the uplink, each with a single antenna, to transmit using the same frequency and time resource, is supported. The signals from the different MU-MIMO users are separated at the base station receiver using accurate channel state information of each user obtained through uplink reference signals that are orthogonal between users.

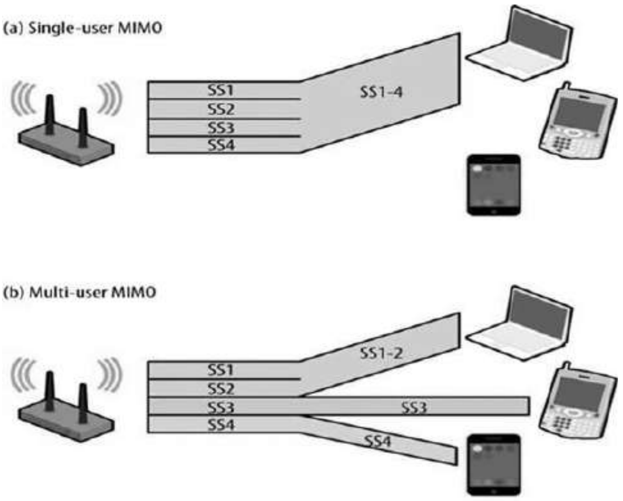


Fig 5: Comparison between Single and multiuser MIMO

1.4.5 IP-Based Flat Network Architecture: The lower infrastructure cost, lower latency and fewer nodes are requirements drove the design toward a flat architecture. It also means fewer interfaces and protocol-related processing, and reduced interoperability testing, which lowers the development and deployment cost. Fewer nodes also allow better optimization of radio interface, merging of some control plane protocols, and short session start-up time. Figure 6 shows how the 3GPP network architecture evolved over a few releases.

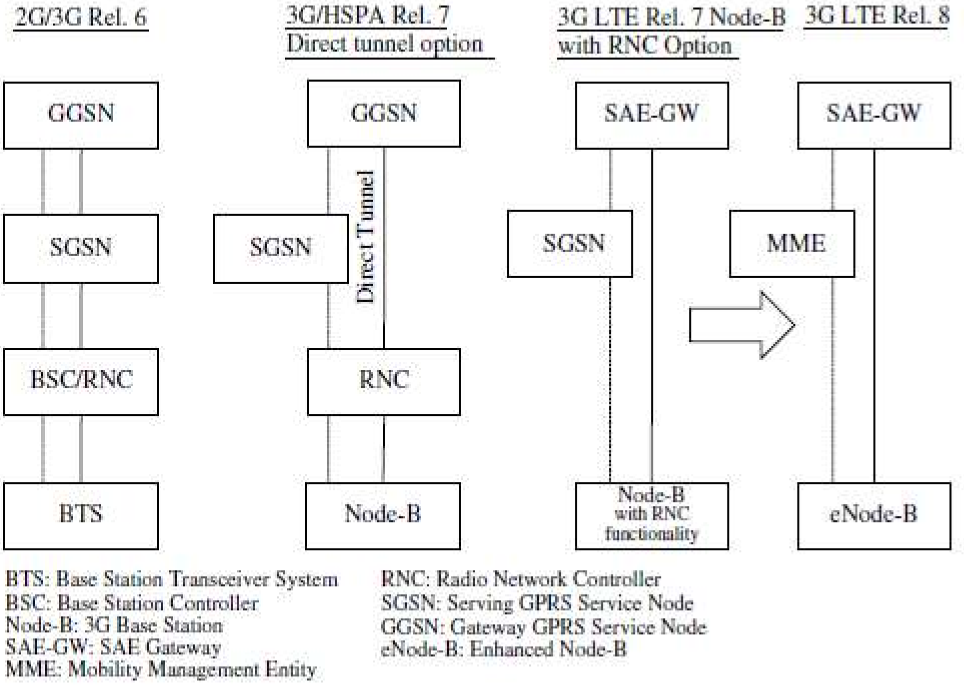


Fig 6: 3GPP evolution toward a flat LTE SAE architecture

- ***Flat LTE architecture description***:
  - 3GPP Release 6 architecture, has four network elements in the data path: *Base station (BS), Radio Network Controller (RNC), Serving GPRS Service Node (SGSN), and Gateway GRPS Service Node (GGSN).*

- Release 7 introduced a direct tunnel option from the RNC to GGSN, which eliminated SGSN from the data path.
- LTE on the other hand, will have only two network elements in the data path: the enhanced Node-B or eNode-B& a System Architecture Evolution Gateway (SAE-GW).
- LTE merges the BS and RNC functionality into a single unit. The control path includes a functional entity called the Mobility Management Entity (MME), which provides control plane functions related to subscriber, mobility, and session management.
- The MME and SAE-GW could be collocated in a single entity called the access gateway (a-GW).
- A key aspect of the LTE flat architecture is that all services, including voice, are supported on the IP packet network using IP protocols. Unlike previous 2g and 3g systems, which had a separate circuit-switched sub-network for supporting voice with their own Mobile Switching Centers (MSC) and transport networks, LTE envisions only a single evolved packet-switched core, the EPC, over which all services are supported, which could provide huge operational and infrastructure cost savings. However, that although LTE has been designed for IP services with a flat architecture, due to backwards compatibility reasons certain legacy, non-IP aspects of the 3GPP architecture such as the GPRS tunneling protocol and PDCP (packet data convergence protocol) still exists within the LTE network architecture.

1.5 LTE Network Architecture***

- *Introduction*: The core network design by 3GPP Release 8 to support LTE is called the Evolved Packet Core (EPC). EPC is designed to provide a high capacity, all IP, reduced latency, flat architecture that dramatically reduces cost and supports advanced real-time and media-rich services with enhanced quality of experience. It is designed not only to support new radio access networks such as LTE, but also provide interworking with legacy 2G GERAN and 3G UTRAN networks connected via SGSN.
- *Functions of LTE architecture:* It include access control, packet routing and transfer, mobility management, security, radio resource management, and network management.
- *LTE architectural elements:* The EPC includes four new elements:
  1. *Serving Gateway (SGW)*
  2. *Packet Data Network Gateway (PGW):*
  3. *Mobility Management Entity (MME):* Which supports user equipment context and identity as well as authenticates and authorizes users.
  4. *Policy and Charging Rules Function (PCRF):* Which manages QoS aspects.

Figure 7 shows the end-to-end architecture including how the EPC supports LTE as well as current and legacy radio access networks. A brief description of each of the four new elements is provided here:

- **Serving Gateway (SGW):**
    - Which terminates the interface toward the 3GPP radio access networks.
    - It acts as a demarcation point between the RAN and core network, and manages user plane mobility.
    - It serves as the mobility anchor when MT move across areas served by different eNode-B elements in E-UTRAN, as well as across other 3GPP radio networks such as GERAN and UTRAN.
    - SGW does downlink packet buffering and initiation of network-triggered service request procedures.
    - Other functions of SGW include:
        - *Lawful interception, packet routing and forwarding.*
        - *Transport level packet marking in the uplink and the downlink.*
        - *Accounting support for per user and inter-operator charging.*

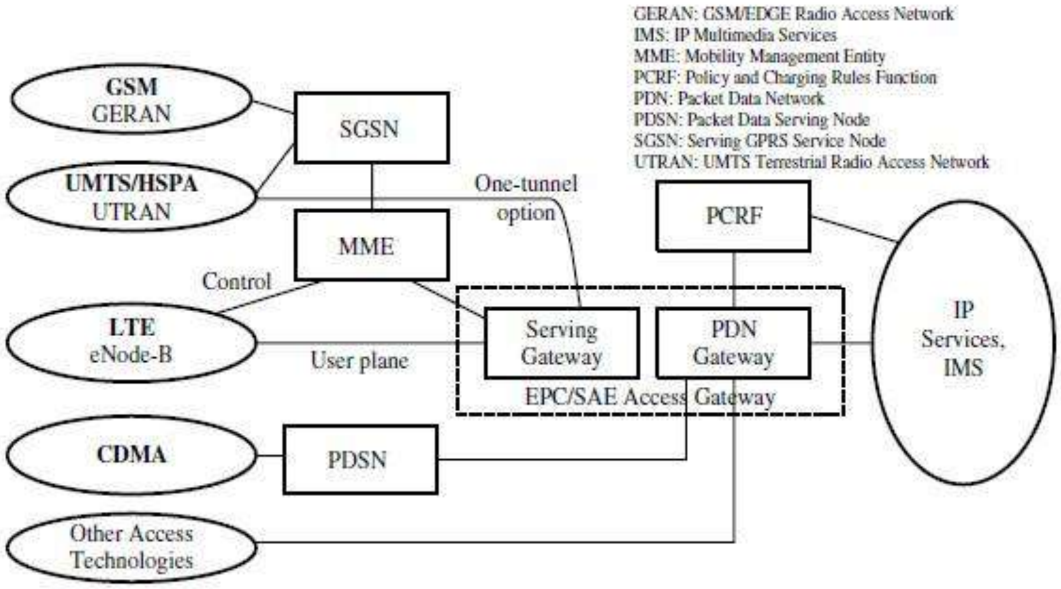

GERAN: GSM/EDGE Radio Access Network
IMS: IP Multimedia Services
MME: Mobility Management Entity
PCRF: Policy and Charging Rules Function
PDN: Packet Data Network
PDSN: Packet Data Serving Node
SGSN: Serving GPRS Service Node
UTRAN: UMTS Terrestrial Radio Access Network
GSM
GERAN
SGSN
UMTS/HSPA
UTRAN
One-tunnel option
PCRF
MME
Control
LTE
eNode-B
User plane
Serving Gateway
PDN Gateway
IP Services, IMS
EPC/SAE Access Gateway
CDMA
PDSN
Other Access Technologies

Fig 7. Evolved Packet Core architecture.

- **Packet Data Network Gateway (PGW):**
    - It controls IP data services, does routing, allocates IP addresses, enforces policy, and provides access for non-3GPP access networks.

  - The PGW acts as the termination point of the EPC toward other Packet Data Networks (PDN) such as the Internet, private IP network, or the IMS network providing end-user services.
  - It serves as an anchor point for sessions toward external PDN and provides functions such as user IP address allocation, policy enforcement, packet filtering, and charging support.
  - Policy enforcement includes operator-defined rules for resource allocation to control data rate, QoS, and usage.
  - Packet filtering functions include deep packet inspection for application detection.
- **Mobility Management Entity (MME):**
  - The MME performs the signaling and control functions to manage the user terminal access to network connections, assignment of network resources.
  - Mobility management function such as idle mode location tracking, paging, roaming, and handovers.
  - MME controls all control plane functions related to subscriber and session management.
  - The MME provides security functions such as providing temporary identities for user terminals, interacting with Home Subscriber Server (HSS) for authentication, and negotiation of ciphering and integrity protection algorithms.
  - It is also responsible for selecting the appropriate serving and PDN gateways, and selecting legacy gateways for handovers to other GERAN or UTRAN networks.
  - MME manages thousands of eNode-B elements, which is one of the key differences from 2G or 3G.
- **Policy and Charging Rules Function (PCRF):**
  - It is a concatenation of Policy Decision Function (PDF) and Charging Rules Function (CRF).
  - The PCRF interfaces with the PDN gateway and supports service data flow detection, policy enforcement, and flow-based charging.

Chapter 2: Wireless Fundamentals

2.1 Cellular System

2.1.1 The Cellular Concept:

- AT&T proposed a core idea of cellular system in 1971.
- In cellular systems, the service area is subdivided into smaller geographic areas called *cells*. Each cell served by their own lower-power Base Station (BS).
- Neighboring cells do not use same set of frequencies to prevent interference.
- In order to minimize interference between cells, the transmit power level of each base station is regulated to be just enough to provide the required signal strength at the cell boundaries.
- *Core cellular Principles*: Small cells tessellate overall coverage area. User's "*handoff*" as they move from one cell to another. The same frequency channels can be reassigned to different cells, as long as those cells are spatially isolated called "*frequency reuse"* concept. It increases the cellular system capacity.

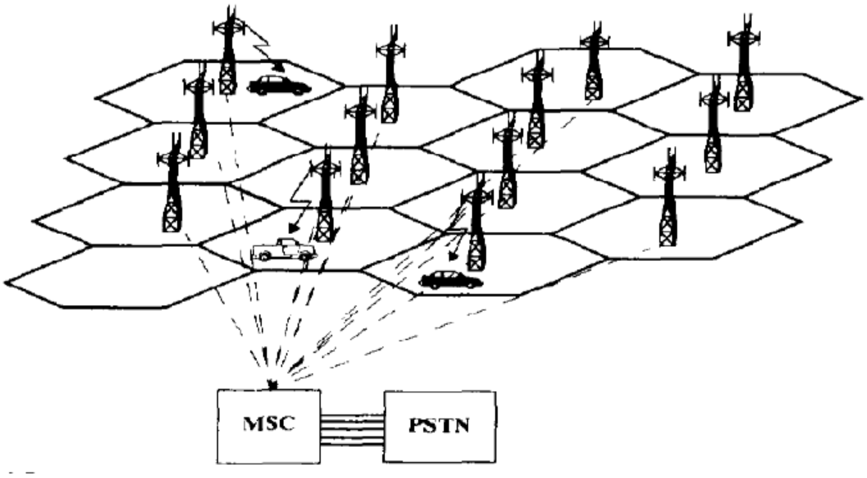
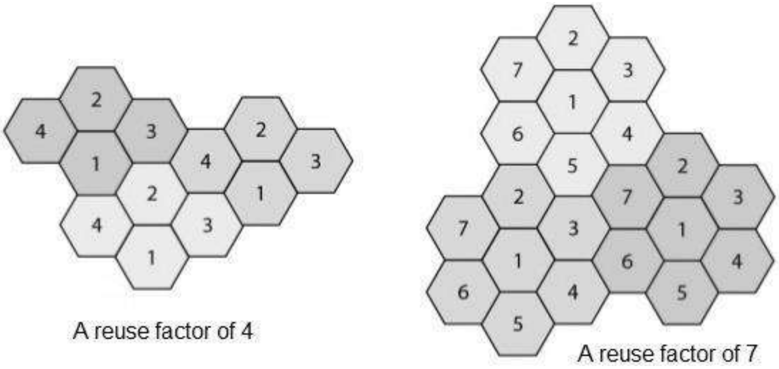


Fig 8.Simple cellular system architecture.

- ***Frequency planning**:* It is required to determine a proper frequency reuse factor and a geographic reuse pattern. Frequencies can be reused should be determined such that the interference between base stations is kept to an acceptable level. The frequency reuse factor f is defined as f ≤ 1, where f = 1 means that all cells reuse all the frequencies. Accordingly, f = 1/3 implies that a given frequency band is used by only 1 out of every 3 cells.



A reuse factor of 4

A reuse factor of 7

Fig 9.Frequency reuse pattern.

- ***Co-cells and cluster:*** Co-cells are cells in cellular system which uses the *same frequency channel se*t. The reuse of the same frequency channels should be intelligently planned in order to maximize the geographic distance between the co-channel base stations. Figure 10 shows an example of hexagonal cellular system model with frequency reuse factor f = 1/7. The group of cells which are using entire frequency channels set are called "*clusters*"

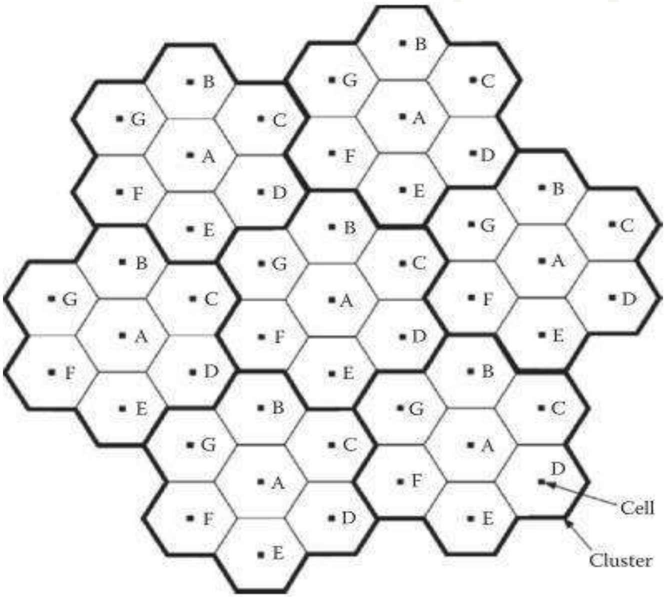


Figure 10: Standard figure of a hexagonal cellular system with f=1/7.

- ***Cellular system capacity***: The overall system capacity can increase by simply making the cells smaller and turning down the power. In this manner, cellular systems have a very desirable scaling property. As the cell size decreases, the transmit power of each base station also decreases correspondingly. For example, if the radius of a cell is reduced by half when the propagation path loss exponent is 4, the transmit power level of a base station is reduced by 12 dB (=l0log16 dB).

- ***Handoff:*** Since cellular systems support user mobility, seamless call transfer from one cell to another should be provided. The handoff process provides a means of the seamless transfer of a connection from one base station to another. Achieving smooth handoffs is a challenging aspect of cellular system design.
- ***Advantages of cellular concept:*** Small cells give a large capacity advantage and reduce power consumption and allows frequency reuse.
- ***Drawback of cellular system:*** As cell size decreases, the number of cells for the same service area need more base stations and their associated hardware costs also increases. It leads to frequent handoffs. Interference level increases and effect on service efficiency.

2.1.2 Analysis of Cellular Systems

- The performance of wireless cellular systems is significantly limited by Co-channel interference (CCI) and other cell interference (OCI) which comes from other users in the same cell or from other cells.
- The cellular systems performance (capacity, reliability) is measured by SIR of the desired cell, i.e., the amount of desired power to the amount of transmitted power.
- The spatial isolation between co-channel cells can be measured by defining the parameter Z, called co-channel reuse ratio is given by

$$= \frac{D}{R} \equiv \sqrt{3/f} \qquad (1)$$

Where D = distance between the co-cells

R = radius of the desired cell

1/f = size of the cluster and inverse of the frequency reuse factor N, therefore equation (1) becomes

$$= \frac{D}{R} \equiv \sqrt{3N} \qquad (2)$$

- *Conclusion:* As the cluster size N increases, CCI decreases, so that it improves the quality of communication link and capacity. However, the overall spectral efficiency decreases with the size of a cluster, so f should be chosen just small enough to keep the received signal-to-interference-plus-noise ratio (SINR) above acceptable levels.
- *Signal to Noise ratio (SNR) of cellular system:* It is given by

$$\frac{S}{I} = \frac{S}{\sum_{i=1}^{N_l} \mathrm{I}_i} \qquad (3)$$

Where S = Received power of desired signal

I_i= Interference power from the i^th^ co-cell base station

- The received SIR depends on the location of each mobile station, and it should be kept above an appropriate threshold for reliable communication.
- The received SIR at the cell boundaries is of great interest since this corresponds to the worst interference scenario.
- The received SIR for the worst case described in Fig 11 and its empirical path loss formula given as

$$\frac{S}{I} = \frac{\chi_0}{\chi_o+\sum_{i=1}^{2}\chi_i+2^{-\propto}\sum_{i=3}^{5}\chi_i+(2.633)^{-\propto}\sum_{i=6}^{11}\chi_i} \tag{4}$$

Where χ_i denotes the shadowing from the ith base station

$\propto$ = path loss components.

χ_o= lognormal distribution for the shadowing value.

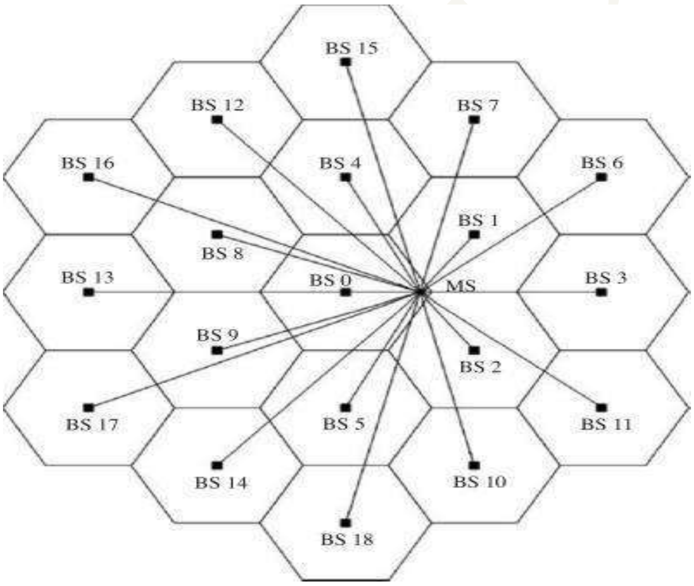


Figure 11: Forward link interference in a hexagonal cellular system (worst case).

- *Outage probability (P_0):* The outage probability that the received SIR falls below a threshold can be derived from the distribution. If the mean and standard deviation of the lognormal distribution areα and σ in dB, the outage probability is derived in the form of Q function is given by

$$\mathbf{P_{OUT=}}\ P\ [\text{SIR}<\gamma] = Q\ (\gamma-\ \mu/\ \sigma) \qquad [5]$$

Where γ= threshold SIR level in dB

- Lower frequency reuse factor is typically adopted in the system design to satisfy the target outage probability at the sacrifice of spectral efficiency

2.1.3 Sectoring:

- It is a capacity expansion technique by keep the cell radius unchanged and seek methods to decrease the D /R ratio.
- It is desirable a techniques to improve SIR without sacrificing so much bandwidth.
- Uses directional antennas by replacing a single Omni-directional antenna at the base station. It provides interference reduction, hence S/I ratio increases.
- No capacity is lost from sectoring because each sector can reuse time and code slots, so each sector has the same nominal capacity as an entire cell.
- The capacity in each sector is actually higher than that in a non-sectored cellular system because the interference is reduced by sectoring. An illustration of sectoring is shown in Figure 12.

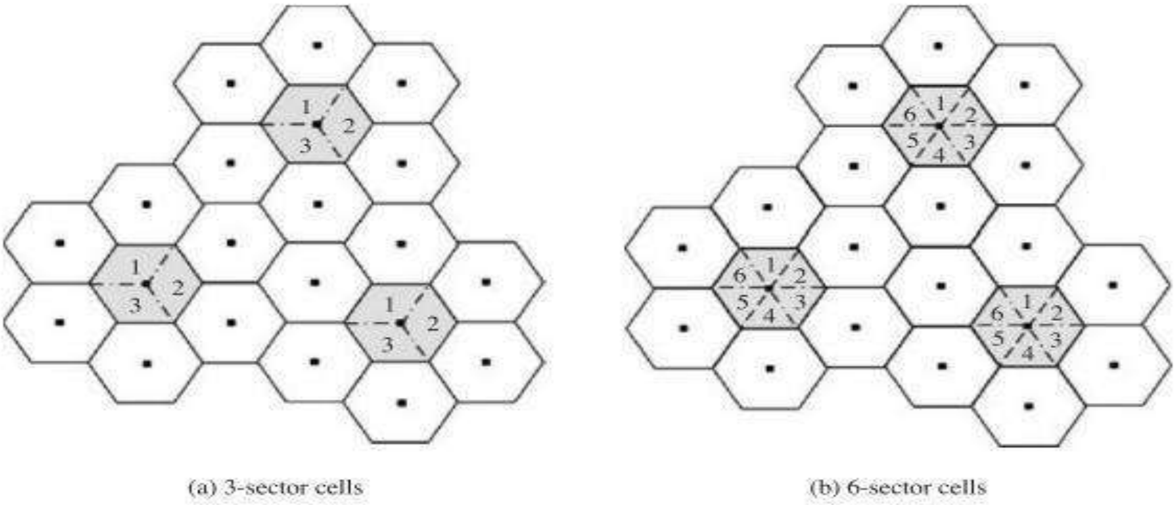


Figure 12: Three-sector (120-degree) and Six-sector (60-degree) cells.

- In Figure 12a, if each sector 1 points the same direction in each cell, then the interference caused by neighboring cells will be dramatically reduced.
- An alternative way to use sectors is to reuse frequencies in each sector and the time/code/frequency slots can be reused in each sector, but there is no reduction in the experienced interference.
- As the number of sectors per cell increases the SIR also increases, thus the capacity of cellular system increases.

- *Advantages of sectoring*:
  1. It is an effective and practical approach to the OCI problem.
  2. It is an antenna technique to increase the system capacity.

- *Drawback*:
    1. Sectoring increases the number of antennas at each base station, hence it increases the implantation cost and the number of handoffs increases
    2. It reduces trunking efficiency due to channel sectoring at the base station.
    3. It also increases the overhead due to the increased number of inter sector handoffs.
    4. It causes inter sector interference as well as power loss.

- **New Approaches to other Cell Interference**. Following are other approaches to reduces cell interference
    1. Use advanced signal processing techniques at the receiver and/or transmitter as a means of reducing or cancelling the perceived interference.
    2. Use network-level approaches such as cooperative scheduling or encoding across Base station. Adopt multi-cell power control and distributed antenna technique.

2.2 The Broadband Wireless Channel: Fading

- ***Introduction:*** Path loss and shadowing attenuation effects due to distance or obstacles. Fading is severe attenuation phenomenon in wireless channels likely for short distance caused by the reception of multiple versions of the same signal. The multiple received versions are caused by reflections that are referred to as multipath. The reflections may arrive at very close to the same time. The multiple different paths between the transmitter and receiver visualization shown in Figure 2.13

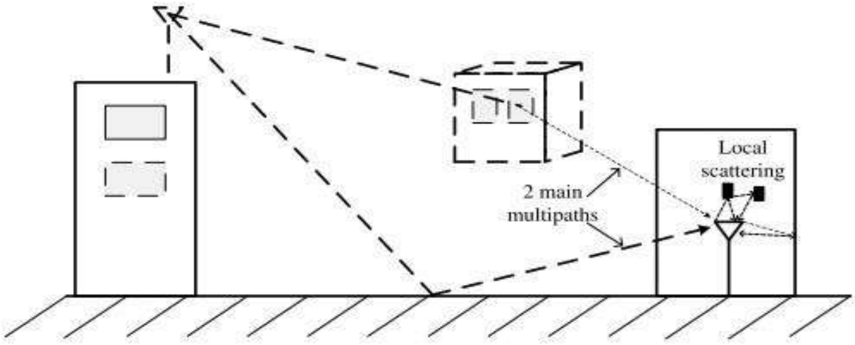


Figure 2.13: The channel may have a few major paths with quite different lengths, and then the receiver may see a number of locally scattered versions of those paths.

- *Fading effect*: When some of the reflections arrive at nearly the same time, the combined effect of those reflections shown in Figure 2.14. Depending on the phase difference between the arriving signals, the interference can be either constructive or destructive.

It causes a very large observed difference in the amplitude of the received signal even over very short distances.

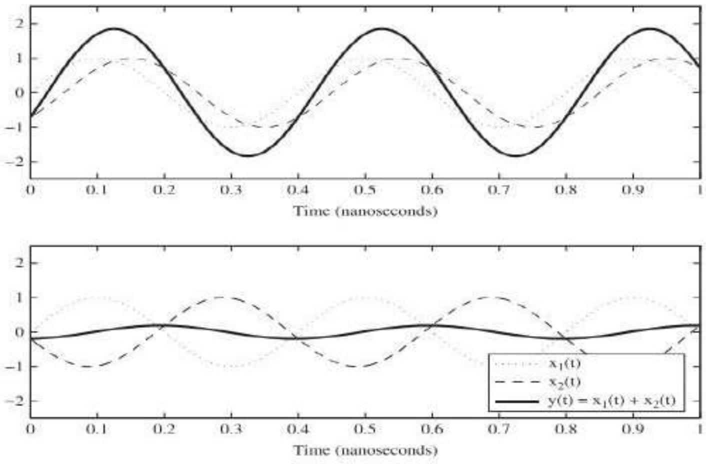


Figure 2.14: The difference between constructive interference (top) and destructive interference (bottom) at 4 = 2.5GHz is less than 0.1 nanoseconds in phase, which corresponds to about 3 cm

- The moving the transmitter or receiver even a very short distance can have a dramatic effect on the received amplitude, even though the path loss and shadowing effects may not have changed at all.
- *Time-varying tapped-delay line channel model of fading:* Either the transmitter or receiver move relative to each other, the channel response h(t) will change. This channel response can be thought of as having two dimensions as shown in Figure 2.15:
  1. *Delay dimension*(τ)
  2. Time-dimension(t).

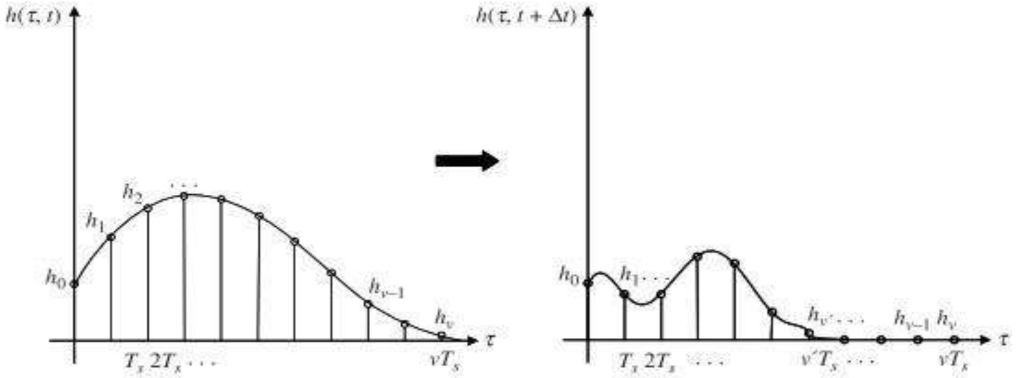


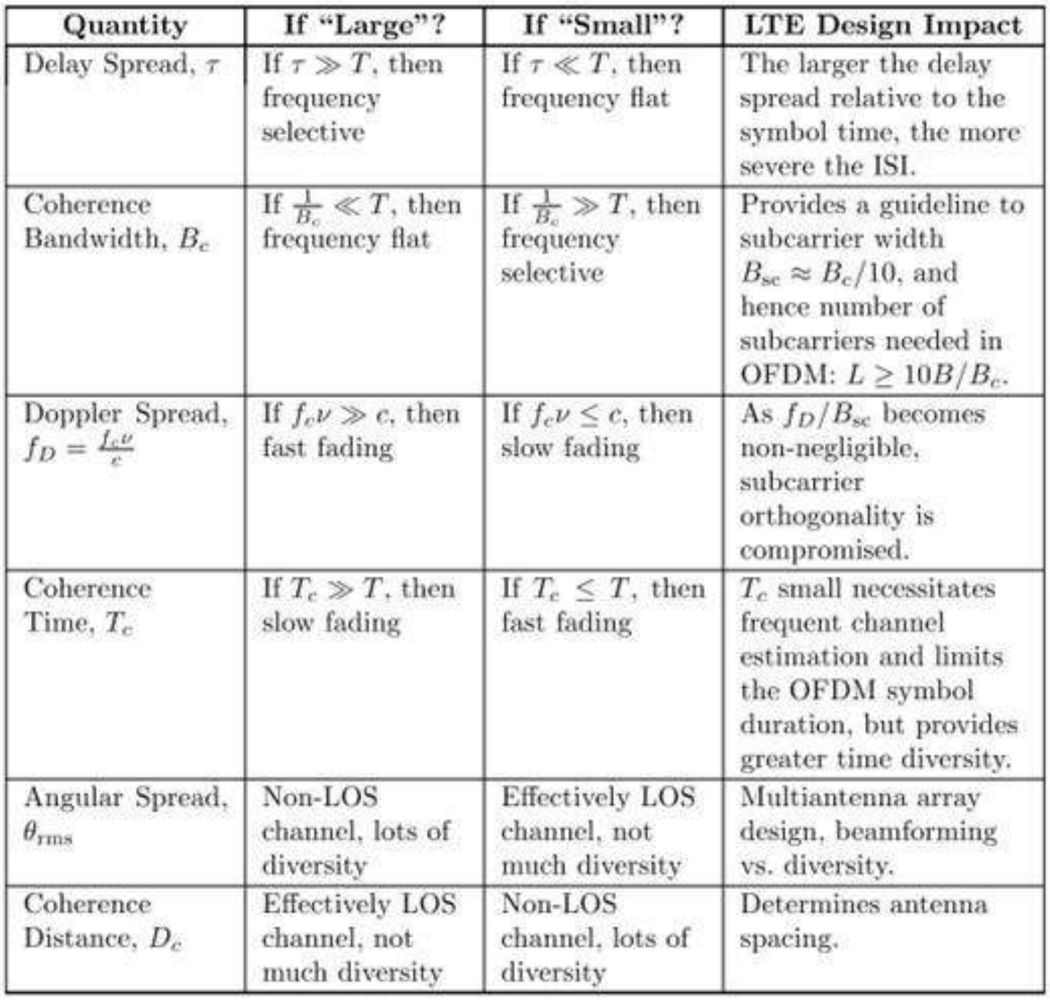
Figure 2.15: The delay τ corresponds to how long the channel impulse response lasts. The channel is time varying, so the channel impulse response is also a function of time, i.e., h $(\tau$, t), and can be quite different at time $(t + \Delta t)$ than it was at time t.

- Since the channel changes over distance (and hence time), the values of h_0, h_1, ... h_v may be totally different at time t vs. time$t + \Delta t$. Because the channel is highly variant in both the τ and t dimensions.
- The fundamental function used to statistically describe broadband fading channels is the two-dimensional autocorrelation function, A($\Delta\tau$, Δt). The autocorrelation function is defined as

$$\begin{aligned} A(\Delta\tau, \Delta t) &= E\big[h(\tau_1, t_1)h^*(\tau_2, t_2)\big] \\ &= E\big[h(\tau_1, t)h^*(\tau_2, t + \Delta t)\big] \\ &= E\big[h(\tau, t)h^*(\tau + \Delta\tau, t + \Delta t)\big] \end{aligned} \tag{6}$$

- The above equation (6) is referred to as Wide Sense Stationary Uncorrelated Scattering (WSSUS), which is the most popular model for wideband fading channels.

Table 2.2 Summary of Broadband Fading Parameters, with Rules of Thumb



Quantity	If "Large"?	If "Small"?	LTE Design Impact
Delay Spread, τ	If $\tau \gg T$, then frequency selective	If $\tau \ll T$, then frequency flat	The larger the delay spread relative to the symbol time, the more severe the ISI.
Coherence Bandwidth, B_c	If $\frac{1}{B_c} \ll T$, then frequency flat	If $\frac{1}{B_c} \gg T$, then frequency selective	Provides a guideline to subcarrier width $B_{sc} \approx B_c/10$, and hence number of subcarriers needed in OFDM: $L \geq 10B/B_c$.
Doppler Spread, $f_D = \frac{f_c\nu}{c}$	If $f_c\nu \gg c$, then fast fading	If $f_c\nu \leq c$, then slow fading	As f_D/B_{sc} becomes non-negligible, subcarrier orthogonality is compromised.
Coherence Time, T_c	If $T_c \gg T$, then slow fading	If $T_c \leq T$, then fast fading	T_c small necessitates frequent channel estimation and limits the OFDM symbol duration, but provides greater time diversity.
Angular Spread, θ_{rms}	Non-LOS channel, lots of diversity	Effectively LOS channel, not much diversity	Multiantenna array design, beamforming vs. diversity.
Coherence Distance, D_c	Effectively LOS channel, not much diversity	Non-LOS channel, lots of diversity	Determines antenna spacing.

- **Wireless channel Parameters**: *The key parameters to evaluate the wireless channels are*

  2.2.1 Delay Spread and Coherence Bandwidth

  2.2.2 Doppler Spread and Coherence Time

  2.2.3 Angular Spread and Coherence Distance

2.2.1 Delay Spread and Coherence Bandwidth:

- ***Delay Spread:***
  - The delay spread is mostly used in the characterization of wireless channels.
  - It is a measure of the multipath richness of a communications channel.
  - It specifies the duration of the channel impulse response h $(\tau$, t).

$$v \approx \frac{\tau_{max}}{T_s} \qquad (8)$$

Where T_s is the sampling time

- Delay spread can be quantified through different metrics, although the most common one is the root mean square (rms) delay spread.

$$\mu_\tau = \frac{\int_0^\infty \Delta\tau A_\tau(\Delta\tau)d(\Delta\tau)}{\int_0^\infty A_\tau(\Delta\tau)d(\Delta\tau)}$$

$$\tau_{\text{rms}} = \sqrt{\frac{\int_0^\infty (\Delta\tau - \mu_\tau)^2 A_\tau(\Delta\tau)d(\Delta\tau)}{\int_0^\infty A_\tau(\Delta\tau)d(\Delta\tau)}}$$

- The formula above is also known as the root of the second central moment of the normalized delay power density spectrum.
- The importance of delay spread is how it affects the Inter Symbol Interference (ISI).
- τ_{rms} gives a measure of the "width" or "spread" of the channel response in time.
- A general rule of thumb is that $\tau_{max} \approx 5\tau_{rms}$

- ***Coherence Bandwidth(B_c):***
  - It is a statistical measurement of the range of frequencies over which the channel can be considered "flat"
  - The B_c is the frequency domain dual of the channel delay spread.
  - The coherence bandwidth gives a rough measure for the maximum separation between a frequency f_1 and a frequency f_2 where the channel frequency response is correlated. That is

$$|f_1 - f_2| \le B_c \Rightarrow \quad H(f_1) \approx H(f_2)$$
$$|f_1 - f_2| > B_c \Rightarrow \quad H(f_1) \text{ and } H(f_2) \text{ are uncorrelated}$$

  - τ_{max} is a value describing the channel duration, B_c is a value describing the range of frequencies over which the channel stays constant. Given the channel delay spread, it can be shown that

$$B_c \approx \frac{1}{5\tau_{\text{rms}}} \approx \frac{1}{\tau_{\max}}.$$

  - The important and prevailing feature is that B_c and τ_r are inversely related.
  -

2.2.2 *Doppler Spread and Coherence Time:*

- Delay spread and coherence bandwidth are parameters which describe the time dispersive nature of the channel in a local area. However, they do not offer information about the time varying nature of the channel caused by either relative motion between the mobile and base station
- Doppler spread and coherence time are parameters which describe the time varying nature of the channel in a small-scale region.

- ***Doppler Spread(B_D):***
  - Doppler spread is a measure of the spectral broadening caused by the time rate of change of the mobile radio channel and is defined as the range of frequencies over which the received Doppler spectrum is essentially non-zero.
  - The Doppler power spectrum gives the statistical power distribution of the channel versus frequency for a signal transmitted at just one exact frequency.
  - Whereas the power delay profile was caused by multipath between the transmitter and receiver, the Doppler power spectrum is caused by motion between the transmitter and receiver.

- The Doppler power spectrum is the Fourier transform of $A_t(\Delta t)$ is given by

$$\rho_t(\Delta f) = \int_{-\infty}^{\infty} A_t(\Delta t) e^{-\Delta f.\Delta}(d\Delta t) \qquad (9)$$

- When a pure sinusoidal tone of frequency fc is transmitted, the received signal spectrum, called the Doppler spectrum.
- The spectrum components in the range $f_c - f_d$ to $f_c + f_d$, where f_d is the Doppler shift.

- The amount of spectral broadening depends on f_d which is a function of the relative velocity of the mobile, and the angle θ between the direction of motion of the mobile and direction of arrival of the scattered waves.
- Maximum Doppler spread f_D is given by

$$f_D = \frac{vf_c}{c} \qquad (10)$$

Where $v =$ maximum speed between the transmitter and receiver,
f_c= the carrier frequency
c = the speed of light.

- As long as the communication bandwidth B << f_c, the Doppler power spectrum can be treated as approximately constant.

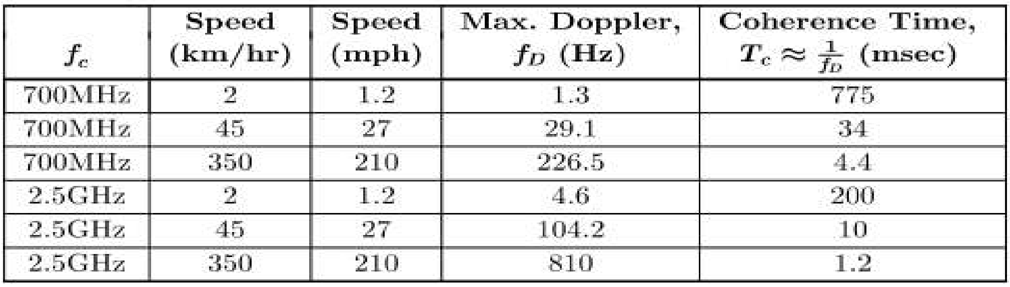
- ***Coherence Time(T_C):***
  - Coherence time Tc is used to characterize the time varying nature of the frequency depressiveness of the channel in the time domain
  - Coherence time is actually a statistical measure of the time duration over which the channel impulse response is essentially invariant, In other words, coherence time is the time duration over which two received signals have a strong potential for amplitude correlation. Mathematically

$$\begin{aligned} |t_1 - t_2| \le T_c &\Rightarrow \mathbf{h}(t_1) \approx \mathbf{h}(t_2) \\ |t_1 - t_2| > t_c &\Rightarrow \mathbf{h}(t_1) \text{ and } \mathbf{h}(t_2) \text{ are uncorrelated} \end{aligned} \qquad (11)$$

  - The coherence time and Doppler spread are also inversely related

$$T_C \approx \frac{1}{f_D} \qquad (12)$$

Values for the Doppler spread and the associated channel coherence time for LTE at Pedestrian, Vehicular, and Maximum Speeds are given in Table below for two possible LTE frequency bands.



f_c	Speed (km/hr)	Speed (mph)	Max. Doppler, f_D (Hz)	Coherence Time, $T_c \approx \frac{1}{f_D}$ (msec)
700MHz	2	1.2	1.3	775
700MHz	45	27	29.1	34
700MHz	350	210	226.5	4.4
2.5GHz	2	1.2	4.6	200
2.5GHz	45	27	104.2	10
2.5GHz	350	210	810	1.2

- *Conclusion:*
  - If the transmitter and receiver are moving fast relative to each other and hence the Doppler is large, the channel will changes its behavior much more quickly than if the transmitter and receiver are stationary.
  - At high frequency and mobility, the channel may change up to 1000 times per second, it results placing a large burden on
    - *Overhead channel and Channel estimation algorithms*
    - *Making the assumption of accurate transmitter channel knowledge questionable.*
    - *Additionally, the large Doppler at high mobility and frequency can also degrade the OFDM subcarrier orthogonally*

2.2.3 Angular Spread and Coherence Distance:

- ***Angular Spread(θ_{rms}):***
  - It refers to the statistical distribution of the angle of the arriving energy.
  - A large θ_r implies that channel energy is coming in from many directions, whereas a small θ_{rms} implies that the received channel energy is more focused.
  - A large angular spread generally occurs when there is a lot of local scattering, and this results in more statistical diversity in the channel.
- ***Coherence Distance(D_C):***
  - The coherence distance is the spatial distance over which the channel does not change appreciably. The dual of angular spread is coherence distance.
  - As the angular spread increases, the coherence distance decreases, and vice versa.
  - An approximate rule of thumb between angular spread and coherence distance is

$$D_C \approx \frac{2\lambda}{\theta_{rms}} \qquad (13)$$

- *Conclusion:*
  - Angular spread and coherence distance are particularly important in multiple antenna (MIMO) systems.
  - The coherence distance gives a rule of thumb for how far antennas should be spaced apart, in order to be statistically independent.
  - If the coherence distance is very small, antenna arrays can be effectively employed to provide rich diversity

2.3 Modelling Broadband Fading Channel :

- Ideally, modeling a channel is calculating all the physical processing effecting a signal from the transmitter to the receiver.
- The two major classes of models are
  1. ***Statistical models****:* These models are simpler, and are useful for analysis and simulations.
  2. ***Empirical models***: These are more complicated but usually represent a specific type of channel more accurately.

2.3.1 Statistical models:

- *Introduction*: These models are used to characterize the amplitude and power of a received signal r(t) when all the reflections arrive at about the same time. This is only true when the symbol time is much greater than the delay spread, i.e., T >>τ_{max} so these models are often said to be valid for "narrowband fading channels.
- Some of the popular statistical models are:
  1. *Rayleigh Fading*
  2. *Ricean Distribution*
  3. *Nakagami-m fading*

1. ***Rayleigh Fading:*** Rayleigh fading is a reasonable model when there are many objects in the environment that scatter the radio signal before it arrives at the receiver. The central limit theorem holds that, if there is sufficiently much scatter, the channel impulse response will be well-modelled as a Gaussian process irrespective of the distribution of the individual components. The envelope of the channel response will be Rayleigh distributed.

   Consider a snapshot value of received signal r(t) at time t = 0, and r(0) = $r_i(0)$ + $r_Q(0)$. Where $r_i(0)$ is in-phase component and $r_Q(0)$ is quadrature components of a Gaussian random variables. The distribution of the envelope amplitude |r| =$\sqrt{r^2{}_i + r^2{}_Q}$is Rayleigh, and the receivedpower $|r| = r^2{}_i + r^2{}_Q$ is exponentially distributed.

Formally

$$f_{|r|}(x) = \frac{2x}{P_r} e^{-x^2/P_r}, \quad x \geq 0,$$

$$f_{|r|^2}(x) = \frac{1}{P_r} e^{-x/P_r}, \quad x \geq 0,$$

Where P_r is the average received power due to shadowing and path loss

- The path loss and shadowing determine the mean received power and the total received power fluctuates around this mean due to the fading.

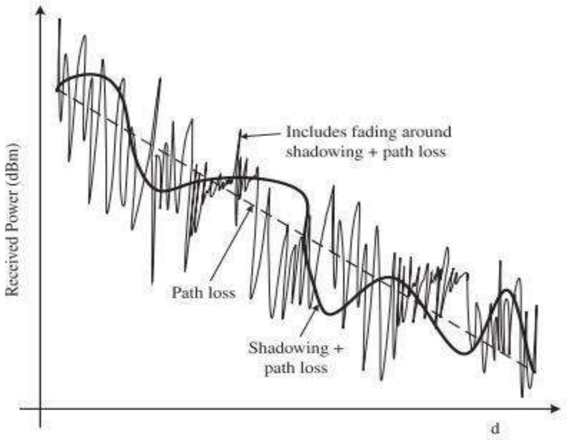


Figure 2.16: The three major channel attenuation factors are shown in terms of their relative spatial (and hence temporal) scales

- The phase of r(t) uniformly distributed from 0 to 2π is defined as

$$\theta_r = \tan^{-1}\left(\frac{r_Q}{r_i}\right) \qquad (14)$$

2. *Ricean Distribution:*

- An important assumption in the Rayleigh fading model is that all the arriving reflections have a mean of zero.

- In Rician fading, a strong dominant component is present for example, a line-of-sight (LOS) path between the transmitter and receiver.
- For a LOS signal, the received envelope distribution is more accurately modelled by a Ricean distribution, which is given by

$$f_{|r|}(x) = \frac{x}{\sigma^2} e^{-(x^2+\mu^2)/2\sigma^2} I_0(\frac{x\mu}{\sigma^2}), \quad x \geq 0, \qquad (15)$$

Where $\mu^2$the power of the LOS component and I_0 is the 0th order

- Ricean distribution reduces to the Rayleigh distribution in the absence of a LOS component
- Since the Ricean distribution depends on the LOS component's powerμ^2, a common way to characterize the channel is by the relative strengths of the LOS and scattered paths. This factor K is quantified as

$$K = \frac{\mu^2}{2\sigma_2} \qquad (16)$$

- The above equation describes how strong the LOS component is relative to the non-LOS (NLOS) components. For K = 0, again the Ricean distribution reduces to Rayleigh, and as K=∞, the physical meaning is that there is only a single LOS path and no other scattering.
- The average received power under Ricean fading is just the combination of the scattering power and the LOS power.
- The Ricean distribution is usually a more accurate depiction of wireless broadband systems, which typically have one or more dominant components.

3. Nakagami-m fading:

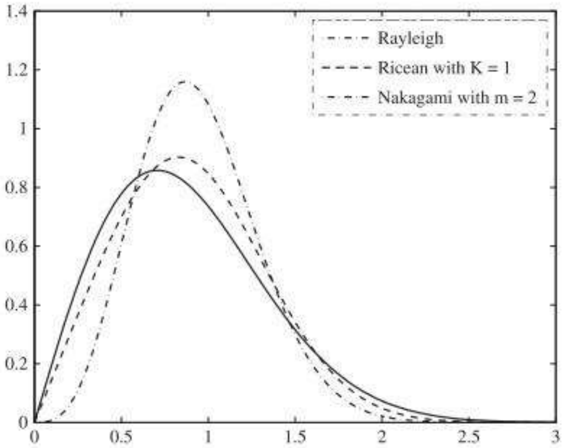
- It is a general model for wireless channel. The probability density function (PDF) of Nakagami fading is parameterized by m and given as

$$f_{|r|}(x) = \frac{2m^m x^{2m-1}}{\Gamma(m) P_r^m} e^{-mx^2/P_r}, \quad m \geq 0.5.$$

- The Nakagami distribution can in many cases be used in tractable analysis of fading channel performance. The power distribution for Nakagami fading is

$$f_{|r|^2}(x) = (\frac{m}{P_r})^m \frac{x^{m-1}}{\Gamma(m)} e^{-mx/P_r}, \quad m \geq 0.5.$$

- Figure below shows comparison of the most popular fading distributions with probability distributions $f_{|r|}(x)$ for Rayleigh, Ricean w/K = 1, and Nakagami with m =2. All have average received power P_r =1.



2.3.2 Statistical Correlation of the Received Signal

- Specific statistical models like Rayleigh, Ricean, and Nakagami-m provided the probability density functions (PDFs) that gave the likelihoods of the received signal envelope and power at a given time instant.
- Use these PDF functions with the channel autocorrelation function, $(\Delta\tau,\ \Delta t)$ in order to understand how the envelope signal r(t) evolves over time, or changes from one frequency or location to another.
- Analysis of statistical correlation of received signal in different domains are
  1. ***Time correlation***
  2. ***Frequency correlation***
  3. ***The Dispersion selectivity duality***
  4. ***Multi-dimensional correlation***

1. *Time correlation:*

- In the time domain, the channel h (τ = 0, t) get one new sample from a Rayleigh distribution for every T_c sec & interpolated with the autocorrelation function of $A_t\,(\Delta t)$.
- The autocorrelation function $A_t\,(\Delta t)$ describes how the channel is correlated in time (see Figure 2.17).
- Its frequency domain Doppler power spectrum $(\Delta f\ )$ provides a band-limited description of the same correlation since it is simply the Fourier transform of $A_t\,(\Delta t)$. In other words, the power spectral density of the channel h(τ = 0, t) should be $\rho_t(\Delta f)$.

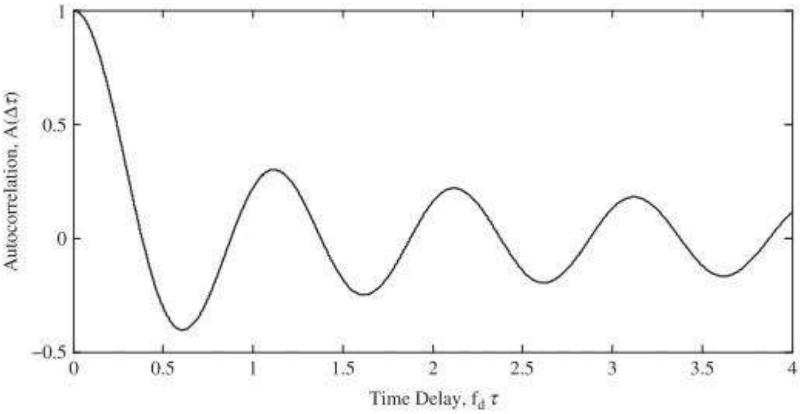


Figure 2.17 Autocorrelation of the signal envelope in time, $A_c(\Delta t)$ which here is normalized by the Doppler f_D. For example, from this figure it can be seen that for Δt = to $0.4/f_D$, which means that after $0.4/f_D$ seconds, the fading value is uncorrelated with the value at time 0.

- For the specific case of uniform scattering, it can been shown that the Doppler power spectrum shown in below equation

$$\rho_t(\Delta f) = \begin{cases} \frac{P_r}{4\pi} \frac{1}{f_D\sqrt{1-\left(\frac{\Delta f}{f_D}\right)^2}}, & |\Delta f| \le f_D \\ 0, & \Delta_f > f_D \end{cases}$$

- A plot of this realization of $(\Delta f\ )$ is shown in Figure 2.18. Which is often used to model the time autocorrelation function A_c ( ), and hence predict the time correlation properties of narrowband fading signals.

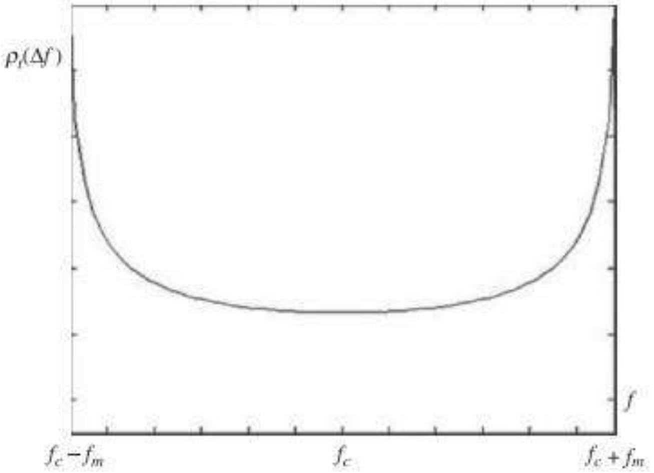


Fig 2.18: The spectral correlation due to Doppler, $(\Delta f\ )$ for uniform scattering

2. *Frequency Correlation*

- Similar to time correlation, fading in frequency is that the channel in the frequency domain, H (f, t = 0), can be thought of as consisting of approximately one new random sample every B_c Hz, with the values in between interpolated.
- The correlated Rayleigh frequency envelope |H (f)| shown in Figure 2.19.

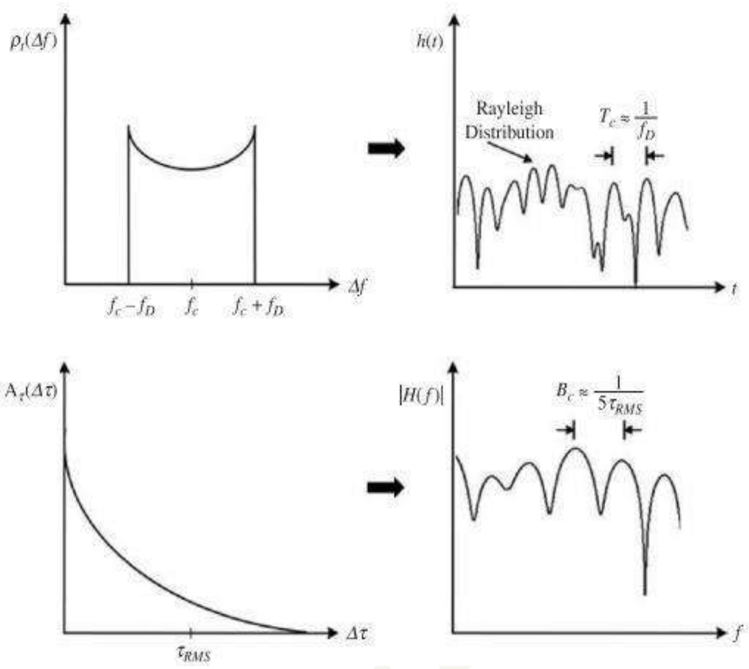


Figure 2.19: The shape of the Doppler power spectrum $(\Delta f\ )$, determines the correlation envelope of the channel in time (top). Similarly, the shape of the Multipath Intensity Profile $(\Delta\tau\ )$, determines the correlation pattern of the channel frequency response (bottom).

- The correlation function that maps from uncorrelated time domain (τ domain) random variables to a correlated frequency response is the Multipath Intensity Profile, $(\Delta\tau\ )$.
- *Conclusion:*
    1. $(\Delta f\ )$ describes the channel time correlation in the frequency domain.
    2. $(\Delta\tau\ )$, describes the channel frequency correlation in the time domain.
    3. The values of |H(f)| are correlated over all frequencies are refer to as "flat fading," i.e., $\tau_{max} \ll T$).

3. *The Dispersion selectivity duality:*

- Selectivity and dispersion are two quite different effects from fading.
- Selectivity means that the signal's received value is changed by the channel over time or frequency.
- Dispersion means that the channel is spread out over time or frequency.

- Selectivity and dispersion are time-frequency duals of each other. This is illustrated in Figure 2.20.

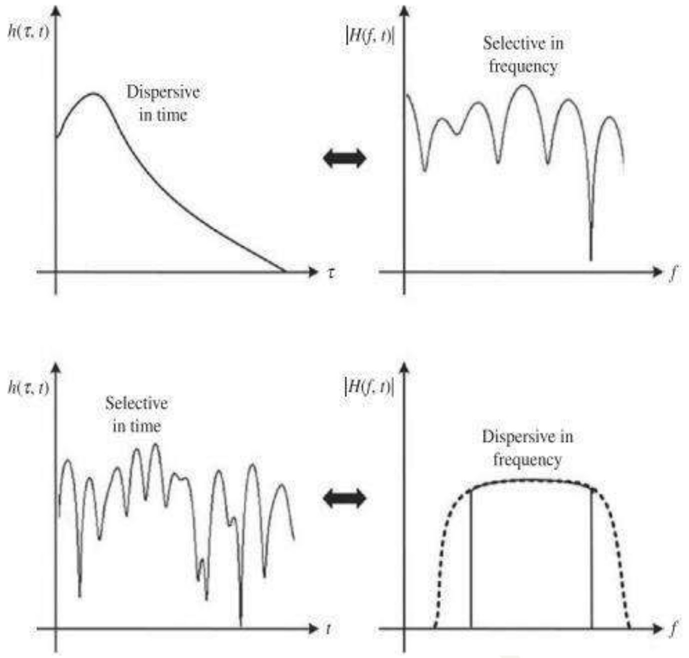


Figure 2.20: The dispersion-selectivity duality: Dispersion in time causes frequency selectivity, while dispersion in frequency causes time selectivity

4. ***Multidimensional Correlation:***

- In reality, signals are correlated in time, frequency, and spatial domains.
- A broadband wireless data system with mobility and multiple antennas is an example of a system where all three types of fading will play a significant role.
- The concept of doubly selective (in time and frequency) fading channels has received recent attention for OFDM.
- Highly frequency-selective channel (long multipath channel) as in a wide area wireless broadband network requires a large number of potentially closely spaced subcarriers to effectively combat the ISI and small coherence bandwidth.
- On the other hand, a highly mobile channel with a large Doppler causes the channel to fluctuate over the resulting long symbol period, which degrades the subcarrier orthogonally.
- In the frequency domain, the Doppler frequency shift can cause significant ISI as the carriers become more closely spaced.
- The mobility and multipath delay spread must reach fairly severe levels before this doubly selective effect becomes significant.

2.3.2 Empirical Channel Models:

- Statistical channel models not considering specific wireless propagation environments.
- Empirical and semi-empirical wireless channel models are the specific models, which have been developed to accurately estimate the path loss, shadowing, and small-scale fast fading.
- Empirical models are based on extensive measurement of various propagation environments, and they specify the parameters and methods for modeling the typical propagation scenarios in different wireless systems.
- These models take into account realistic factors such as angle of arrival (AoA), angle of departure (AoD), antenna array fashion, angle spread (AS), and antenna array gain pattern and other real time factors.
- Different empirical channel models exist for different wireless scenarios, such as sub-urban macro, urban macro, urban micro cells, and so on.
- For channels experienced in different wireless standards, the empirical channel models are also different. Some of the empirical models for LTE as follows

1. LTE Channel Models for Path Loss

- These models are widely used in modeling the outdoor macro and micro cell wireless environments. These are also referred to as "3GPP" channel models
- First, we need to specify the environment where an empirical channel model is used, e.g., suburban macro, urban macro, or urban micro environment.
- The BS to BS distance is typically larger than 3 km for a macro-cell environment and less than 1 km for an urban micro-cell environment.
- For the 3GPP macro-cell environment, the path loss is given by the so-called COST Hata model, which is given by the following easily computable formula

$$\boldsymbol{PL_c}\,[\boldsymbol{dB}] = (\mathbf{44.9} - \mathbf{6.55}\boldsymbol{log_{10}}(\boldsymbol{h_b}))\boldsymbol{log_{10}}(\boldsymbol{d}) + \mathbf{46.3} + \mathbf{33.9}\boldsymbol{log_{10}}(\boldsymbol{f_c})$$
$$- \mathbf{13.82}\boldsymbol{log_{10}}(\boldsymbol{h_b}) - \boldsymbol{a}(\boldsymbol{h_m}) + \boldsymbol{C_o}$$

Where h_b= Base station antenna height

f_c= Carrier frequency in MH

d = Distance between the BS and MS in kilometer

(h_m) = relatively negligible correction function for the mobile height defined as

$$(h_m) = (1.1log_{10}(f_c) - 0.7)\ h_m - 1.56log_{10}(f_c) - 0.8.$$

$$Where\ h_m = mobile\ antenna\ height.$$

- *Conclusion*: COST Hata model is considered to be accurate when d = 100mt to 20 km and $f_c = 1500$ to 2000MHz.

- LTE system also operate with below 1500Mhz, for example 700MHz, the empirical channel model used in such scenarios is the Hata model, which is closely related to the COST Hata model, but with slightly different parameters.
- Several slightly different Hata models exist, depending on whether the environment is urban, suburban, or for open areas. The Hata Model for Urban Areas is:

$$PL_u[\text{dB}] = (44.9 - 6.55\log_{10}(h_b))\log_{10}(d) + 69.55 + 26.16\log_{10}(f_c) - 13.82\log_{10}(h_b) + C_1,$$

where C_1 is a corrective factor that further varies depending on the size of the city, but for a medium or small city is

$$C_1 = 0.8 + (1.1\log_{10}(f_c) - 0.7)h_m - 1.56\log_{10}(f_c)$$

The Hata Model for both Suburban and Open Areas derives from the Urban model. The Suburban path loss is given as

$$PL_s[\text{dB}] = PL_u - 2\left(\log_{10}\left(\frac{f_c}{28}\right)\right)^2 - 5.4$$

while the Open Area Hata Model is

$$PL_o[\text{dB}] = PL_u - 4.78(\log_{10} f_c)^2 + 18.33\log_{10}(f_c) - 40.94$$

2. LTE Channel Models for Multipath

- The received signal at the mobile receiver consists of N time-delayed versions of the transmitted signal. Example as shown in figure 2.21

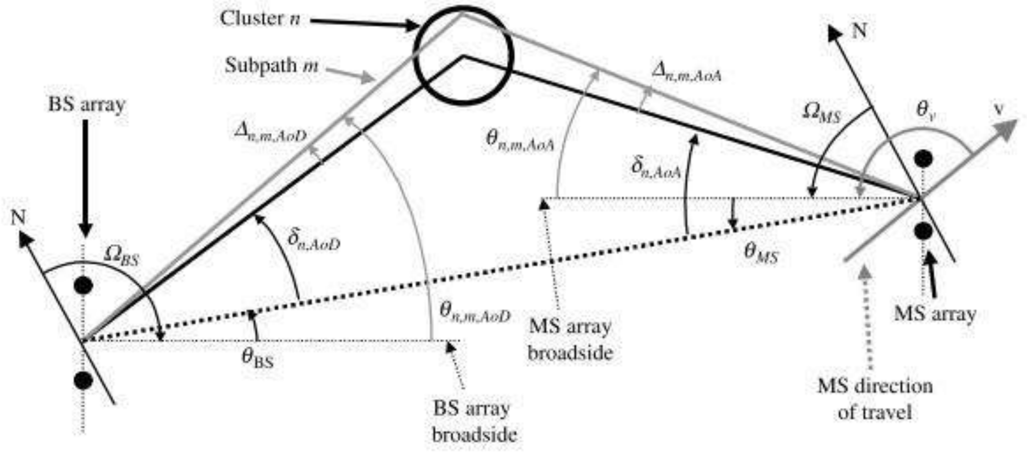


Figure 2.21: 3GPP channel model for MIMO simulations.

- The N paths are characterized by powers and delays that are chosen according to prescribed channel generation procedures, as follows.

i. The number of paths N ranges from 1 to 20 and is dependent on the specific channel models. For example, the 3GPP channel model has N = 6 multipath components. The power distribution normally follows the exponential profile.

ii. Each multipath component further corresponds to a cluster of M subpaths, where each subpath characterizes the incoming signal from a scatter.

iii. The M subpaths have random phases and subpath gains, specified by the given procedure in different stands.

iv. For 3GPP, the phases are random variables uniformly distributed from 0 to 360 degrees, and the subpath gains are given .In the 3GPP channel model, the nth multipath component from the u^th^ transmit antenna to the s^th^ receive antenna, is given as

$$h_{u,s,n}(t) = \sqrt{\frac{P_n \sigma_{SF}}{M}} \sum_{m=1}^{M} \left(\begin{array}{l} \sqrt{G_{BS}(\theta_{n,m,AoD})} \exp\left(j\left[k d_s \sin\left(\theta_{n,m,AoD} + \Phi_{n,m}\right)\right]\right) \\ \times \sqrt{G_{BS}(\theta_{n,m,AoA})} \exp\left(j k d_u \sin\left(\theta_{n,m,AoA}\right)\right) \\ \times \exp\left(j k \|\mathbf{v}\| \cos\left(\theta_{n,m,AoA} - \theta_v\right) t\right) \end{array} \right)$$

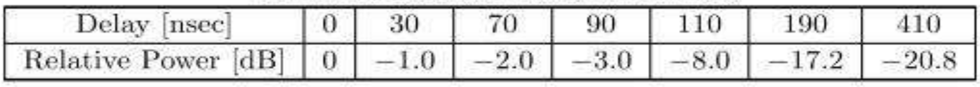
- P_n is the power of the nth path, following exponential distribution.
- σ_{SF} is the lognormal shadow fading, applied as a bulk parameter to the n paths. The shadow fading is determined by the delay spread (DS), angle spread (AS), and shadow fading (SF) parameters, which are correlated random variables generated with specific procedures.
- M is the number of subpaths per path.
- $\theta_{n,m,AoD}$ is the AoD for the mth subpath of the nth path.
- $\theta_{n,m,AoA}$ is the AoA for the mth subpath of the nth path.

- $G_{BS}(\theta_{n,m,AoD})$ is the BS antenna gain of each array element.
- $G_{BS}(\theta_{n,m,AoA})$ is the MS antenna gain of each array element.
- k is the wave number $\frac{2\pi}{\lambda}$ where λ is the carrier wavelength in meters.
- d_s is the distance in meters from BS antenna element s from the reference $(s = 1)$ antenna.
- d_u is the distance in meters from MS antenna element u from the reference $(u = 1)$ antenna.
- $\Phi_{n,m}$ is the phase of the mth subpath of the nth path, uniformly distributed between 0 and 360 degrees.
- $\|\mathbf{v}\|$ is the magnitude of the MS velocity vector, which consists of the velocity of the MS array elements.
- θ_v is the angle of the MS velocity vector.

3. ***LTE Semi-Empirical Channel Models:*** Constructing a fully empirical channel model is relatively time-consuming and computationally intensive due to the huge number of parameters involved. Therefore semi-empirical channel models are provide the accurate inclusion of the practical parameters in a real wireless system, while maintaining the simplicity of statistical channel models. Well-known examples of the simpler multipath channel models include the 3GPP2 Pedestrian A, Pedestrian B, Vehicular A, and Vehicular B models, suited for low-mobility pedestrian mobile users and higher mobility vehicular mobile users. The power delay profile of the channel is determined by the number of multipath taps and the power and delay of each multipath component. Each multipath component is modelled as independent Rayleigh fading with a different power level, and the correlation in the time domain is created according to a Doppler spectrum corresponding to the specified speed. The Pedestrian A is a flat fading model corresponding to a single Rayleigh fading component with a speed of 3 km/hr. Pedestrian B model corresponds to a power delay profile with four paths of delays (0 .11, .19, .41] µs and the power profile given as [1, 0.1071, 0.0120, 0.0052] at 3 km/hr. Vehicular A model, the mobile speed is specified at 30 km/hr. Four multipath components exist, each with delay profile [0, 0.11, 0.19, and 0.41] microseconds and power profile [1, 0.1071, 0.0120, and 0.0052]. For the vehicular B model, the mobile speed is 30 Km/h, with six multipath components, delay profile [0, 0.2, 0.8, 1.2, 2.3, 3.7] microseconds and power profile [1, 0.813, 0.324 0.158, 0.166, 0.004]. These models are often referred to as Ped A/B and Veh A/B.

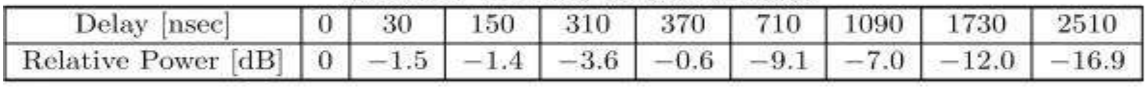
- LTE standard additionally defined extended delay profile with increased multipath resolution known as Extended Pedestrian A, Extended Vehicular A, and Extended Typical Urban. These profiles are given in Tables 2.4, 2.5, and 2.6.

Table 2.4 Extended Pedestrian A Model



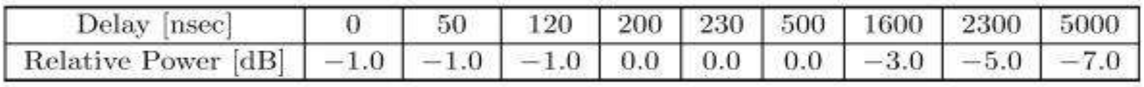
Delay [nsec]	0	30	70	90	110	190	410
Relative Power [dB]	0	−1.0	−2.0	−3.0	−8.0	−17.2	−20.8

Table 2.5 Extended Vehicular A Model



Delay [nsec]	0	30	150	310	370	710	1090	1730	2510
Relative Power [dB]	0	−1.5	−1.4	−3.6	−0.6	−9.1	−7.0	−12.0	−16.9

Table 2.6 Extended Typical Urban Model



Delay [nsec]	0	50	120	200	230	500	1600	2300	5000
Relative Power [dB]	−1.0	−1.0	−1.0	0.0	0.0	0.0	−3.0	−5.0	−7.0

2. 4 Mitigation of N arrow band Fading * * *

2 .4 .1 .The Effects of Unmitigated Fading

- The probability of bit error rate (BER) is the principle metric of interest for the physical layer (PHY) of a communication system.
- For a QAM-based modulation system, the BER in an additive white Gaussian noise (AWGN, no fading) can accurately be approximated by the following relation

$$P_b \ \le 0.2e^{-1.5SNR/(M-1)}$$

If M ≥ 4 is the M-QAM, the probability of error decreases very rapidly (exponentially) with the SNR. Since the channel is constant, the BER is constant over time.

- However, in a fading channel, the BER become a random variable that depends on the instantaneous channel strength and M level modulation, it given as

$$\bar{P}_b = \frac{M}{SNR}$$

  - For fading channel, BER goes down very slowly with SNR, only inversely. This trend is captured plainly in Figure 2.22.

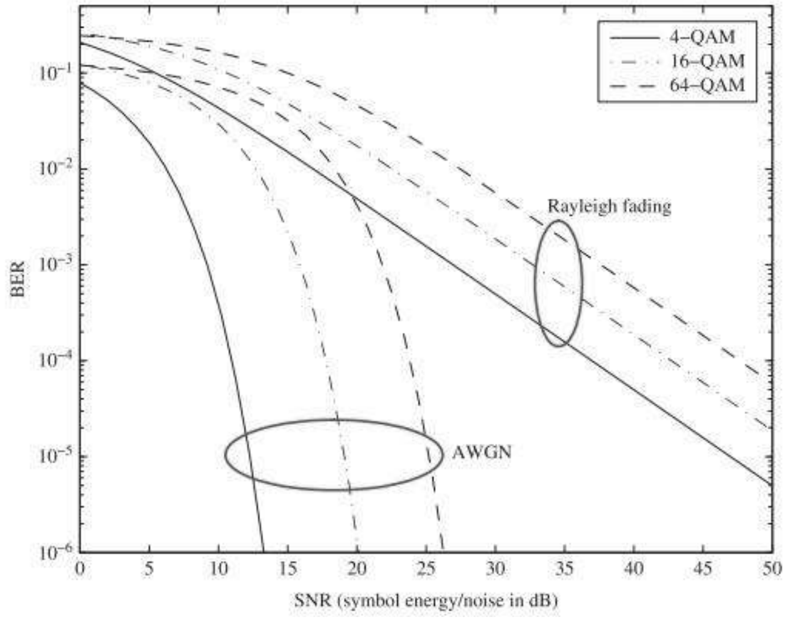


Figure 2.22: Flat fading causes a loss of at least 20-30 dB at reasonable BER values.

- ***Conclusion:***
  - From the figure 2.22, at reasonable system BERs like 10^{-6}, the required SNR is over 30 dB higher in fading. Clearly, it is not desirable, or even possible, to increase the power by over a factor of 1000 to overcome occasional deep fades.

- ***The main techniques for mitigation of narrowband fading are******
  1. *Spatial Diversity*
  2. *Coding and Interleaving*
  3. *Automatic Repeat Request (ARQ)*
  4. *Adaptive Modulation and Coding (AMC)*
  5. *Combining Narrowband Diversity Techniques*

1. *Spatial Diversity:*

- Diversity is the key and potential technique to overcoming the fading problems in wireless channels and to improving PER and BER.
- It is also known as antenna diversity and it usually is achieved by having two or more antennas at the receiver and/or the transmitter (see figure2.24).
- Spatial diversity is a powerful form of diversity, and particularly desirable since it does not necessitate redundancy in time or frequency.
- The simplest form of space diversity consists of two receive antennas, where the stronger of the two signals is selected. As long as the antennas are spaced sufficiently, the two received signals will undergo approximately uncorrelated fading.
- This type of diversity is sensibly called selection diversity, and is illustrated in Fig 2.24.
- More sophisticated forms of spatial diversity include receive antenna arrays (two or more antennas) with maximal ratio combining, transmit diversity using space-time codes, transmit pre-coding, and other combinations of transmit and receive diversity.
- Spatial signaling techniques are so important, the ultimate success of LTE.

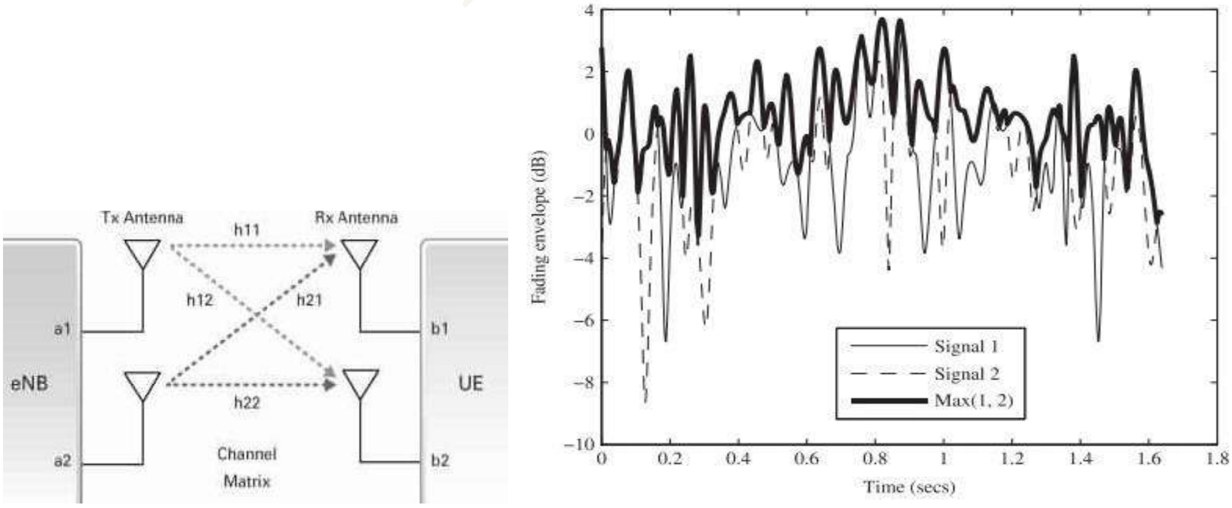


Figure 2.24: Simple two-branch selection diversity eliminates most deep fades.

2. *Coding and Interleaving:***

- Coding and interleaving provides ubiquitous form of diversity for all wireless communication systems.
- Traditionally it is a form of time diversity, in a multicarrier system they also can capture frequency diversity.
- ***Coding***:
  - Usually refer to as channel coding /Error Correction Codes (ECCs), which is also known as forward error correction (FEC).
  - ECCs efficiently introduce redundancy at the transmitter to allow the receiver to recover the input signal even if the received signal is significantly degraded by attenuation, interference, and noise.
  - Coding techniques can be categorized by their coding rate r ≤ 1, which is the inverse of the redundancy added.
  - Code rates are the ratio of information bits to a coding process to the total number of bits created by the coding process. A coding rate of ¼ indicates for each information bit into the coding process there will be 4 bits created for transmission. The higher the code rate, the higher percentage of error detection/correction overhead. Higher the coding rate gives higher transmission reliability gain.
- In Figure 2.25 shows convolutional encoder defined by LTE for use in the Broadcast Channel (BCH).

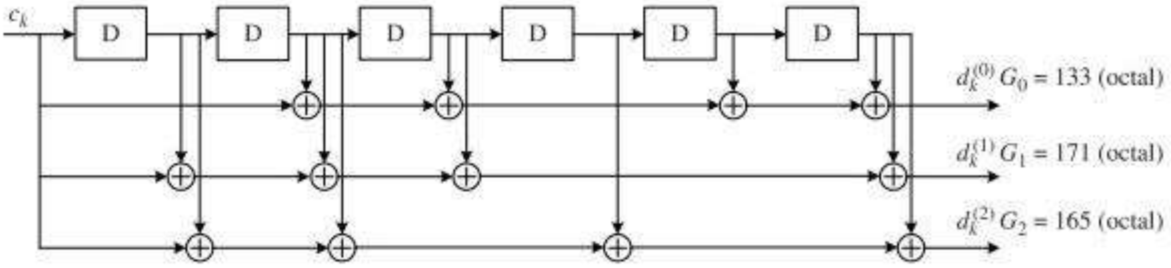

c_k
D
D
D
D
D
D
$d_k^{(0)}$ $G_0 = 133$ (octal)
$d_k^{(1)}$ $G_1 = 171$ (octal)
$d_k^{(2)}$ $G_2 = 165$ (octal)

Figure 2.25: The rate $r = 1/3$ convolutional encoder de fined by LTE for use in the Broadcast Channel (BCH)

- The above figures clearly a rates 1/3 code since there is one input bit (C_k) and 3 outputs (d_k).
- The constraint length of this code is 7; equivalently, there are 6 delay elements or 64 possible states. The most common decoding technique for convolutional codes

- ***Turbo codes***: It class of high-performance forward error correction (FEC) codes. It is sometimes built using two identical convolutional codes of special type, such as, recursive systematic (RSC) type with parallel concatenation. It provide increased resilience to errors through iterative decoding. A rate turbo code is also deployed by LTE as shown in Figure 2.26

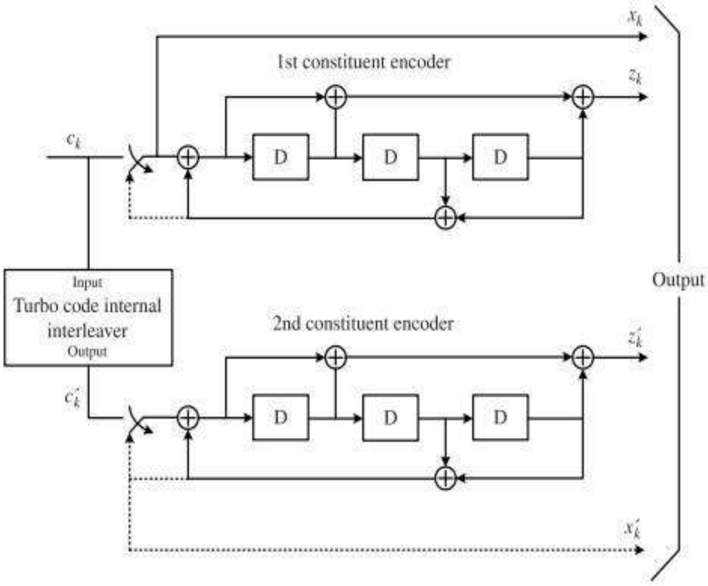


Figure 2.26: The rate parallel concatenated turbo encoder defined by LTE for use in the uplink and downlink shared channels, among others.

- In particular, the encoder is a parallel concatenated convolutional code that comprises an 8-state rate I systematic encoder and an 8-state rate 1 systematic encoder that operates on an interleaved input sequence, for a net coding rate of 1/3.
- By systematic, we mean that one output is generated by a linear modulo-2 sum of the current encoder state that is a function of both the input bit(s) and the previous states (i.e., there is feedback in the state machine), while the other outputs are simply passed through to the output, like X_k in Figure 2.26.
- Codes in LTE can also be punctured, which means that some of the output coded bits are simply dropped, in order to lower the transmission rate.
- For example, if the output of a rate 1/2 convolutional code had a puncturing factor of 1/4 , this means that out of every four output bits, one is dropped. Hence, the effective code rate would become 2/3, since only three coded bits are transmitted for every two information bits. At the decoder, a random or fixed coded his is inserted in the decoding process.
- Puncturing the code to achieve lower the coding rates allows the decoder structure to remain the same regardless of the code rate.

- ***Interleaving:***
  - Interleaving is a process or methodology to make a system more efficient, fast and reliable by arranging data in a noncontiguous manner.
  - Interleaving, a technique for making forward error correction more robust with respect to burst errors.
  - Interleaving is typically used in both convolutional coding and turbo coding. For use with a conventional convolutional code, the interleaver shuffles coded bits to provide robustness to burst errors that can be caused by either bursty noise and interference, or a sustained fade in time or frequency.
  - Interleaving seeks to spread out coded bits so that the effects of a burst error, after de-interleaving, are spread roughly evenly over a frame, or block.
  - For both conventional convolutional codes and turbo codes, the interleaver block size would, from a data reliability standpoint, ideally be quite large.
  - The interleaver block size is usually constrained to be at most over a single packet, and often much less than that. De-interleaving delays have been one of the primary impediments to turbo-coding since they cause considerable latency.
  - Nevertheless, interleaving has proven very effective in allowing ECCs designed for constant, time-invariant additive noise channels to also work well on fading, time-variant noisy channels.

3. *Automatic Repeat Request (ARQ):*

- LTE uses is ARQ (automatic repeat request) and Hybrid-ARQ technique for flow and error control.
- ARQ simply is a MAC layer retransmission protocol that allows erroneous packets to be quickly retransmitted.
- These protocol works in conjunction with PHY layer ECCs and parity checks to ensure reliable links even in hostile channels.
- Since a single bit error causes a packet error, with ARQ the entire packet must be retransmitted even when nearly all of the bits already received were correct, which is clearly inefficient.
- Hybrid-ARQ combines the two concepts of ARQ and FEC to avoid such waste, by combining received packets.
- Hybrid-ARQ, therefore, is able to extract additional time diversity in a fading channel as well.

- In H-ARQ a channel encoder such as a convolution encoder or turbo encoder is used to generate additional redundancy to the information bits.
- However, instead of transmitting all the encoded bits (systematic bits + redundancy bits), only a fraction of the encoded bits are transmitted.
- This is achieved by puncturing some of the encoded bits to create an effective code rate greater than the native code rate of the encoder.
- After transmitting the encoded and punctured bits, the transmitter waits for an acknowledgment from the receiver telling it whether the receiver was able to successfully decode the information bits from the transmission.
- If the receiver was able to decode the information bits, then nothing else needs to be done. If, on the other hand, the receiver was unable to decode the information bits, then the transmitter can resend another copy of the encoded bits.

4. *Adaptive Modulation and Coding (AMC):*

- LTE systems employ adaptive modulation and coding (AMC) in order to take advantage of fluctuations in the channel over time and frequency.
- *The basic idea of AMC*:
    - Transmit as high a data rate as possible when and where the channel is good, and transmit at a lower rate when and where the channel is poor in order to avoid excessive dropped packets.
    - Lower data rates are achieved by using a small constellation such as QPSK and low rate error correcting codes such as rate 1/3 turbo codes.
    - The higher data rates are achieved with large constellations such as 64QAM and less robust error correcting codes.
- To perform AMC, the transmitter must have some knowledge of the instantaneous Channel gain. Once it does, it can choose the modulation technique that will achieve the highest possible data rate while still meeting a BER or packet error rate (PER) requirement.
- An alternative objective is to pick the modulation and/or coding combination that simply maximizes the successful throughput.
- A block diagram of an AMC system is given in Figure 2.27. For simplicity, consider just a single user system attempting to transmit as quickly as possible through a channel with a variable SINR, for example, due to fading.

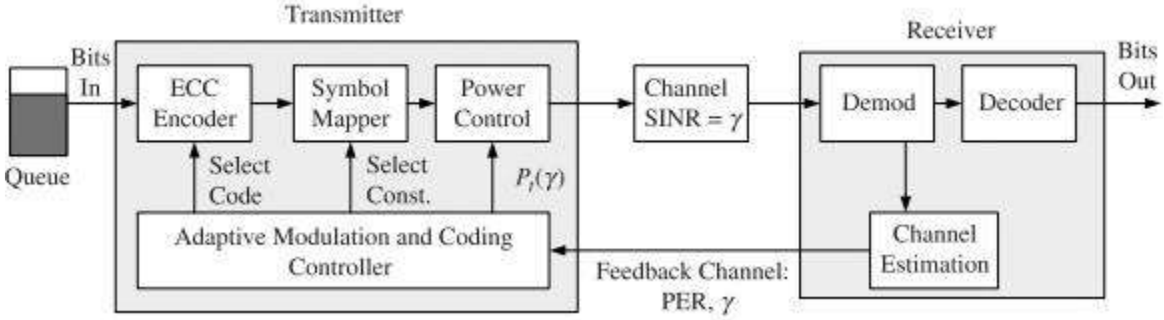


Figure 2.27: Adaptive modulation and coding block diagram

- The goal of the transmitter is to transmit data from its queue as rapidly as possible, subject to the data being demodulated and decoded reliably at the receiver.
- Feedback is critical for adaptive modulation and coding: the transmitter needs to know the "channel SINR".
- ***A Practical Example of AMC***: Figure 2.28 shows a possible realization of AMC, using three different code rates (1/2, 2/3, 3/4), and three different modulation types (QPSK, 16QAM, 64QAM).

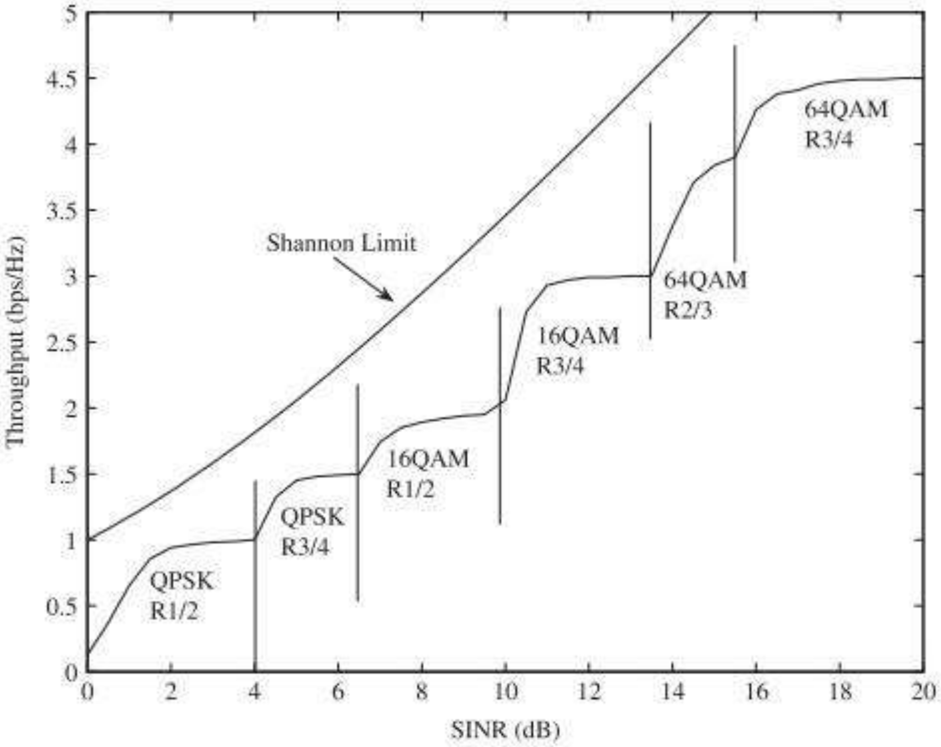


Figure 2.28 Throughput vs. SINR, assuming the best available constellation and coding configuration is chosen for each SINR.

2. 5 Mitigation of Broadband Fading***

- In LTE broadband channel Inter Symbol Interference (ISI) is very serious problem. This is due to frequency-selective fading cause dispersion in time.
- Choosing a technique to effectively combat ISI is a central design decision for any high data rate system.
- OFDM is the most popular choice for combatting ISI in a range of high rate systems.
- Other main techniques for ISI mitigation are
    1. ***Spread Spectrum and RAKE Receivers***
    2. ***Equalization***
    3. ***Multicarrier Modulation: OFDM***
    4. ***Single-Carrier Modulation with Frequency Domain Equalization***

1. Spread Spectrum and RAKE Receivers:

- It is a technique of transmitting of narrowband data signal in a wideband channel called spread spectrum. Spread spectrum techniques are generally broken into two quite different categories:
    1. *Direct Sequence Spread Spectrum(DSSS):* It also known as Code Division Multiple Access (CDMA), is used widely in cellular voice networks and is effective at multiplexing a large number of variable rate users in a cellular environment
    2. *Frequency hopping Spread Spectrum (FHSS)*: Frequency hopping is used in low-rate wireless LANs like Bluetooth, and also for its interference averaging properties in GSM cellular networks.
- Spread spectrum techniques is not an appropriate technology for high data rates due self-interference. In short, spread spectrum is not a natural choice for wireless broadband networks.
- Although this self-interference can be corrected with an equalizer this largely defeats the purpose of using spread spectrum to help with ISI.

2. Equalization

- Equalizers are most logical alternative for ISI-suppression since they don't require additional antennas or bandwidth, and have moderate complexity.
- Equalizers are implemented at the receiver, and attempt to reverse the distortion introduced by the channel.

- Equalizers are broken into two classes: linear and decision-directed (nonlinear).

1. *Linear Equalizers:*
   - A linear equalizer simply runs the received signal through a filter that roughly models the inverse of the channel.
   - The problem with this approach is that it inverts not only the channel, but also the received noise.
   - This noise enhancement can severely degrade the receiver performance, especially in a wireless channel with deep frequency fades.
   - Linear receivers are relatively simple to implement, but achieve poor performance in a time-varying and severe-ISI channel.
2. *Nonlinear Equalizers:*
   - A nonlinear equalizer uses previous symbol decisions made by the receiver to cancel out their subsequent interference, and so is often called a decision feedback equalizers (DFE).
   - One problem with this approach is that it is common to make mistakes about what the prior symbols were (especially at low SNR), which causes "error propagation."
   - Nonlinear equalizers pay for their improved performance relative to linear receivers with sophisticated training and increased computational complexity.

3. Multicarrier Modulation: OFDM:

- Multicarrier modulation is that rather than fighting the time-dispersive ISI channel
- For a large number of subcarriers (L) are used in parallel, so that the symbol time for each goes from T to LT.
- In other words, rather than sending a single signal with data rate R and bandwidth B, why not send L signals at the same time, each having bandwidth B/L and data rate R/L.
- In this way, if $B/L \ll B_c$, each of the signals will undergo approximately flat fading and the time dispersion for each signal will be negligible.
- As long as the number of subcarriers L is large enough, the condition $B/L \ll B_{c,}$ can be met.

4. Single-Carrier Modulation with Frequency Domain Equalization:

- A primary drawback of the OFDM approach has a high peak-to-average ratio (PAR).
- The dynamic range of the transmit power is too large, which results in either significant clipping or distortion, or in a requirement for highly linear power amplifier.

- One can transmit a single carrier signal with a cyclic prefix, which has a low PAR, and then do all the processing at the receiver.
- Said processing consists of a Fast Fourier Transform (FFT) to move the signal into the frequency domain, a 1-tap frequency equalizer (just like in OFDM), and then an Inverse FFT to convert back to the time domain for decoding and detection.
- In addition to eliminating OFDM's PAR problem, an additional advantage of this approach for the uplink is the potential to move the FFT and IFFT operations to the base station.
- In LTE, however, because multiple uplink users share the frequency channel at the same time, the mobile station still must perform FFT and IFFT operations.
- The resulting approach, known in LTE as Single-Carrier Frequency Division Multiple Access (SC-FDMA).

Module – 3

Overview and Channel Structure of LTE:

- Introduction to LTE
- Channel Structure of LTE
- Downlink OFDMA Radio Resource
- Uplink SC-FDMA Radio Resource

Downlink Transport Channel Processing:

- Overview
- Downlink shared Channels
- Downlink Control Channels
- Broadcast channels
- Multicast channels
- Downlink physical channels
- H-ARQ on Downlink

Write the full form of
i). RNC : Radio Network Controller
ii) UTRAN : UMTS Terrestrial Radio Access Network
iii) E-UTRAN: Evolved UMTS Terrestrial Radio Access Network
iv). GERAN: GSM/EDGE Radio Access Network

6.1 Overview of the LTE radio interface:

- The radio interface of a wireless network is the interface between the Mobile Terminal (MT) and the Base Station (BS)
- 3GPP divides the whole LTE network into a radio access network and a core network.
- 3GPP focuses to develop UTRA, i, e 3G RAN developed within 3GPP, and on optimizing 3GPP's overall radio access architecture.
- Another parallel project in 3GPP is the Evolved Packet Core (EPC), which focuses on the Core Network evolution with a flatter all-IP, packet-based architecture.
- The complete packet system consisting of LTE and EPC is called the Evolved Packet System (EPS).
- LTE is also referred to as Evolved UMTS Terrestrial Radio Access (E-UTRA), and the RAN of LTE is also referred to as Evolved UMTS Terrestrial Radio Access Network (E-UTRAN).
- The RAN architectures of UTRAN (3G) and E-UTRAN (LTE) are shown in Figure 6.1

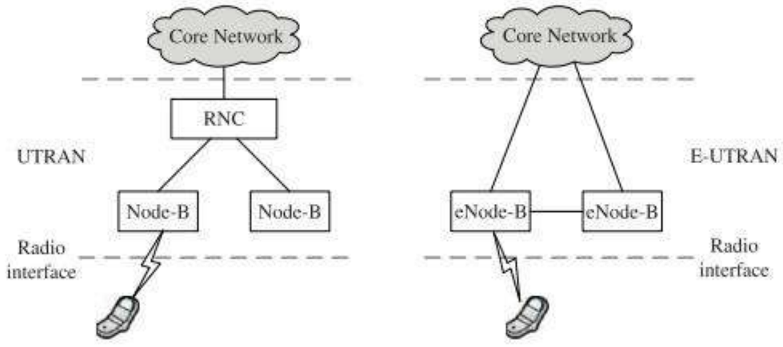


Figure 6.1 Radio interface architectures of UTRAN and E-UTRAN.

- The main architectural difference are, in E-UTRAN (4G) eNode-B is composed of RNC and Node-B of UTRAN (3G) and eNode-Bs are interconnected.
- The eNode-B supports additional features, such as
  1. *Radio resource control*
  2. *Admission control and*
  3. *Mobility management*
- The above three functions were originally performed in the RNC of UTRAN. This simpler structure simplifies the network operation and allows for higher throughput and lower latency over the radio interface.
- The LTE radio interface aims for a long-term evolution, so it is designed with a clean slate approach add-on to UMTS in order to increase throughput of packet switched services.

6.2 Introduction to LTE

- LTE was designed primarily for high-speed data services, which is why LTE is a packet-switched network from end to end and has no support for circuit-switched services.
- The low latency of LTE and its sophisticated quality of service (QoS) architecture allow a network to emulate a circuit-switched connection on top of the packet-switched framework of LTE. For example voice over LTE or VoLTE.

6.1.1 Design Principles of LTE ***

List and briefly explain the design principles of LTE.

- Following are the basic design principles that were agreed upon and followed in 3GPP while designing the LTE specifications. It includes

1. ***Network Architecture***
2. ***Data Rate and Latency***
3. ***Performance Requirements:*** *Spectrum Efficiency, Mobility, Coverage, MBMS service*
4. ***Radio Resource Management***
5. ***Deployment Scenario and Co-existence with 3G***
6. ***Flexibility of Spectrum and Deployment***
7. ***Interoperability with 3G and 2G Networks***

1. ***Network Architecture***:
   - Basically LTE has flat network architecture. It was designed to support purely packet- switched traffic with support for various QoS classes of services.
   - LTE is different by use of clean slate design and supports packet switching for high data rate services from the start.
   - The LTE radio access network, E-UTRAN, was efficiently designed to have the minimum number of interfaces and support for traffic belonging to all the QoS classes such as conversational, streaming, real-time, non-real-time, and background classes.
2. ***Data Rate and Latency***:
   - *Data rate*: The design peak data rate target in LTE for downlink 100 Mbps and uplink 50 Mbps, when operating at the 20MHz channel size.
   - *Latency*: The one-way latency in the user plane is 5 ms in an unloaded network, that is, if only a single UE is present in the cell. For the control-plane latency, the transition time from a camped state to an active state is less than 100 ms, while the transition time between a dormant state and an active state should be less than 50 ms.
3. ***Performance Requirements***:
   - The performance requirements for LTE are specified in terms of

     i. Spectrum efficiency ii. Mobility iii. Coverage. iv. MBMS Service

i. *Spectrum Efficiency:* The average downlink user data rate and spectrum efficiency target is 3 to 4 times that of HSDPA (3G) network. For uplink the average user data rate and spectrum efficiency target is 2 to 3 times that of HSUPA network. The cell edge throughput should be 2 to 3 times that of HSDPA and HSUPA.

ii. *Mobility*: The mobility requirement for LTE is to be able to support mobility at different mobile terminal speeds. Maximum performance at lower mobile speeds of 0 to 15 km/hr. With minor degradation in performance at higher mobile speeds up to 120 km/hr. LTE is also expected to be able to sustain a connection for mobile speeds up to 350 km/hr but with significant degradation in the system performance.

iii. *Coverage*: Good performance should be met up to 5 km. Slight degradation of the user throughput is tolerated cell ranges up to 30 km. Cell ranges up to 100 km should not be precluded by the specifications. The above coverage performance depends on user mobility.

iv. *MBMS Service:* LTE should also provide enhanced support for the Multimedia Broadcast and Multicast Service (MBMS) compared to UTRA (3G) operation.

4. ***Radio Resource Management(RRM)***: RRM requirements cover various aspects such as
   - Enhanced support for end-to-end QoS
   - Efficient support for transmission of higher layers
   - Support for load sharing/balancing and policy management/enforcement across different access technologies.

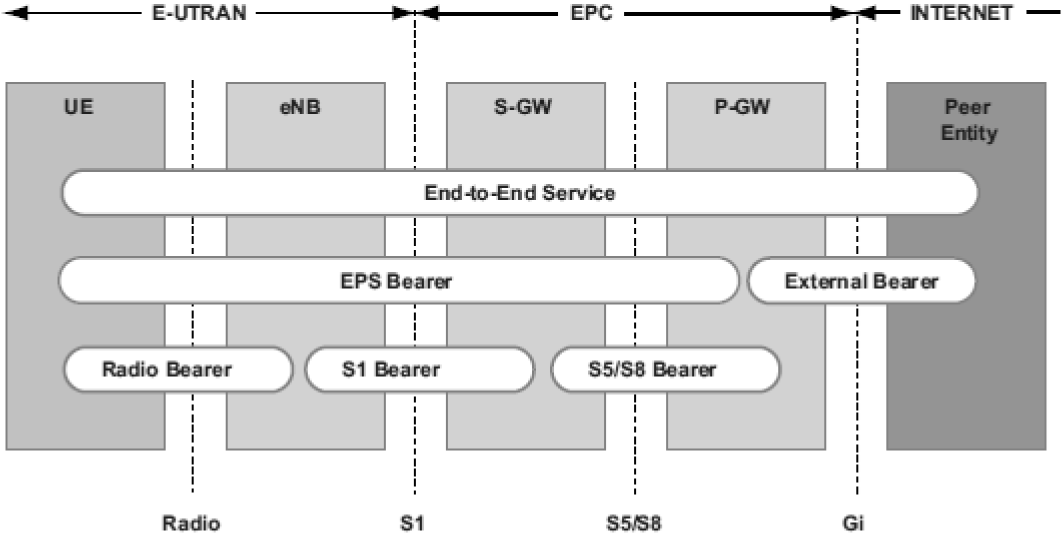
5. ***Deployment Scenario and Co-existence with 3G:*** LTE shall support the following two deployment scenarios:

   i. *Standalone deployment scenario*: where the operator deploys LTE either with no previous network deployed in the area or with no requirement for interworking with 2g and 3g networks.

   ii. *Integrating with existing UTRAN and/or GERAN deployment scenario*: where the operator already has either a UTRAN (3g) and/or a GERAN (2g) network deployed with full or partial coverage in the same geographical area.

6. ***Flexibility of Spectrum and Deployment:***
   - LTE was designed to be operable under a wide variety of spectrum scenarios, including its ability to coexist and share spectrum with existing 3G technologies.
   - LTE was designed to have a scalable bandwidth from 1.4MHz to 20MHz.
   - LTE was designed to operate in both FDD and TDD modes.

7. ***Interoperability with 3G and 2G Networks:***
    - Multimode LTE terminals, which support UTRAN and/or GERAN operation with acceptable terminal complexity and network performance.



6.1.2 Network Architecture***

Build the LTE end-to-end network architecture and explain each components. OR List the components of LTE architecture

- Figure 6.2 shows the end-to-end network architecture of LTE and the various components of the network.
- The entire LTE network is composed of
    - *The radio access network (E-UTRAN) and*
    - *The core network (EPC).*

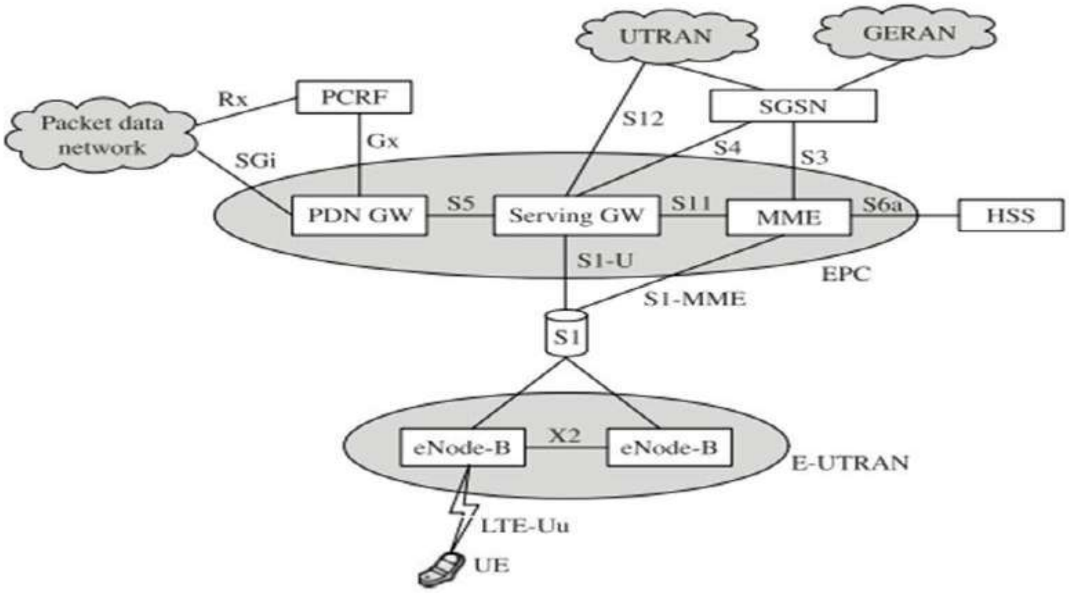


Figure 6.2 LTE end-to-end network architecture.

- The main components of the E-UTRAN and EPC are

1. ***UE (user Equipment)***: It I also called mobile terminal. It is an access device for user. Provides measurements that indicate channel conditions to the network.
2. ***ENode-B:*** It also called the base station. It interface UE to EPC and is the first point of contact for the UE. The eNode-B is the only logical node in the E-UTRAN, so it includes some functions such as

a. *Radio bearer management,*

b. *Uplink and downlink dynamic radio resource management*

c. *Data packet scheduling*

d. *Mobility management.*

3. ***Mobility Management Entity (MME):*** MME is similar in function to the control plane of legacy Serving GPRS Support Node (SGSN). It manages mobility aspects such as gateway selection and tracking area list management.
4. ***Serving Gateway (Serving GW):*** It terminates the interface toward E-UTRAN, and routes data packets between E-UTRAN and EPC. It perform local mobility anchor point for inter-eNode-B handovers and also provides an anchor for inter-3GPP mobility. The Serving GW and the MME may be implemented in one physical node or separate physical nodes. Other responsibilities include
   - *Lawful intercept.*
   - *Charging, and some policy enforcement.*
5. ***Packet Data Network Gateway (PDN GW):*** Following are the responsibilities of PDN GW
   - *It terminates the SGi interface toward the Packet Data Network (PDN).*
   - *It routes data packets between the EPC and the external PDN, and is the key node for policy enforcement and charging data collection.*
   - *It also provides the anchor point for mobility with non-3GPP accesses.*
   - *The external PDN can be any kind of IP network as well as the IP Multimedia Subsystem (IMS) domain.*
   - *The PDN GW and the Serving GW may be implemented in one physical node or separated physical nodes.*
6. ***S1 Interface***: The S1 interface is the interface that separates the E-UTRAN and the EPC. It is split into two parts:
   i. *The SI-U*: It carries traffic data between the eNode-B and the Serving GW.
   ii. *The S1-MME*: It is a signaling-only interface between the eNode-B and the MME.

7. ***X2 Interface:*** The X2 interface is the interface between eNode-Bs. It always exists between eNode-Bs that need to communicate with each other, for example, for support of handover. It consisting of two parts:
   i. *The X2-C:* It is the control plane interface between eNode-Bs.
   ii. *The X2-U:* It is the user plane interface between eNode-Bs.
8. ***Policy and Charging Rules Function (PCRF)****:* It is for policy and charging control.
9. ***Home Subscriber Server (HSS):*** It is responsible for the service authorization and user authentication
10. ***Serving GPRS Support Node (SGSN)*****:** It is for controlling packet sessions and managing the mobility of the UE for GPRS networks.

6.1.3 Radio Interface Protocols**

Classify the LTE radio interface protocols and identify different layers present.

- The LTE radio interface is designed based on a layered protocol stack, which can be divided into ***Control Plane (CP)*** and ***User Plane (UP)*** protocol stacks and is shown in Figure 6.3.

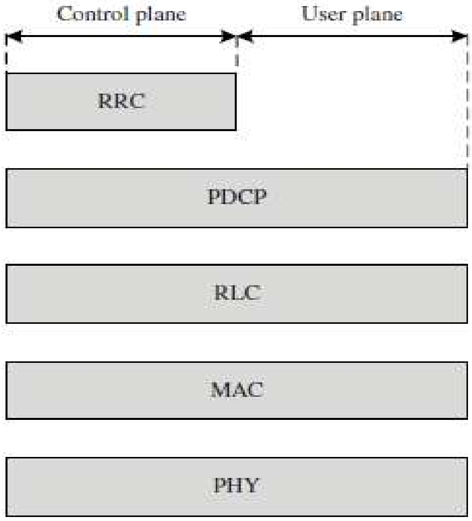


Figure 6.3 The LTE radio interface protocol stack.

- The LTE radio interface protocol is composed of the following layers:

1. ***Radio Resource Control (RRC):*** This layer performs the control plane functions including
   - *Paging*
   - *Maintenance and release of an RRC connection*
   - *security handling*
   - *mobility and QoS management*
2. ***Packet Data Convergence Protocol (PDCP):*** There is only one PDCP entity at the eNode-B and the UE per bearer. The main functions of the PDCP sublayer include
   - *IP packet header compression and decompression based on the Robust Header Compression (ROHC) protocol*
   - *Ciphering of data and signaling*
   - *Integrity protection for signaling*

3. ***Radio Link Control (RLC):*** The main functions of the RLC sublayer are
   - Segmentation and concatenation of data units.
   - Error correction through the Automatic Repeat request (ARQ) protocol.
   - In-sequence delivery of packets to the higher layers.

- It operates in three modes:

  i. ***The Transparent Mode (TM):*** The TM mode is the simplest one, without RLC header addition, data segmentation or concatenation and it is used for specific purposes such as random access.

  ii. ***The Unacknowledged Mode (UM):*** This mode allows the detection of packet loss and provides packet reordering and reassembly, but does not require retransmission of the missing protocol data units (PDUs).

  iii. ***The Acknowledged Mode (AM):*** The AM mode is the most complex one, and it is configured to request retransmission of the missing PDUs in addition to the features supported by the UM mode. There is only one RLC entity at the eNode-B and the UE per bearer.

4. ***Medium Access Control (MAC):*** There is only one MAC entity at the eNode-B and at the UE. The main functions of the MAC sublayer include
   - *Error correction through the Hybrid-ARQ (H-ARQ) mechanism*
   - *Mapping between logical channels and transport channels*
   - *Multiplexing/demultiplexing of RLC PDUs on to transport blocks,*
   - *Priority handling between logical channels of one UE*
   - *Priority handling between UEs by means of dynamic scheduling.*
   - It responsible for transport format selection of scheduled UEs , which includes
     i. Selection of modulation format
     ii. Code rate
     iii. MIMO rank and power level.

With diagram, explain packet flow in the user plane in LTE radio interface protocol stack.

5. ***Physical Layer (PHY):*** The main function of PHY is the actual transmission and reception of data in forms of transport blocks. The PHY is also responsible for various control mechanisms such as
   - *Signaling of H-ARQ feedback*
   - *Signaling of scheduled allocations*
   - *Channel measurements.*

- The packet flow in the user plane is shown in figure below

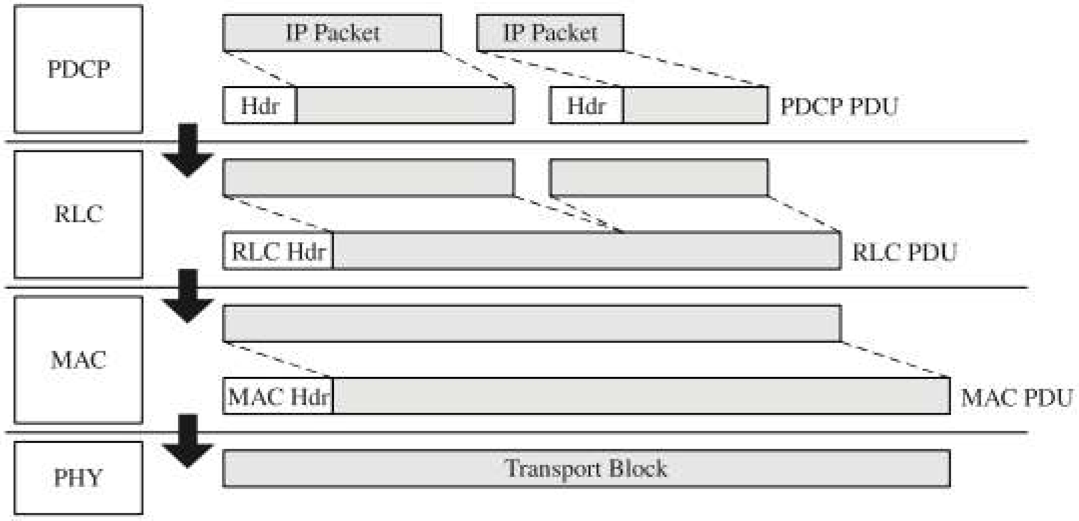
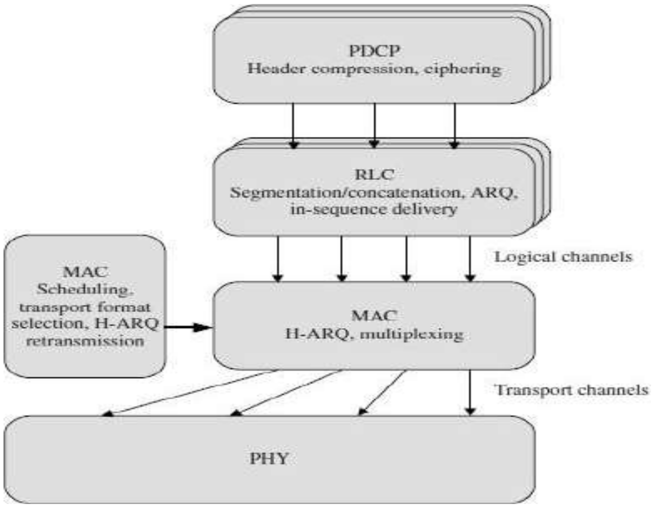


Figure 6.4: The packet flow in the user plane.

6.2 Hierarchical Channel Structure of LTE

- LTE adopts a hierarchical channel structure to efficiently support various QoS classes of services. 1. ***List out the different channel types defined in LTE.*** OR **2.*With the LTE channel structure explain list out the three classes of channels.*3.*Construct the LTE radio interface protocol stack and architecture and SAPs between layers.***
- There are three different channel types defined in LTE ***
  1. ***Logical channels***
  2. ***Transport channels***
  3. ***Physical channels***
- Each channel type associated with a service access point (SAP) between different layers.
- These channels are used by the lower layers of the protocol stack to provide services to the higher layers. The radio interface protocol architecture and the SAPs between different layers are shown in Figure 6.5:

Figure 6.5 The LTE radio interface protocol stack and architecture and the SAPs between different layers between different layers.


PDCP
Header compression, ciphering
RLC
Segmentation/concatenation, ARQ, in-sequence delivery
Logical channels
MAC
Scheduling, transport format selection, H-ARQ retransmission
MAC
H-ARQ, multiplexing
Transport channels
PHY

The radio interface protocol architecture and the SAPs between different layers.

- Logical channels provide services at the SAP between MAC and RLC layers
- Transport channels provide services at the SAP between MAC and PHY layers
- Physical channels are the actual implementation of transport channels over the radio interface.

6.3 LTE Communication Channel*** :

- The information flows between the different protocols layers are known as *channels*. These are used to segregate the different types of data and allow them to be transported across different layers.
- These channels provide interfaces to each layers within the LTE protocol stack and enable an orderly and defined segregation of the data.
- Channels are distinguished based on kind of information they carry and by the way in which the information is processed.
- LTE uses three classes of channels(see fig 6.6):
    1. *Logical channels*: Define ***what type*** of information is transmitted.
    2. *Transport channels:* Define ***how this*** *information* transmitted.
    3. *Physical channels*: Define ***where to*** *send this information.*

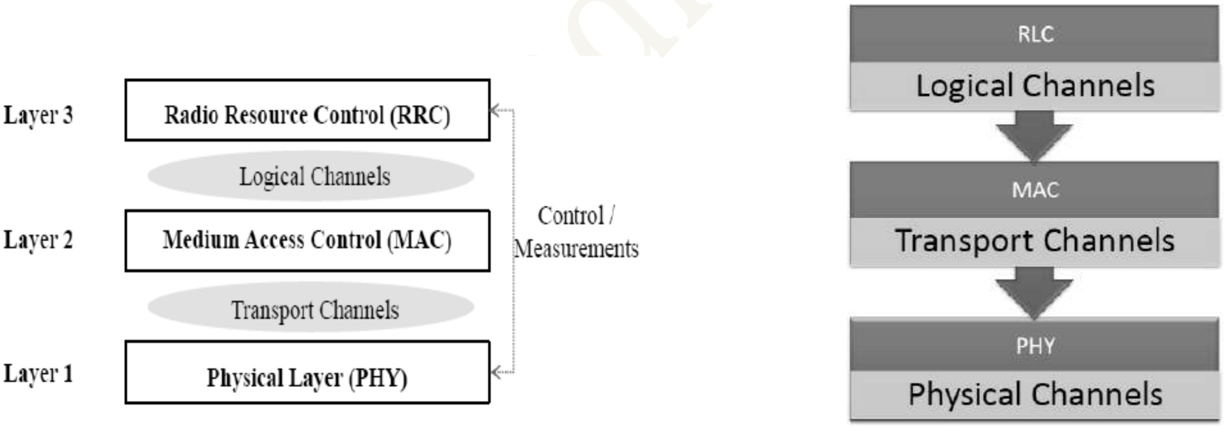


Figure 6.6: LTE channel structure

6.3.1 Logical Channels: Describe Logical channels.

What to Transmit

- Logical channels are used by the MAC to provide services to the RLC.
- Each logical channel is defined based on the type of information it carries.
- In LTE, there are two categories of logical channels depending on the service they provide:
    1. Logical Control Channels*: Which carries the signaling information in control plane*
    2. Logical Traffic Channels*: Which carries the date information in user plane*

1. **The Logical Control Channels (LCC):** which are used to transfer control plane information. Control Channel can be either common channel or dedicated channel. A common channel means common to all users in a cell Point to multipoint while dedicated channels means channels can be used only by one user Point to Point. It include the following types:
   a. ***Broadcast Control Channel (BCCH):*** These channels are used to broadcast system control information to the mobile terminals in the cell, including downlink system bandwidth, antenna configuration, and reference signal power. Due to the large amount of information carried on the BCCH, it is mapped to two different transport channels: the Broadcast Channel (BCH) and the Downlink Shared Channel (DL-SCH).
   b. ***Multicast Control Channel (MCCH):*** A point-to-multipoint downlink channel used for transmitting control information to UEs in the cell. It is only used by UEs that receive multicast/broadcast services.
   c. ***Paging Control Channel (PCCH):*** A downlink channel that transfers paging information to registered UEs in the cell, for example, in case of a mobile-terminated communication session.
   d. ***Common Control Channel (CCCH):*** A bi-directional channel for transmitting control information between the network and UEs when no RRC connection is available, implying the UE is not attached to the network such as in the idle state. Most commonly the CCCH is used during the random access procedure.
   e. ***Dedicated Control Channel (DCCH):*** A point-to-point, bi-directional channel that transmits dedicated control information between a UE and the network. This channel is used when the RRC connection is available, that is, the UE is attached to the network.

- The logical traffic channels, which are to transfer user plane information, include:
  a. ***Dedicated Traffic Channel (DTCH):*** A point-to-point, bi-directional channel used between a given UE and the network. It can exist in both uplink and downlink.
  b. ***Multicast Traffic Channel (MTCH):*** A unidirectional, point-to-multipoint data channel that transmits traffic data from the network to UEs. It is associated with the multicast/broadcast service.

6.3.2 Transport Channels: *Explain the classification of Transport channels briefly*

How to Transmit

- The transport channels are used by the PHY to offer services to the MAC.
- These channel is basically characterized by how and with what characteristics data is transferred over the radio interface, that is, the *channel coding scheme, the modulation scheme, and antenna mapping*.
- Transport channels are classified in to
  1. *Downlink Transport Channels*

2. *Uplink Transport Channels*

1. ***Downlink Transport Channels***

a. Downlink Shared Channel (DL-SCH):

- These channel are used for transmitting the downlink data, including both control and traffic data, and thus it is associated with both logical control and logical traffic channels.
- It supports H-ARQ, dynamic link adaption, dynamic and semi-persistent resource allocation, UE discontinuous reception, and multicast/broadcast transmission.
- By sharing the radio resource among different UEs the DL-SCH is able to maximize the throughput by allocating the resources to the optimum UEs.

b. Broadcast Channel (BCH):

- A downlink channel associated with the BCCH logical channel and is used to broadcast system information over the entire coverage area of the cell.
- It has a fixed transport format defined by the specifications.

c. Multicast Channel (MCH):

- These channels are associated with MCCH and MTCH logical channels for the multicast/broadcast service.
- It supports Multicast/Broadcast Single Frequency Network (MBSFN) transmission, which transmits the same information on the same radio resource from multiple synchronized base stations to multiple UEs.

d. Paging Channel (PCH):

- These are associated with the PCCH logical channel.
- It is mapped to dynamically allocate physical resources, and is required for broadcast over the entire cell coverage area.
- It is transmitted on the Physical Downlink Shared Channel (PDSCH), and supports UE discontinuous reception.

2. **Uplink Transport Channels**

a. Uplink Shared Channel (UL-SCH):

- It can be associated to CCCH, DCCH, and DTCH logical channels.
- It supports H-ARQ, dynamic link adaption, and dynamic and semi-persistent resource allocation.

b. Random Access Channel (RACH):

- A specific transport channel that is not mapped to any logical channel.
- It transmits relatively small amounts of data for initial access or, in the case of RRC, state changes.

- The data on each transport channel is organized into transport blocks.
- The transmission time of each transport block, also called Transmission Time Interval (TTI).

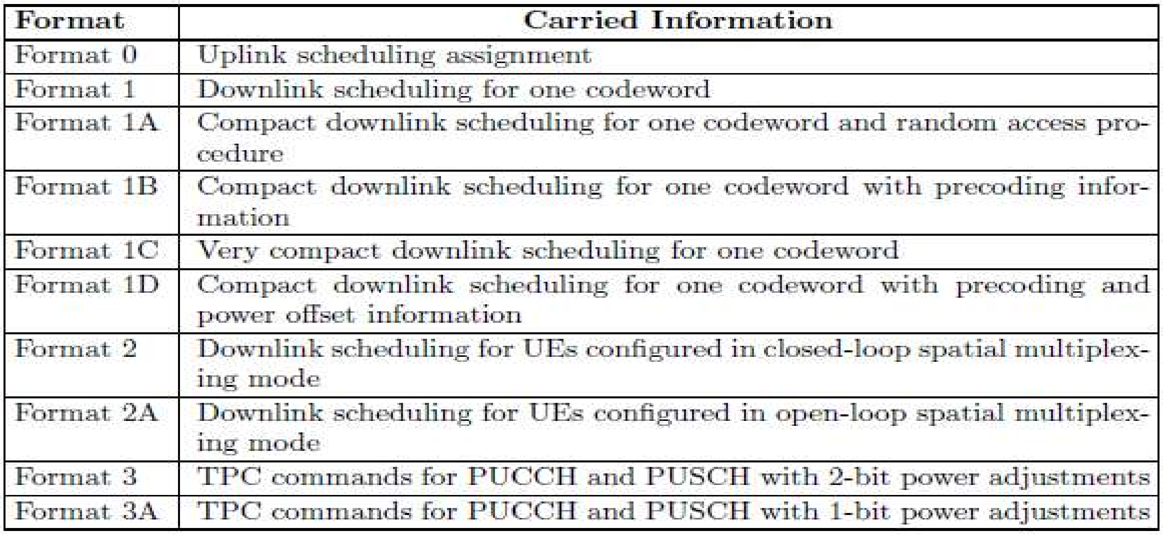
- In LTE TTI is 1 ms. TTI is also the minimum interval for link adaptation and scheduling decision.
- Without spatial multiplexing, at most one transport block is transmitted to a UE in each TTI; with spatial multiplexing, up to two transport blocks can be transmitted in each TTI to a UE.
- Besides transport channels, there are different types of control information defined in the MAC layer, which are important for various physical layer procedures. The defined control information includes

1. ***Downlink Control Information (DCI):***

○ It carries information related to down-link/uplink scheduling assignment, modulation and coding scheme, and Transmit Power Control (TPC) command, and is sent over the Physical Downlink Control Channel (PDCCH).

○ The DCI supports 10 different formats, listed in Table 6.1.

Table 6.1 DCI Formats



Format	Carried Information
Format 0	Uplink scheduling assignment
Format 1	Downlink scheduling for one codeword
Format 1A	Compact downlink scheduling for one codeword and random access procedure
Format 1B	Compact downlink scheduling for one codeword with precoding information
Format 1C	Very compact downlink scheduling for one codeword
Format 1D	Compact downlink scheduling for one codeword with precoding and power offset information
Format 2	Downlink scheduling for UEs configured in closed-loop spatial multiplexing mode
Format 2A	Downlink scheduling for UEs configured in open-loop spatial multiplexing mode
Format 3	TPC commands for PUCCH and PUSCH with 2-bit power adjustments
Format 3A	TPC commands for PUCCH and PUSCH with 1-bit power adjustments

2. ***Control Format Indicator (CFI):***

○ It indicates how many symbols the DCI spans in that subframe.

○ It takes values CFI = 1, 2, or 3, and is sent over the Physical Control Format Indicator Channel (PCFICH).

3. ***H-ARQ Indicator (HI):***

○ It carries H-ARQ acknowledgment in response to uplink transmissions, and is sent over the Physical Hybrid ARQ Indicator Channel (PHICH).

○ HI = 1 for a positive acknowledgment (ACK) and HI = 0 for a negative acknowledgment (NAK).

4. ***Uplink Control Information (UCI):***

○ It is for measurement indication on the downlink transmission, scheduling request of uplink, and the H-ARQ acknowledgment of downlink transmissions.

○ The UCI can be transmitted either on the Physical Uplink Control Channel (PUCCH) or the Physical Uplink Shared Channel (PUSCH).

6.3.3 Physical Channels:

6.3.4 Explain special added channel feature of physical channels

Actual Transmission

- Each physical channel corresponds to a set of resource elements in the time-frequency grid that carry information from higher layers.
- The basic entities that make a physical channel are resource elements and resource blocks.
- Physical channels are classified into
  1. *Downlink Physical Channels*
  2. *Uplink Physical Channels*

1. Downlink Physical Channels

a. ***Physical Downlink Control Channel (PDCCH):***
   - It carries information about the transport format and resource allocation related to the DL-SCH and PCH transport channels, and the H-ARQ information related to the DL-SCH.
   - It also informs the UE about the transport format, resource allocation, and H-ARQ information related to UL-SCH. It is mapped from the DCI transport channel.

b. ***Physical Downlink Shared Channel (PDSCH):***
   - This channel carries user data and higher-layer signaling. It is associated to DL-SCH.

c. ***Physical Broadcast Channel (PBCH):***
   - It corresponds to the BCH transport channel and carries system information.

d. ***Physical Multicast Channel (PMCH):***
   - It carriers multicast/broadcast information for the MBMS service.

e. ***Physical Hybrid-ARQ Indicator Channel (PHICH):***
   - This channel carries H-ARQ ACK/NAKs associated with uplink data transmissions. It is mapped from the HI transport channel.

f. ***Physical Control Format Indicator Channel (PCFICH):***
   - It informs the UE about the number of OFDM symbols used for the PDCCH. It is mapped from the CFI transport channel.

2. Uplink Physical Channels

a. ***Physical Uplink Control Channel (PUCCH):***
   - It carries uplink control information including Channel Quality Indicators (CQI), ACK /NAKs for H-ARQ in response to downlink transmission, and uplink scheduling requests.

b. ***Physical Uplink Shared Channel (PUSCH):***
   - It carries user data and higher layer signaling. It corresponds to the UL-SCH transport channel.

c. ***Physical Random Access Channel (PRACH):***

  - This channel carries the random access preamble sent by UEs.

- Besides physical channels, there are signals embedded in the downlink and uplink physical layer, which do not carry information from higher layers. The physical signals defined in the LTE specifications are
  1. ***Reference signal***: It is defined in both downlink and uplink for channel estimation that enables coherent demodulation and for channel quality measurement to assist user scheduling.
  2. ***Synchronization signal***: It is split into a primary and a secondary synchronization signal, and is only defined in the downlink to enable acquisition of symbol timing and the precise frequency of the downlink signal

6.3.5 Channel Mapping

Briefly explain channel mapping with the neat diagram

- These all three types of channel are present in Downlink as well as Uplink direction. Mapping of these channels is shown in below pictures.
- Need to exist a good correlation based on the purpose and the content between channels in different layers. This is achieved by
  1. Mapping between the logical channels and transport channels at the MAC SAP.
  2. Mapping between transport channels and physical channels at the PHY SAP.
- The allowed mapping between different channel types is shown in Figure 6.6 and mapping between control information and physical channels is shown in Figure 6.7.

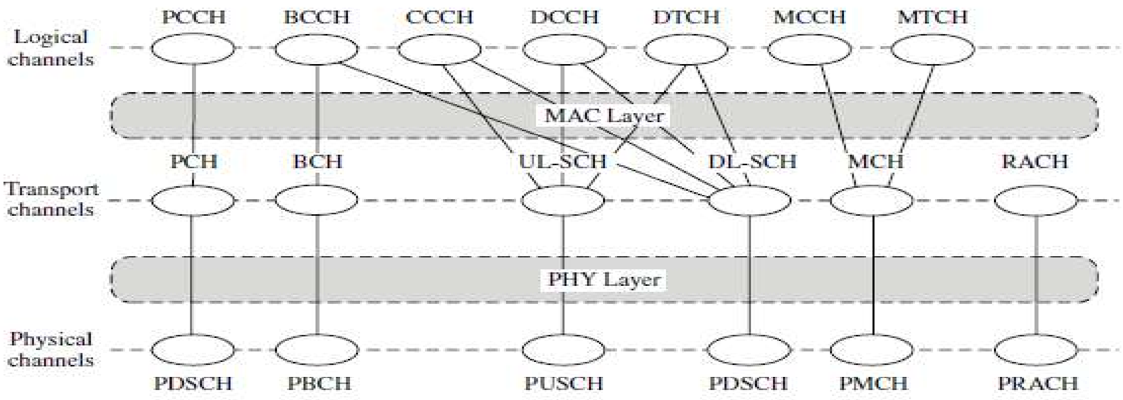


Figure 6.6 Mapping between different channel types.

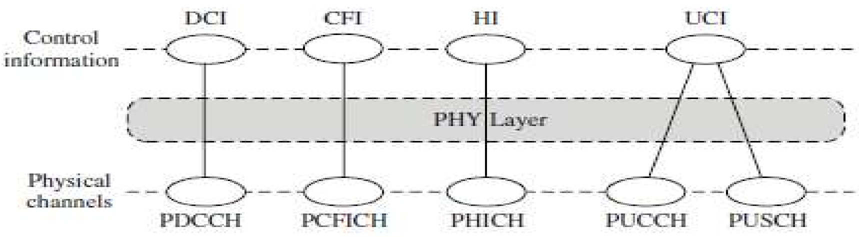


Figure 6.7 Mapping of control information to physical channels.

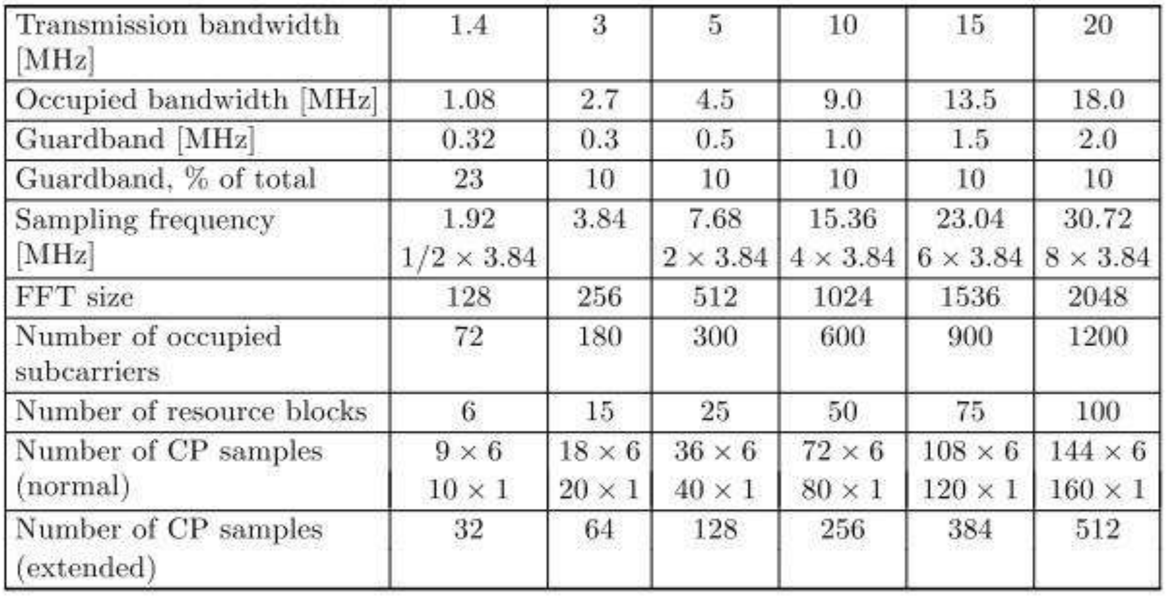
6.4 Downlink OFDMA Radio Resources***

- In LTE, the downlink and uplink use different transmission schemes due to different considerations.
- The multiple access in the downlink is based on OFDMA. In each TTI, a scheduling decision is made where each scheduled UE is assigned a certain amount of radio resources in the time and frequency domain.
- The radio resources allocated to different UEs are orthogonal to each other, which means there is no intra-cell interference
- The following describes the frame structure and the radio resource block structure in the downlink, as well as the basic principles of resource allocation and the supported MIMO modes.

6.4.1 Frame Structure:

- Frames are the common time domain elements shared by both downlink and uplink in LTE.
- Typical parameters used in LTE specification for down link as shown in table 6.2

Table 6.2 Typical Parameters for Downlink Transmission



Transmission bandwidth [MHz]	1.4	3	5	10	15	20
Occupied bandwidth [MHz]	1.08	2.7	4.5	9.0	13.5	18.0
Guardband [MHz]	0.32	0.3	0.5	1.0	1.5	2.0
Guardband, % of total	23	10	10	10	10	10
Sampling frequency [MHz]	1.92 1/2 × 3.84	3.84	7.68 2 × 3.84	15.36 4 × 3.84	23.04 6 × 3.84	30.72 8 × 3.84
FFT size	128	256	512	1024	1536	2048
Number of occupied subcarriers	72	180	300	600	900	1200
Number of resource blocks	6	15	25	50	75	100
Number of CP samples (normal)	9 × 6 10 × 1	18 × 6 20 × 1	36 × 6 40 × 1	72 × 6 80 × 1	108 × 6 120 × 1	144 × 6 160 × 1
Number of CP samples (extended)	32	64	128	256	384	512

- T_s is the basic time unit for LTE. T_s can be regarded as the sampling time of an FFT-based OFDM transmitter/receiver implementation with FFT size N_{FFT} = 2048.
- As the normal subcarrier spacing is defined to be $\Delta f = 15kHz$
- T_s is defined as $T_s = \frac{1}{(\Delta f \text{ x } N_{FFT})} = \frac{1}{(15000 \text{ x } 2048)}$ seconds or about 32.6 nanoseconds.
- Downlink and uplink transmissions are organized into frames of duration

  T_f = 307200 x T_s = 10ms
- The 10 ms frames divide into 10 subframes. Each subframe divides into 2 slots of 0.5 ms.
- For flexibility, LTE supports both FDD and TDD modes, but most of the design parameters are common to FDD and TDD in order to reduce the terminal complexity.

- LTE supports two kinds of frame structures:
  1. *Frame structure type 1*: It is for the FDD mode.
  2. *Frame structure type 2*: It is for the TDD mode.

1. **Frame Structure Type 1** ***Describe the Frame structure type 1 of Downlink OFDMA radio resources with relevant neat diagram***

   Frame structure type 1 is applicable to both full duplex and half duplex FDD**.**

- There are three different kinds of units specified for this frame structure, illustrated in Fig 6.8.

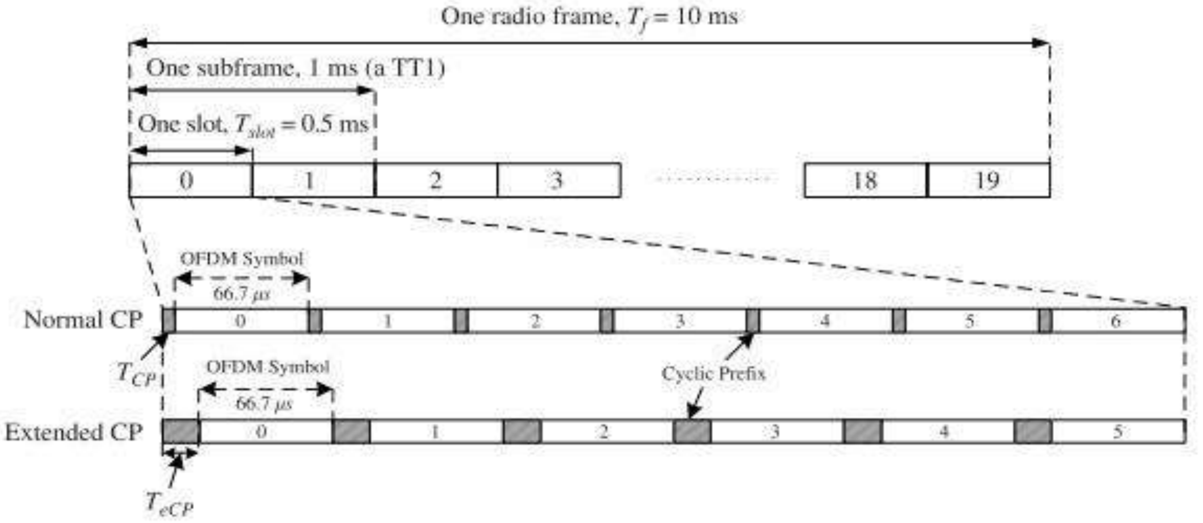


Figure 6.8 Frame structure type 1. For the normal CP, $T_{CP} = 160 \cdot T_s \approx 5.2\mu s$ for the first OFDM symbol, and $T_{CP} = 144 \cdot T_s \approx 4.7\mu s$ for the remaining OFDM symbols, which together fill the entire slot of 0.5 ms. For the extended CP, $T_{eCP} = 512 \cdot T_s \approx 16.7\mu s$.

- ***Description of the frame:***
  - *The smallest time unit is called a "slot" of length* $T_{slot} = 15360 \times T_s = 0.5ms$.
  - *Two consecutive slots are defined as a "subframe" of length* $1ms$.
  - *Ten subframes or 20 slots, numbered from 0 to 19, constitute a one radio frame of 10 ms.*
  - *Channel-dependent scheduling and link adaptation operate on a subframe level.*
  - *The subframe duration corresponds to the minimum downlink TTI, which is of 1 ms duration, compared to a 2 ms TTI for the UMTS (3G).*
  - *A shorter TTI is for fast link adaptation and is able to reduce delay and better exploit the time-varying channel through channel-dependent scheduling.*
  - *Each slot carries a number of OFDM symbols including Cyclic prefix (CP). With subcarrier spacing* $\Delta f = 15kHz$ *, OFDM symbol time is* $\frac{1}{\Delta f} \approx 66.7\mu s$.

  - LTE defines two different CP lengths (see Fig 6.8):
    1. **Normal CP:**
       - *It corresponds to seven OFDM symbols per slot.*

- *The normal CP is suitable for urban environment and high data rate applications.*

- *The normal CP lengths are different for the first ($T_{CP} = 160 \times T_s \approx 5.2\mu s$) and subsequent OFDM symbols $T_{CP} = 144 \times T_s \approx 4.7\mu s$) which is to fill the entire slot of 0.5 ms.*
- *The numbers of CP samples for different bandwidths are shown in Table 6.2. For example, with 10MHz bandwidth, the sampling time is 1/(15000 x 1024) sec*

2. **Extended CP:**

- *It corresponding to six OFDM symbols per slot.*
- *The extended CP is for multicell multicast/broadcast and very-large-cell scenarios with large delay spread at a price of bandwidth efficiency.*
- *The extended CP lengths $T_{eCP} = 512 \times T_s \approx 16.7\mu s$.*
- *The number of CP samples for the extended CP is 256, which provides the required CP length of 256/ (15000 x 1024) =* $1.67\mu s$.
- *In case of 7.5 kHz subcarrier spacing, there is only a single CP length, corresponding to 3 OFDM symbols per slot.*

2. **Frame Structure Type 2**

Build the Frame structure type 2 of Downlink OFDMA radio resources and explain special sub frames

- Frame structure type 2 is applicable to the TDD mode. Type 2 structure shown in fig 6.9
- It is designed for coexistence with legacy systems such as the 3GPP TD-SCDMA-based standard.

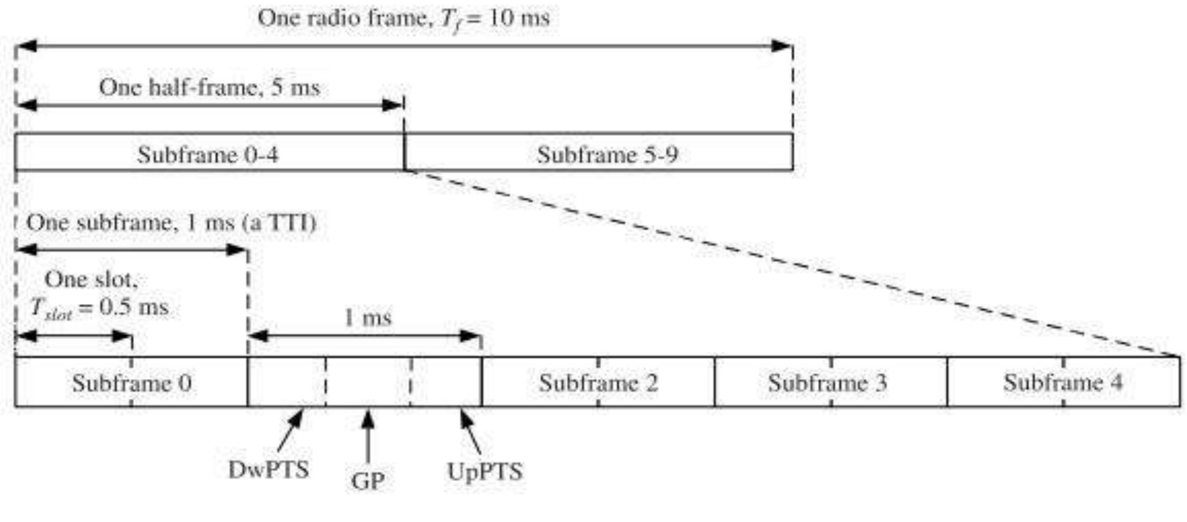
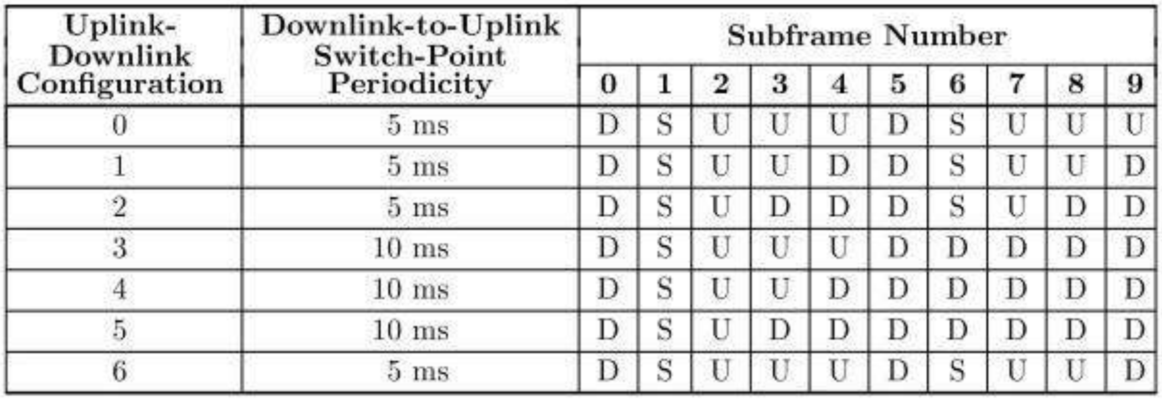


Figure 6.9 Frame structure type 2.

- ***Description of the frame type 2:***
- *Frame structure type 2 is of length $T_f = 30720 \times T_s = 10ms$.*
- *Each frame consists of two half-frames of length 5 ms each.*
- *Each half-frame is divided into five subframes with 1 ms duration.*

- *There are special subframes, which consist of three fields:*
  1. ***Downlink Pilot TimeSlot (DwPTS):** It is a shorter downlink subframe for downlink data transmission. Its length can be varied from three up to twelve OFDM symbols.*
  2. ***Uplink Pilot TimeSlot (UpPTS):** This is the uplink part of the special subframe, and has a short duration with one or two OFDM symbols. It can be used for transmission of uplink sounding reference signals and random access preambles.*
  3. ***Guard Period (GP):** GP field used to provide the guard period for the downlink-to-uplink and the uplink-to-downlink switch.*
- The total length of these three special fields has a constraint of 1 ms.
- LTE supports a guard period ranging from two to ten OFDM symbols, sufficient for cell size up to and beyond 100 km.
- All other subframes are defined as two slots, each with length $T_{slot} = 0.5\,ms$.
- Uplink and down link configurations are Illustrated in Table 6.3. where "D" and "U" denote subframes reserved for downlink and uplink, respectively, and "S" denotes the special subframe.
- In the case of 5 ms switch-point periodicity, the special subframe exists in both half-frames, and the structure of the second half-frame is the same as the first one depicted in Figure 6.9.
- In the case of 10 ms switch-point periodicity, the special subframe exists in the first half-frame only.
- Subframes 0, 5, and the field DwPTS are always reserved for downlink transmission, while UpPTS and the subframe immediately following the special subframe are always reserved for uplink transmission.
- Table 6.3 Uplink-Downlink Configurations for the LTE TDD Mode



Uplink-Downlink Configuration	Downlink-to-Uplink Switch-Point Periodicity	Subframe Number									
		0	1	2	3	4	5	6	7	8	9
0	5 ms	D	S	U	U	U	D	S	U	U	U
1	5 ms	D	S	U	U	D	D	S	U	U	D
2	5 ms	D	S	U	D	D	D	S	U	D	D
3	10 ms	D	S	U	U	U	D	D	D	D	D
4	10 ms	D	S	U	U	D	D	D	D	D	D
5	10 ms	D	S	U	D	D	D	D	D	D	D
6	5 ms	D	S	U	U	U	D	S	U	U	D

6.3.2 Physical Resource Blocks for OFDMA

Draw the structure of downlink resource grid and explain the parameters of resource grid.

- The physical resource in the downlink in each slot is described by a time-frequency grid, called *a" resource grid"*, as illustrated in Figure 6.10.
- Each column and each row of the resource grid correspond to one OFDM symbol and one OFDM subcarrier, respectively.
- The duration of the resource grid in the time domain corresponds to one slot in a radio frame.
- The smallest time-frequency unit in a resource grid is denoted as a "*resource element*"
- Each resource grid consists of a number of "*resource blocks*", which describe the mapping of certain physical channels to resource elements.

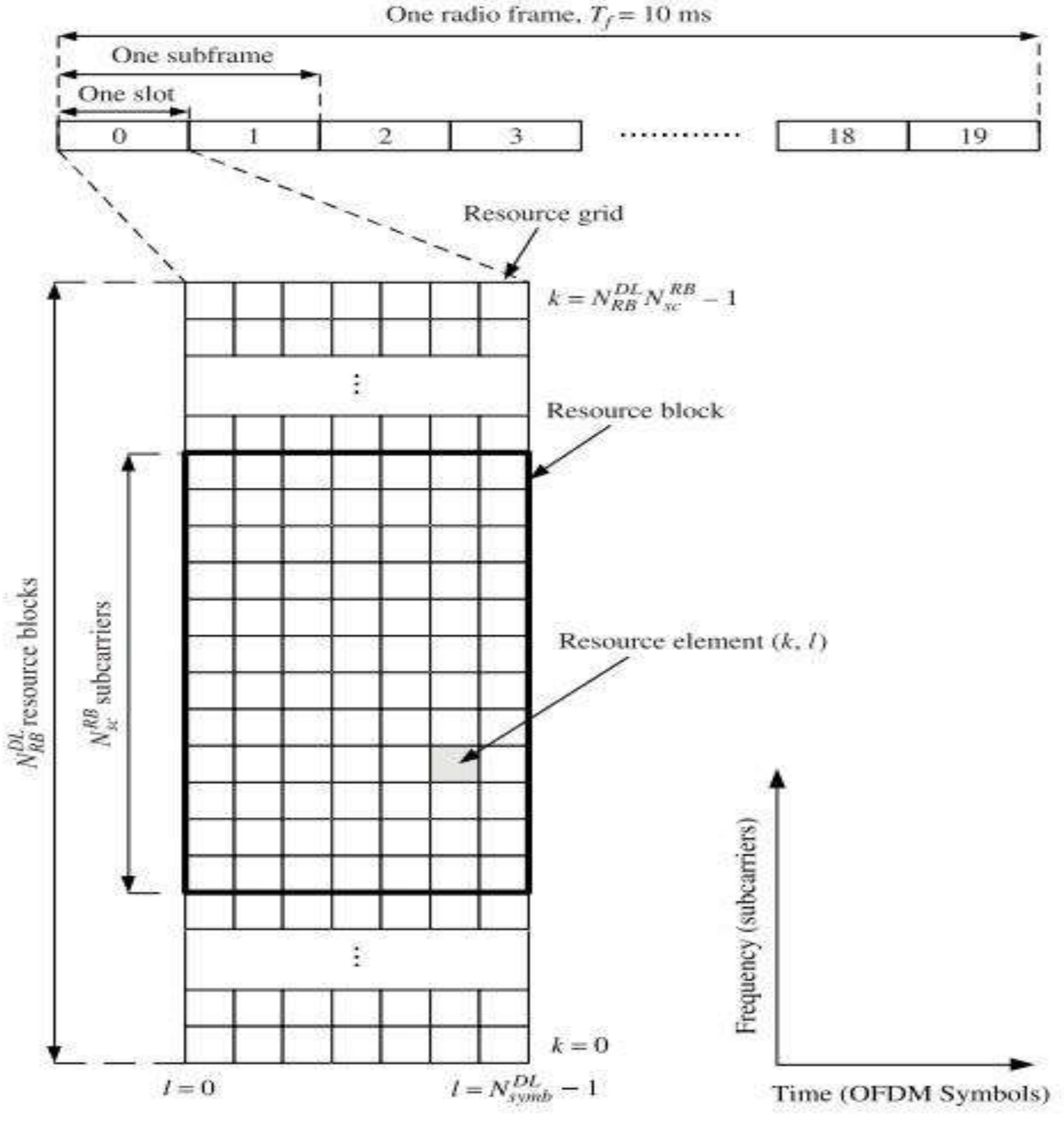


Fig 6.10: The structure of downlink resource grid

- **Resource Grid** : The structure of each resource grid is characterized by the following three parameters:

  1. ***The number of downlink resource blocks*** N_{RB}^{D}: It depends on the transmission bandwidth and shall fulfill $N_{RB}^{min.DL} \leq N_{RB}^{DL} \leq N_{RB}^{max.DL}$, where $N_{RB}^{min.DL} = 6$ and $N_{RB}^{max.DL} = 110$

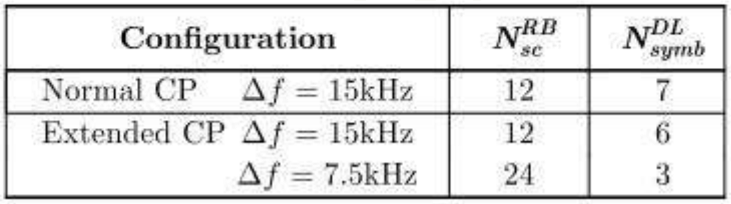
     are for the smallest and largest downlink channel bandwidth, respectively. The values of N_{RB}^{DL} for several current specified bandwidths are listed in Table 6.2.

  2. ***The number of subcarrier in resource blocks*** N_{SC}^{R}:It depends on the subcarrier spacing Δf, satisfying $N_{SC}^{RB}\ \Delta f$ = 180 kHz, that is, each resource block of 180 kHz wide in the frequency domain. The values of N_{SC}^{RB} for different subcarrier spacing are shown in Table 6.4.
     There are a total of $N_{RB}^{DL} \times N_{SC}^{RB}$ subcarriers in each resource grid.

  3. ***The number of OFDM symbols in each block*** N_{symb}^{DL}: It depends on both the CP length and the subcarrier spacing, specified in Table 6.4.

- Each downlink resource grid has $N_{RB}^{DL} \times N_{SC}^{RB} \times N_{symb}^{DL}$ resource elements.

- For example, with 10MHz bandwidth, Δf = 15 kHz, and normal CP, we get N_{RB}^{DL}= 50 from Table 6.2, N_{SC}^{RB} = 12 and N_{symb}^{DL} = 7 from Table 6.4, so there are 50 x 12 x 7 = 4200 resource elements in the downlink resource grid.



Configuration		N_{sc}^{RB}	N_{symb}^{DL}
Normal CP	$\Delta f = 15\text{kHz}$	12	7
Extended CP	$\Delta f = 15\text{kHz}$	12	6
	$\Delta f = 7.5\text{kHz}$	24	3

Table 6.4 Physical Resource Block Parameters for the Downlink

- In case of multi-antenna transmission, there is one resource grid defined per antenna port.
- An antenna port is defined by its associated reference signal, which may not correspond to a physical antenna.
- The set of antenna ports supported depends on the reference signal configuration in the cell.
- there are three different reference signals defined in the downlink, and the associated antenna ports are as follows:
  - Cell-specific reference signals support a configuration of 1, 2, or 4 antenna ports and the antenna port number p shall fulfill p = 0, p ∈ {0, 1}, and p ∈{0,1,2,3}, respectively.
  - MBSFN reference signals are transmitted on antenna port p = 4.
  - UE-specific reference signals are transmitted on antenna port p = 5.

- **Resource Element**
  - Each resource element in the resource grid is uniquely identified by the index pair (k, l) in a slot, where k = 0,1,... , N $N_{RB}^{DL}N_{SC}^{RB} - 1$ and l = 0,1, ... , $N_{symb}^{DL} - 1$ are indices in the frequency and time domains, respectively. The size of each resource element depends on the subcarrier spacing Δf and the CP length.
- **Resource Block**
  - The resource block is the basic element for radio resource allocation.
  - The minimum size of radio resource that can be allocated is the minimum TTI in the time domain, that is, one subframe of 1 ms, corresponding to two resource blocks.
  - The size of each resource block is the same for all bandwidths, which is 180 kHz in the frequency domain.
  - There are two kinds of resource blocks defined for
  - LTE: physical and virtual resource blocks, which are defined for different resource allocation schemes.

6.4.3 Resource Allocation

What do you mean by resource allocation? With an example explain the types of resource allocation in Downlink OFDMA radio resource

- Resource allocation's role is to dynamically assign available time-frequency resource blocks to different UEs in an efficient way to provide good system performance.
- In LTE, channel-dependent scheduling is supported, and transmission is based on the shared channel structure where the radio resource is shared among different UEs.
- Multiuser diversity can be exploited by assigning resource blocks to the UEs with favorable channel qualities.
- Resource allocation in LTE is able to exploit the channel variations in both the time and frequency domain, which provides higher multiuser diversity gain.
- With OFDMA, the downlink resource allocation is characterized by the fact that each scheduled UE occupies a number of resource blocks while each resource block is assigned exclusively to one UE at any time.
- Physical Resource Blocks (PRBs) and Virtual Resource Blocks (VRBs) are defined to support different kinds of resource allocation types.
- The VRB is introduced to support both block-wise transmission (localized) and transmission on non-conse0cutive subcarriers (distributed) as a means to maximize frequency diversity.
- The downlink scheduling is performed at the eNode-B based on the channel quality information fed back from UEs, and then the downlink resource assignment information is sent to UEs on the PDCCH channel.
- A PRB is defined as N_{symb}^{DL} consecutive OFDM symbols in the time domain and N^{RB} consecutive

subcarriers in the frequency domain, as demonstrated in Figure 6.10.

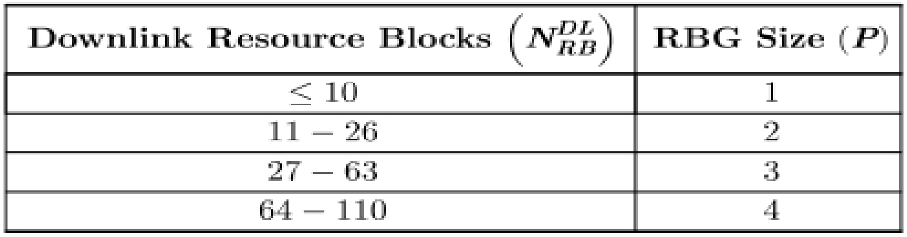
- Each PRB corresponds to one slot in the time domain (0.5 ms) and 180 kHz in the frequency domain.
- PRBs are numbered from 0 to $N_{RB}^{DL} - 1$ in the frequency domain.
- The PRB number ηPRB of a resource element (k, l) in a slot is given by:

$$n_{PRB} = \left\lfloor \frac{k}{N_{sc}^{RB}} \right\rfloor.$$

- ***Resource Allocation Type***: It specifies the way in which the scheduler allocate resource blocks for each transmission. Just in terms of flexibility, the way to give the maximum flexibility of resource block allocation would be to use a string of a bit map (bit stream), each bit of which represent each resource block. This way you would achieve the maximum flexibility, but it would create too much complication of resource allocation process or too much data (too long bit map) to allocate the resources
- The LTE downlink supports three resource allocation types: type 0, 1, and 2.

1. ***Resource Allocation Type 0:*** *This is the simplest way of allocation resources. First it divides resource blocks into multiples of groups. This resource block group is called RBG (Resource Block Group). The number of resource block in each group varies depending on the system band width. It means RBG size gets different depending on the system bandwidth. The relationship between RBS size (the number of resource block in a RBG) and the system bandwidth as shown in Table 6.5.*

Table 6.5 Resource Allocation RBG Size vs. Downlink System Bandwidth



Downlink Resource Blocks $\left(N_{RB}^{DL}\right)$	RBG Size (P)
≤ 10	1
$11 - 26$	2
$27 - 63$	3
$64 - 110$	4

- An exp of type 0, resource allocation is shown in Figure 6.11, where P = 4 and RBGs 0, 3, 4, ... , are allocated to a particular UE.

2. ***Resource Allocation Type 1:*** *Here all the RBGs are grouped into a number of RBG subsets, and certain PRBs inside a selected RBG subset are allocated to the UE. There are a total of P RBG subsets, where P is the RBG size. An RBG subset p, where* $0 \leq p \leq P$ *consists of every P*th *RBG starting from RBG p. Therefore, the resource assignment information consists of three fields:*
    1. *The first field indicates the selected RBG subset*
    2. *The second field indicates whether an offset is applied, and*

- *The third field contains the bitmap indicating PRBs inside the selected RBG subset. This type of resource allocation is more flexible and is able to provide higher frequency diversity, but it also requires a larger overhead.*

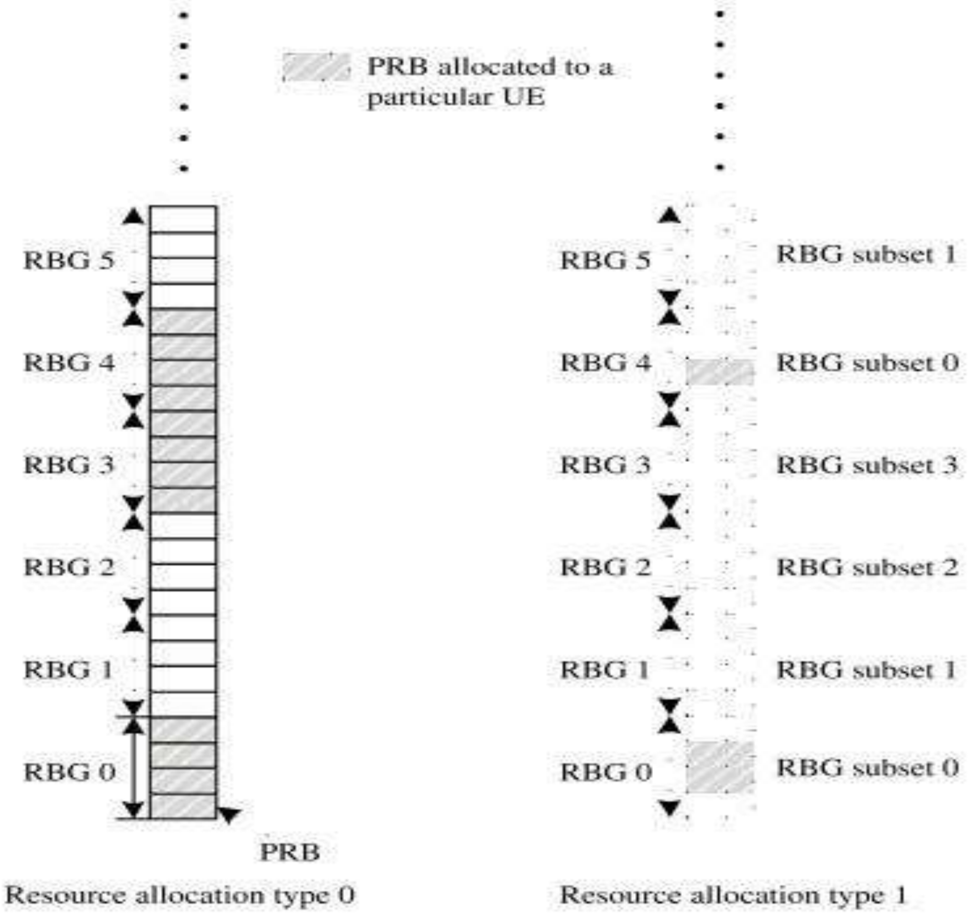


Figure 6.11 Examples of resource allocation type 0 and type 1, where the RBG size P= 4.

3. ***Resource Allocation Type 2:*** *In type 2 resource allocations that are defined for the DCI format 1A, 1B, 1C, and 1D, PRBs are not directly allocated. Instead, VRBs are allocated, which are then mapped onto PRBs. A VRB is of the same size as a PRB. There are two types of VRBs: VRBs of the localized type and VRBs of the distributed type. For each type of VRB, a pair of VRBs over two slots in a subframe are assigned together with a single VRB number,* ηVRB. VRBs of the localized type are mapped directly to physical resource blocks such that the VRB number ηVRB corresponds to the PRB number $\eta PRB = \eta VRB$. *For resource allocations of type 2, the resource assignment information indicates a set of contiguously allocated localized VRBs or distributed VRBs. A one-bit flag indicates whether localized VRBs or distributed VRBs are assigned.*

6.4.4 Supported MIMO Modes

List out the supported SU-MIMO modes of Downlink OFDMA radio resources.

- The downlink transmission supports both single-user MIMO (SU-MIMO) and multiuser MIMO (MU-MIMO).
- For SU-MIMO, one or multiple data streams are transmitted to a single DE through space-time processing; for MU-MIMO, modulation data streams are transmitted to different UEs using the same time-frequency resource.
- The supported SU-MIMO modes are listed as follows:
  1. Transmit diversity with space frequency block codes (SFBC)
  2. Open-loop spatial multiplexing supporting four data streams
  3. Closed-loop spatial multiplexing, with closed-loop preceding as a special case when channel rank = 1
  4. Conventional direction of arrival (DOA)-based beamforming
- The supported MIMO mode is restricted by the UE capability.
- The PDSCH physical channel supports all the MIMO modes, while other physical channels support transmit diversity except PMCH, which only supports single-antenna—port transmission.

6.5 Uplink SC-FDMA Radio Resources

- For the I.TE uplink transmission, SC-FDMA with a CP is adopted.
- Nevertheless, the uplink transmission has its own properties. Different from the downlink, only localized resource allocation on consecutive subcarriers is allowed in the uplink.

6.5.1 Frame Structure

- *Frame structure type 1*: Uplink radio frame consists of 20 slots of 0.5 ms each, and one subframe consists of two slots, as in Figure 6.8.
- *Frame structure type 2*: It consists of ten subframes, with one or two special subframes including DwPTS, GP, and UpPTS fields, as shown in Figure 6.9.
- A CP is inserted prior to each SC-FDMA symbol. Each slot carries seven SC-FDMA symbols in the case of normal CP, and six SC-FDMA symbols in the case of extended CP.

6.5.2 Physical Resource Blocks for SC-FDMA

Build the structure of the uplink resource grid and briefly explain about the physical resource blocks for SC- FDMA

- Figure 6.12, illustrated a number of resource blocks in the time-frequency plane.
- The number of resource blocks in each resource grid, N_{RB}^{UL} , depends on the uplink transmission bandwidth configured in

$$N_{RB}^{min.UL} \leq N_{RB}^{UL} \leq N_{RB}^{max.UL}$$

Where $N^{min.UL} = 6\ and\ N^{max.UL} = 110$ correspond to the smallest and largest uplink bandwidth.

RB RB

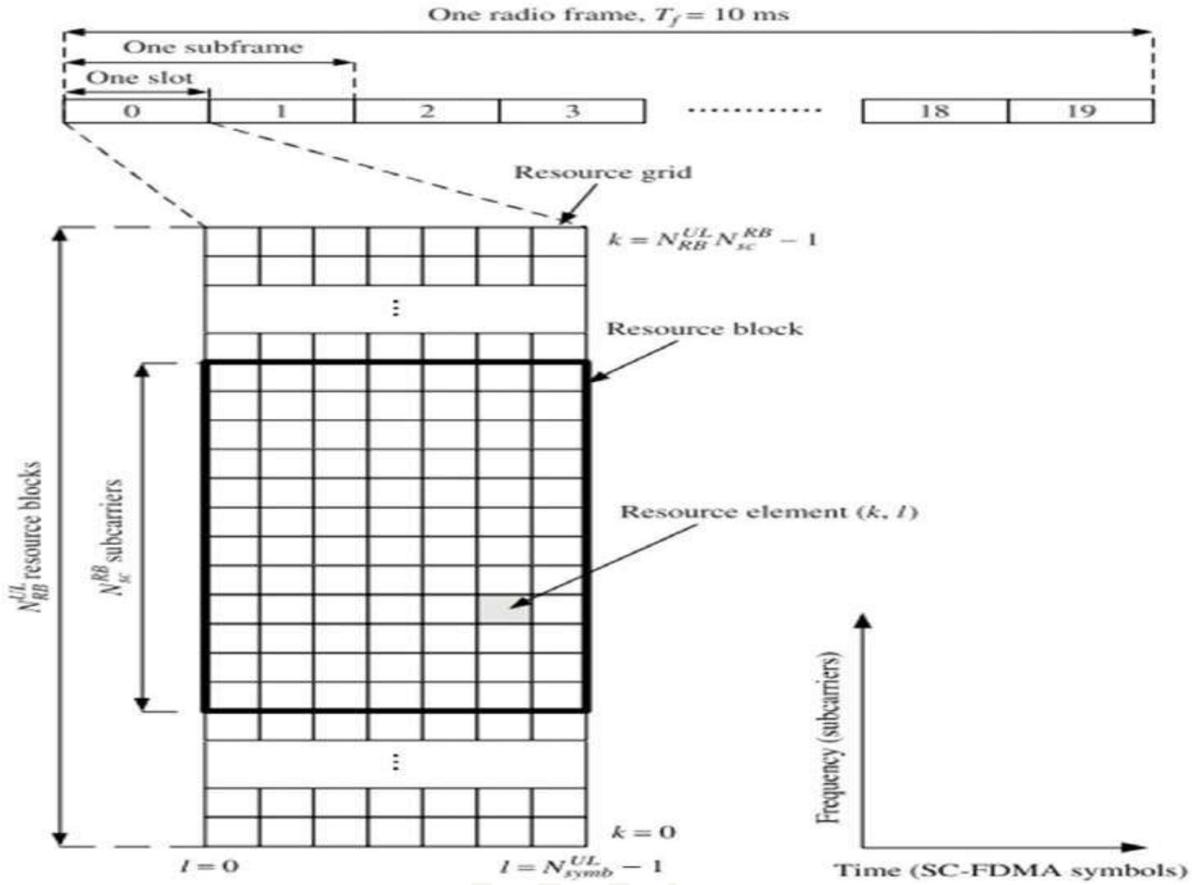
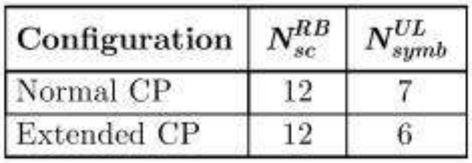

One radio frame, $T_f = 10$ ms
One subframe
One slot
0
1
2
3
18
19
Resource grid
$k = N_{RB}^{UL} N_{sc}^{RB} - 1$
Resource block
Resource element (k, l)
N_{RB}^{UL} resource blocks
N_{sc}^{RB} subcarriers
$k = 0$
$l = 0$
$l = N_{symb}^{UL} - 1$
Frequency (subcarriers)
Time (SC-FDMA symbols)

Figure 6.12: The structure of the uplink resource grid.

- There are $N^{RB}_{SC} \times N^{RB}_{symb}$ resource elements in each resource block. The values of N^{RB}_{SC} and N^{UL}_{symb} for normal and extended CP are given in Table 6.6.

Table 6.6 Physical Resource Block Parameters for Uplink



Configuration	N_{sc}^{RB}	N_{symb}^{UL}
Normal CP	12	7
Extended CP	12	6

- There is only one subcarrier spacing supported in the uplink, which is Δf = 15 kHz.
- The DC subcarrier is used in the uplink, as the DC interference is spread over the modulation symbols due to the DFT-based pre-coding.

- As for the downlink, each *resource element* in the resource grid is uniquely defined by the index pair (k, l) a slot, where k = 0, … , $N^{UL}_{RB} \times N^{RB}_{SC}$— 1 and l = 0………. N^{UL}_{symb} − 1 are the indices in the frequency and time domain, respectively.
- For the uplink, no antenna port is defined, as only single antenna transmission is supported in the current specifications.
- A PRB in the uplink is defined as N^{UL}_{symb} consecutive SC-FDMA symbols in the time domain and N^{RB}_{SC} consecutive subcarriers in the frequency domain, corresponding to one slot in the time domain and 180 kHz in the frequency domain.
- The relation between the PRB number ling in the frequency domain and resource elements (k, l) in a slot is given by:

$$n_{PRB} = \left\lfloor \frac{k}{N^{RB}_{sc}} \right\rfloor.$$

6.5.2 Resource Allocation

- Resource allocation in the uplink is performed at the eNode-B.
- The eNode-B assigns a unique time-frequency resource to a scheduled UE based on the channel quality measured on the uplink sounding reference signals and the scheduling requests sent from UEs.
- Using timing advance such that the transport blocks of different UEs are received synchronously at the eNode-B.
- SC-FDMA is able to support both localized and distributed resource allocation.
- In the current specification, only localized resource allocation is supported in the uplink, which preserves the single-carrier property and can better exploit the multiuser diversity gain in the frequency domain.
- Compared to distributed resource allocation, localized resource allocation is less sensitive to frequency offset and also requires fewer reference symbols.
- The resource assignment information for the uplink transmission is carried on the PDCCH with DCI format 0, indicating a set of contiguously allocated resource blocks.

6.5.3 Supported MIMO Modes

- The terminal complexity and cost are the major concerns in MIMO modes support in uplink.
- SC-FDMA support MU-MIMO, which allocates the same time and frequency resource to two UEs with each transmitting on a single antenna. This is also called Spatial Division Multiple Access (SDMA). The advantage is that only one transmit antenna per UE is required.

- To separate streams for different UEs, channel state information is required at the eNode-B, which is obtained through uplink reference signals that are orthogonal between UEs.
- Uplink MU-MIMO also requires power control, as the near-far problem arises when multiple UEs are multiplexed on the same radio resource.
- For UEs with two or more transmit antennas, closed-loop adaptive antenna, resource allocation transmit diversity shall be supported.

Downlink Transport Channel Processing

7.1 Introduction:

LTE uses a channels to provide effective, efficient data transport over the LTE radio interface.

- There are three categories into which the various data channels may be grouped.
  1. *Physical channels*: These are transmission channels that carry user data and control messages.
  2. *Logical channels*: Provide services for the Medium Access Control (MAC) layer within the LTE protocol structure.
  3. *Transport channels*: The physical layer transport channels offer information transfer to Medium Access Control (MAC) and higher layers. The PHY layer provides services to the MAC layer through transport channels.

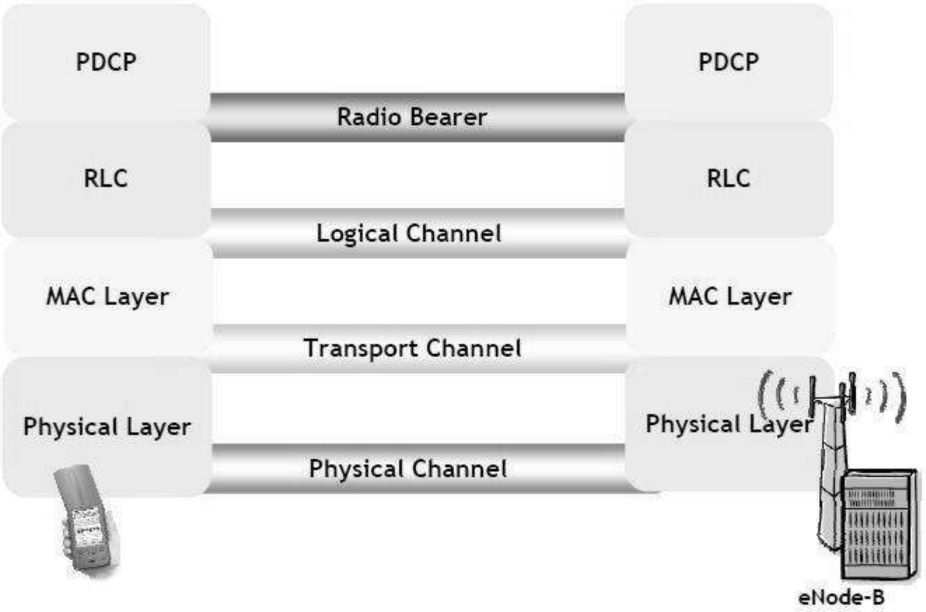


Fig 7.1 LTE channel Structure

- Following are Downlink Transport Channels:

  1. ***Broadcast Channel (BCH) characterized by:***
     - *Fixed, pre-defined transport format*
     - *Requirement to be broadcast in the entire coverage area of the cell.*
  2. ***Downlink Shared Channel (DL-SCH) characterized by:***
     - *Support for HARQ*
     - *Support for dynamic link adaptation by varying the modulation, coding and transmit power*
     - *Possibility to be broadcast in the entire cell*
     - *Possibility to use beamforming*
     - *Support for both dynamic and semi-static resource allocation*
     - *Support for UE discontinuous reception (DRX) to enable UE power saving.*
  3. ***Paging Channel (PCH) characterized by***:
     - *Support for UE discontinuous reception (DRX) to enable UE power saving (DRX cycle is indicated by the network to the UE)*
     - *Requirement to be broadcast in the entire coverage area of the cell*
     - *Mapped to physical resources which can be used dynamically also for traffic or other control channels.*
  4. ***Multicast Channel (MCH) (from Release 9) characterized by:***
     - *Requirement to be broadcast in the entire coverage area of the cell*
     - *Support for MBSFN combining of MBMS transmission on multiple cells*
     - *Support for semi-static resource allocation e.g., with a time frame of a long cyclic prefix.*

- ***Transport Blocks:*** Data and control streams coming from the MAC layer are organized in the form of transport blocks. Each transport block is a group of resource blocks with a common modulation and coding scheme. Downlink Shared Channel (DL_ SCH) are used to transmit transport block.
- ***The physical layer processing***: It mainly consists of coding and modulation, which maps each transport block to specific physical time-frequency resources.

7.2 Downlink Transport Channel Processing Overview

Draw the neat block diagram of downlink transport channel processing and identify the 2different process.

- The downlink physical layer processing mainly consists of
  1. *Channel coding process :* It involves mapping the incoming transport blocks from the MAC layer into different code words
  2. *Modulation process*: Modulation generates complex-valued OFDM baseband signals for each antenna port, which are then up converted to the carrier frequency.

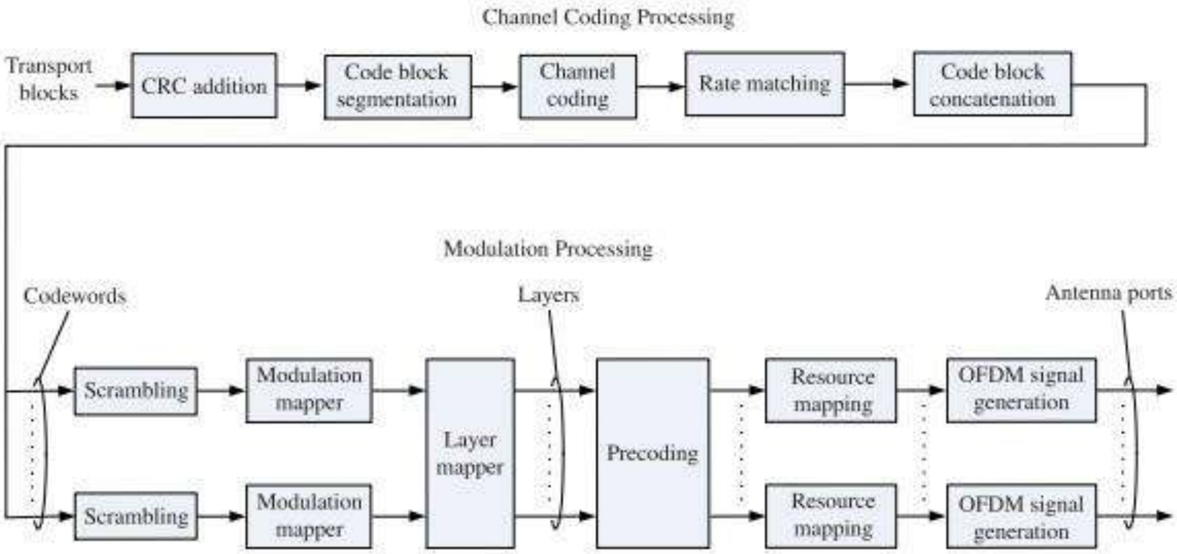


Figure 7.2 Overview of downlink transport channel processing.

7.2.1 Channel Coding Processing: The channel coding processing steps as shown in figure 7.2. The Channel Coding Processing procedure includes

1. ***CRC Addition***
2. ***Code Block segmentation***
3. ***Channel coding****: Tail-Biting Convolutional, Convolution Turbo Coding*
4. ***Rate Matching****: Sub-block interleaving, Bit collection and Bit selection*
5. ***Code Block Concatenation***

- The downlink channel coding processing is shown in Figure 7.2. Channel coding provides an error-control mechanism for data transmission using forward error correction (FEC) code and error detection based on cyclic redundancy check (CRC). In LTE, the coding rate at the channel encoder is fixed, and different effective coding rates for the whole transport block are achieved by repetition/puncturing during the rate matching procedure.

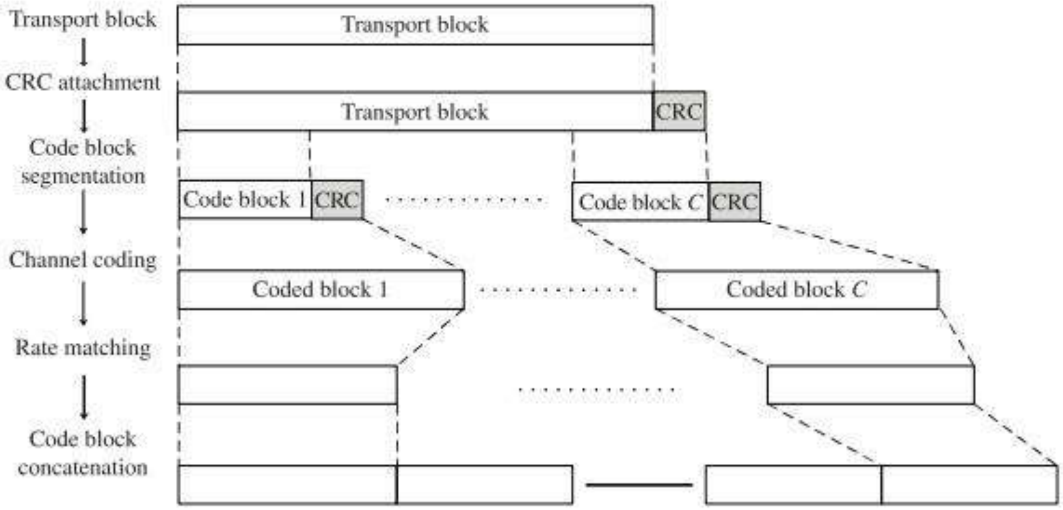


Figure 7.2 Channel coding processing.

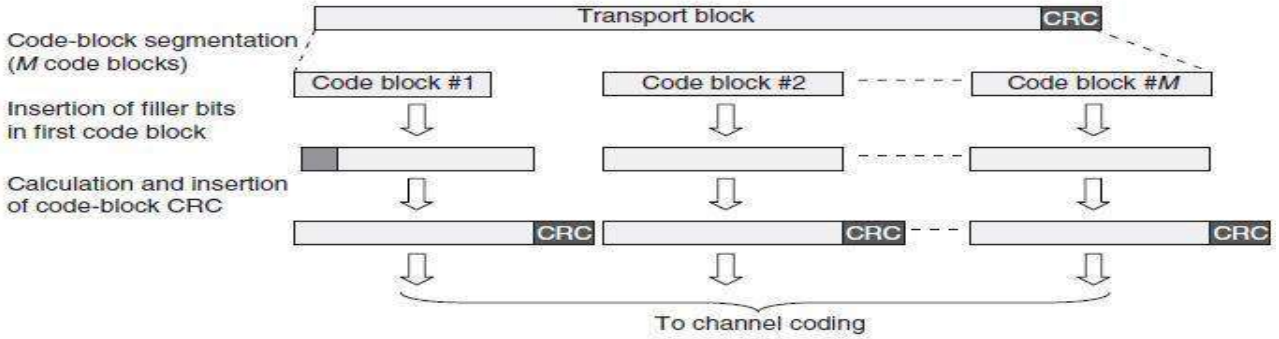
1. CRC Addition :

- The CRC is used to provide error detection on the transport block.
- It generates cyclic generator polynomials, which are then added at the end of the transport block.
- The 24-bit CRC is added to the each transport block for the downlink shared channel.
- The CRC allows for receiver side detection errors in the decoded transport block.
- The corresponding error indication is then used by the down link hybrid- ARQ protocol.

Transport Block	CRC

2. Code Block Segmentation:

- Transport block is divided into smaller size code blocks in LTE, which is referred as code block segmentation in the LTE physical layer.



- In LTE there are two sizes defined for code block i.e. minimum and maximum code block size. These block sizes are based on block sizes as supported by the turbo interleaver module of CTC Encoder. They are as follows:
  - 40 bits of minimum code block size
  - 6144 bits of maximum code block size
- If input transport block length B is greater than the maximum code block size as supported by encoder then the input block is segmented into the one supported. This segmented block is referred as code blocks (c) and it is given by

$$C = \begin{cases} 1 & if \;\; \leq Z \\ \dfrac{B}{(Z-L)} & if \; B > Z \end{cases}$$

Where L is the number of CRC parity bits. Each of these C code blocks is then encoded independently. This is to prevent excessive complexity and memory requirement for de-coding at the receiver

- Each of these code blocks has a 24 bit CRC attached. This CRC is calculated similar to Transport Block CRC calculation.
- Filler bits are appended at the start of segment, this helps code block size to match a set of valid turbo interleaver block sizes.

3. Channel Coding

(List out the channel coding processing steps. Explain each processing procedure)

- In LTE, the channel encoders applied to transport channels include
  1. *Tail-biting convolutional coding*
  2. *Convolutional turbo coding*.
- The usage of channel coding schemes and coding rates for different downlink transport channels is specified in Table below

Transport Channel	Coding Scheme	Coding Rate
DL-SCH, PCH, MCH	Turbo coding	1/3
BCH	Tail-biting convolutional coding	1/3

- For control information, other channel coding schemes are supported, including block coding and repetition coding, specified in Table below

Control Information	Coding Scheme	Coding Rate
DCI	Tail-biting convolutional coding	1/3
CFI	Block coding	1/16
HI	Repetition coding	1/3

A. *Tail-Biting Convolutional Coding*:

Explain tail biting convolutional coding with data rate 1/3 encoder diagram

- The convolutional encoder used in LTE is a rate 1/3 encoder with a constraint length of 7 as shown in Figure 7.3.

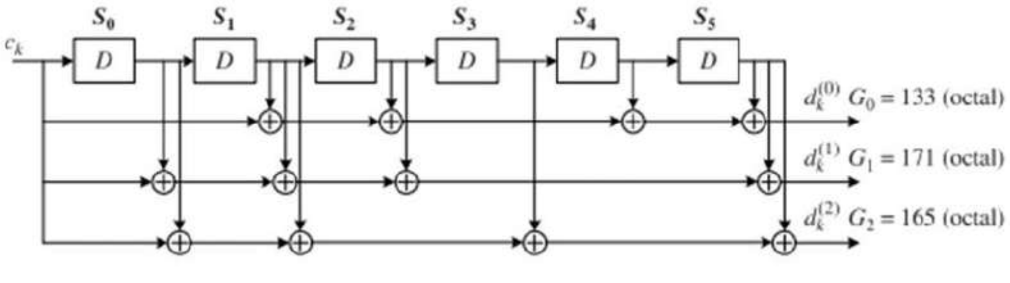

c_k
S_0
S_1
S_2
S_3
S_4
S_5
D
D
D
D
D
D
$d_k^{(0)}$ $G_0 = 133$ (octal)
$d_k^{(1)}$ $G_1 = 171$ (octal)
$d_k^{(2)}$ $G_2 = 165$ (octal)

Figure 7.3 Rate 1/3 tail-biting convolutional encoder.

- Trellis termination must be performed at the end of each code block in order to restore the state of the encoder to the initial state for the next code block.
- If the initial and the final states of the encoder are known, then a lower block error rate can be achieved at the decoder while using a Viterbi algorithm.

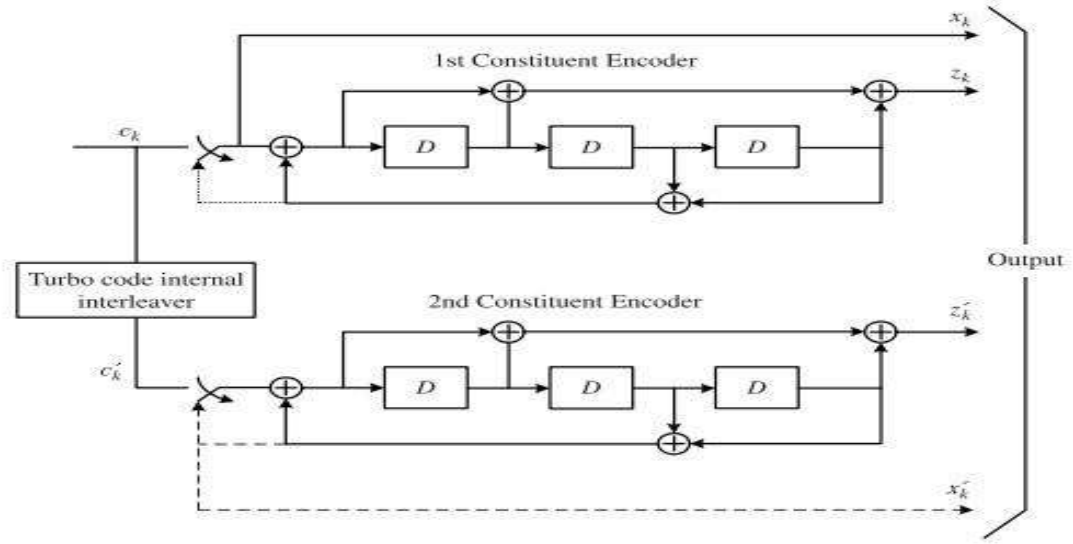
- Two of the most common approaches for trellis termination are
  a. Padding: *Here the end of the code block is padded with zeros. This forces the encoder to state '0' at the end of the code block, which is the starting state for the next code block*. *Main drawbacks of this method is that additional bandwidth is wasted due to the extra zeros that are added to the end of each code block.*
  b. Tail biting: *It is more efficient method, where the information bits from the end of each code block are appended to the beginning of the code block. Once these appended bits are passed through the encoder, it ensures that the start and end states of the encoder are the same. With tail biting, all the input bits are afforded the same amount of error protection, and there is no code-rate loss compared to zero padding, but the decoding algorithm becomes more complicated.*

B. *Convolution Turbo Coding:*
What are tubo codes? Explain with relevant diagram

- It is a Parallel Concatenated Convolutional Code (PCCC) with two eight-state constituent encoders and one turbo code internal interleaver, with a coding rate of 1/3.
- The encoder used for the turbo codes is systematic and therefore recursive in nature.
- LTE employs a new contention-free internal interleaver based on Quadrature Permutation Polynomial (QPP)
- The QPP interleaver requires a small parameter storage and allows highly flexible parallelization due to its maximum contention-free property, which substantially reduces the encoder-decoder complexity
- The structure of the encoder is illustrated in Figure 7.4.

Figure 7.4 Structure of rate 1/3 turbo encoder (dotted lines apply for trellis termination only)

.



- The transfer function of the eight-state constituent code for the PCCC is

$$G(D) = \left[1, \frac{g_1(D)}{g_0(D)}\right],$$

where

$$g_0(D) = 1 + D^2 + D^3,$$
$$g_1(D) = 1 + D + D^3.$$

- The initial values of the shift registers shall be all zeros when starting to encode the input bits.
- Due to the recursive nature of the encoder, the trellis termination is performed by taking the recursive bit and performing a modulo 2 addition with itself as shown in Figure 7.4.
- For each K-bit input code block, the output of the turbo encoder consists of three K-bit data streams:
  a. *One systematic bit stream*
  b. *Two parity bit streams*.
- 12 tail bits due to trellis termination are added to the end of the output streams, so each bit stream has K + 4 bits. Therefore, the actual coding rate is slightly lower than 1/3.

4. Rate Matching

What is rate matching? What are different ways to achieve rate matching?

- The main task of the rate-matching is to extract the exact set of bits to be transmitted within a given TTI.
- The rate-matching for Turbo coded transport channels is defined for each code block: there are three basic steps composing a rate-matching, As illustrated in Figure 7.5.
- Rate matching is defined per coded block and consists of the following stages:
  a. *Interleaving b. Bit collection c. Bit selection*

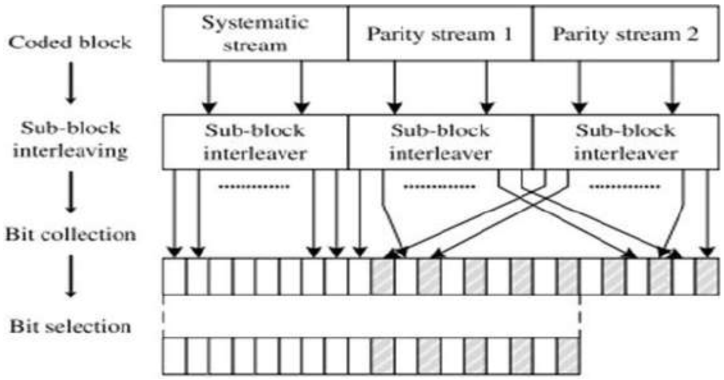


Figure 7.5 Rate matching for coded transport channels.

a. Interleaving: *It is performed at Sub-block level in order to spread out the occurrence of bursty errors across the code block, which improves the overall performance of the decoder. It is performed independently for each bit stream, done by a block interleaver with inter-column permutations. The inter-column permutation patterns are different for turbo coding and convolutional coding.*

b. Bit Collection: *Bit collection stage is required to place the systematic and parity bits in the right order as needed by the decoder. A virtual circular buffer is formed by collecting bits from the interleaved streams. The systematic bits are placed at the beginning, followed by bit-by-bit interlacing of the two interleaved parity streams, as shown in Figure 7.5.*

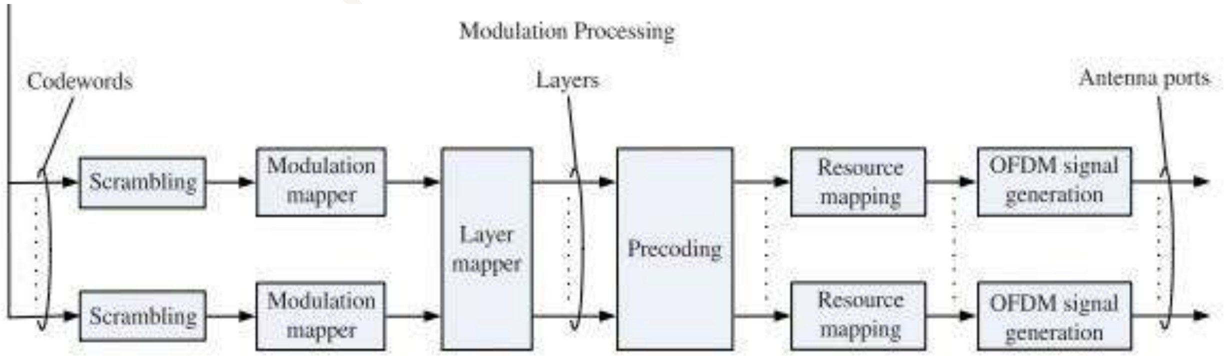
c. Bit Selection: *The bit selection extracts consecutive bits from the circular buffer to the extent that fits into the assigned physical resource. To select the output bit sequence, the sequence length L should first be determined, Then L bits are read from the virtual circular buffer. The starting point of the bit selection depends on the redundancy version of the current transmission, which is different for different retransmissions associated with the H-ARQ process. This means that from one H-ARQ transmission to the next even though the number of bits L is the same, the parity bits that are punctured or repeated can be different. During bit selection if the end of the buffer is reached, the reading continues by wrapping around to the beginning of the buffer*. *With K input bits to the channel encoder, the effective coding rate is K/L, which can achieve any continuum of coding rates.*

d. Code Block Concatenation: *It is needed only for turbo coding when the number of code blocks is larger than one. It consists of sequentially concatenating the rate matching outputs for different code blocks, forming the code word input to the modulation processing.*

7.1.2 Modulation Processing

With the neat block diagram explain modulation processing by listing out the processing blocks.

- Modulation takes in one or two code words, depending on whether spatial multiplexing is used, and converts them to complex-valued OFDM baseband signals for each antenna port.



- The modulation processing consists of
  - ***Scrambling***
  - ***Modulation Mapping***
  - ***Layer Mapping and Pre-coding***
  - ***Resource Mapping***
  - ***OFDM Signal Generation.***

1. Scrambling : A scrambler (or randomizer) is an algorithm that converts an input string into a seemingly random output string of the same length , thus avoiding long sequences of bits of the same value

- There are two main reasons scrambling is used:

1. *To enable accurate timing recovery on receiver equipment without resorting to redundant line coding. It facilitates the work of a timing recovery circuit, an automatic gain control and other adaptive circuits of the receiver*.
2. *For energy dispersal on the carrier, reducing inter-carrier signal interference.*

- Before modulation, the code word is scrambled by a bit-level scrambling sequence.
- The block of bits for code word q is denoted as $b^{(q)}(0), \dots \dots b^{(q)}\ (M^{(q)}_q - 1)$, *Where* $M^{(q)}_q$ is the number of bits transmitted in one sub-frame.
- The scrambling sequence $^{(q)}$is a pseudo-random sequence defined by a length-31 Gold sequence [3J. The scrambled bits are generated using a modulo 2 addition as:

$$\tilde{b}^{(q)}(i) = \left(b^{(q)}(i) + c^{(q)}(i)\right) \bmod 2, \quad i = 0, 1, \dots, M_b^{(q)} - 1.$$

- Up to two codewords can be transmitted in the same subframe, so $q = 0$ if spatial multiplexing is not used or q ∈{0,1} if spatial multiplexing is used.
- Except the multicast channel, for all other downlink transport channels and control information, the scrambling sequences are different for neighboring cells so that inter-cell interference is randomized, which is one of the approaches for interference mitigation.

2. Modulation Mapping:

What are steps involved in modulation mapping?

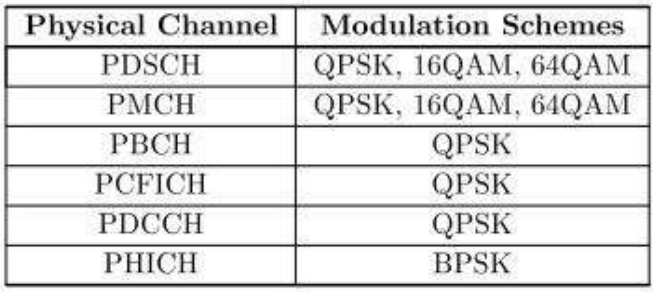
- For each codeword q, the block of scrambled bits $b^{(q)}(0), \dots \dots b^{(q)}\ (M^{(q)}_q - 1)$ are modulated into a block of complex-valued modulation symbols $d^{(q)}(0), \dots \dots d^{(q)}\ (M^{(q)}_s - 1)$ where $M^{(q)}_s$ is the number of the modulation symbols in each codeword and depends on the modulation scheme. The relation between $^{(q)}_s$and $M^{(q)}_q$ is as follows:

$$M_s^{(q)} = \frac{M_b^{(q)}}{Q_m},$$

- Where is the number of bits in the modulation constellation, with Qin = 2 for QPSK, Q_m = 4 for 16QAM, and Q„, = 6 for 64QAM.
- The supported data-modulation schemes in LTE include QPSK, 16QAM, and 64QAM, and BPSK is applied for the PHICH physical channel.

- ○ Different physical channels employ different modulation listed in Table 7.3.

Table 7.3 Modulation Schemes for Different Physical Channels



Physical Channel	Modulation Schemes
PDSCH	QPSK, 16QAM, 64QAM
PMCH	QPSK, 16QAM, 64QAM
PBCH	QPSK
PCFICH	QPSK
PDCCH	QPSK
PHICH	BPSK

3. Layer Mapping and Precoding

- Mapping and pre-coding are associated with MIMO. An illustrated in figure 7.6

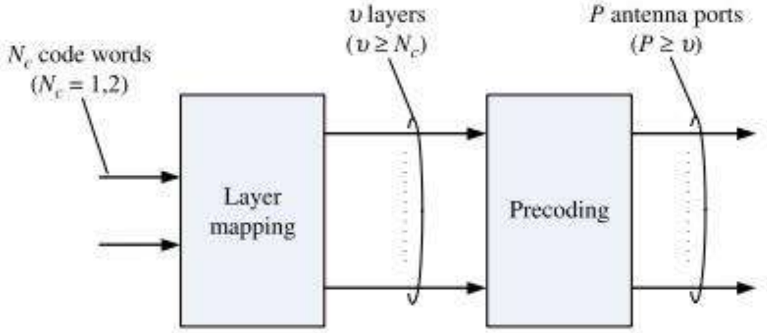


Figure 7.6 Layer mapping and precoding.

Layer Mapping:

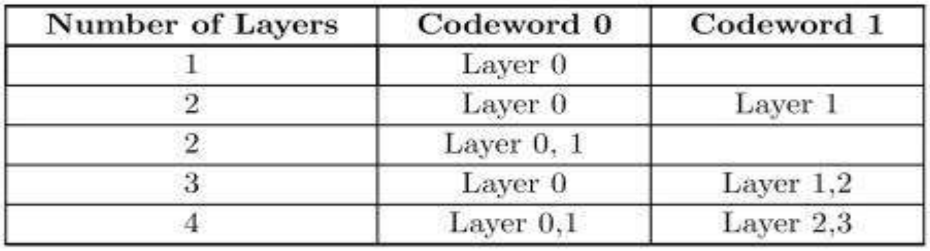
- This is the process where each *codeword* is mapped to one or multiple *layers*. A *codeword* is defined as the output of each channel coding associated with a single transport block coming from the MAC layer. For MIMO transmission with multiple codewords on different spatial channels. In LTE, up to four transmit/receive antennas are supported, the number of codewords is limited to two. *A layer* corresponds to a data stream of the spatial multiplexing channel. Each codeword is mapped into one or multiple layers
- ***Pre-coding***: This is process where the layer data are allocated to multiple antenna ports. An antenna port is defined by its associated reference signal. The number of transmit antenna ports at the eNode-B is sent to UEs through the PBCH channel, which can be 1, 2, or 4 in LTE. Antenna ports are divided into three groups:
    1. *Antenna ports 0-3:* These ports are cell specific, which are used for downlink MIMO transmission.
    2. *Antenna port 4:* It is MBSFN specific and is used for MBSFN transmission.
    3. *Antenna port 5:* It is UE specific, which is used for beamforming to a single UE using all physical antennas.

- Cell-specific ports and the UE-specific port cannot be simultaneously used.

Layer mapping is different for different MIMO modes, described as follows.

1. *Single antenna port*: One codeword is mapped to a single layer.
2. *Transmit diversity*: One codeword is mapped to two or four layers.
3. *Spatial multiplexing*: Are codewords are mapped to v layers, the detailed mapping is in Table 7.4. Note that the case of a single codeword mapped to two layers occurs only when the initial transmission contains two codewords and a codeword mapped onto two layers needs to be retransmitted. Both open-loop (OL) and closed-loop (CL) spatial multiplexing modes are supported in LTE.

Table 7.4 Codeword-to-Layer Mapping for Spatial Multiplexing



Number of Layers	Codeword 0	Codeword 1
1	Layer 0	
2	Layer 0	Layer 1
2	Layer 0, 1	
3	Layer 0	Layer 1,2
4	Layer 0,1	Layer 2,3

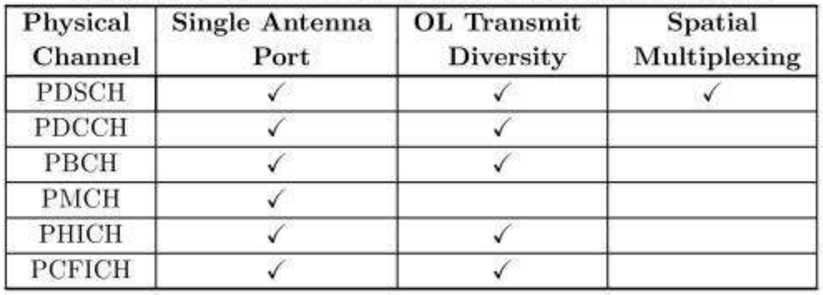
- The precoder is either fixed or selected from a predefined codebook based on the feedback from UEs. The general form for precoding is

$$y(i) = W(i) * x(i)$$

Where (i) is the precoding matrix of size $P \times v$.

- Different physical channels support different MIMO modes, specified in Table 7.5. The PDSCH channel supports all the specified MIMO modes, while the PMCH channel only supports single-antenna-port transmission(antenna port 4).

Table 7.5 Supported MIMO Modes for Different Physical Channels



Physical Channel	Single Antenna Port	OL Transmit Diversity	Spatial Multiplexing
PDSCH	✓	✓	✓
PDCCH	✓	✓	
PBCH	✓	✓	
PMCH	✓		
PHICH	✓	✓	
PCFICH	✓	✓	

4. Resource Mapping

Explain the concept of resource mapping.

- For each of the antenna ports used for transmission of physical channels.
- The block of complex-valued symbols $y_{p}(0), \dots\dots y_{p}(M_{s}^{(ap)} - 1)$ shall be mapped in sequence.
- Starting with $y(0)$, to resource blocks assigned for transmission.
- The mapping to resource element (k, l) on antenna port p not reserved for other purposes.

5. OFDM Baseband Signal Generation

- The continuous-time signal $s_l^{(p)}(t)$ on antenna port p in OFDM symbol l in a downlink slot is generated as:

$$s_l^{(p)}(t) = \sum_{k=-\lfloor N_{RB}^{DL} N_{sc}^{RB}/2 \rfloor}^{-1} a_{k^{(-)},l}^{(p)} \cdot e^{j2\pi k \Delta f (t - N_{CP,l} T_s)} + \sum_{k=1}^{\lceil N_{RB}^{DL} N_{sc}^{RB}/2 \rceil} a_{k^{(+)},l}^{(p)} \cdot e^{j2\pi k \Delta f (t - N_{CP,l} T_s)} \tag{7.4}$$

for $0 \leq t \leq (N_{CP,l} + N) \times T_s$, where $k^{(-)} = k + \lfloor N_{RB}^{DL} N_{sc}^{DL}/2 \rfloor$ and $k^{(+)} = k + \lfloor N_{RB}^{DL} N_{sc}^{DL}/2 \rfloor - 1$, and for 20MHz bandwidth the value of N is given by:

$$N = \begin{cases} 2048, & \text{if } \Delta f = 15\text{kHz} \\ 4096, & \text{if } \Delta f = 7.5\text{kHz}. \end{cases} \tag{7.5}$$

The cyclic prefix (CP) length $N_{CP,l}$ depends on the CP type and the subcarrier spacing, listed in Table 7.6.

Table 7.6 Values of $N_{CP,l}$

Configuration		CP Length $N_{CP,l}$
Normal CP	$\Delta f = 15$kHz	160 for $l = 0$ 144 for $l = 1, 2, \ldots, 6$
Extended CP	$\Delta f = 15$kHz $\Delta f = 7.5$kHz	512 for $l = 0, 1, \ldots, 5$ 1024 for $l = 0, 1, 2$

- The OFDM signal generation with multiple users are illustrated in figure 7.8

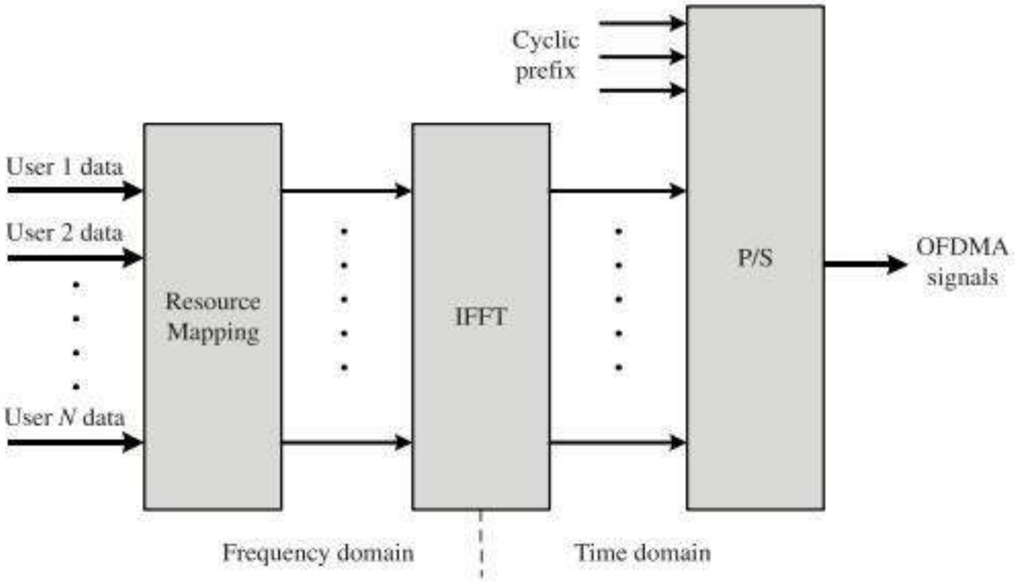


Figure 7.8 OFDMA signal generation with N users, where P/S denotes the parallel-to-serial converter.

7.2 Downlink Shared Channels (DL-SCH)

- The DL-SCH is carried on the Physical Downlink Shared Channel (PDSCH).
- Data transmission in the PDSCH is based on the concept of shared-channel transmission, where the resource blocks available for PDSCH, is treated as a common resource that can be dynamically shared among different UEs.
- The dynamic multiplexing of LTEs on the PDSCH is done by the scheduler on 1ms interval.
- The channel mapping around the DL-SCH is shown in Figure 7.9.

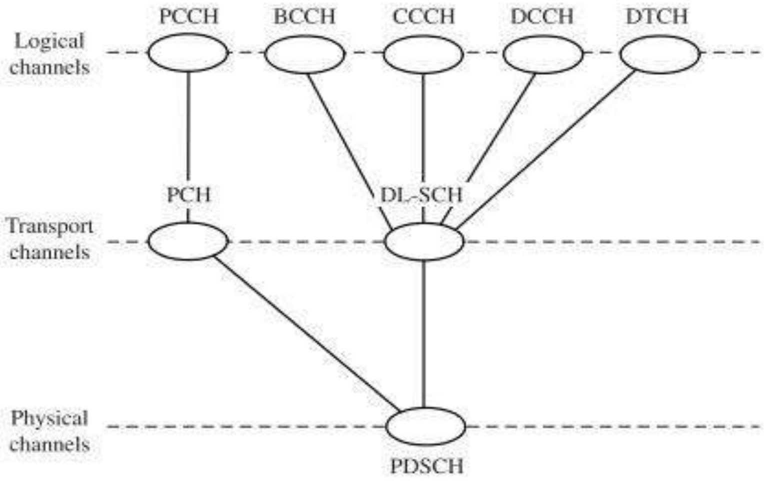


Figure 7.9 Channel mapping around the downlink shared channel.

- DL-SCHs carry both traffic and control data from logical channels, and the Paging Channel (PCH) is also carried on the PDSCH (See figure 7.9).

7.2.1 Channel Encoding and Modulation

Describe the channel encoding and modulation for downlink shared channels.

- *Channel Coding of DL-SCH:*
  - It uses the rate 1/3 convolutional turbo code.
  - Rate matching is used in order to achieve an effective channel coding rate that matches the payload capacity.
  - For MIMO spatial multiplexing with two code words, different modulation and coding can be used for each code word, which requires individual signaling.

- *Modulation scheme of DL-SCH*:
  - It includes QPSK, 16QAM, and 64QAM and is chosen based on the Channel Quality Indicator (CQI) provided by the UE and various other parameters.
  - The transport block size, the redundancy version, and the modulation order are indicated in the Downlink Control Information (DCI).
  - Channel coding for the PCH transport channel is the same as that for the DL-SCH channel. Both of which are mapped to the PDSCH physical channel.

7.2.2 Multi-antenna Transmission

(List the seven different transmission modes defined for data transmission on the PDSCH channel.)

- The PDSCH supports all the MIMO modes specified in LTE.***
- There are seven transmission modes defined for data transmission on the PDSCH channel:
  1. ***Single-antenna port (port 0):*** One transport block is transmitted from a single physical antenna corresponding to antenna port 0.
  2. ***Transmit diversity***: One transport block is transmitted from more than one physical antenna, that is, ports 0 and 1.
  3. ***Open-loop (OL) spatial multiplexing***: One or two transport blocks are transmitted from two or four physical antennas. In this case, precoding is fixed based on RI feedback.
  4. ***Closed-loop (CL) spatial multiplexing***: One or two transport blocks are transmitted from two or four physical antennas. The precoding is adapted based on the Precoding Matrix Indicator (PMI) feedback from the UE.
  5. ***Multiuser MIMO:*** Two UEs are multiplexed onto two or four physical antennas with one transport block to each UE.
  6. ***Closed-loop rank-1 precoding***: It is a special case of the Closed Loop spatial multiplexing with single-layer transmission, that is, a P x 1 precoder is applied.
  7. ***Single-antenna port (port 5):*** A single transport block is transmitted from two or more physical antennas. The eNode-B performs beamforming to a single UE using all physical antennas. Beamforming can be used to improve the received signal power and/or reduce the interference signal power, which is especially important for cell edge users.
- Transmission mode 1 can be classified as a Single-Input-Single-Output (SISO) mode that does not require any layer mapping and precoding.
- Transmission modes 2-6 can be classified as MIMO modes, which require explicit layer mapping and precoding.
- Transmission MIMO modes classified into
  - *i.* Open Loop(OL) Transmission MIMO modes: *OL MIMO technique requires no feedback from UEs, so it is suitable for scenarios where accurate feedback is difficult to obtain or the channel changes rapidly enough, such as the high mobility scenario. This mode includes*

    (a). OL transmit diversity (b) OL Spatial multiplexing
  - *ii.* Closed Loop (CL) Transmission MIMO modes: *CL MIMO transmission requires explicit feedback from UEs. UE determines precoding matrix based on its current MIMO channel and sends this information to the eNode-B using the uplink control channel. This mode includes*

    (a) CL Spatial Multiplexing (RI > 1)

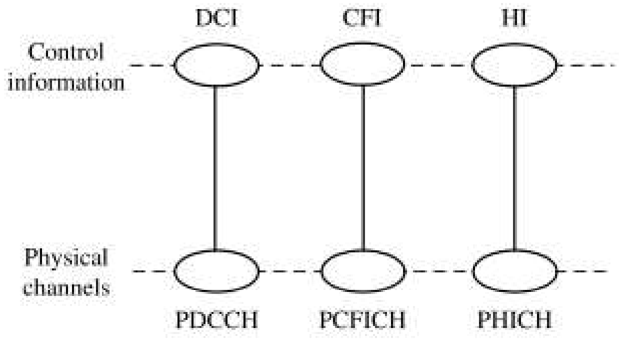
    (b) CL Rank-1 Precoding (RI = 1)

7.3 Downlink Control Channels

Explain down link control channel and its mapping.

- Downlink control channels are carried over the Physical Downlink Control Channel (PDCCH).
- Control information from the MAC layer, including
  1. *Downlink Control Information (DCI).*
  2. *Control Format Indicator (CFI).*
  3. *H-ARQ Indicator (HI).*

- **Channel mapping** between control information and physical channels in the downlink is shown in Figure 7.11.
- There is a specific physical channel for each type of control information. On the physical layer the PDCCH and the PDSCH are time multiplexed and
  - PDCCH is carried over the first few OFDM symbols of each subframe
  - PDSCH is carried over the rest of the OFDM symbols.
  - The number of OFDM symbols allocated for PDCCH can vary from one to four and is conveyed by the CFI.
  - The CFI is carried on yet another control channel known as the Physical Control Format Indicator Channel (PCFICH), which is always carried in a predetermined format over the first OFDM symbol of each subframe.
  - This predetermined format of PCFICH allows each UE to decode the CFI without ambiguity and thus determine the number of OFDM symbols in the beginning of earth subframe that are used as the control region
  -



7.11 Channel mapping for control information in the downlink

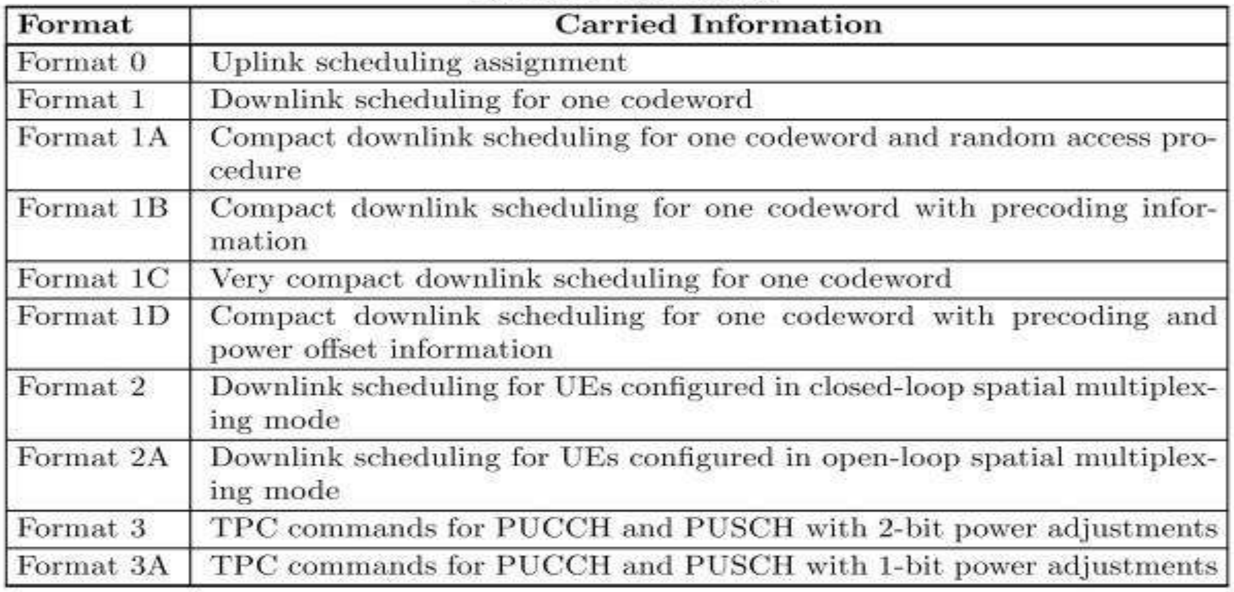
7.3.1 Downlink Control Information (DCI) Formats:

- DCI is the most important as it carries detailed control information for both downlink and uplink transmissions.
- The DCI carries the downlink scheduling assignments, uplink scheduling grants, power control commands, and other information necessary for the scheduled UEs to decode and demodulate data symbols in the downlink or encode and modulate data symbols in the uplink.

Describe the ten different DCI formats for different transmission scenarios

- In Table 6.1, LTE defines ten different DCI formats for different transmission scenarios, summarized as follows:

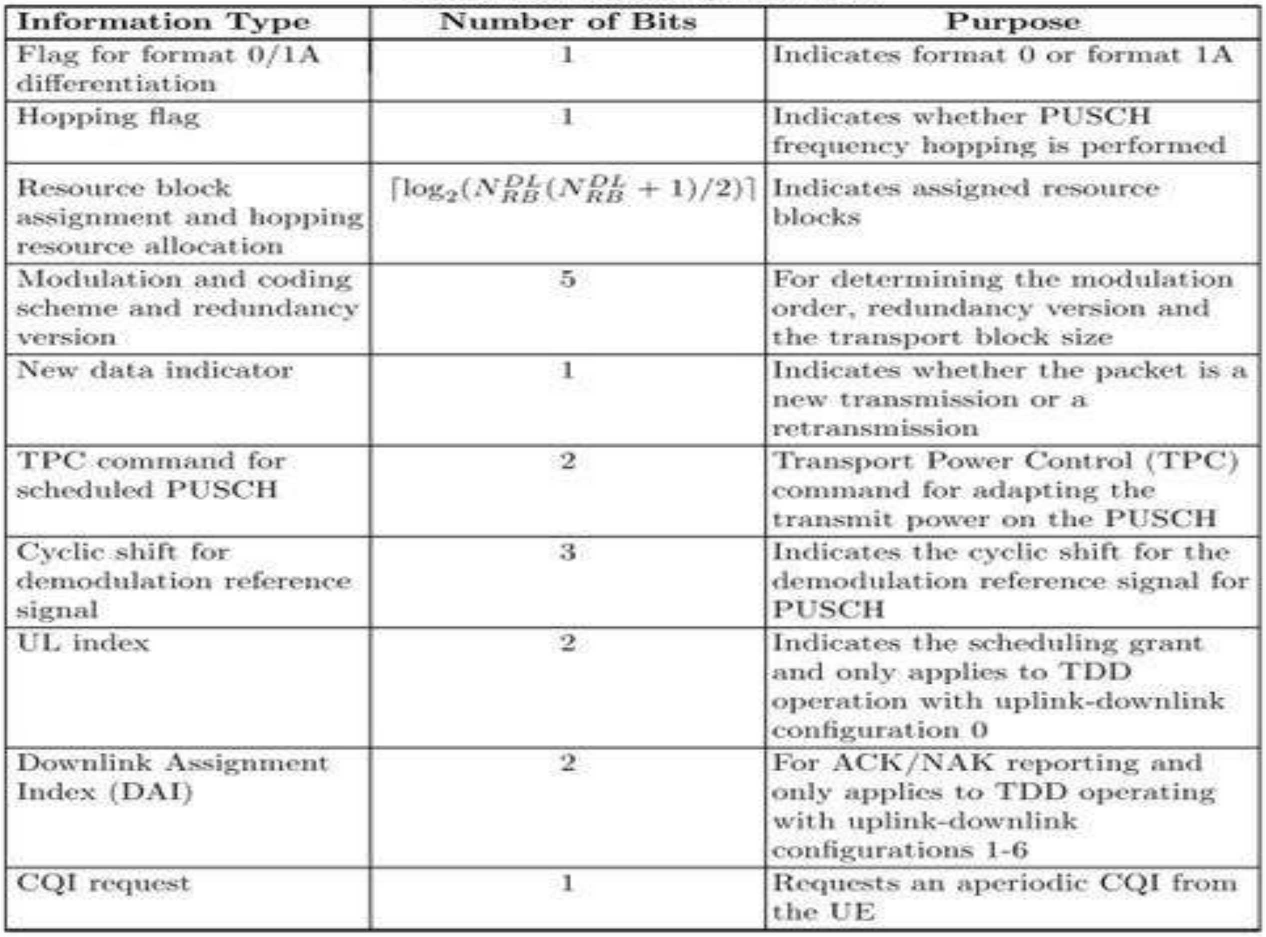
Table 6.1 DCI Formats



Format	Carried Information
Format 0	Uplink scheduling assignment
Format 1	Downlink scheduling for one codeword
Format 1A	Compact downlink scheduling for one codeword and random access procedure
Format 1B	Compact downlink scheduling for one codeword with precoding information
Format 1C	Very compact downlink scheduling for one codeword
Format 1D	Compact downlink scheduling for one codeword with precoding and power offset information
Format 2	Downlink scheduling for UEs configured in closed-loop spatial multiplexing mode
Format 2A	Downlink scheduling for UEs configured in open-loop spatial multiplexing mode
Format 3	TPC commands for PUCCH and PUSCH with 2-bit power adjustments
Format 3A	TPC commands for PUCCH and PUSCH with 1-bit power adjustments

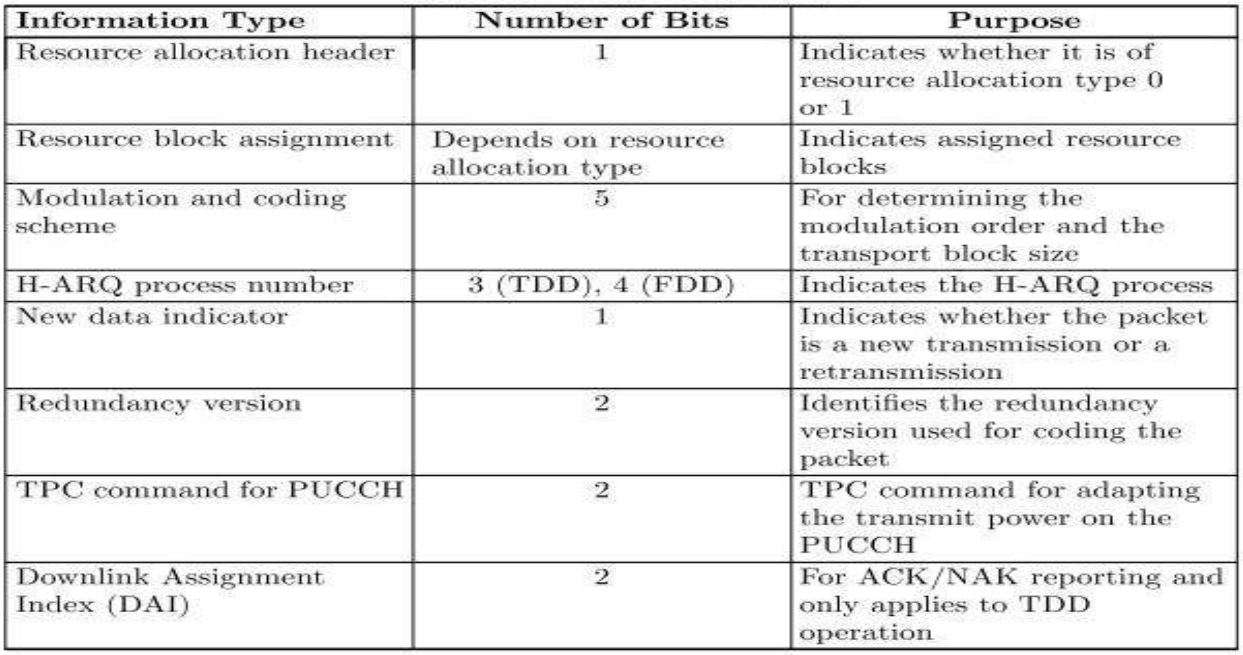
- By considering format 0 and format 1 as examples, the different fields of DCI format are explained in Table 7.10 and Table 7.11, respectively.

Table 7.10 Fields of DCI Format 0



Information Type	Number of Bits	Purpose
Flag for format 0/1A differentiation	1	Indicates format 0 or format 1A
Hopping flag	1	Indicates whether PUSCH frequency hopping is performed
Resource block assignment and hopping resource allocation	$\lceil \log_2(N_{RB}^{DL}(N_{RB}^{DL}+1)/2) \rceil$	Indicates assigned resource blocks
Modulation and coding scheme and redundancy version	5	For determining the modulation order, redundancy version and the transport block size
New data indicator	1	Indicates whether the packet is a new transmission or a retransmission
TPC command for scheduled PUSCH	2	Transport Power Control (TPC) command for adapting the transmit power on the PUSCH
Cyclic shift for demodulation reference signal	3	Indicates the cyclic shift for the demodulation reference signal for PUSCH
UL index	2	Indicates the scheduling grant and only applies to TDD operation with uplink-downlink configuration 0
Downlink Assignment Index (DAI)	2	For ACK/NAK reporting and only applies to TDD operating with uplink-downlink configurations 1-6
CQI request	1	Requests an aperiodic CQI from the UE

Table 7.11 Fields of DCI Format 1

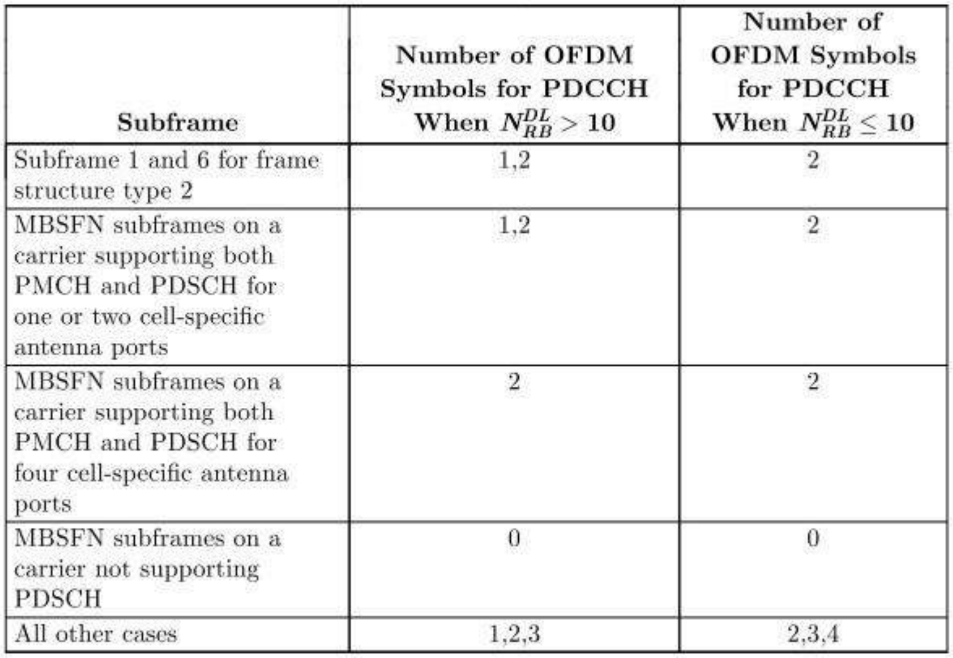


Information Type	Number of Bits	Purpose
Resource allocation header	1	Indicates whether it is of resource allocation type 0 or 1
Resource block assignment	Depends on resource allocation type	Indicates assigned resource blocks
Modulation and coding scheme	5	For determining the modulation order and the transport block size
H-ARQ process number	3 (TDD), 4 (FDD)	Indicates the H-ARQ process
New data indicator	1	Indicates whether the packet is a new transmission or a retransmission
Redundancy version	2	Identifies the redundancy version used for coding the packet
TPC command for PUCCH	2	TPC command for adapting the transmit power on the PUCCH
Downlink Assignment Index (DAI)	2	For ACK/NAK reporting and only applies to TDD operation

7.3.2 Control Format Indicator (CFI).

- The CFI is a parameter used on the LTE air interface. It defines the amount of symbols in each subframe allocated to PDCCH. The CFI takes values CFI = 1, 2 or 3 OFDM symbols as shown in Table 7.13

Table 7.13 Number of OFDM Symbols Used for PDCCH



Subframe	Number of OFDM Symbols for PDCCH When $N_{RB}^{DL} > 10$	Number of OFDM Symbols for PDCCH When $N_{RB}^{DL} \leq 10$
Subframe 1 and 6 for frame structure type 2	1,2	2
MBSFN subframes on a carrier supporting both PMCH and PDSCH for one or two cell-specific antenna ports	1,2	2
MBSFN subframes on a carrier supporting both PMCH and PDSCH for four cell-specific antenna ports	2	2
MBSFN subframes on a carrier not supporting PDSCH	0	0
All other cases	1,2,3	2,3,4

- For example system bandwidths $N_{RB}^{DL} > 10$, the DCI spans 1, 2, or 3 OFDM symbols, given by the value of the CFI; for system bandwidths $N_{RB}^{DL} \leq 10$, the DCI spans 2, 3, or 4 OFDM symbols, given by CFI+1.
- Finally, the CFI is mapped to the PCFICH physical channel carried on specific resource elements in the first OFDM symbol of the subframe.
- The PCFICH is transmitted when the number of OFDM symbols for PDCCH is greater than zero. The PCFICH shall be transmitted on the same set of antenna ports as the PBCH.

7.3.3 H-ARQ Indicator (HI)

- LTE uses a hybrid automatic repeat request (HARQ) scheme for error correction.
- The eNodeB sends a HARQ indicator to the UE to indicate a positive acknowledgement (ACK) or negative acknowledgement (NACK) for data sent using the uplink shared channel.
- The channel coded HARQ indicator codeword is transmitted through the Physical Hybrid Automatic Repeat Request Indicator Channel (PHICH).
- H-ARQ Indicator: H-ARQ indicator of '0' represents a NACK and a '1' represents an ACK.
- A repetition code with rate 1/3 and BPSK modulation is applied used for encoding and mapping the H-ARQ Indicator.
- Multiple PHICHs mapped to the same set of resource elements constitute a PHICH group, where PHICHs within the same group are separated through different orthogonal sequences with a spreading factor of four.

7.4 Broadcast Channels (PBCH)***

Explain the broadcast and multicast channels.

- Broadcast channels carry *system information* such as downlink system bandwidth, antenna configuration, and reference signal power.
- Due to the large size of the *system information field*, it is divided into two portions:
  1. *Master Information Block (MIB)*: It is transmitted on the PBCH. The PBCH contains basic system parameters necessary to demodulate the PDSCH. The transmission of the PBCH is characterized by a fixed pre-determined transport format and resource allocation
  2. *System Information Blocks (SIB)*: It is transmitted on the PDSCH. Which contains the remaining SIB.
- *Coding and Modulation types for PBCH*:
  - Error detection is provided through a 16-bit CRC.
  - The tail-biting convolutional coding with rate 1/3 is used, and the coded bits are rate matched to 1920 bits for the normal CP and to 1728 bits for the extended CP.
  - The modulation scheme is QPSK. No H-ARQ is supported.
  - PBCH supports single-antenna transmission and OL transmit diversity.

7.5 Multicast Channels

- Multimedia Broadcast and Multicast Services (MBMS), introduced in 3GPP supports multicast /broadcast services in a cellular system.
- MBMS is a point-to-multipoint service in which data is transmitted from a single source entity to multiple recipients. Transmitting the same data to multiple recipients allows network resources to be shared.
- The MBMS bearer service offers two modes:

  1. Broadcast Mode. 2. Multicast Mode.
- In principle, the MBMS transmission can originate from a single base station or multiple base stations, but multicell transmission is preferred as large gains can be achieved through soft combining of transmissions from multiple base stations.
- One major design requirement for LTE is to provide enhanced support for the MBMS transmission, which is called Enhanced MBMS (E-MBMS) and is achieved through the so- called Single-Frequency Network (SFN) operation.
- Combining of multicast/broadcast transmissions from multiple base stations is possible in LTE with an extended CP.
- The extended CP is used as the propagation delay from multiple cells and will typically be larger than the delay spread in a single cell.
- A longer CP can ensure that signals from different base stations still fall within the CP at the receiver, which avoids inter-symbol interference at the cost of a slight reduction in peak data rate.
- The E-MBMS transmission in LTE occurs on the MCH transport channel, along with the 7.5 kHz subcarrier spacing and the extended CP. There are two types of E-MBMS transmissions:
  1. *Single-cell transmission (non-MBSFN operation):* The MBMS service (MTCH and MCCH) is transmitted on the NKR, and combining of MBMS transmission from multiple cells is not supported.
  2. *Multi-cell transmission (MBSFN operation):* The MBMS service (MTCH and MCCH) is transmitted synchronously on the RICH, and combining is supported with the SFN operation.
- The PMCH and DL-SCH can be multiplexed with the following rules:
  - The MBSFN and DL-SCH transmission can be multiplexed in a time-division multiplexing (TDM) manner on a subcarrier basis, but cannot be transmitted within the same subframe.

- In the subframes where PMCH is transmitted on a carrier supporting a mix of PDSCH and PMCH transmissions, up to two of the first OFDM symbols of a subframe can be reserved for non-MBSFN transmission and shall not be used for PMCH transmission.
- In a cell with four cell-specific antenna ports, the first OFDM symbols of a subframe are reserved for non-MBSFN transmission in the subframes in which the PMCH is transmitted.
- The non-MBSFN symbols shall use the same CP as used for subframe 0.
- PMCH shall not be transmitted in subframes 0 and 5 on a carrier supporting a mix of PDSCH and PMCH transmissions.

7.6 Downlink Physical Signals: ***Which are the down link physical signals, explain each.***

It including downlink *reference signals* and *synchronization signals*.

7.6.1 Downlink Reference Signals:

- Downlink *reference signals* consist of known reference symbols that are intended for downlink channel estimation at the UE needed to perform coherent demodulation.
- To facilitate the channel estimation process, scattered reference signals are inserted in the resource grid at pre-determined intervals.
- The time and frequency intervals are mainly determined by the characteristics of the channels, and should make a tradeoff between the estimation accuracy and the overhead.
- There are three different types of downlink reference signals:
  1. Cell-specific reference signals
  2. MBSFN reference signals
  3. UE-specific reference signals.

1. Cell-Specific Reference Signals:

- The reference sequence is generated from a pseudo-random sequence, with different initializations for different types of reference signals
- Cell-specific reference signals are transmitted in all downlink subframes in a cell supporting non-MBSFN transmission.
- There is one reference signal transmitted per downlink antenna port.
- Cell-specific reference signals are defined separately for antenna ports 0, 1, 2, and 3 as shown in Figure 7.12.
- Only the first two OFDM symbols can be used for cell-specific reference symbols. Therefore, in LTE a maximum of four antennas can be used while transmitting the cell specific reference signal.
- Cell specific reference signal are defined only for normal subcarrier spacing of $\Delta f = 15kHz$.

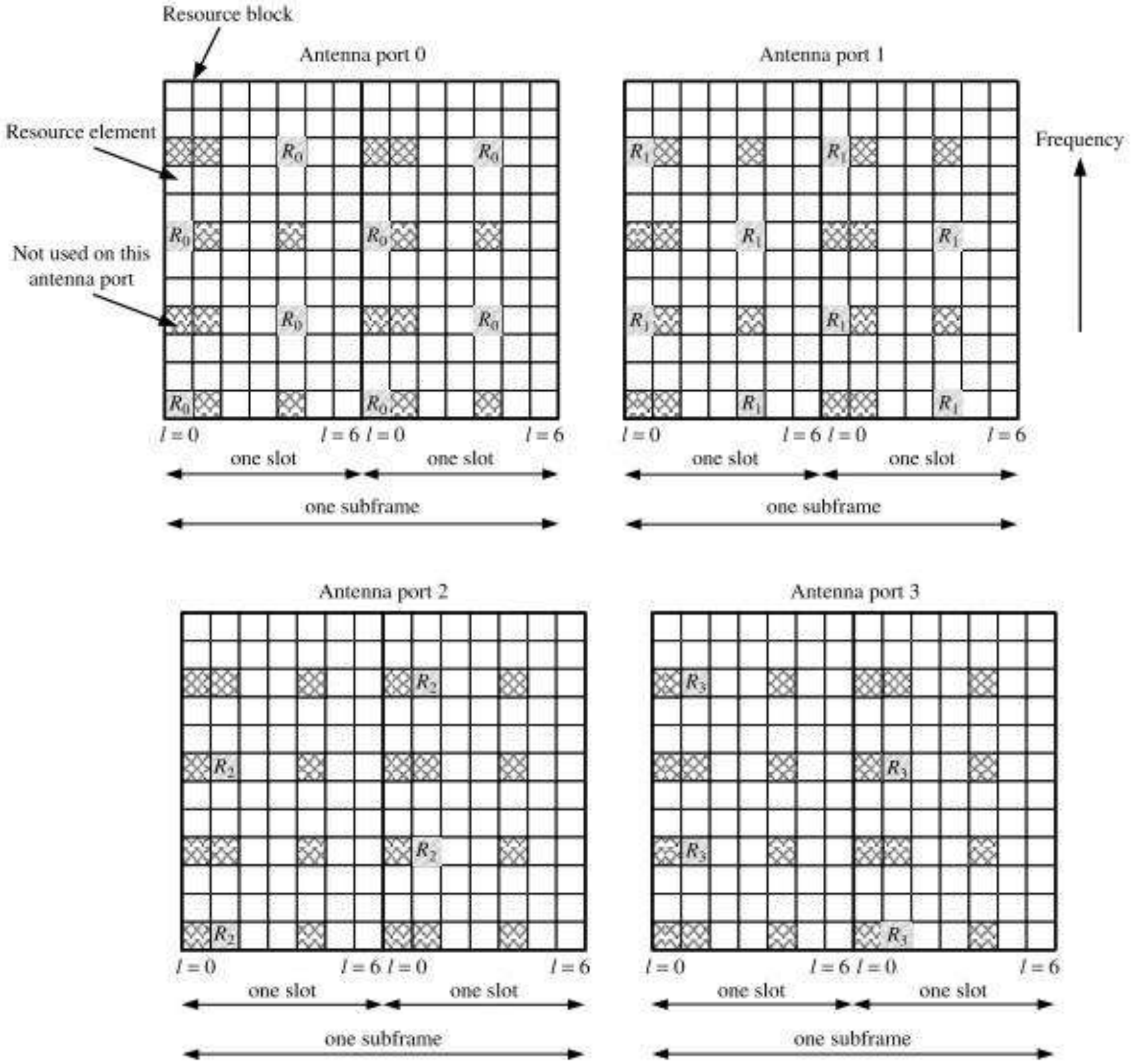


Figure 7.12 An example of mapping of downlink cell-specific reference signals, with four antenna ports and the normal CP. R_p denotes the resource element used for reference signal transmission on antenna port p.

- *Reference Signal(RS) mapping in time domain*:
  - For the antenna port p ∈ {0, 1}, the RS are inserted within the first and the third last OFDM symbols in each slot, which are the 1st and 5th OFDM symbols for the normal CP and the 1st and 4th OFDM symbols for the extended CP.
  - For p ∈ (2, 3), the RSs are only inserted in the 2nd OFDM symbol. So antenna ports 0 and 1 have twice as many reference symbols as antenna ports 2 and 3. This is to reduce the reference signal overhead but also causes an imbalance in the quality of the respective channel estimates.
- *Reference Signal(RS) mapping in time domain:*
  - The spacing between neighboring reference symbols in the same OFDM symbol is five subcarriers, that is, the reference symbols are transmitted every six subcarriers.
  - There is a staggering of three subcarriers between the 1st and 2nd reference symbols.

2. *MBSFN Reference Signals*

- MBSFN RSs are only transmitted in subframes allocated for MBSFN transmission, which is only defined for extended CP and transmitted on antenna port 4.
- *In the time domain*: For even-numbered slots, the RSs are inserted in the 3rd OFDM symbol for $\Delta f = 15kHz$ and in the second OFDM symbol for $\Delta f = 7.5\ kHz$. For odd-numbered slots, the reference symbols are inserted in the 1st and 5th OFDM symbols for $\Delta f = 15kHz$ and in the first and third OFDM symbols for $\Delta f = 7.5\,kHz$.
- *In the frequency domain*: The RSs are transmitted every two subcarriers for $\Delta f = 15\ kHz$ and every four subcarriers for $\Delta f = 7.5\ kHz$. In the 0th OFDM symbols, the reference symbols are transmitted from the 2nd and the 3rd subcarrier for $\Delta f = 15\ kHz$ and $\Delta f = 7.5\ kHz$.
- Based on these rules, an example of the resource mapping of MBSFN reference signals is shown in Figure 7.13 with the extended CP, and $\Delta f = 15\,kHz$.
- *Note:* The density of the MBSFN reference signal in the frequency domain is three times higher than that of the cell-specific reference signal.

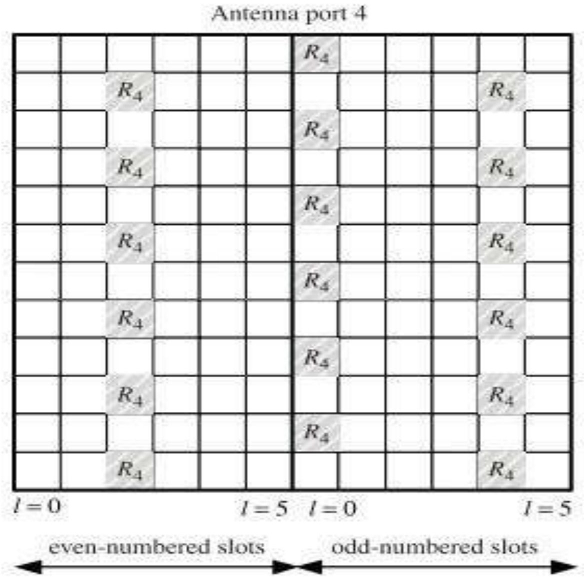


Figure 7.13 An example of mapping of MBSFN reference signals, with the extended CP and $\Delta f =$ 15kHz.

3. *UE- Specific Reference Signals*

- UE-specific reference signals support single-antenna-port transmission with beam forming for the PDSCH and are transmitted on antenna port 5.
- They are transmitted only on the resource blocks upon which the corresponding PDSCH is mapped.
- The UE-specific signal is not transmitted in resource elements in which one of the other physical signals or physical channels is transmitted.

- An example of resource mapping of UE-specific reference signals is shown in Figure 7.14 with the normal CP. In the even-numbered slots, the reference symbols are inserted in the fourth and seventh OFDM symbols: in the odd-number slots, the reference symbols are inserted in the third and sixth OFDM symbols. There is a frequency shift of two subearriers in neighboring reference symbols.

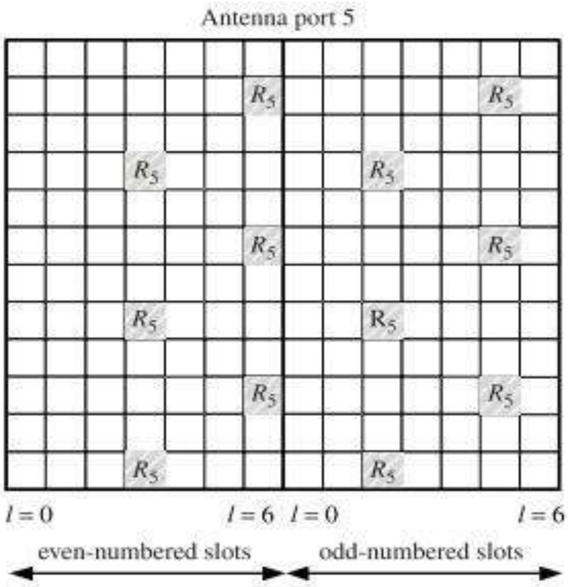


Figure 7.14 An example of mapping of UE-specific signals, with the normal CP.

7.6.2 Synchronization Signals

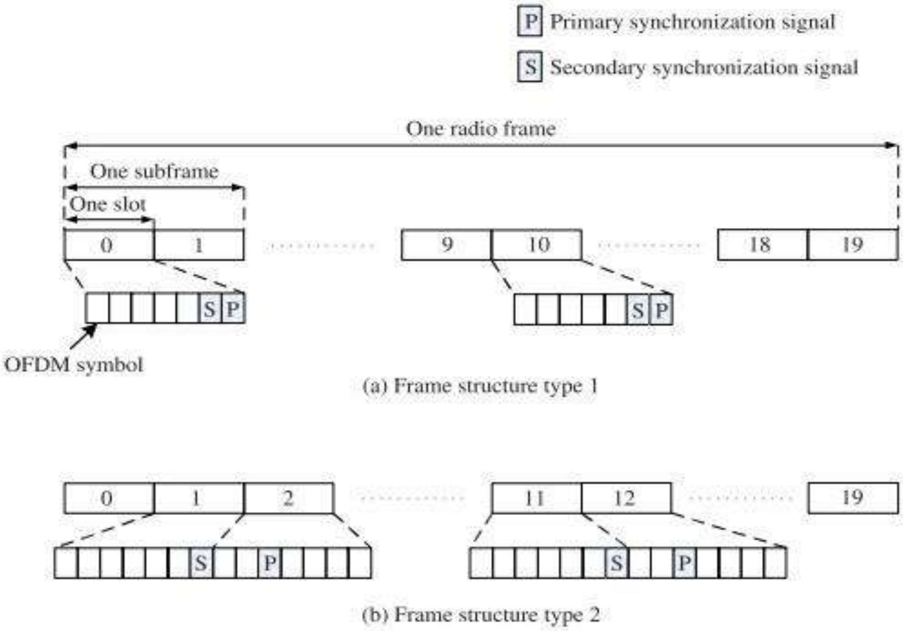
- The downlink synchronization signals are sent to facilitate the cell search procedure, during which process the time and frequency synchronization between the UE and the eNode-B is achieved and the cell ID is obtained.
- There are a total of 504 unique physical-layer cell IDs, which are grouped into 168 physical-layer cell-ID groups. A physical-layer cell ID is uniquely defined as:

$$^{(1)}_{ID} = 3N^{(1)}_{ID} + N^{(2)}_{ID}$$

Where $^{(1)}_{ID}$ = 0,1 … … 167 represents the physical-layer cell-ID group and $N^{(2)}_{ID}$ = 0, 1, 2 represents the physical-layer ID within the cell-ID group. Each cell is assigned a unique physical-layer cell ID.

- The synchronization signals are classified as
  1. *Primary synchronization signals (P-SS):* P-SS signals identify the *symbol timing* and the cell ID index $^{(2)}_{ID}$
  2. *Secondary synchronization signals(S-SS)*. These signals are used for detecting the cell-ID group index $N^{(1)}_{ID}$and the frame timing.
- The secondary synchronization signal can only be detected after detecting the primary synchronization signal.

- The synchronization signals are designed in such a way to make the cell search procedure fast and of low complexity.
- The sequence used for the primary synchronization signal is generated from a frequency-domain Zadoff-Chu sequence.
- The Zadoff-Chu sequence possesses the Constant Amplitude Zero Auto-Correlation (CAZAC) property, which means low peak-to-average power ratio (PAPR). This property is desirable for synchronization signals as it improves coverage, which is an important design objective.
- Both primary and secondary synchronization signals are transmitted on the 62 sub-carriers centered on the DC subcarrier, with five reserved subcarriers on either side in the frequency domain, so there are a total of 72 subcarriers occupied by synchronization signals, corresponding to the narrowest bandwidth supported by LTE (1.4MHz).
- In the time domain, both primary and secondary synchronization signals are transmitted twice per 10 ms in predefined slots.
- For frame structure type 1, the primary and secondary synchronization signals are mapped to the last and the OFDM symbols in slot 0 and 10.
- For frame structure type 2, the primary synchronization signal is mapped to the third OFDM symbol in slot 2 and 12 and the secondary synchronization signal is mapped to the last OFDM symbol in slot 1 and 11.
- The difference in the location of the synchronization signal enables the UE to detect the duplex mode of the cell.
- The resource mapping for synchronization signals is illustrated in Figure 7.15.



(a) Frame structure type 1

(b) Frame structure type 2

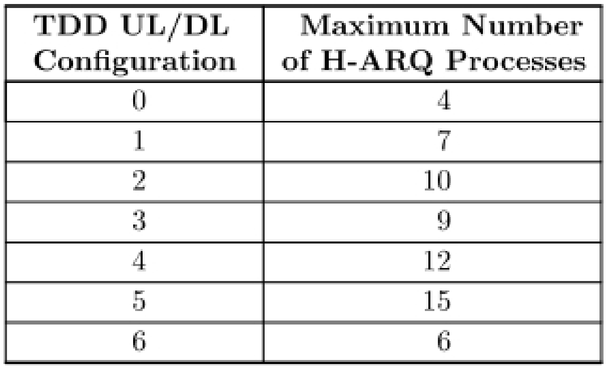
Figure 7.15 The mapping of primary and secondary synchronization signals to OFDM symbols for frame structure type 1 and type 2, with the normal CP. 'P' and 'S' denote primary and secondary synchronization signals, respectively.

7.7 H-ARQ in the Downlink

Brief the concept of H-ARQ in the down link.

- It is an acknowledgement processes in LTE for a received error packet.
- In the case of LTE both Type I Chase Combining (CC) H-ARQ and Type II Incremental Redundancy (IR) H-ARQ schemes have been defined.
- The H-ARQ operation is part of the MAC layer, while the PHY layer handles soft combining.
- ***At the receiver:*** Turbo decoding is first applied on the received code block. If this is a retransmission, which is indicated in the DCI, the code block will be combined with the previously received versions for decoding. If there is no error detected in the output of the decoder, an ACK signal is fed back to the transmitter through the PUCCH physical channel and the decoded block is passed to the upper layer; otherwise, an NAK signal is fed back and the received code block is stored in the buffer for subsequent combining.
- ***At the transmitter:*** For each (re)transmission, the same turbo-encoded data is transmitted with different puncturing, so each of these (re)transmissions has a different redundancy version and each is self-decodable. Puncturing is performed during the rate matching process. The rate matcher can produce four different redundancy versions of the original coded block. H-ARQ transmissions are indexed with the redundancy version rv_{idx}, which indicates whether it is a new transmission (rv_{idx} =0) or the $rv_{idx}th$ retransmission (rv_{idx} = 1, 2, or 3).
- Time interval between two successive H-ARQ transmissions, which is typically 8 ms in LTE.
- N-channel Stop-and-Wait protocol is used for downlink H-ARQ operation. An N-channel Stop-and-Wait protocol consists of N parallel H-ARQ processes. When one or more of the processes are busy waiting for the H-ARQ ACK /NAK, the processes that are free can be used to transmit other transport blocks.
- The maximum number of H-ARQ processes in the downlink is determined by the UL/DL configuration, specified in Table 7.17, which ranges from 4 to 15.

Table 7.17 Maximum Number of Downlink H-ARQ Processes for TDD



TDD UL/DL Configuration	Maximum Number of H-ARQ Processes
0	4
1	7
2	10
3	9
4	12
5	15
6	6

- Figure 7.16 an example of a 10-msec frame with eight H-ARQ processes.

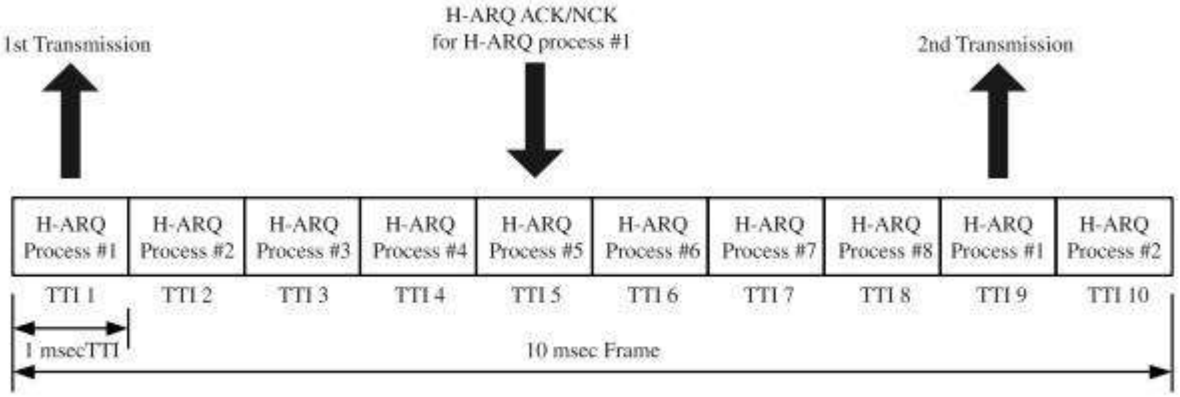


Figure 7.16 An example of a 10-msec frame with eight H-ARQ processes. The H-ARQ process 1 is transmitted in the first TTI, for which the H-ARQ ACK/NAK is received in the 5-th TTI, and then the H-ARQ process 1 is transmitted again in the 9-th TTI.

- The H-ARQ process 1 is transmitted in the first TTI, for which the H-ARQ ACK/NAK is received in the 5-th TTI, and then the H-ARQ process 1 is transmitted again in the 9-th TTI.
- Each H-ARQ process is associated with an 11-ARQ process ID.
- When spatial multiplexing is used, both transport blocks are associated with the same H-ARQ process.
- Figure 7.16 shows a 10 msec frame with TTI index 1 transmitting the H-ARQ process 1, TTI index 2 transmitting the H-ARQ process 2, and so on.
- The H-ARQ ACK/NAK for the 11-ARQ process 1 is received in TTI index 5.2 , Then in TTI index 9 the H-ARQ process 1 is transmitted again, either a new transmission if an ACK is received or a retransmission if an NAK is received.
- LTE downlink applies the asynchronous H-ARQ protocol, where the H-ARQ processes can be transmitted in any order without fixed timing. Therefore, in the example in Figure 7.16, the retransmission of H-ARQ process 1 does not necessarily occur in the 9th TTI.
- The asynchronous H-ARQ makes it possible to reflect channel quality measurements at the instance of retransmission, which is able to provide a higher throughput with re-scheduling or changing the modulation and coding scheme, called adaptive RQ.
- In addition, asynchronous operation makes it possible for the eNode-B to avoid potential collision of H-ARQ retransmissions with other high priority scheduled transmissions such as persistent scheduling.
- Meanwhile, the asynchronous 11-ARQ requires more overhead, as the receiver does not know ahead of time what is being transmitted and when the retransmission occurs.

- To support asynchronous H-ARQ in the downlink, PDCCH contains fields indicating the H-ARQ process number and the current redundancy version (see Table 7.11 for an example with DCI format 1).
- The maximum number of H-ARQ retransmissions of each transport block is configured by the Radio Resource Control (RRC) layer.
- When this maximum number is reached without a successful transmission of the transport block or the transmission is in error due to the error in H-ARQ-ACK signaling, a Radio Link Control (RLC) layer ARQ protocol will be triggered to handle the error event.

Write a short note on i) H- ARQ Indicator (HI). ii) Broadcast channels (PBCH) iii) Multicast channels

Explain the coding and modulation process involved in DCI.

REAL – TIME SYSTEMS

Module – 1

Introduction to Real – Time Systems:

Historical Background, RTS Definition, Classification of Real – Time Systems, Time constraints, Classification of programs.

Concepts of Computers Control:

Introduction, Sequence Control, Loop Control, Supervisory Control, Centralized Computer Control, Distributed System, Human-Computer interface, Benefits of Computer Control Systems.

Module - 2

Computer Hardware Requirements for RTS:

Introduction, General Purpose Computer, Single Chip Microcontroller, Specialized Processors, Process –Related Interfaces, Data Transfer Techniques, Communications, Standard Interface.

Module- 3

Languages For Real –Time Applications:

Introduction, Syntax Layout and Readability, Declaration and Initialization of Variables and Constants, Modularity and Variables, Compilation , Data Type, Control Structure, Exception Handling, Low –Level Facilities, Co routines, Interrupts and Device Handling, Concurrency, Real – Time Support, Overview of Real –Time Languages.

PART –B

Module-4

Operating Systems:

Introduction, Real –Time Multi –Tasking OS, Scheduling Strategies, Priority Structures, Task Management, Scheduler and Real –Time Clock Interrupt Handles, Memory Management ,Code Sharing, Resource control, Task Co-operation and Communication, Mutual Exclusion

Module-5

Design of RTSS General Introduction:

Introduction, Specification documentation, Preliminary design, Single –Program Approach, Foreground /Background, Multi- Tasking approach, Mutual Exclusion Monitors.

RTS Development Methodologies:

Introduction, Yourdon Methodology, Requirement definition For Drying Oven, Ward and Mellor Method, Hately and Pirbhai Method.

Text Books:

1. **Real –Time Computer control –An Introduction**, Stuart Bennet, 2nd Edn. Pearson Education 2005.

Reference: Books:

1. **Real-Time Systems Design and Analysis**, Phillip. A. Laplante, Second Edition, PHI, 2005.

2. **Embedded Systems**, Raj kamal, Tata MC Graw Hill, INDIA, 2005.

CONTENTS

MODULE – 1

Introduction to Real – Time Systems

Historical Background, RTS Definition, Classification of Real – Time Systems, Time constraints, Classification of programs.

Recommended book for reading:

1. **Real –Time Computer control –An Introduction**, Stuart Bennet, 2nd Edn. Pearson Education 2005.
2. **Real-Time Systems Design and Analysis**, Phillip. A. Laplante, Second Edition, PHI, 2005.

Introduction to Real –Time Systems.

1.1 Historical Background:

The origin of the term Real –Time Computing is unclear. It was probably first used either with project whirlwind, a flight simulator developed by IBM for the U.S. Navy in 1947, or with SAGE, the Semiautomatic Ground Environment air defense system developed for the U.S. Air force in the early 1950s. Modern real-time systems, such as those that control Nuclear Power stations, Military Aircraft weapons systems, or Medical Monitoring Equipment, are complex and they exhibit characteristics of systems developed from the 1940s through the 1960s. Moreover, today’s real time systems exist because the computer industry and systems requirements grew.

The earliest proposal to use a computer operating in real time as part of a control system was made in a paper by Brown and Campbell in 1950. It shows a computer in both the feedback and feed forward loops. The diagram is shown below:

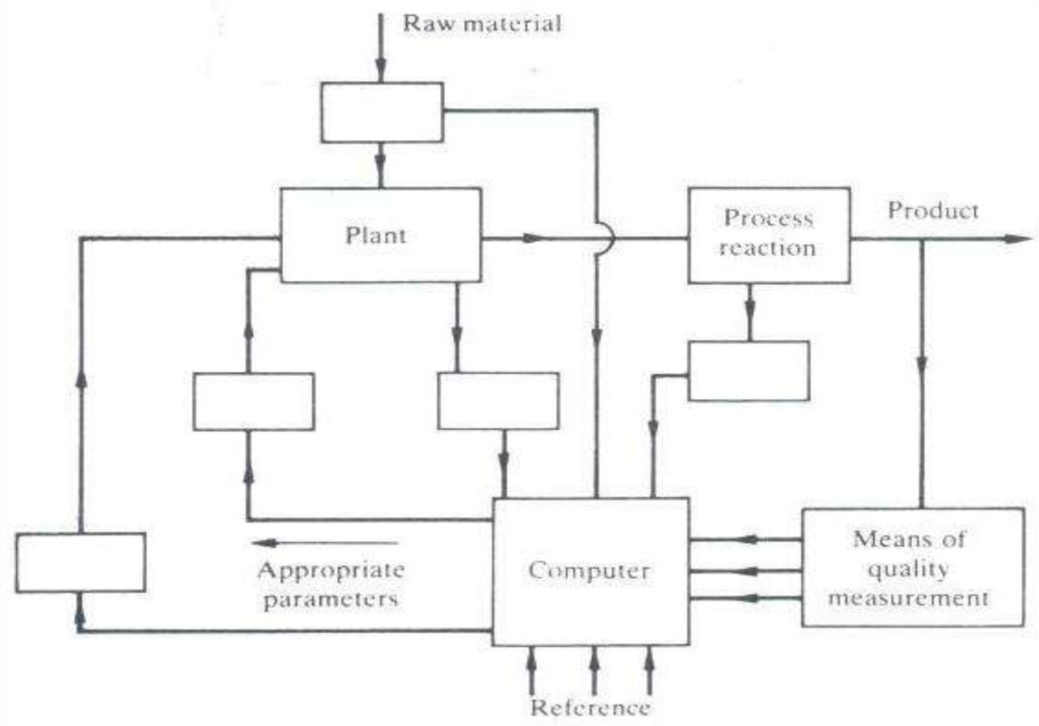


Figure: Computer used in control of plant.

The first digital Computers developed specifically for real time control were for airborne operation, and in 1954 a digitrac digital computer was successfully used to provide an automatic flight and weapons control system.

The application of digital computers to industrial control began in the late 1950s.

* The first industrial installation of a computer system was in September 1958. When the Louisiana Power and Light Company installed a Day Storm Computer system for plant monitoring at their power station in sterling, Louisiana.

* The first industrial Computer Control installation was made by the Texaco Company who installed an RW-300 (Ramo -Wooldridge Company) system at their Port Arthur refinery in Texas.

* During 1957-8 the Monsanto Chemical Company, in co-operation with the Ramo-Wooldridge company, studied the possibility of using computer control and in October 1958 decided to implement a scheme on the ammonia plant at luling, Louisiana.

* The same system was installed by the B.F. Goodrich Company on their acrylanite plant at Calvert City, Kentucky, in 1959-60.

* The first direct digital control (DDC) computer system was the Ferranti Argus 200 system installed in November 1962 at the ICI ammonia – soda plant at Fleetwood Lancashire.

The computers used in the early 1960s combined magnetic core memories and drum stores, the drum eventually giving way to hard disk drives. They included the General Electric 4000 series, IBM 1800, CDC 1700, Foxboro Fox 1 and 1A, the SDS and Xerox SIGMA Series, Ferranti Argus and Elliot Automation 900 series. The attempt to resolve some of the problems of the early machines led to an increase in the cost of systems.

The consequence of the generation of further problems particularly in the development of the software. The increase in the size of the programs meant that not all the code could be stored in core memory; provision to be made for the swapping of code between the drum memory and core. The solution appeared to lie in the development general purpose real-time operating systems and high –level languages.

In the late 1960s real time operating system were developed and various PROCESS FORTRAN Compilers made their appearance. The problems and the costs involved in attempting to do everything in one computer led users to retreat to smaller system for which the newly developing minicomputer (DEC PDP-8, PDP-11, Data General Nova, Honey well 316 etc.) was to prove ideally suited.

1.2 REAL-TIME SYSTEM DEFINITION:

Real- time processing normally requires both parallel activities and fast response. In fact, the term ‘real –time’ is often used synonymously with ‘multi – tasking’ or ‘multi- threading’.

Although there is no clear dividing line between real-time and non-real-time Systems, there are a set of distinguishing features:

The oxford Dictionary of computing offers the definition:

Any system in which the time at which the output is produced is significant. This usually because the input corresponded to some movement in the physical world, and output has to relate to that same movement. The lag from input time to output time must be sufficiently small for acceptable timeliness.

This definition covers a very wide range of systems; for examples, from workstations running under the UNIX operating system from which the user expects to receive a response within a few seconds through to aircraft engine control systems which must respond within a specified time and failure to do so could cause the loss of control and possibly the loss of many lives.

Latter type of system cooling (1991) offers the definition:

Real- time systems are those which must produce correct responses within a definite time limit.

An alternative definition is:

A real- time system read inputs from the plant and sends control signals to the plant at times determined by plant operational considerations.

We can therefore define a real –time program as:

A program for which the correctness of operation depends both on the logical results of the computation and the time at which the results are produced. One of the classification schema to identify real-time.

Timing:

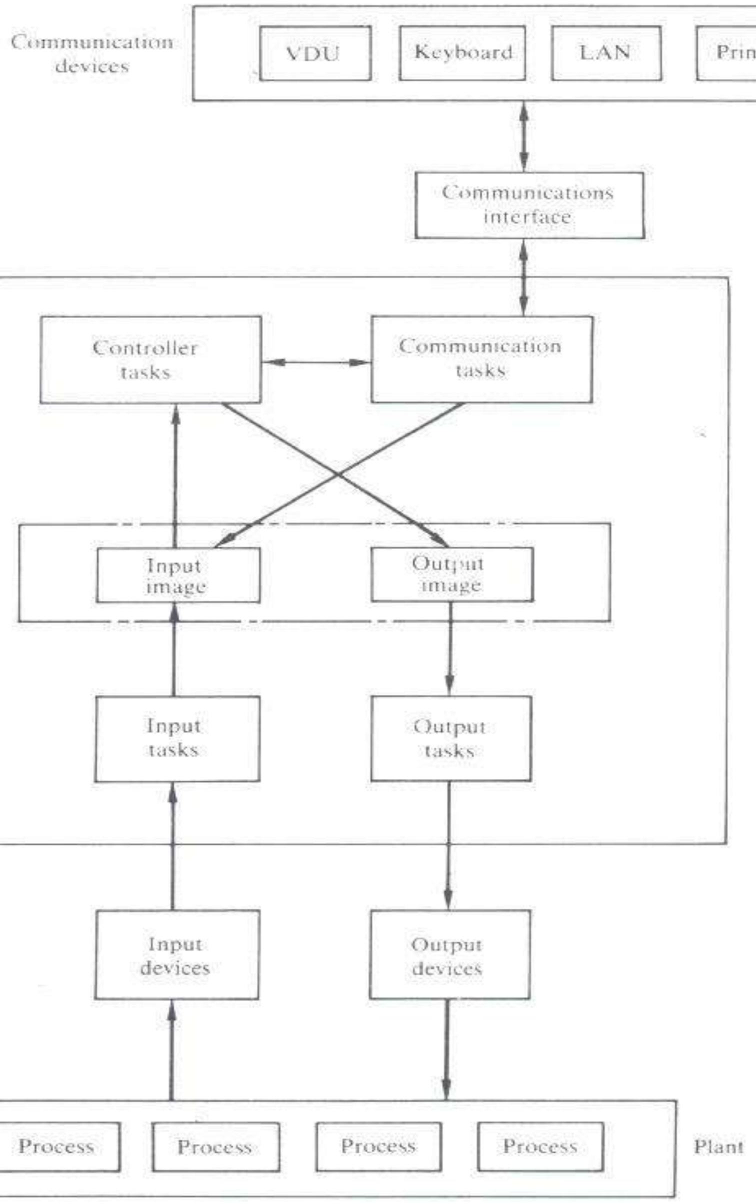
The most common definition of a real-time system involves a statement similar to 'Real-time systems are required to compute and deliver correct results within a specified period of time'. Does this mean that a non-real-time system such as a payroll program, could point salary cheques two years late, and be forgiven because it was not a real-time system? Hardly so! Obviously there are time constraints on non-real-time systems too.

1.3 CLASSIFICATION OF REAL-TIME SYSTEM

A common feature of real-time systems and embedded computers is that the computer is connected to the environment within which it is working by a wide range of interface device and receives and sends a variety of stimuli. For example, the plant input, plant output, and communication tasks shown in figure:

In below figure one common feature they are connected by physical devices to processor which are external to computer. External processes all operate in their own time-scale and the computer is said to operate in real time if actions carried out in the computer relate to the time –scales of the external processes. Synchronization between the external processes and the internal actions (tasks) carried out by the computer may be defined in term of the passage of time, or the actual time of day, in which case the system is said to be clock based the operations are carried out according to a time schedule.

Other category, interactive, in which the relationship between the actions in the computer and the system is much more loosely defined. The control tasks, although not obviously and directly connected to the external environment, also need to operate in real -time, since time is usually involved in determining the parameters of the algorithms.



1.3.1 CLOCK – BASED TASKS (CYCLIC, PERIODIC):

Clock – based tasks are typically referred to as cyclic or periodic tasks where the terms can imply either that the task is to run once per time period T (or cycle time T), or is to run at exactly T unit intervals.

The completion of the operations within the specified time is dependent on the number of operations to be performed and the speed of the computer.

Synchronization is usually obtained by adding a clock to the computer system- referred as a real-time clock that uses signal from this clock to interrupt the operations of the computer at some predetermined fixed time interval.

For example in process plant operation, the computer may carry out the plant input, plant output and control tasks in response to the clock interrupt or, if the clock interrupt has been set at a faster rate than the sampling rate, it may count each interrupt until it is time to run the tasks.

In larger system the tasks may be subdivided into groups for controlling different parts of the plant and these may need to run a different sampling rate. A tasks or process comprises some code, its associated data and a control block data structure which the operating system uses to define and manipulate the task.

1.3.2 EVENT – BASED TASKS (APERIODIC):

Events occurring at non-deterministic interval and event-based tasks are frequently referred as aperiodic tasks. Such tasks may have deadlines expressed in term of having start times or finish times (or even both).

For example, a task may be required to start within 0.5 seconds of an event occurring, or alternatively it may have to produce an output (end time) within 0.5 seconds of the events. For many system where actions have to be performed not at particular times or time intervals but in response to some event.

Examples: Turning off a pump or closing a value when the level of a liquid in a tank reaches a predetermined valve; or switching a motor off in response to the closure of a micro switch indicating that some desired position has been reached.

Event based systems are also used extensively to indicate alarm conditions and initiate alarm actions.

1.3.3 INTERACTIVE SYSTEM:

Interactive systems probably represent the largest class of real-time systems and cover such systems as automatic bank tellers; reservation systems for hotels, airlines and car rental companies; computerized tills, etc. The real-time requirement is usually expressed in terms such as 'the average response time must not exceed ... ‘

For example, an automatic bank teller system might require an average response time not exceeding 20 seconds. Superficially this type of system seems similar to the event-based system in that it apparently responds to a signal from the plant (in this case usually a person), but it is different because it responds at a time determined by the internal state of the computer and without any reference to the environment. An automatic bank teller does not know that you are about to miss a train, or that it is raining hard and you are getting wet: its response depends on how busy the communication lines and central computers are (and of course the *wire* of your account).

Many interactive systems give the impression that they are clock based in that they are capable of displaying the date and time; they do indeed have a real-time clock which enables them to keep track of time.

1.4 TIME CONSTRAINTS:

Real time systems are divided into two categories:

- Hard real-time: these are systems that must satisfy the deadlines on each and every occasion.
- Soft real-time: these are systems for which an occasional failure to meet a deadline does not comprise the correctness of the system.

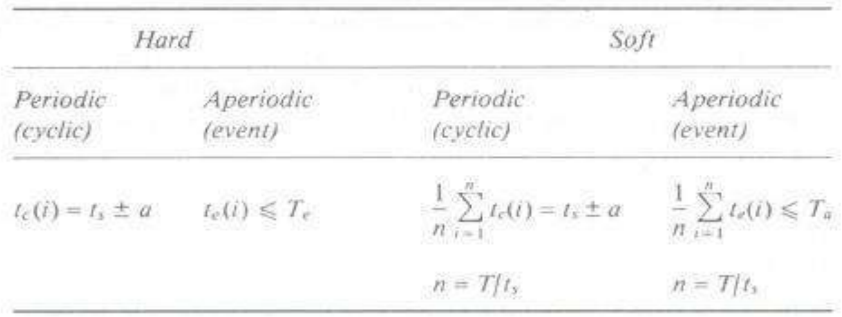
A typical example of a hard real-time control system is the temperature control loop of the hot-air blower system described above. In control terms, the temperature loop is a *sampled data* system. Design, of a suitable control algorithm for this system involves the choice of the sampling interval Ts. If we assume that a suitable sampling interval is 10 ms, then at 10 ms intervals the input value must be read, the control calculation carried out and the output value calculated, and the output value sent to the heater drive.

As an example of hard time constraints associated with event-based tasks let us assume that the hot-air blower is being used to dry a component which will be damaged if exposed to temperatures greater than 50°C for more than 10 seconds. Allowing for the time taken for the air to travel from the heater to the outlet and the cooling time of the heater element - and for a margin of safety - the alarm response requirement may be, say, that overt temperature should be detected and the heater switched off within seven seconds of the over temperature occurring. The general form of this type of constraint is that the computer must respond to the event within some specified maximum time.

An automatic bank teller provides an example of a system with a soft time constraint. A typical system is event initiated in that it is started by the customer placing their card in the machine. The time constraint on the machine responding will be specified in terms of an average response time of, say, 10 seconds, with the average being measured over a 24 hour period. (Note that if the system has been carefully specified there will also be a maximum time; say 30 seconds, within which the system should respond.) The actual response time will vary: if you have used such a system you will have learned that the response time obtained between 12 and 2 p.m. on a Friday is very different from that at 10 a.m. on a Sunday.

A hard time constraint obviously represents a much more severe constraint on the performance of the system than a soft time constraint and such systems present a difficult challenge both to hardware and to software designers. Most real-time systems contain a mixture of activities that can be classified as clock based, event based, and interactive with both hard and soft time constraints (they will also contain activities which are not real time). A system designer will attempt to reduce the number of activities (tasks) that are subject to a hard time constraint.

Formally the constraint is defined as follows:



Hard		*Soft*	
Periodic (cyclic)	*Aperiodic (event)*	*Periodic (cyclic)*	*Aperiodic (event)*
$t_c(i) = t_s \pm a$	$t_e(i) \leqslant T_e$	$\frac{1}{n}\sum_{i=1}^{n} t_c(i) = t_s \pm a$	$\frac{1}{n}\sum_{i=1}^{n} t_e(i) \leqslant T_a$
		$n = T/t_s$	$n = T/t_s$

tc (i) the interval between the i and i-I cycles,

te (i) the response time to the ith occurrence of event e,

ts the desired periodic (cyclic) interval, Te the maximum permitted response time to event e,

Ta the average permitted response time to event e measured over some time interval T,

n the number of occurrences of event e within the time interval T, or the number of cyclic repetitions during the time interval T,

a a small timing tolerance.

For some systems and tasks the timing constraints may be combined in some form or other, or relaxed in some way.

1.5 CLASSIFICATION OF PROGRAMS:

The importance of separating the various activities carried out by computer control systems into real-time and non-real-time tasks, and in subdividing real-time tasks into the two different types, arises from the different levels of difficulty of designing and implementing the different types of computer program. Experimental studies have shown clearly that certain types of program, particularly those involving real time and interface operations, are substantially more difficult to construct than, for instance, standard data processing programs (Shooman, 1983; Pressman, 1992).The division of software into small, coherent *modules* is an important design technique and one of the guidelines for module division that we introduce is to put activities with different types of time constraints into separate modules.

Theoretical work on mathematical techniques for proving the correctness of a program, and the development of formal specification languages, such as 'z' and VOM, has clarified the understanding of differences between different types of program. Pyle (1979), drawing on the work of Wirth (1977), presented definitions identifying three types of programming:

- Sequential;
- Multi-tasking; and
- Real-time.

The definitions are based on the kind of arguments which would have to be made in order to verify, that is to develop a formal proof of correctness for programs of each type.

1.5.1 SEQUENTIAL:

In classical sequential programming *actions* are strictly ordered as a time sequence: the behavior of the program depends only on the effects of the individual *actions* and their *order;* the time taken to perform the action is not of consequence. Verification, therefore, requires two kinds of argument:

1. That a particular statement defines a stated action; and
2. That the various program structures produce a stated sequence of events.

1.5.2 MULTI-TASKING:

A multi-task program differs from the classical sequential program in that the actions it is required to perform are not necessarily disjoint in time; it may be necessary for several actions to be performed in parallel. Note that the *sequential relationships* between the actions may still be important. Such a program may be built from a number of parts (processes or tasks are the names used for the parts) which are themselves partly sequential, but which are executed concurrently and which communicate through shared variables and synchronization signals.

Verification requires the application of arguments for sequential programs with some additions. The task (processes) can be verified separately only if the constituent variables of each task (process) are distinct. If the variables are shared, the potential concurrency makes the effect of the program unpredictable (and hence not capable of verification) unless there is some further rule that governs the sequencing of the several actions of the tasks (processes). The task can proceed at any speed: the correctness depends on the actions of the synchronizing procedure.

1.5.3 REAL-TIME:

A real-time program differs from the previous types in that, in addition to its actions not necessarily being disjoint in time, the sequence of some of its actions is not determined by the designer but by the environment - that is, by events occurring in the outside world which occur in real time and without reference to the internal operations of the computer. Such events cannot be made to conform to the intertask synchronization rules.

A real-time program can still be divided into a number of tasks but communication between the tasks cannot necessarily wait for a synchronization signal: the environment task cannot be delayed. (Note that in process control applications the main environment task is usually that of keeping real time, that is a real-time clock task. It is this task which provides the timing for the scanning tasks which gather information from the outside world about the process.) In real-time programs, in contrast to the two previous types of program, the *actual time taken* by an action is an essential factor in the process of verification. We shall assume that we are concerned with real-time software and references to sequential and multi-tasking programs should be taken to imply that the program is real time. Non-real-time programs will be referred to as standard program.

Consideration of the types of reasoning necessary for the verification of programs is important, not because we, as engineers, are seeking a method of formal proof, but because we are seeking to understand the factors which need to be considered when designing real-time software. Experience shows that the design of real-time software is significantly more difficult than the design of sequential software. The problems of real-time software design have not been helped by the fact that the early high-level languages were sequential in nature and they did not allow direct access to many of the detailed features of the computer hardware.

As a consequence, real-time features had to be built into the operating system which was written in the assembly language of the machine by teams of specialist programmers. The cost of producing such operating systems was high and they had therefore to be general purpose so that they could be used in a wide range of applications in order to reduce the unit cost of producing them. These operating systems could be *tailored,* that is they could be reassembled to exclude or include certain features, for example to change the number of tasks which could be handled, or to change the number of input/output devices and types of device. Such changes could usually only be made by the supplier.

Excepted Question:

1. Explain the difference between a real-time program and a non-real-time program.
   Why are real-time programs more difficult to verify than non-real-time programs?

2 To design a computer-based system to control all the operations of a retail petrol
(gasoline) station (control of pumps, cash receipts, sales figures, deliveries, etc.).
What type of real-time system would you expect to use?

3. Classify any of the following systems as real-time?
   In each case give reasons for your answer and classify the real-time systems
   as hard or soft.
   (a) A simulation program run by an engineer on a personal computer.

(b) An airline seat-reservation system with on-line terminals.

(c) A microprocessor-based automobile ignition and fuel injection system.

(d) A computer system used to obtain and record measurements of force and strain from a tensile strength testing machine.

e) An aircraft autopilot.

4 An automatic bank teller works by polling each teller in turn. Some tellers are located outside buildings and others inside. How the polling system could be organized to ensure that the waiting time at the outside locations was less than at the inside locations?

5 .Explain the precision required for the analog-to-digital and digital-to-analog converters taking hot-air blower as an example?

MODULE– 1

Concepts of Computers Control

Introduction, Sequence Control, Loop Control, Supervisory Control, Centralized Computer Control, Distributed System, Human-Computer interface, Benefits of Computer Control Systems.

Recommended book for reading:

1. **Real –Time Computer control –An Introduction**, Stuart Bennet, 2nd Edn. Pearson Education 2005.
2. **Real-Time Systems Design and Analysis**, Phillip. A. Laplante, Second Edition, PHI, 2005.

2.1 Concepts of computers control:

Introduction:

Computers are now used in so many different ways that we could take it up by simply describing various applications. However, when we examine the applications closely we find that there are many common features. The basic features of computer control systems are illustrated in the following sections using examples drawn from industrial process control. In this field applications are typically classified under the following headings:

- Batch;
- Continuous; and
- Laboratory (or test).

The categories are not mutually exclusive: a particular process may involve activities which fall into more than one of the above categories; they are, however, useful for describing the general character of a particular process.

BATCH:

The term *batch* is used to describe processes in which a sequence of operations are carried out to produce a quantity of a product - the batch - and in which the sequence is then repeated to produce further batches. The specification of the product or the exact composition may be changed between the different runs.

A typical example of batch production is rolling of sheet steel. An ingot is passed through the rolling mill to produce a particular gauge of steel; the next ingot may be either of a different composition or rolled to a different thickness and hence will require different settings of the rolling mill.

An important measure in batch production is *set-up* time (or *change-over* time), that is, the time taken to prepare the equipment for the next production batch. This is *wasted* time in that no output is being produced; the ratio between *operation* time (the time during which the product is being produced) and set-up time is important in determining a suitable batch size.

In mechanical production the advent of the NC (Numerically Controlled) machine tool which can be set up in a much shorter time than the earlier automatic machine tool has led to a reduction in the size of batch considered to be economic.

CONTINUOUS:

The term *continuous* is used for systems in which production is maintained for long periods of time without interruption, typically over several months or even years. An example of a continuous system is the catalytic cracking of oil in which the crude oil enters at one end and the various products - fractionates – are removed as the process continues. The ratio of the different fractions can be changed but this is done without halting the process.

Continuous systems may produce batches, in that the product composition may be changed from time to time, but they are still classified as continuous since the change in composition is made without halting the production process.

A problem which occurs in continuous processes is that during change-over from one specification to the next, the output of the plant is often not within the product tolerance and must be scrapped. Hence it is financially important that the change be made as quickly and smoothly as possible. There is a trend to convert processes to continuous operation - or, if the whole process cannot be converted, part of the process.

For example, in the baking industry bread dough is produced in batches but continuous ovens are frequently used to bake it whereby the loaves are placed on a conveyor which moves slowly through the oven. An important problem in mixed mode systems, that is systems in which batches are produced on a continuous basis, is the tracking of material through the process; it is obviously necessary to be able to identify a particular batch at all times.

LABORATORY SYSTEMS:

Laboratory-based systems are frequently of the operator-initiated type in that the computer is used to control some complex experimental test or some complex equipment used for routine testing. A typical example is the control and analysis of data from a vapour phase chromatograph.

Another example is the testing of an audiometer, a device used to lest hearing. The audiometer has to produce sound levels at different frequencies; it is complex in that the actual level produced is a function of frequency since the sensitivity of the human ear varies with frequency. Each audiometer has to be tested against a sound-level meter and a test certificate produced. This is done by using a sound-level meter connected to a computer and using the output from the computer to drive the audiometer through its frequency range. The results printed out from the test computer provide the test certificate.

As with attempts to classify systems as batch or continuous so it can be difficult at times to classify systems solely as laboratory. The production of steel using the electric arc furnace involves complex calculations to determine the appropriate mix of scrap, raw materials and alloying additives. As the melt progresses samples of the steel are taken and analyzed using a spectrometer. Typically this instrument is connected to a computer which analyses the results and calculates the necessary adjustment to the additives. The computer used may well be the computer which is controlling the arc furnace itself.

In whatever way the application is classified the activities being carried out will include:

- Data acquisition;
- Sequence control;
- Loop control (DDC);
- Supervisory control;
- Data analysis;
- Data storage; and

- Human-computer interfacing (HCI).

- Efficiency of operation;
- Ease of operation;
- Safety;
- Improved products;
- Reduction in waste;
- Reduced environmental impact; and
- A reduction in direct labour.

GENERAL EMBEDDED SYSTEMS:

In the general range of systems which use embedded computers – from domestic appliances, through hi-fi systems, automobile management systems, intelligent instruments, active control of structures, to large flexible manufacturing systems and aircraft control systems - we will find that the activities that are carried out in the computer and the objectives of using a computer are similar to those listed above. The major differences will lie in the balance between the different activities, the time-scales involved, and the emphasis given to the various objectives.

2.2 SEQUENCE CONTROL:

Although sequence control occurs in some part of most systems it often predominates in batch systems and hence a batch system is used to illustrate it. Batch systems are widely used in the food processing and chemical industries where the operations carried out frequently involve mixing raw materials, carrying out some process, and then discharging the product. A typical reactor vessel for this purpose is shown in Figure 2.1 below.

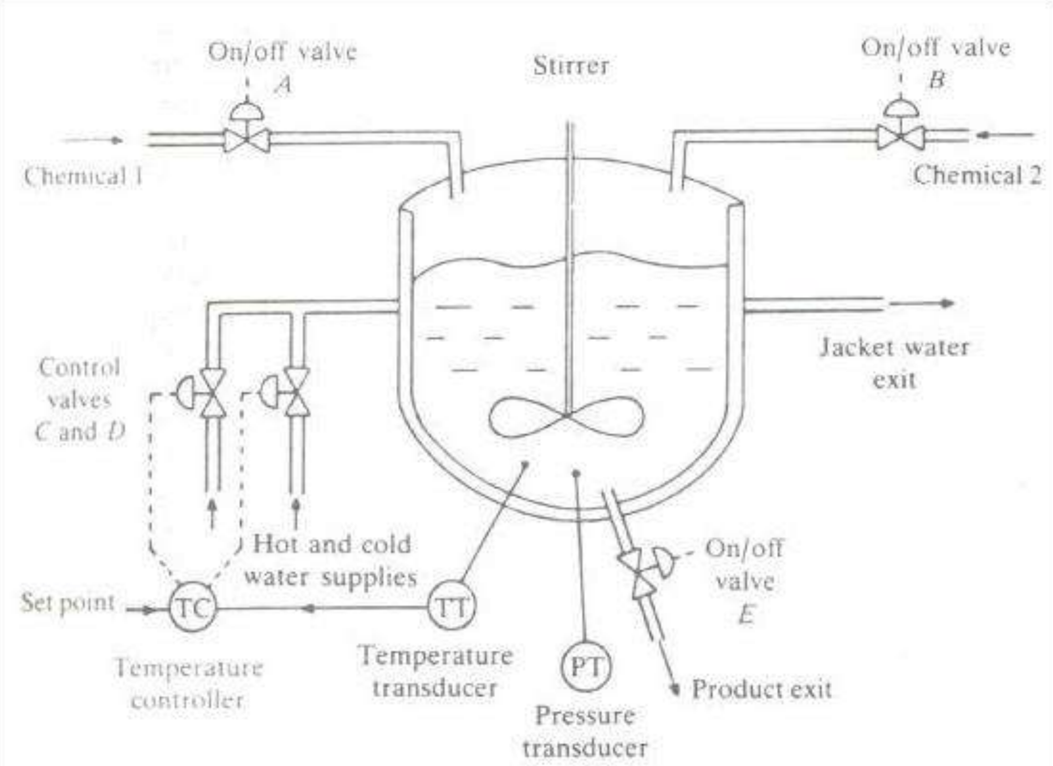


Figure2.1: A simple chemical reactor vessel

A chemical is produced by the reaction of two other chemicals at a specified temperature. The chemicals are mixed together in a sealed vessel (the reactor) and the temperature of the reaction is controlled by feeding hot or cold water through the water jacket which surrounds the vessel.

The water flow is controlled by adjusting valves C and D. The flow of material into and out of the vessel is regulated by the valves A, Band E. The temperature of the contents of the vessel and the pressure in the vessel are monitored.

The procedure for the operation of the system may be as follows:

1. Open valve A to charge the vessel with chemical 1.
2. Check the level of the chemical in the vessel (by monitoring the pressure in the vessel); when the correct amount of chemical has been admitted, close valve A.
3. Start the stirrer to mix the chemicals together.
4. Repeat stages 1 and 2 with valve B in order to admit the second chemical.
5. Switch on the three-term controller and supply a set point so that the chemical mix is heated up to the required reaction temperature.
6. Monitor the reaction temperature; when it reaches the set point, start a timer to

time the duration of the reaction.

7. When the timer indicates that the reaction is complete, switch off the controller and open valve C to cool down the reactor contents. Switch off the stirrer.
8. Monitor the temperature; when the contents have cooled, open valve E to remove the product from the reactor.

When implemented by computer all of the above actions and timings would be based upon software. For a large chemical plant such sequences can become very lengthy and intricate and, to ensure efficient operating, several sequences may take place in parallel.

The processes carried out in the single reactor vessel shown in Figure 2.1 are often only part of a larger process as is shown in Figure 2.2. In this plant two reactor vessels (R 1 and R2) are used alternately, so that the processes of preparing for the next batch and cleaning up after a batch can be carried out in parallel with the actual production. Assuming that R 1 has been filled with the mixture and the catalyst, and the reaction is in progress, there will be for R 1: loop control of the temperature and pressure; operation of the stirrer; and timing of the reaction (and possibly some in process measurement to determine the state of the reaction). In parallel with this, vessel R2 will be cleaned - the wash down sequence - and the next batch of raw material will be measured and mixed in the mixing tank.

Meanwhile, the previous batch will be thinned down and transferred to the appropriate storage tank and, if there is to be a change of product or a change in product quality, the thin-down tank will be cleaned. Once this is done the next batch can be loaded into R2 and then, assuming that the reaction in R1 is complete, the contents of R1 will be transferred to the thin-down tank and the wash down procedure for R1 initiated. The various sequences of operations required can become complex and there may also be complex decisions to be made as to when to begin a sequence. The sequence initiation may be left to a human operator or the computer may be programmed to supervise the operations (supervisory control - see below). The decision to use human or computer supervision is often very difficult to make.

The aim is usually to minimize the time during which the reaction vessels are idle since this is unproductive time. The calculations needed and the evaluation of options can be complex, particularly if, for example, reaction times vary with product mix, and therefore it would be expected that decisions made using computer supervisory control would give the best results. however, it is difficult using computer control to obtain the same flexibility that can be achieved using a human operator (and to match the ingenuity of good operators). As a consequence many supervisory systems are mixed; the computer is programmed to carry out the necessary supervisory calculations and to present its decisions for confirmation or rejection by the operator, or alternatively it presents a range of options to the operator.

In most batch systems there is also, in addition to the sequence control, some continuous feedback control: for example, control of temperatures, pressures, flows, speeds or currents. In process control terminology continuous feedback control is referred to as loop control or modulating control and in modern systems this would be carried out using DOC.

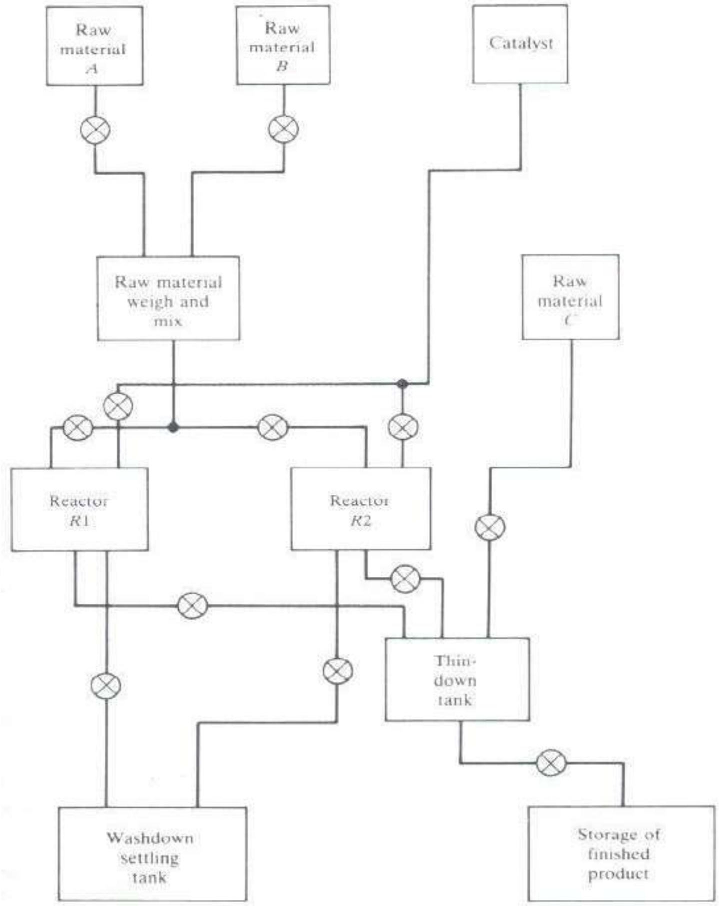


Figure2.2: Typical chemical batch process.

A similar mixture of sequence, loop and supervisory control can be found in continuous systems. Consider the float glass process shown in Figure 2.3. The raw material - sand, powdered glass and fluxes (the frit) - is mixed in batches and fed into the furnace. It melts rapidly to form a molten mixture which flows through the furnace. As the molten glass moves through the furnace it is refined. The process requires accurate control of temperature in order to maintain quality and to keep fuel costs to a minimum - heating the furnace to a higher temperature than is necessary wastes energy and increases costs. The molten glass flows out of the furnace and forms a ribbon on the float bath; again, temperature control is important as the glass ribbon has to cool sufficiently so that it can pass over rollers without damaging its surface.

The continuous ribbon passes into the lehr where it is annealed and where temperature control is again required. It then passes under the cutters which cut it into sheets of the required size; automatic stackers then lift the sheets from the production line. The whole of this process is controlled by several computers and involves loop, sequence and supervisory control. Sequence control systems can vary from the large - the start-up of a large boiler turbine unit in a power station when some 20000 operations and checks may have to be made - to the small - starting a domestic washing machine. Most sequence control systems are simple and frequently have no loop control. They are systems which in the past would have been controlled by relays, discrete logic, or integrated circuit logic units. Examples are simple presses where the sequence might be: locate blank, spray lubricant, lower press, raise press, remove article, spray lubricant. special computer systems known as *programmable logic controllers* (PLCs).

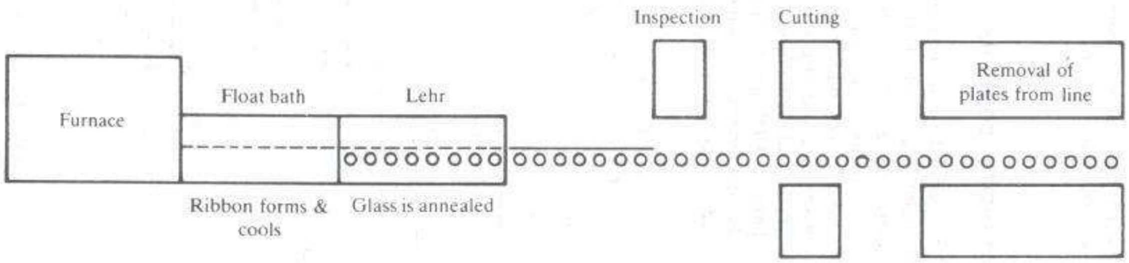


Figure 2.3: Schematic of float glass process.

2.3 LOOP CONTROL (DIRECT DIGITIAL CONTROL):

In direct digital control (DDC) the computer is in the feedback loop as is shown in Figure 2.4., the system shown in Figure 2.4 is assumed to involve several control loops all of which are handled within one computer.

A consequence of the computer being in the feedback loop is that it forms a *critical* component in terms of the reliability of the system and hence great care is needed to ensure that, in the event of the failure or malfunctioning of the computer, the plant remains in a safe condition. The usual means of ensuring safety are to limit the DDC unit to making *incremental* changes to the actuators on the plant; and to limit the rate of change of the actuator settings (the actuators are labeled *A* in Figure 2.4).

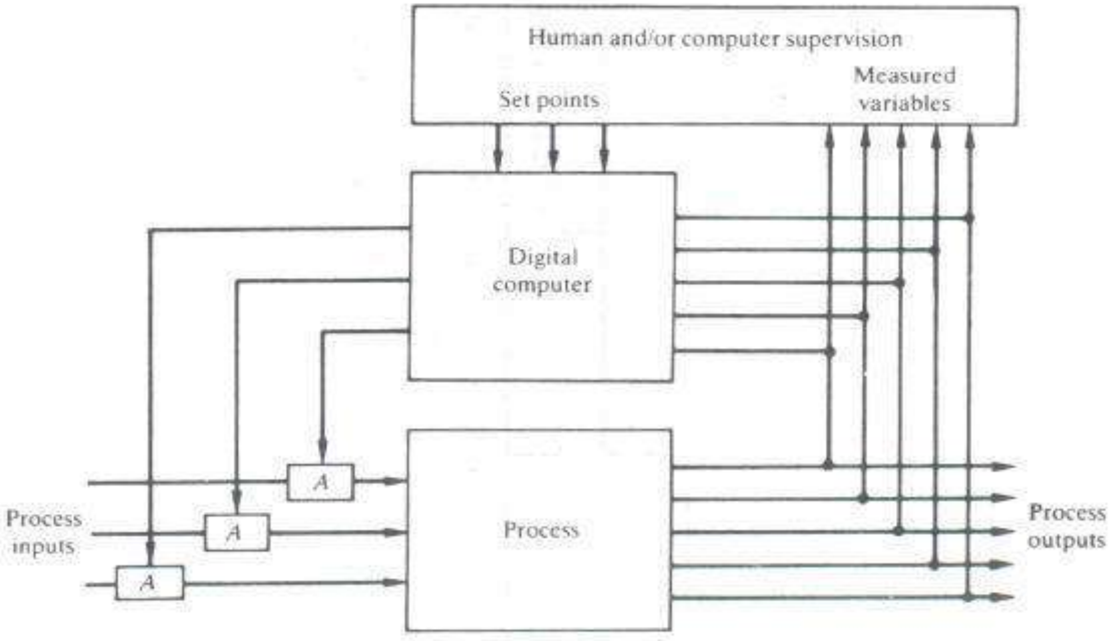


Figure 2.4: Direct digital control.

The advantages claimed for DDC over analog control are:

1. Cost - a single digital computer can control a large number of loops. In the early days the break-even point was between 50 and 100 loops, but now with the introduction of microprocessors a single-loop DDC unit can be cheaper than an analog unit.

2. Performance - digital control offers simpler implementation of a wide range of control algorithms, improved controller accuracy and reduced drift.

3. Safety - modern digital hardware is highly reliable with long mean-time between- failures and hence can improve the safety of systems. However, the software used in programmable digital systems may be much less reliable than the hardware.

The development of integrated circuits and the microprocessor have ensured that in terms of cost the digital solution is now cheaper than the analog. Single-loop controllers used as stand-alone controllers are now based on the use of digital techniques and contain one or more microprocessor chips which are used to implement DDC algorithms. The adoption of improved control algorithms has, however, been slow. Many computer control implementations have simply taken over the well-established analog PID (Proportional + Integral + Derivative) algorithm.

PID CONTROL:

The PID control algorithm has the general form

$$m(t) = Kp\ [e(t) + 1/Ti \int_0^1 e(t)dt + T_d\ de(t)/dt]$$

Where $e(t) = r(t) - c(t)$ and $c(t)$ is the measured variable, $r(i)$ is reference value or set point, and $e(t)$ is error; Kp is the overall controller gain; T; is the integral action time; and T_d is the derivative action time.

For a wide range of industrial processes it is difficult to improve on the control performance that can be obtained by using either PI or PID control (except at considerable expense) or it is for this reason that the algorithms are widely used. For the majority of systems PI control is all that is necessary. Using a control signal that is made proportional to the error between the desired value of an output and the actual value of the output is an obvious and (hopefully) a reasonable strategy. Choosing the value of Kp involves a compromise: a high value of Kp gives a small steady-state error and a fast response, but the response will be oscillatory and may be unacceptable in many applications; a low value gives a slow response and a large steady-state error. By adding the integral action term the steady-state error can be reduced to zero since the integral term, as its name implies, integrates the error signal with respect to time. For a given error value the rate at which the integral term increases is determined by the integral action time Ti. The major advantage of incorporating an integral term arises from the fact that it compensates for changes that occur in the process being controlled.

A purely proportional controller operates correctly only under one particular set of process conditions: changes in the load on the process or some environmental condition will result in a steady-state error; the integral term compensates for these changes and reduces the error to zero. For a few processes which are subjected to sudden disturbances the addition of the derivative term can give improved performance. Because derivative action produces a control signal that is related to the rate of change of the error signal, it anticipates the error and hence acts to reduce the error that would otherwise arise from the disturbance.

In fact, because the PID controller copes perfectly adequately with 90070 of all control problems, it provides a strong deterrent to the adoption of new control system design techniques. DDC may be applied either to a single-loop system implemented on a small microprocessor or to a large system involving several hundred loops. The loops may be cascaded, that is with the output or actuation signal of one loop acting as the set point for another loop, signals may be added together (ratio loops) and conditional switches may be used to alter signal connections.

A typical industrial system is shown in Figure 2.5. This is a steam boiler control system. The steam pressure is controlled by regulating the supply of fuel oil to the burner, but in order to comply with the pollution regulations a particular mix of air and fuel is required. We are not concerned with how this is achieved but with the elements which are required to implement the chosen control system.

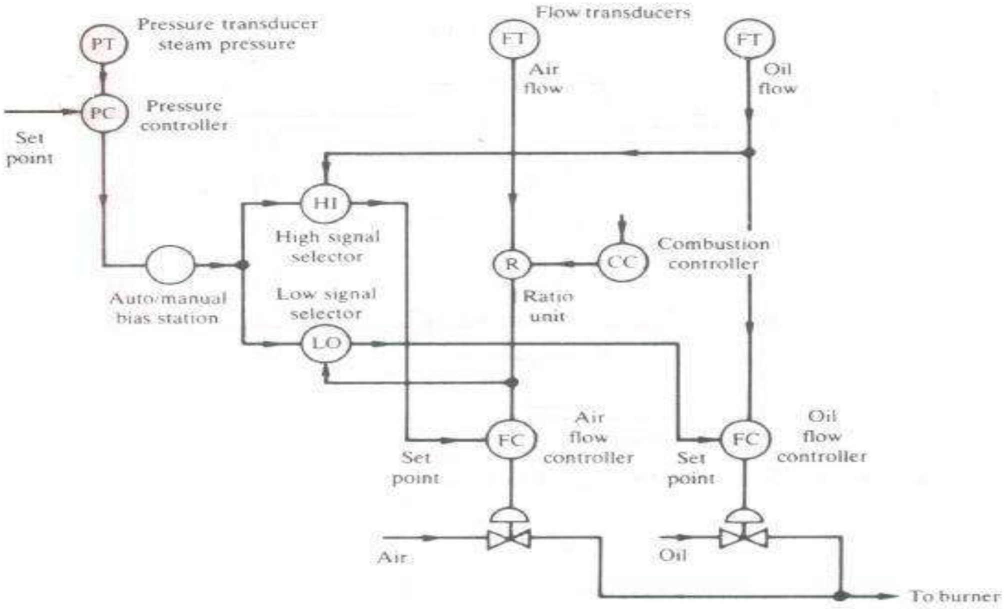


Figure 2.5: A boiler control scheme.

DDC APPLICATIONS:

DDC may be applied either to a single-loop system implemented on a small microprocessor or to a large system involving several hundred loops. The loops may be cascaded, that is with the output or actuation signal of one loop acting as the set point for another loop, signals may be added together (ratio loops) and conditional switches may be used to alter signal connections. A typical industrial system is shown in Figure 2.5. This is a steam boiler control system.

The steam pressure control system generates an actuation signal which is fed to an auto/manual bias station. If the station is switched to auto then the actuation signal is transmitted; if it is in manual mode a signal which has been entered manually (say, from keyboard) is transmitted. The signal from the bias station is connected to two units, a high signal selector and a low signal selector each of which has two inputs and one output. The signal from the low selector provides the set point for the DDC loop controlling the oil flow, the signal from the high selector provides the set point for the air flow controller (two cascade loops). A ratio unit is installed in the air flow measurement line.

DDC is not necessarily limited to simple feedback control as shown in Figure 2.6. It is possible to use techniques such as inferential, feed forward and adaptive or self-tuning control. Inferential control, illustrated in Figure 2.7, is the term applied to control where the variables on which the feedback control is to be based cannot be measured directly, but have to be 'inferred' from measurements of some other quantity.

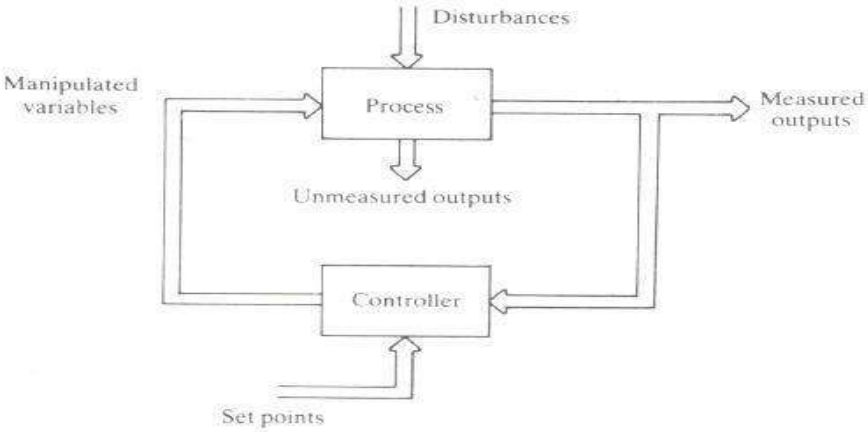


Figure 2.6: General structure of feedback control configuration.

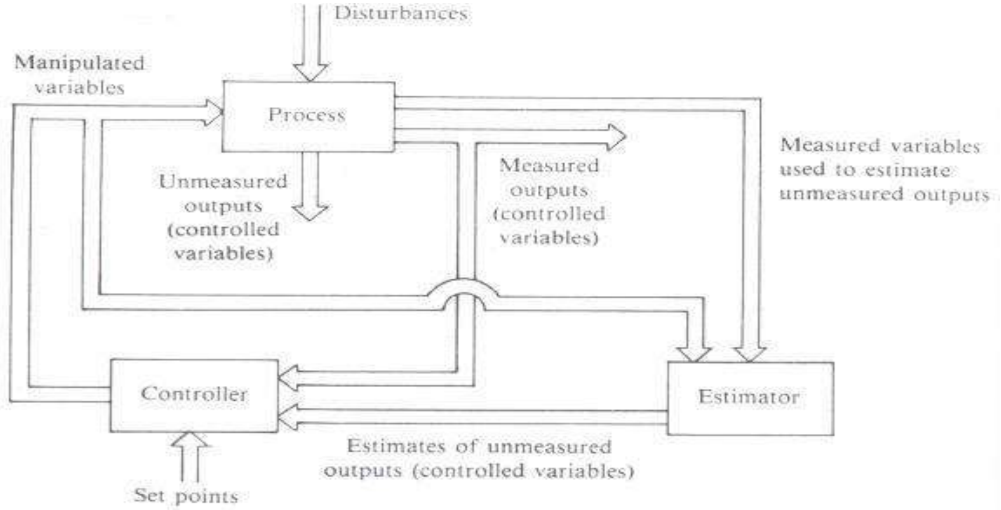

Disturbances
Manipulated variables
Process
Unmeasured outputs (controlled variables)
Measured outputs (controlled variables)
Measured variables used to estimate unmeasured outputs
Controller
Estimator
Estimates of unmeasured outputs (controlled variables)
Set points

Figure 2.7: General control of inferential control configurations.

ADAPTIVE CONTROL:

Adaptive control can take several forms. Three of the most common are:

- Preprogrammed adaptive control (gain 5cheduled control);
- Self-tuning; and
- Model-reference adaptive control.

Programmed adaptive control is illustrated in Figure 2.10a. The adaptive, or adjustment, mechanism makes preset changes on the basis of changes in auxiliary process measurements. For example, in a reaction vessel a measurement of the level of liquid in the vessel (an indicator of the volume of liquid in the vessel) might be used to change the gain of the temperature controller; in many aircraft controls the measured air speed is used to select controller parameters according to a preset schedule.

An alternative form is shown in Figure 2.10b in which measurements of changes in the external environment are used to select the gain or other controller parameters. For example, in an aircraft auto stabilizer, control parameters may be changed according to the external air pressure.

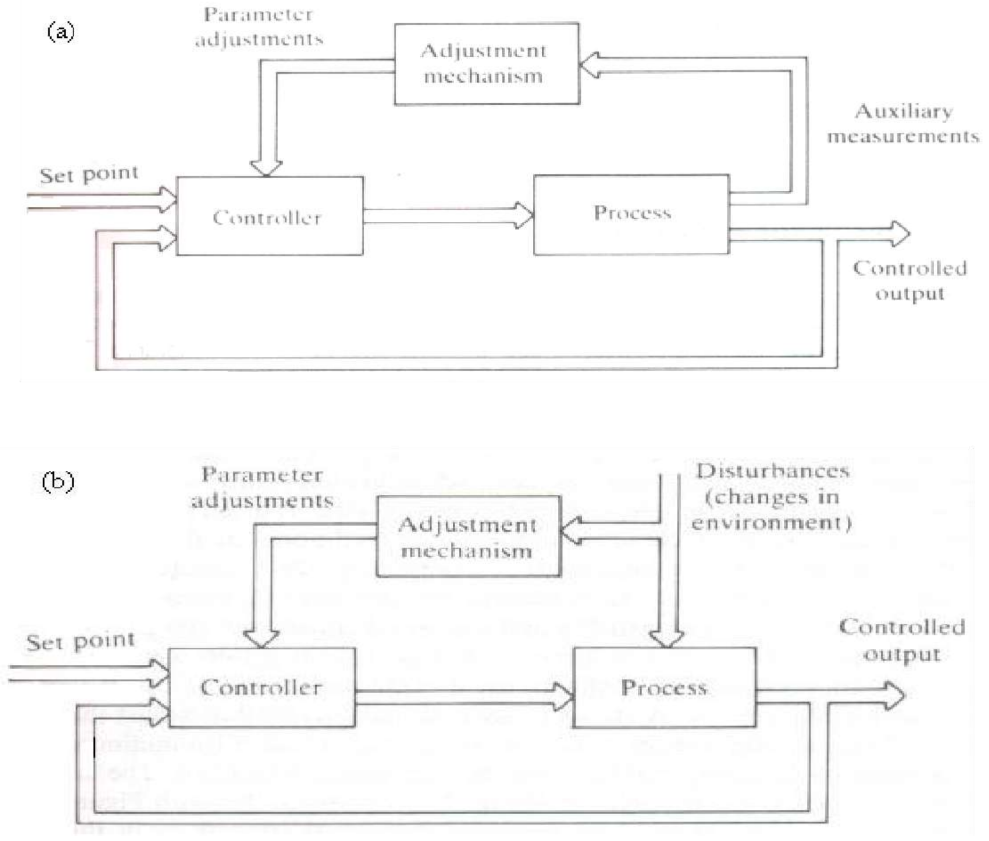


Figure2.10 Programmed adaptive control (gain scheduled):

(a) Auxiliary process measurements; (b) External environment (open loop).

Another example is the use of measurements of external temperature and wind velocities to adjust control parameters for a building environment control system. Adaptive control using self-tuning is illustrated in Figure 2.11 and uses identification techniques to achieve continual determination of the parameters of the process being controlled; changes in the process parameters are then used to adjust the actual controller. An alternative form of self-tuning is frequently found in commercial PID controllers (usually called auto tuning). The comparison may be based on a simple measure such as percentage overshoot or some more complex comparators. The model reference technique is illustrated in Figure 2.12; it relies on the ability to construct an accurate model of the process and to measure the disturbances which affect the process.

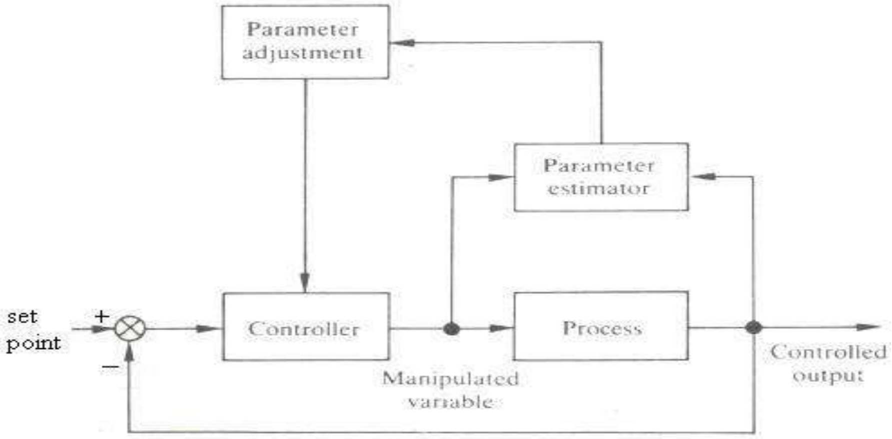


Figure 2.11: Self-tuning adaptive control.

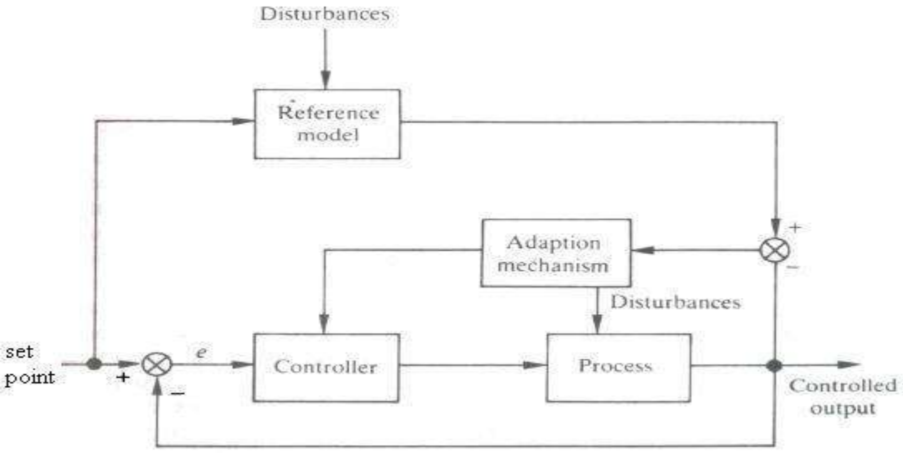


Figure 2.12: Model-reference adaptive control.

2.4 SUPERVISORY CONTROL:

The adoption of computers for process control has increased the range of activities that can be performed, for not only can the computer system directly control the operation of the plant, but also it can provide managers and engineers with a comprehensive picture of the status of the plant operations. It is in this supervisory role and in the presentation of information to the plant operator - large rooms full of dials and switches have been replaced by VDUs and keyboards - that the major changes have been made: the techniques used in the basic feedback control of the plant have changed little from the days when pneumatically operated three-term controllers were the norm. Direct digital control (DDC) is often simply the computer implementation of the techniques used for the traditional analog controllers.

Many of the early computer control schemes used the computer in a supervisory role and not for DDC. The main reasons for this were (a) computers in the early days were not always very reliable and caution dictated that the plant should still be able to run in the event of a computer failure; (b) computers were very expensive and it was not economically viable to use a computer to replace the analog control equipment in current use. A computer system that was used to adjust the set points of the existing analog control system in an optimum manner (to minimize energy or to maximize production) could perhaps be economically justified. The basic idea of supervisory control is illustrated in Figure 2.13 (compare this with Figure 2.4).

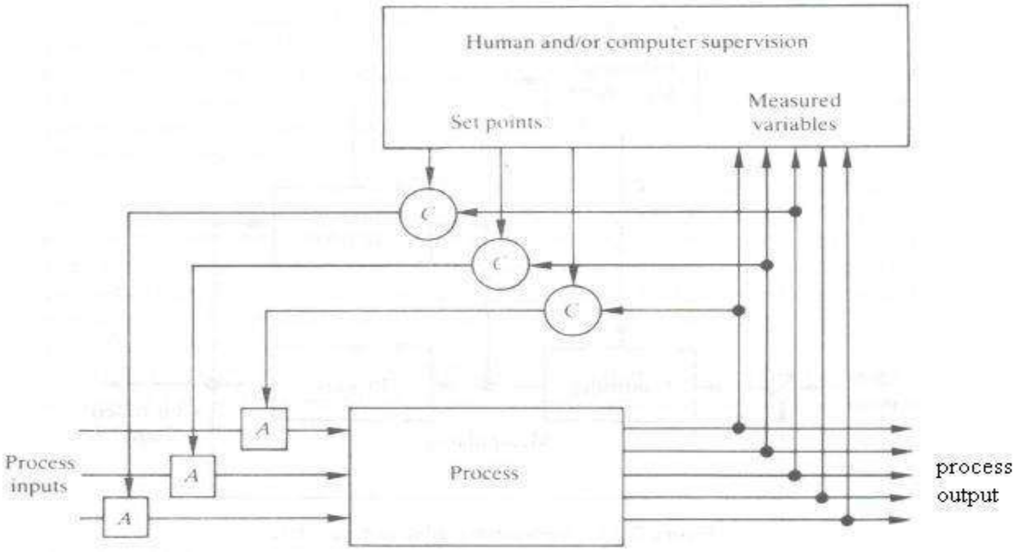


Figure 2.13: Supervisory control.

An example of supervisory control is shown in Figure 2.14. Two evaporators are connected in parallel and material in solution is fed to each unit. The purpose of the plant is to evaporate as much water as possible from the solution. Steam is supplied to a heat exchanger linked to the first evaporator and the steam for the second evaporator is supplied from the vapours boiled off from the first stage. To achieve maximum evaporation the pressures in the chambers must be as high as safety permits. However, it is necessary to achieve a balance between the two evaporators; if the first is driven at its maximum rate it may generate so much steam that the safety thresholds for the second evaporator are exceeded.

A supervisory control scheme can be designed to balance the operation of the two evaporators to obtain the best overall evaporation rate. Most applications of supervisory control are very simple and are based upon knowledge of the steady-state characteristics of the plant. In a few systems complex control algorithms have been used and have been shown to give increased plant profitability.

The techniques used have included optimization based on hill climbing, linear programming and simulations involving complex non-linear models of plant dynamics and economics.

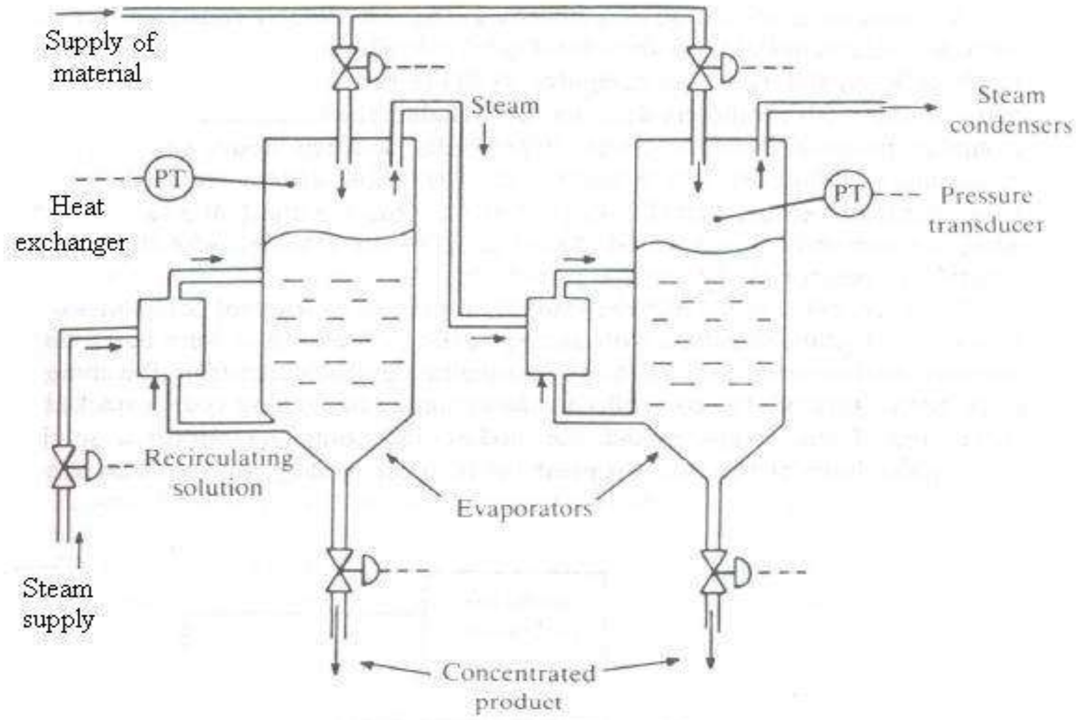


Figure 2.14: An evaporation plant.

2.5 CENTRALISED COMPUTER CONTROL:

Throughout most of the 1960s computer control implied the use of one central computer for the control of the whole plant. The reason for this was largely financial: computers were expensive. From the previous sections it should now be obvious that a typical computer-operation process involves the computer in performing many different types of operations and tasks. Although a general purpose computer can be programmed to perform all of the required tasks the differing time-scales and security requirements for the various categories of task make the programming job difficult, particularly with regard to the testing of software. For example, the feedback loops in a process may require calculations at intervals measured in seconds while some of the alarm and switching systems may require a response in less than 1 second; the supervisory control calculations may have to be repeated at intervals of several minutes or even hours; production management will want summaries at shift or daily intervals; and works management will require weekly or monthly analyses. Interrelating all the different time-scales can cause serious difficulties.

A consequence of centralized control was the considerable resistance to the use of DOC schemes in the form shown in Figure 2.4; with one central computer in the feedback loop, failure of the computer results in the loss of control of the 'whole plant. In the 1960s computers were not very reliable: the mean-time-to-failure of the computer hardware was frequently of the order of a few hours and to obtain a mean-time-to-failure of 3 to 6 months for the whole system required defensive programming to ensure that the system could continue running in a safe condition while the computer was repaired. Many of the early schemes were therefore for supervisory control as shown in Figure 2.13. However, in the mid 1960s the traditional process instrument companies began to produce digital controllers with analog back-up. These units were based on the standard analog controllers but allowed a digital control signal from the computer to be passed through the controller to the actuator: the analog system tracked the signal and if the computer did not update the controller within a specified (adjustable) interval the unit dropped on to local analog control. This scheme enabled DDC to be used with the confidence that if the computer failed, the plant could still be operated. The cost, however, was high in that two complete control systems had to be installed.

By 1970 the cost of computer hardware had reduced to such an extent that it became feasible to consider the use of dual computer systems (Figure 2.15). Here, in the event of failure of one of the computers, the other takes over. In some schemes the change-over is manual, in others automatic failure detection and change-over is incorporated. Many of these schemes are still in use. They do, however, have a number of weaknesses: cabling and interface equipment is not usually duplicated, neither is the software - in the sense of having independently designed and constructed programs - so that the lack of duplication becomes crucial. Automatic failure and change-over equipment when used becomes in itself a critical component. Furthermore, the problems of designing, programming, testing and maintaining the software are not reduced: if anything they are further complicated in that provision for monitoring ready for change-over has to be provided. The continued reduction of the cost of hardware and the development of the microprocessor has made multi-computer systems feasible. These fall into two types:

1. Hierarchical - Tasks are divided according to function, for example with one computer performing DDC calculations and being subservient to another which performs supervisory control.

2. Distributed - Many computers perform essentially similar tasks in parallel.

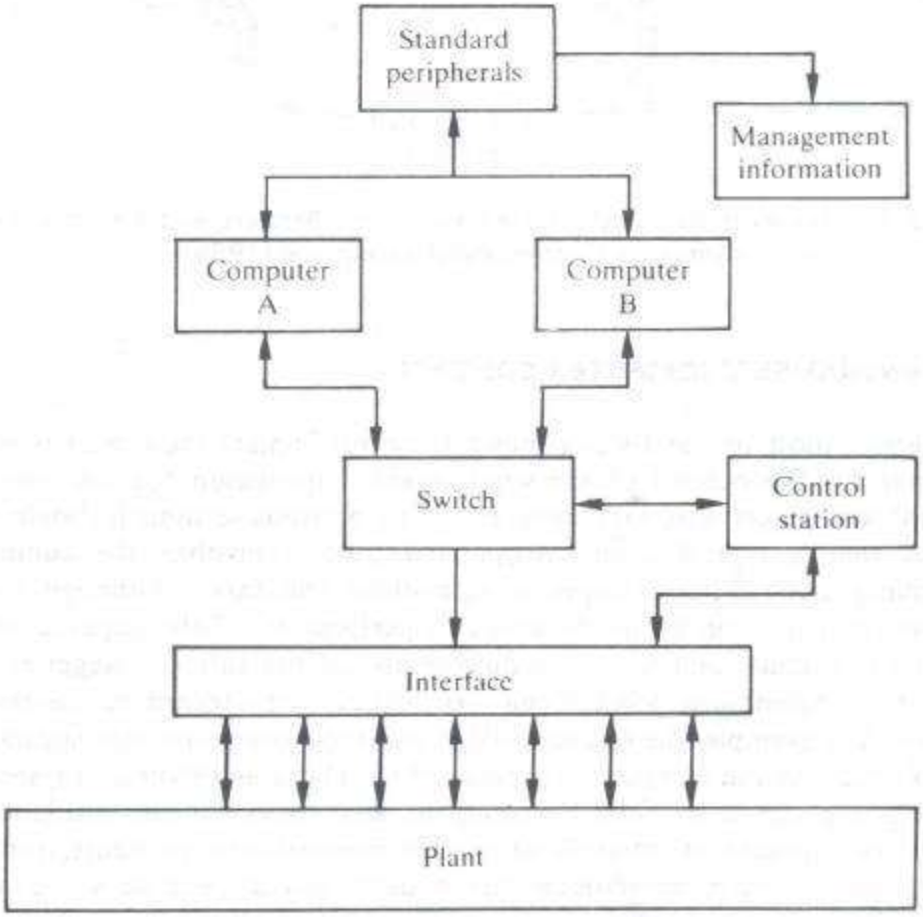


Figure 2.15: Dual computer scheme.

2.6 DISTRIBUTED SYSTEMS:

The underlying assumptions of the distributed approach are:

1. Each unit is carrying out essentially similar tasks to all the other units; and

2. In the event of failure or overloading of a particular unit all or some of the work can be transferred to other units.

In other words, the work is not divided by function and allocated to a particular computer as in hierarchical systems: instead, the total work is divided up and spread across several computers. This is a conceptually simple and attractive approach - many hands make light work - but it poses difficult hardware and software problems since, in order to gain the advantages of the approach, allocation of the tasks between computers has to be dynamic, that is there has to be some mechanism which can assess the work to be done and the present load on each computer in order to allocate work. Because each computer needs access to all the information in the system, high-bandwidth data highways are necessary. There has been considerable progress in developing such highways and the various types are discussed below:

Computer scientists and engineers are also carrying out considerable research on multi-processor computer systems and this work could lead to totally distributed systems becoming feasible. There is also a more practical approach to distributing the computing load whereby no attempt is made to provide for the dynamic allocation of resources but

instead a simple ad hoc division is adopted with, for example, one computer performing all non-plant input and output, one computer performing all DDC calculations, another performing data acquisition and yet another performing the control of the actuators. In most modern schemes a mixture of distributed and hierarchical approaches is used as shown in Figure 2.19. The tasks of measurement, DDC, operator communications, etc., are distributed among a number of computers which are linked together via a common serial communications highway and are configured in a hierarchical command structure. Five broad divisions of function are shown:

Level 1 all computations and plant interfacing associated with measurement and actuation. This level provides a measurement and actuation database for the whole system.

Level 2 All DDC calculations.

Level 3 all sequence calculations.

Level 4 Operator communications.

Level 5 Supervisory control

Level 6 Communications with other computer systems.

It is not necessary to preserve rigid boundaries; for example, a DDC unit may perform some sequencing or may interface directly to plant.

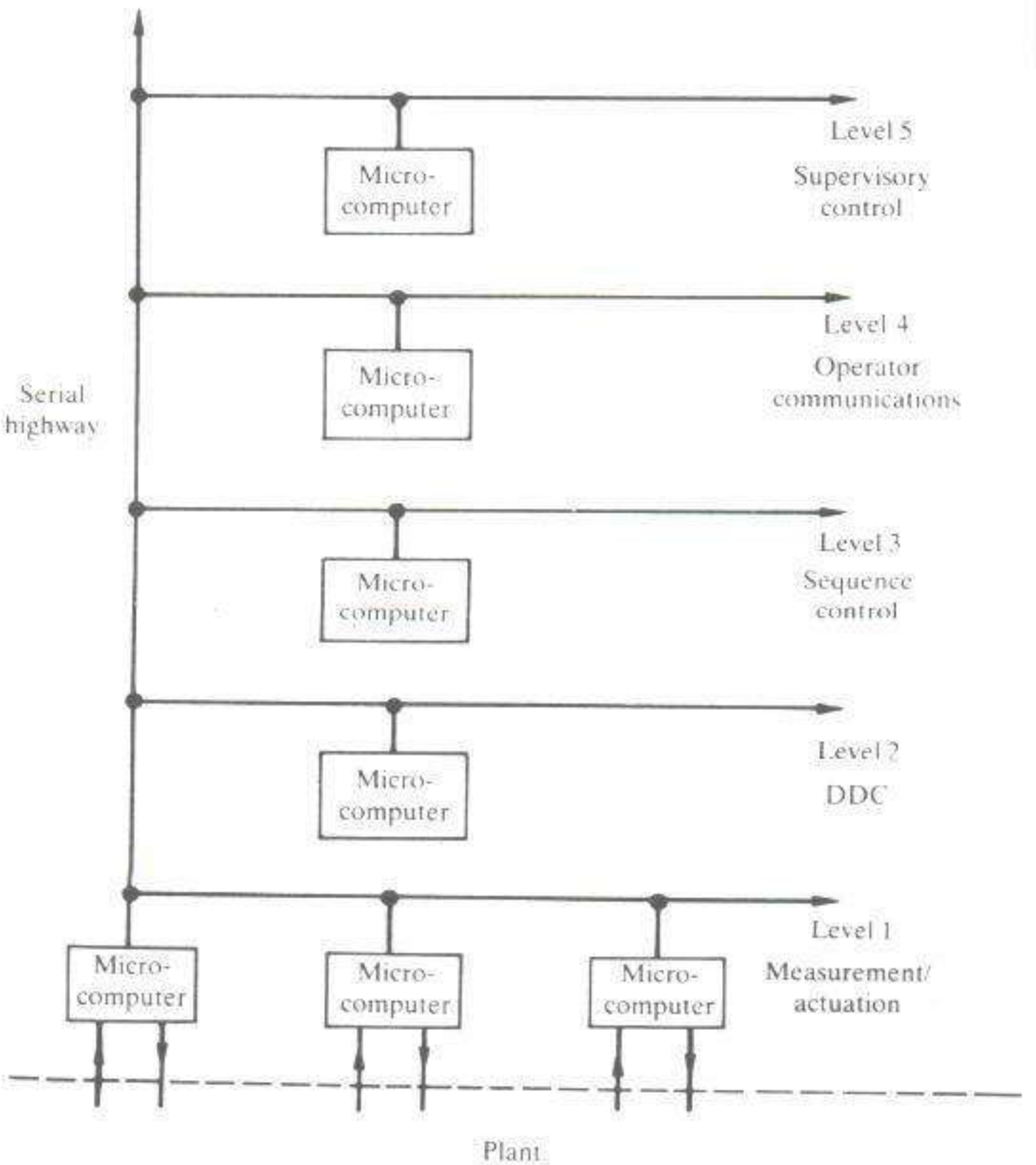


Figure 2.16: A distributed system.

The major advantages of this approach are:

1. The system capabilities are greatly enhanced by the sharing of tasks between processors - the burden of computation for a single processor becomes very great if all of the described control

features are included. One of the main computing loads is that of measurement scanning, filtering and scaling, not because anyone calculation is onerous but because of the large number of signals involved and the frequency at which the calculations have to be repeated. Separation of this aspect from the DDC, even if only into two processors, greatly enhances the number of control loops that can be handled. The DDC computer will collect measurements, already processed, via the communications link at a much lower frequency than that at which the measurement computer operates.

2. The system is much more flexible than the use of a single processor: if more loops are required or an extra operator station is needed, all that is necessary is to add more boxes to the communication link - of course the other units on the link will need to be updated to be aware of the additional items. It also allows standardization, since it is much easier to develop standard units for well-defined single tasks than for overall control schemes.

3. Failure of a unit will cause much less disruption in that only a small portion of the overall system will not be working. Provision of automatic or semiautomatic transfer to a back-up system is much easier.

4. It is much easier to make changes to the system, in the form of either hardware replacements or software changes. Changing large programs is hazardous because of the possibility of unforeseen side-effects: with the use of small modules such effects are less likely to occur and are more easily detected and corrected.

5. Linking by serial highway means that the computer units can be widely dispersed: hence it is unnecessary to bring cables carrying transducer signals to a central control room.

2.7 HUMAN –COMPUTER INTERFACE:

The key to the successful adoption of a computer control scheme is often the facilities provided for the plant operator or user of the system. A simple and clear system for the day-to-day operation of the plant must be provided. All the information relevant to the current state of its operation should be readily available and facilities to enable interaction with the plant - to change set points, to adjust actuators by hand, to acknowledge alarm conditions, etc. - should be provided. A large proportion of the design and programming effort goes into the design and construction of operator facilities and the major process control equipment companies have developed extensive schemes for the presentation of information.

A typical operator station has specially designed keyboards and several display and printer units; extensive use is made of color displays and mimic diagrams; video units are frequently provided to enable the operator to see parts of the plant (Jovic, 1986). The standard software packages typically provide a range of display types: an alarm overview presenting information on the alarm status of large areas of the plant; a number of area displays presenting information on the control systems associated with each area; and loop displays giving extensive information on the details of a particular control loop. The exact nature of the displays is usually determined by the engineer responsible for the plant or part of the plant.

The plant manager requires access to different information: hard copy printouts - including graphs - that summarize the day-to-day operation of the plant and also provide a permanent plant operating history. Data presented to the manager will frequently have been analyzed statistically to provide more concise information and

to make decision-making more straightforward. The manager will be interested in assessing the economic performance of the plant and in determining possible improvements in plant operation. The design of user interfaces is a specialist area. The safe operation of complex systems such as aircraft, nuclear power stations, chemical plants, air traffic control systems and other traffic control systems can be crucially affected by the way in which information is presented to the operator.

2.8 BENEFITS OF COMPUTER CONTROL SYSTEMS:

Before the widespread availability of microprocessors, computer control was expensive and a very strong case was needed to justify the use of computer control rather than conventional instrumentation. In some cases computers were used because otherwise plant could not have been made to work profitably: this is particularly the case with large industrial processes that require complex sequencing operations. The use of a computer permits the repeatability that is essential, for example, in plants used for the manufacture of drugs. In many applications flexibility is important - it is difficult with conventional systems to modify the sequencing procedure to provide for the manufacture of a different product.

Flexibility is particularly important when the product or the product specification may have to be changed frequently: with a computer system it is simple to maintain a database containing the product recipes and thus to change to a new recipe quickly and reliably.

The application of computer control systems to many large plants has frequently been justified on the grounds that even a small increase in productivity (say I or 2070) will more than pay for the computer system. After installation it has frequently been difficult to establish that an improvement has been achieved; sometimes production has decreased, but the computer proponents have then argued that but for the introduction of the computer system production would have decreased by a greater amount! Some of the major benefits to accrue from the introduction of computer systems have been in the increased understanding of the behavior of the process that has resulted from the studies necessary to design the computer system and from the information gathered during running. This has enabled supervisory systems to keep the plant running at an operating point closer to the desired point to be designed.

The other main area of benefit has been in the control of the starting and stopping of batch operations in that computer-based systems have generally significantly reduced the dead time associated with batch operations. The economics of computer control have been changed drastically by the microprocessor in that the reduction in cost and the improvement in reliability have meant that computer-based systems are the first choice in many applications. Indeed, microprocessor-based instrumentation is frequently cheaper than the equivalent analog unit. The major costs of computer control are now no longer the computer hardware, but the system design and the cost of software: as a consequence attention is shifting towards greater standardization of design and of software products and the development of improved techniques for design (particularly software design) and for software construction and testing. The availability of powerful, cheap and highly reliable computer hardware and communications systems makes it possible to conceive and construct large, complex, computer-based control systems. The complexity of such systems raises concern about their dependability and safety.

Recommended Question:

1. List the advantages and disadvantages of using DDC.
2. In the section on human-computer interfacing we made the statement 'the design
   of user interfaces is a specialist area'. Can you think of reasons to support this statement and suggest what sort of background and training a specialist in user interfaces might require?
3. What are the advantages/disadvantages of using a continuous oven? How will the
   control of the process change from using a standard oven on a batch basis to

using an oven in which the batch passes through on a conveyor belt? Which will be the easier to control?

4. List the advantages of using several small computers instead of one large computer in control applications. Are there any disadvantages that arise from using several computers?
5. List the characteristics of Batch process and continuous process.

MODULE- 2

Computer Hardware Requirements for RTS

Introduction, General Purpose Computer, Single Chip Microcontroller, Specialized Processors, Process –Related Interfaces, Data Transfer Techniques, Communications, Standard Interface.

Recommended book for reading:

1. **Real –Time Computer control –An Introduction**, Stuart Bennet, 2nd Edn. Pearson Education 2005.
2. **Real-Time Systems Design and Analysis**, Phillip. A. Laplante, Second Edition, PHI, 2005.

3.1 COMPUTER HARDWARE REQUIREMENTS FOR RTS.

INTRODUCTION:

Although almost any digital computer can be used for real-time computer control and other real-time operations, they are not all equally easily adapted for such work. In the majority of embedded computer-based systems the computer used will be a microprocessor, a microcomputer or a specialized digital processor. Specialized digital processors include fast digital signal processors, parallel computers such as the transputer, and special RISC (Reduced Instruction Set Computers) for use in safety-critical applications (for example, the VIPER (Cullyer and Pygott, 1987).

3.2 GENERAL PURPOSE COMPUTER:

The general purpose microprocessors include the Intel XX86 series, Motorola 680XX series, National 32XXX series and the Zilog Z80 and Z8000 series. A characteristic of computers used in control systems is that they are modular: they provide the means of adding extra units, in particular specialized input and output devices, to a basic unit. The capabilities of the basic unit in terms of its processing power, storage capacity, input/output bandwidth and interrupt structure determine the overall performance of the system.

A simplified block diagram of the basic unit is shown in Figure 3.1; the arithmetic and logic, control, register, memory and input/ output units represent a general purpose digital computer. Of equal importance in a control computer are the input/output channels which provide a means of connecting process instrumentation to the computer, and also the displays and input devices provided for the operator. The instruments are not usually connected directly but by means of interface units.

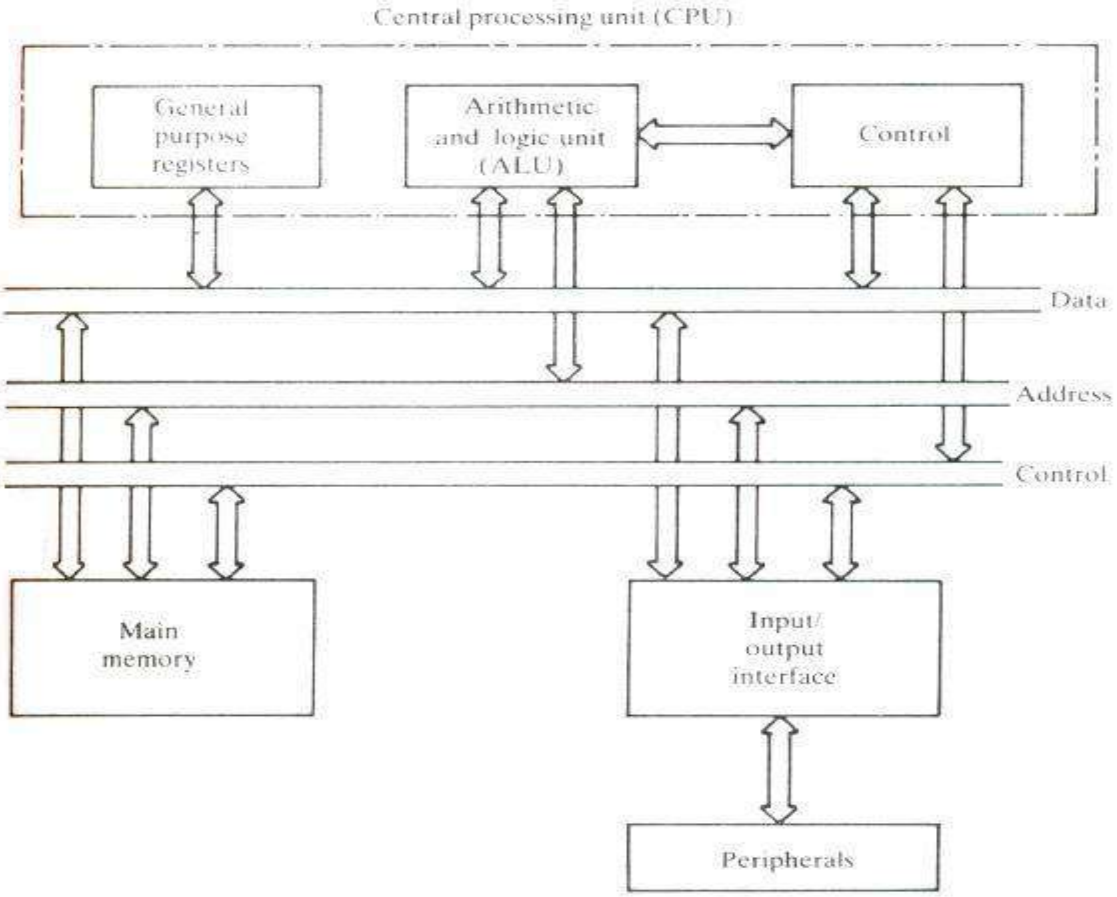

Central processing unit (CPU)
General purpose registers
Arithmetic and logic unit (ALU)
Control
Data
Address
Control
Main memory
Input/ output interface
Peripherals

Figure: 3.1 Schematic diagram of a general purpose digital computer.

CENTRAL PROCESSING UNIT:

The arithmetic and logic unit (ALU) together with the control unit and the general purpose registers make up the central processing unit (CPU). The ALU contains the circuits necessary to carry out arithmetic and logic operations, for example to add numbers, subtract numbers and compare two numbers. Associated with it may be hardware units to provide multiplication and division of fixed point numbers and, in the more powerful computers, a floating point arithmetic unit. The general purpose registers can be used for storing data temporarily while it is being processed. Early computers had a very limited number of general purpose registers and hence frequent access to main memory was required. Most computers now have CPUs with several general purpose registers - some large systems have as many as 256 registers - and for many computations, intermediate results can be held in the CPU without the need to access main memory thus giving faster processing.

The control unit continually supervises the operations within the CPU: it fetches program instructions from main memory, decodes the instructions and sets up the necessary data paths and timing cycles for the execution of the instructions. The features of the CPU which determine the processing power available and hence influence the choice of computer for process control include:

- Wordlength;
- Instruction set;
- Addressing methods;
- Number of registers;
- Information transfer rates; and
- Interrupt structure.

The computer word length is important both in ensuring adequate precision in calculations and in allowing direct access to a large area of main storage within one instruction word. It is possible to compensate for short wordlengths, both for arithmetic precision and for memory access, by using multiple word operations, but the penalty is increased time for the operations. The basic instruction set of the CPU is also important in determining its overall performance. Features which are desirable are:

- Flexible addressing modes for direct and immediate addressing;
- Relative addressing modes;
- Address modification by use of index registers;
- Instructions to transfer variable length blocks of data between storage units or locations within memory; and
- Single commands to carry out multiple operations.

STORAGE:

The storage used on computer control systems divides into two main categories: fast access storage and auxiliary storage. The fast access memory is that part of the system which contains data, programs and results which are currently being operated on. The major restriction with current computers is commonly the addressing limit of the processor. In addition to RAM (random access memory - read/write) it is now common to have ROM (read-only memory), PROM (programmable read-only memory) or EPROM (electronically programmable read only memory) for the storage of critical code or predefined functions. The use of ROM has eased the problem of memory protection to prevent loss of programs through power failure or corruption by the malfunctioning of the software (this can be a particular problem during testing).

An alternative to using ROM is the use of memory mapping techniques that trap instructions which attempt to store in a protected area. This technique is usually only used on the larger systems which use a memory management system to map program addresses onto the physical address space. An extension of the system allows particular parts of the physical memory to be set as read only, or even locked out altogether: write access can be gained only by the use of 'privileged' instructions. The auxiliary storage medium is typically disk or magnetic tape. These devices provide bulk storage for programs or data which are required infrequently at a much lower cost than fast access memory. The penalty is a much longer access time and the need for interface boards and software to connect them to the CPU. In a real-time system use of the CPU to carry out the transfer is not desirable as it is slow and no other computation can take place during transfer. For efficiency of transfer it is sensible to transfer large blocks of data rather than a single word or byte and this can result in the CPU not being available for up to several seconds in some cases. The approach frequently used is direct memory access (DMA). For this the interface controller for the backing memory must be able to take control of the address and data buses of the computer.

INPUT AND OUTPUT:

The input/output (I/O) interface is one of the most complex areas of a computer system; part of the complication arises because of the wide variety of devices which have to be connected and the wide variation in the rates of data transfer. A printer may operate at 300 baud whereas a disk may require a rate of 500 kbaud. The devices may require parallel or serial data transfers, analog-to-digital or digital-to-analog conversion, or conversion to pulse rates. The I/O system of most control computers can be divided into three sections:

- Process I/O;
- Operator I/O; and
- Computer I/O.

BUS STRUCTURE:

Buses are characterized into three ways:

- Mechanical (physical) structure;
- Electrical; and
- Functional.

In mechanical or physical terms a bus is a collection of conductors which carry electrical signals, for example tracks on a printed circuit board or the wires in a ribbon cable. The physical form of the bus represents the *mechanical characteristic* of the bus system. The *electrical characteristics* of the bus are the signal levels, loading (that is, how many loads the line can support), and type of output gates (open-collector, tri-state). The *functional characteristics* describe the type of information which the electrical signals flowing along the bus conductors represent. The bus lines can be divided into three functional groups:

- Address lines;
- Data lines; and
- Control and status lines.

3.3 SINGLE CHIP MICROCONTROLLER:

Many integrated circuit manufacturers produce microcomputers in which all the components necessary for a complete computer are provided on one single chip. A typical single-chip device is shown in Figure 3.2. With only a small amount of EPROM and an even smaller amount of RAM this type of device is obviously intended for small, simple systems. The memory can always be extended by using external memory chips. The microcontroller is similarly a single-chip device that is specifically intended for embedded computer control applications. The main difference between it and a microcomputer is that it typically will have on board the chip a multiplexed ADC and some form of process output, for example a pulse width modulator unit. The chip may also contain a real-time clock generator and a watch-dog timer.

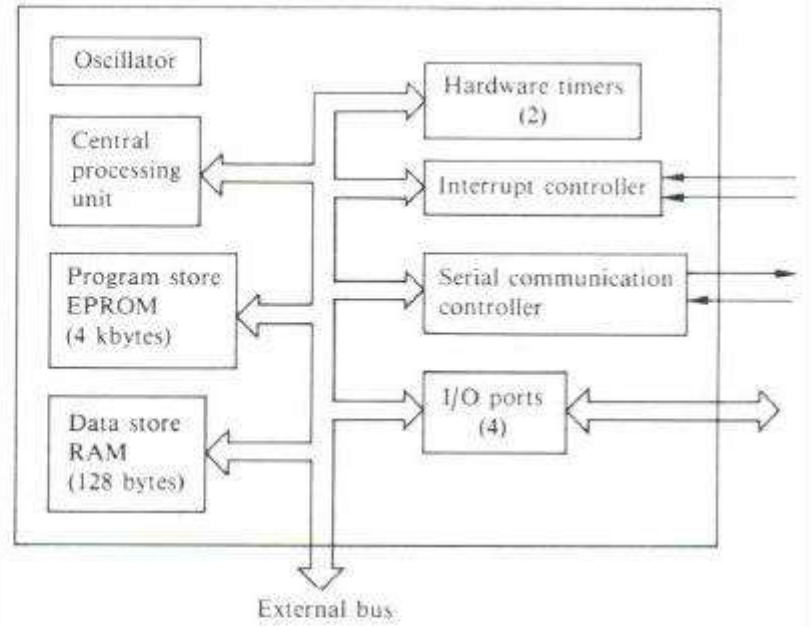


Figure 3.2: A typical single-chip computer.

3.4 SPECIALIZED PROCESSORS:

Specialized processors have been developed for two main purposes:

- Safety-critical applications; and
- Increased computation speed.

For safety-critical applications the approach has been to simplify the instruction set - the so-called reduced instruction set computer (RISC). The advantage of simplifying the instruction set is the possibility of formal verification (using mathematical proofs) that the logic of the processor is correct. The second advantage of the RISC machine is that it is easier to write assemblers and compilers for the simple instruction set. An example of such a machine is the VIPER (Cullyer, 1988; Dettmer, 1986), the main features of which are:

- Formal mathematical description of the processor logic.
- Integer arithmetic (32 bit) and no floating point operations (it is argued that floating point operations are inexact and cannot be formally veri fled).
- No interrupts - all event handling is done using polling (again interrupts make formal verification impossible).
- No dynamic memory allocation.

The traditional Von Neumann computer architecture with its one CPU through which all the data and instructions have to pass sequentially results in a bottleneck. Increasing the processor speed can increase the throughput but eventually systems will reach a physical limit because of the fundamental limitation on the speed at which an electronic signal can travel. The search for increased processing speed has led to the abandonment of the Von Neumann architecture for high-speed computing.

3.5 PARELLEL COMPUTERS:

Many different forms of parallel computer architectures have been devised; however, they can be summarized as belonging to one of three categories: SIMD MISD MIMD

Single instruction stream, multiple data stream.

Multiple instruction stream, single data stream.

Multiple instruction stream, multiple data stream.

These are illustrated in Figure 3.3 where the traditional architecture characterized as SISD (Single instruction stream, single data stream) is also shown. MIMD systems are obviously the most powerful class of parallel computers in that each processor can potentially be executing a different program on a different data set. The most widely available MIMD system is the INMOS transputer.

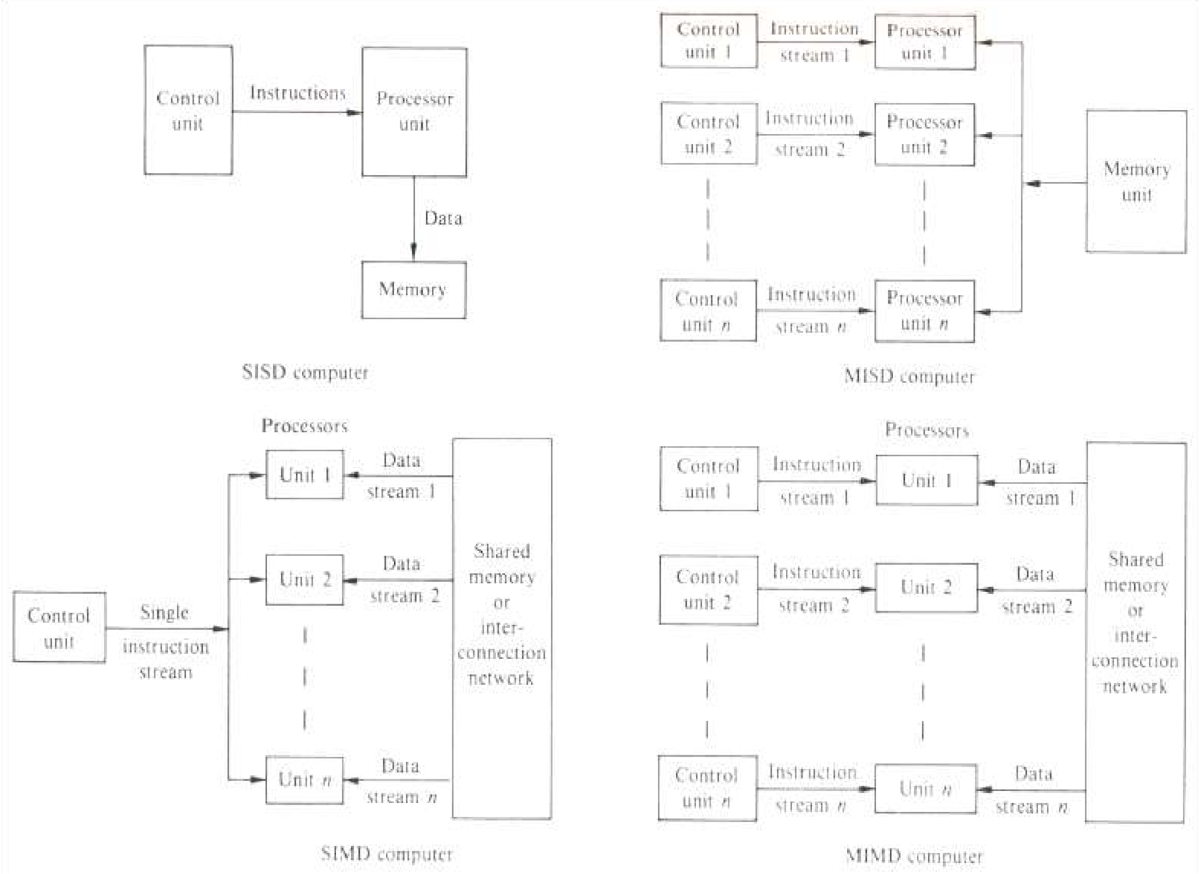


Figure 3.3: Computer system architecture.

An individual chip can be used as a stand-alone computing device; however, the power of the transputer is obtained when several transputers are interconnected to form a parallel processing network. INMOS developed a special programming language, occam, for use with the transputer. Occam is based on the assumption that the application to be implemented on the transputer can be modelled as a set of processes (actions) that communicate with each other via channels. A channel is a unidirectional link between two processes which provides synchronized communication. A process can be a primitive process, or a collection of processes; hence the system supports a hierarchical structure. Processes are dynamic in that they can be created, can die and can create other processes.

3.6 DIGITIAL SIGNAL PROCESSORS:

In applications such as speech processing, telecommunications, radar and hi-fi systems analog techniques have been used for modifying the signal characteristics. There are advantages to be gained if such processing can be done using digital techniques in that the digital devices are inherently more reliable and not subject to drift. The problem is that the bandwidth of the signals to be processed is such as to demand very high processing speeds. Special purpose integrated circuits optimized to meet the signal processing requirements have been developed. They typically use the so-called Harvard architecture in which separate paths are provided for data and for instructions. DSPs typically use fixed point arithmetic and the instruction set contains instructions for manipulating complex numbers. They are difficult to program as few high-level language compilers are available.

3.7 PROCESS-RELATED INTERFACES:

Instruments and actuators connected to the process or plant can take a wide variety of forms: they may be used for measuring temperatures and hence use thermocouples, resistance thermometers, thermistors, etc.; they could be measuring flow rates and use impulse turbines; they could be used to open valves or to control thyristor-operated heaters. In all these operations there is a need to convert a digital quantity, in the form of a bit pattern in a computer word, to a physical quantity, or to convert a physical quantity to a bit pattern. Designing a different interface for each specific type of instrument or actuator is not sensible or economic and hence we look for some commonality between them. Most devices can be allocated to one of the following four categories:

1. Digital quantities: These can be either binary that is a valve is open or closed, a switch is on or off, a relay should be opened or closed, or a generalized digital quantity, that is the output from a digital voltmeter in BCD (binary coded decimal) or other format.

2. Analog quantities: Thermocouples, strain gauges, etc., give outputs which are measured in mill volts; these can be amplified using operational amplifiers to give voltages in the range - 10 to + 10 volts; conventional industrial instruments frequently have a current output in the range 4 to 20 mA (current transmission gives much better immunity to noise than transmission of low-voltage signals). The characteristic of these signals is that they are continuous variables and have to be both sampled and converted to a digital value.

3. Pulses and pulse rates: A number of measuring instruments, particularly flow meters, provide output in the form of pulse trains; similarly the increasing use of stepping motors as actuators requires the provision of pulse outputs. Many traditional controllers have also used pulse outputs: for example, valves controlling flows are frequently operated by switching a dc or ac motor on and off, the length of the on pulse being a measure or

the change in valve opening required .

4. Telemetry: The increasing use of remote outstations, for example electricity substations and gas pressure reduction stations, has increased the use of telemetry. The data may be transmitted by landline, radio or the public telephone net work: it is, however, characterized by being sent in serial form, usually encoded in standard ASCII characters. For small quantities of data the transmission is usually asynchronous. Telemetry channels may also be used on a plant with a hierarchy of computer systems instead of connecting the computers by some form of network. An example of this is the CUTLASS system used by the Central Electricity Generating Board, which uses standard RS232 lines to connect a hierarchy of control computers. The ability to classify the interface requirements into the above categories means that a limited number of interfaces can be provided for a process control computer.

3.8 DIGITIAL SIGNAL INTERFACES:

A simple digital input interface is shown in Figure 3.4. It is assumed that the plant outputs are logic signals which appear on lines connected to the digital input register. It is usual to transfer one word at a time to the computer, so normally the digital input register will have the same number of input lines as the number of bits in the computer word. The logic levels on the input lines will typically be 0 and + 5 V; if the contacts on the plant which provide the logic signals use different levels then conversion of signal levels will be required. To read the lines connected to the digital input register the computer has to place the address of the register on the address bus and decoding circuitry is required in the interface (address decoder) to select the digital input register. In addition to the 'select' signal an 'enable' signal may also be required; this could be provided by the 'read' signal from the computer control bus. In response to both the 'select' and 'enable' signals the digital input register enables its output gates and puts data onto the computer data bus.

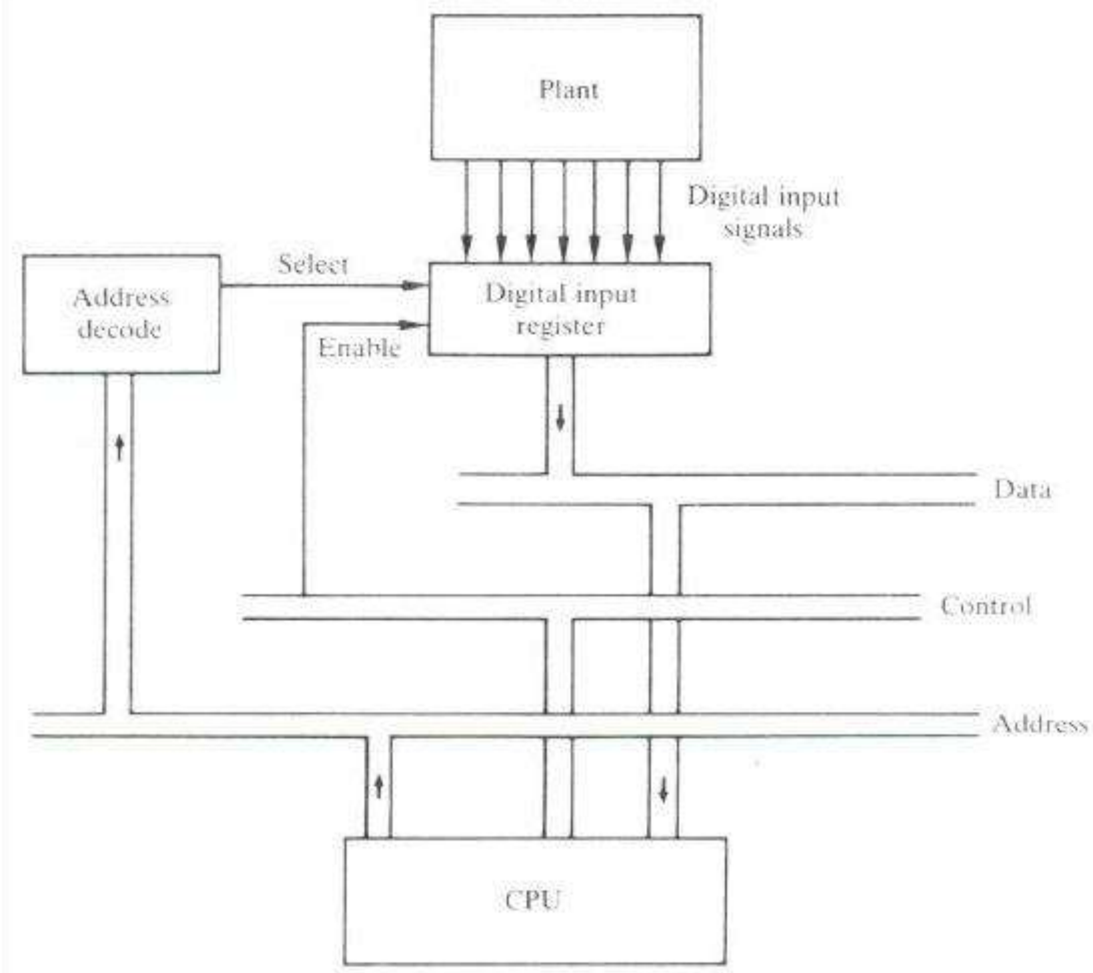


Figure 3.4: A simple digital input interface.

Figure 3.4 shows a system that provides information only on demand from the computer: it cannot indicate 10 the computer that information is waiting. There are many circumstances in which it is useful to indicate a change of status of input lines to the computer. To do this a status line which the computer can test, or which can be used as an interrupt, is needed. A simple digital output interface is shown in Figure 3.6. Digital output is the simplest form of output: all that is required is a register or latch which can hold the data output from the computer. To avoid the data in the register changing when the data on the data bus changes, the output latch must respond only when it is addressed. The 'enable' signal is used to indicate to the device that the data is stable on the data bus and can be read.

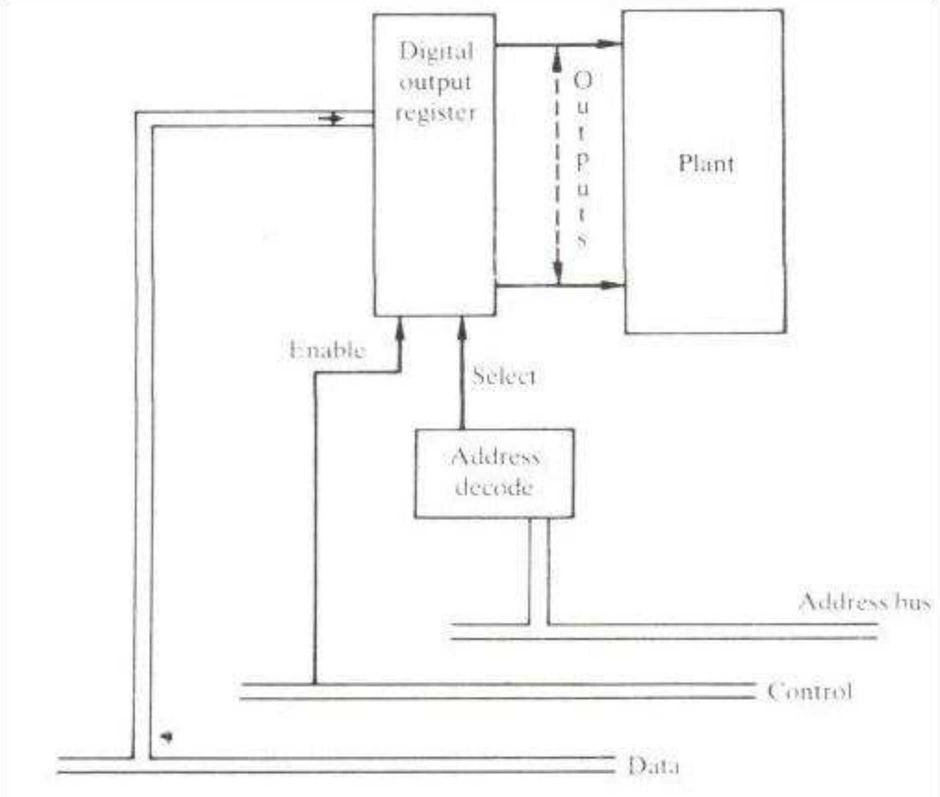


Figure 3.6: A simple digital output interface.

3.9 PULSE INTERFACES:

In its simplest form a pulse input interface consists of a counter connected to a line from the plant. The counter is reset under program control and after a fixed length of time the contents are read by the computer. A typical arrangement is shown in Figure 3.7, which also shows a simple pulse output interface. The transfer of data from the counter to the computer uses techniques similar to those for the digital input described above. The measurement of the length of time for which the count proceeds can be carried out either by a logic circuit in the counter interface or by the computer. If the timing is done by the computer then the 'enable' signal must inhibit the further counting of pulses. If the computing system is not heavily loaded, the external interface hardware required can be reduced by connecting the pulse input to an interrupt and counting the pulses under program control.

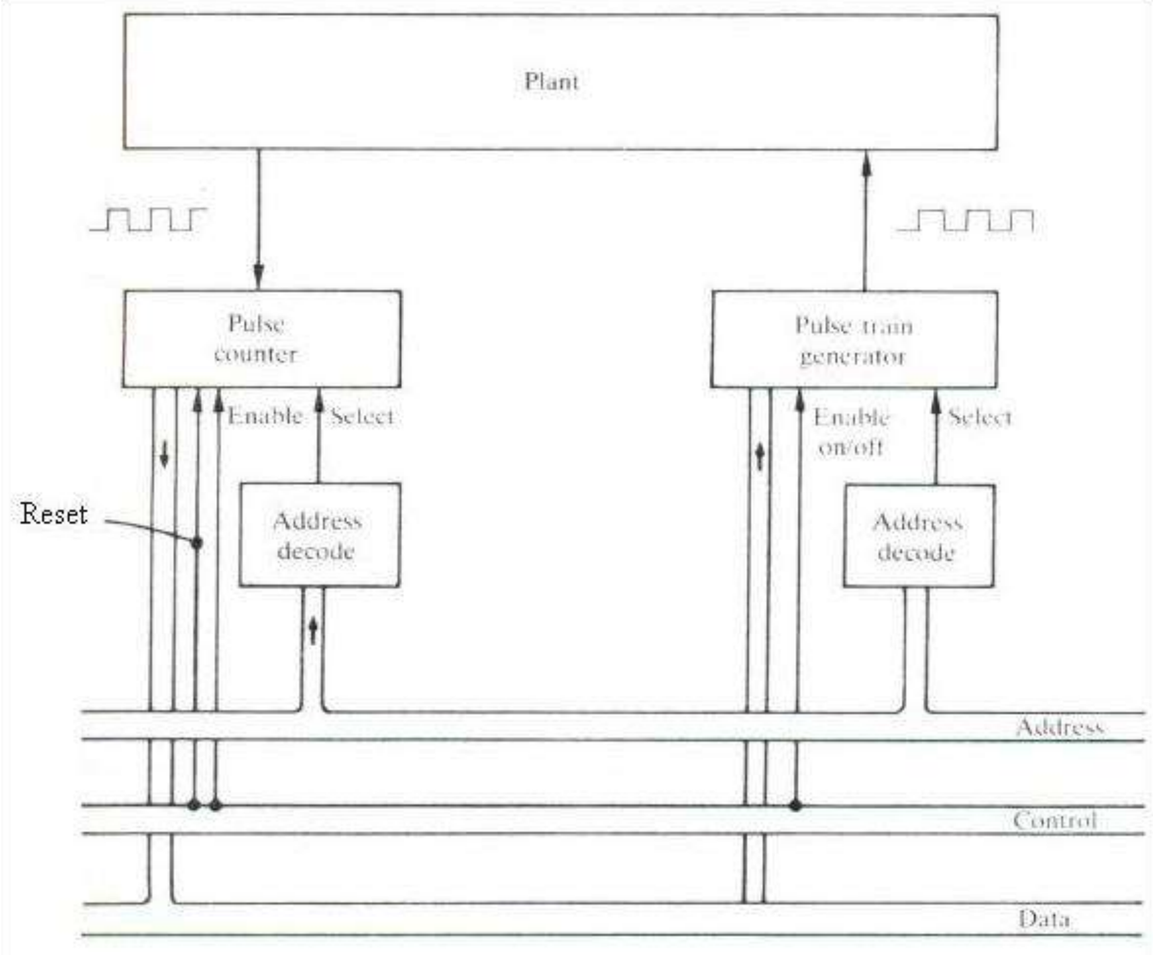


Figure 3.7: Pulse input and output interfaces.

Pulse outputs can take a variety of forms:

1. a series of pulses of fixed duration;
2. a single pulse of variable length (time-proportioned output); and
3. pulse width modulation - a series of pulses of different widths sent at a fixed frequency.

3.10 ANALOG INTERFACES:

The conversion of analog measurements to digital measurements involves two operations: sampling and quantization. The sampling rate necessary for controlling a process is discussed in the next chapter. As is shown in Figure 3.8 many analog-to-digital converters (ADCs) include a 'sample-hold' circuit on the input to the device. The sample time of this unit is much shorter than the sample time required for the process; this sample-hold unit is used to prevent a change in the quantity being measured while it is being converted to a discrete quantity. To operate the analog input interface the computer issues a 'start' or 'sample' signal, typically a short pulse (I microsecond), and in response the ADC switches the 'sample-hold' into SAMPLE for a short period after which the quantization process commences. Quantization may take from a few microseconds to several milliseconds. On completion of the conversion the ADC raises a 'ready' or 'complete' line which is either polled by the computer or is used to generate an
interrupt.

Digital-to-analog conversion is simpler (and hence cheaper) than analog-to-digital conversion and as a consequence it is normal to provide one converter for each output. (It is possible to produce a multiplexer in order to use a single digital-to-analog converter (DAC) for analog output. Why would this solution not be particularly useful?) Figure 3.9 shows a typical arrangement. Each DAC is connected to the data bus and the appropriate channel is selected by putting the channel address on the computer address bus. The DAC acts as a latch and holds the previous value sent to it until the next value is sent. The conversion time is typically from 5 to 20 ms and typical analog outputs are - 5 to + 5 V, - 10 to + 10 V. or a current output of 0 to 20 mA.

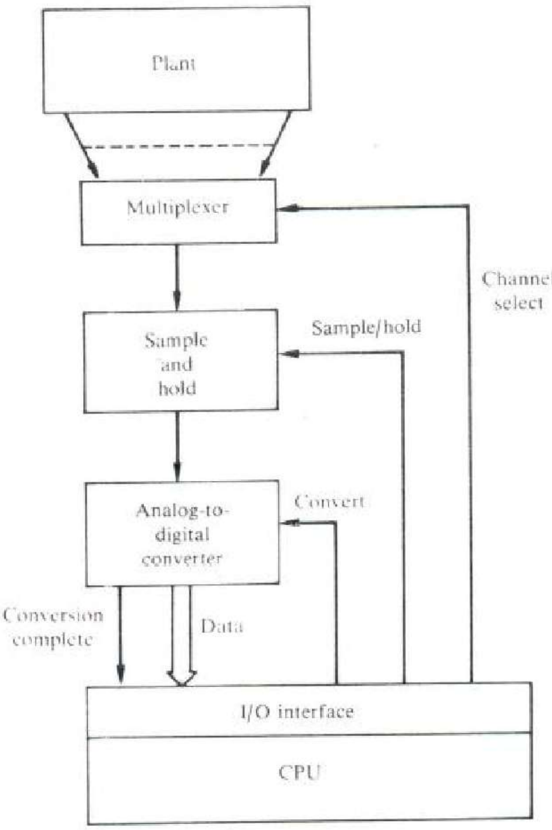


Figure 3.8: Analog input system.

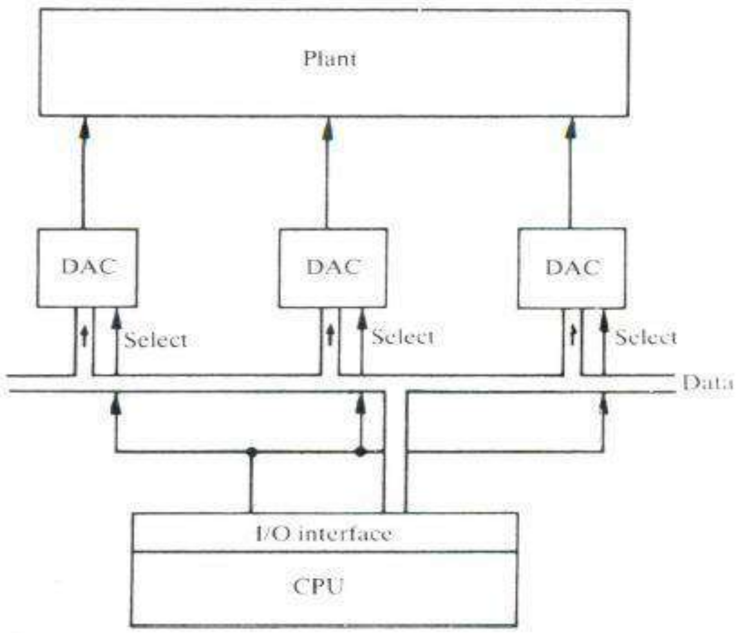


Figure 3.9: Analog output system.

3.11 REAL TIME CLOCK:

A real-time clock is a vital auxiliary device for control computer systems. The hardware unit given the name 'real-time clock' mayor may not be a clock; in many systems it is nothing more than a pulse generator with a precisely controlled frequency. A common form of clock is based on using the ac supply line to generate pulses at 50 (or 60) times per second. By using slightly more complicated circuitry higher pulse rates can be generated, for example 100 (or 120) pulses per second. The pulses are used to generate interrupts and the interrupt handling software counts the interrupts and hence keeps time. If a greater precision in the time measurement than can be provided from the power supply is required then a hardware timer is used. A fixed frequency pulse generator (usually crystal-driven) decrements a counter which, when it reaches zero, generates an interrupt and reloads the count value. The interrupt activates the real-time clock software. The interval at which the timer generates an interrupt, and hence the precision of the clock, is controlled by the count value loaded into the hardware timer.

Real-time clocks are also used in batch processing and on-line computer systems. In the former, they are used to provide date and time on printouts and also for accounting purposes so that a user can be charged for the computer time used; the charge may vary depending on the time of day or day of the week. In on-line systems similar facilities to those of the batch computer system are required, but in addition the user expects the terminal to appear as if it is the only terminal connected to the system. The user may expect delays when the program is performing a large amount of calculation but not when it is communicating with the terminal. To avoid any one program causing delays to other programs, no program is allowed to run for more than a fraction of a second; typically timings are 200 ms or less. If further processing for a particular program is required it is only performed after all other programs have been given the opportunity to run. This technique is known as time slicing.

3.12 DATA TRANSFER TECHNIQUES:

Although the meaning of the data transmitted by the various processes, the operator and computer peripherals differ, there are many common features which relate to the transfer of the data from the interface to the computer. A characteristic of most interface devices is that they operate synchronously with respect to the computer and that they operate at much lower speeds. Direct control of the interface devices by the computer is known as 'programmed transfer' and involves use of the CPU. Programmed transfer gives maximum flexibility of operation but because of the difference in operating speeds of the CPU and many interface devices it is inefficient. An alternative approach is to use direct memory access (DMA); the transfer requirements are set up using program control but the data transfers take place directly between the device and memory without disturbing the operation of the CPU (except that bus cycles are used). With the reduction in cost of integrated circuits and microprocessors, detailed control of the input/output operations is being transferred to I/O processors which provide buffered entry.

For a long time in on-line computing, buffers have been used to collect information (for example, a line) before invoking the program requesting the input. This approach is now being extended through the provision of I/O processors for real-time systems. For example, an I/O processor can be used to control the scanning of a number of analog input channels, only requesting main computer time when it has collected data from all the channels. This can be extended so that the I/O processor checks the data to test if any values are outside preset limits set by the main system. A major problem in data transfer is timing. It may be thought that under programmed transfer, the computer can read or write at any time to a device, that is, can make an unconditional transfer. For some process output devices, for example switches and indicator lights connected to a digital output interface, or for DACs, unconditional transfer is possible since they are always ready to receive data. For other output devices, for example printers and communications channels, which are not fast enough to keep up with the computer but must accept a sequence of data items without missing any item, unconditional transfer cannot be used. The computer must be sure that the device is ready to accept the next item of data; hence either a timing loop to synchronies the computer to the external device or conditional transfer has to be used. Conditional transfer can be used for digital inputs but not usually for pulse inputs or analog inputs.

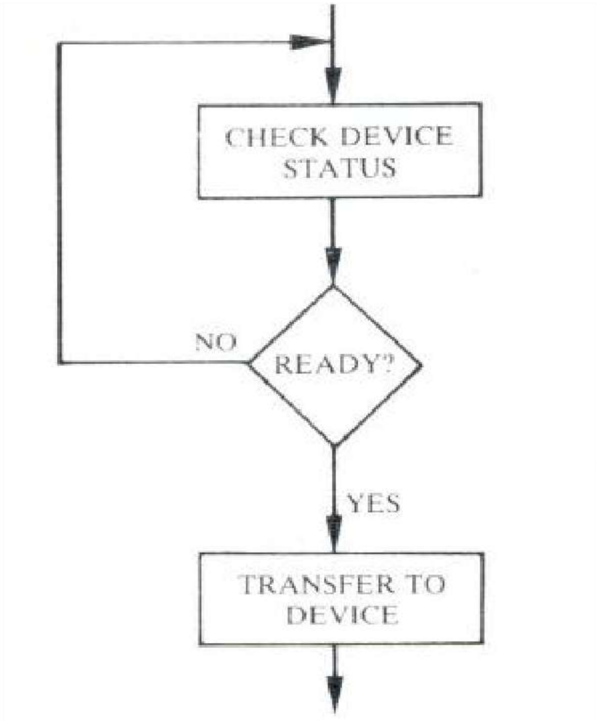


Figure 3.10: Conditional transfer (busy wait).

3.13 COMMUNCIATIONS:

The use of distributed computer systems implies the need for communication: between instruments on the plant and the low-level computers (see Figure 3.20); between the Level land Level 2 computers; and between the Level 2 and the higher level computers. At the plant level communications systems typically involve parallel analog and digital signal transmission techniques since the distances over which communication is required are small and high-speed communication is usually required. At the higher levels it is more usual to use serial communication methods since, as communication distances extend beyond a few hundred yards, the use of parallel cabling rapidly becomes cumbersome and costly. As the distance between the source and receiver increases it becomes more difficult, when using analog techniques, to obtain a high signal-to-noise ratio; this is particularly so in an industrial environment where there may be numerous sources of interference. Analog systems are therefore generally limited to short distances. The use of parallel digital transmission provides high data transfer rates but is expensive in terms of cabling and interface circuitry and again is normally only used over short distances (or when very high rates of transfer are required).

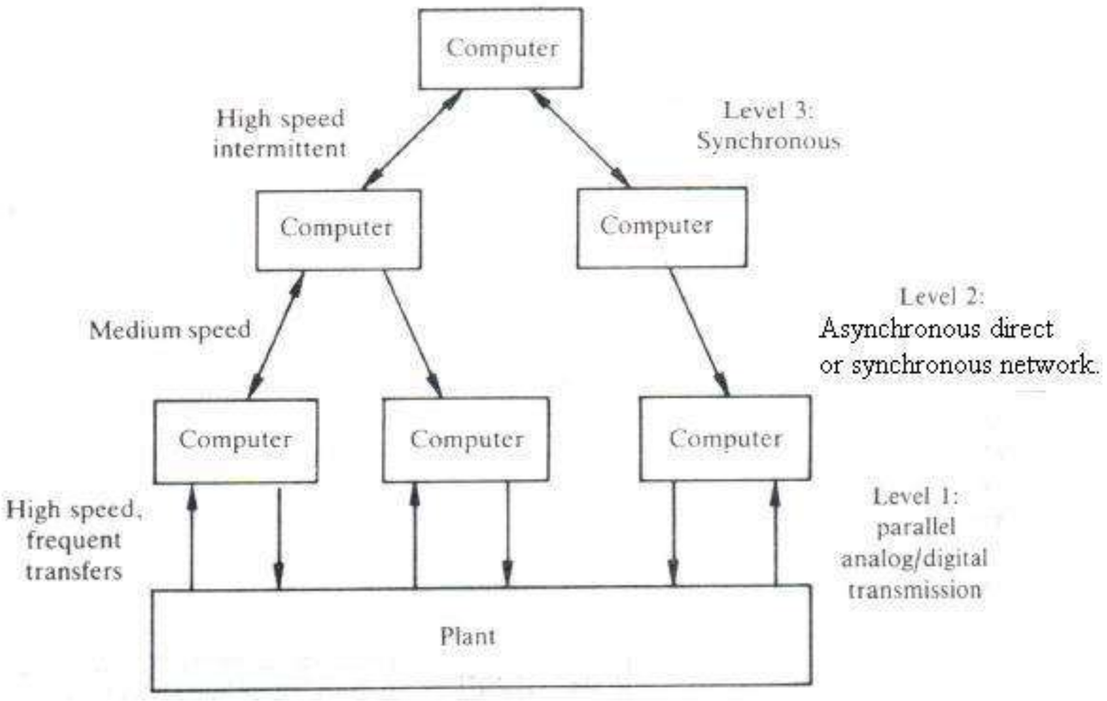

Computer
High speed intermittent
Level 3: Synchronous
Computer
Computer
Medium speed
Level 2: Asynchronous direct or synchronous network.
Computer
Computer
Computer
High speed, frequent transfers
Level 1: parallel analog/digital transmission
Plant

Figure 3.20: Data transmission links.

Serial communications can be characterized in several ways:

1. Mode
   (a) Asynchronous
   (b) Synchronous
2. Quantity
   (a) Character by character
   (b) Block
3. Distance
   (a) Local
   (b) Remote that is wide area
4. Code
   (a) ASCII
   (b) Other

3.14 STANDARD INTERFACES:

Most of the companies which supply computers for real-time control have developed their own 'standard' interfaces, such as the Digital Equipment Corporation's Q-bus for the PDP-ll series, and, typically, they, and independent suppliers, will be able to offer a large range of interface cards for such systems. The difficulty with the standards supported by particular manufacturers is that they are not compatible with each other; hence a change of computer necessitates a redesign of the interface. An early attempt to produce an independent standard was made by the British Standards Institution (BS 4421, 1969). Unfortunately the standard is limited to the concept of how the devices should interconnect and the standard does not define the hardware. It is not widely used and has been overtaken by more recent developments.

An interface which was originally designed for use in atomic energy research laboratories - the computer automated measurement and control (CAMAC) system - has been widely adopted in laboratories, the nuclear industry and some other industries. There are also FORTRAN libraries which provide software to support a wide range of the interface modules. One of the attractions of the system is that the CAMAC data highway) connects to the computer by a special card; to change to a different computer only requires that the one card be changed. A general purpose interface bus (GPIB) was developed by the Hewlett Packard Company in the early 1970s for connecting laboratory instruments to a computer. The system was adopted by the IEEE and standardized as the IEEE 488 bus system.

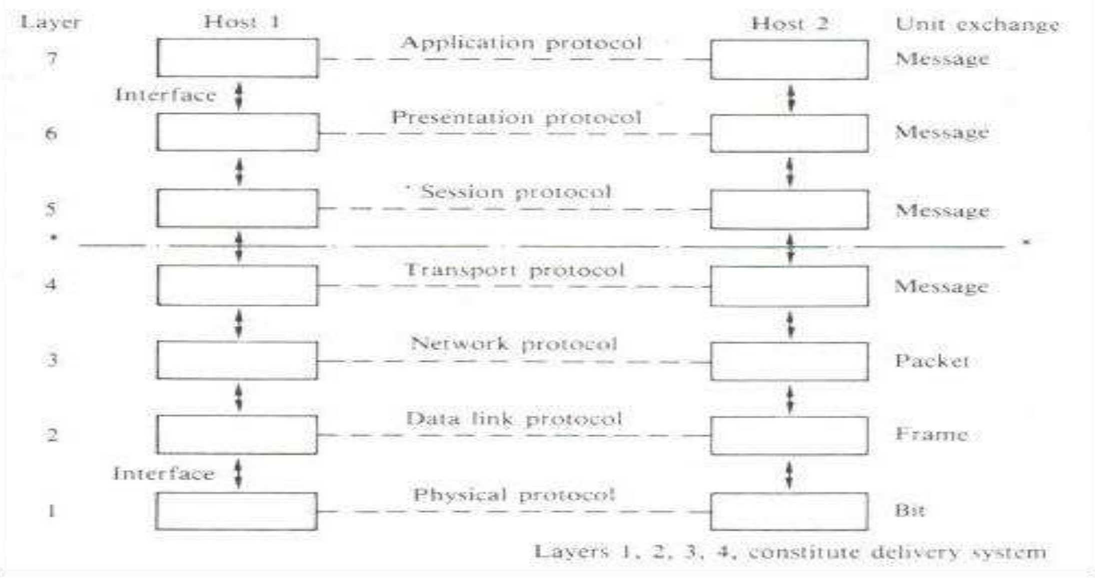
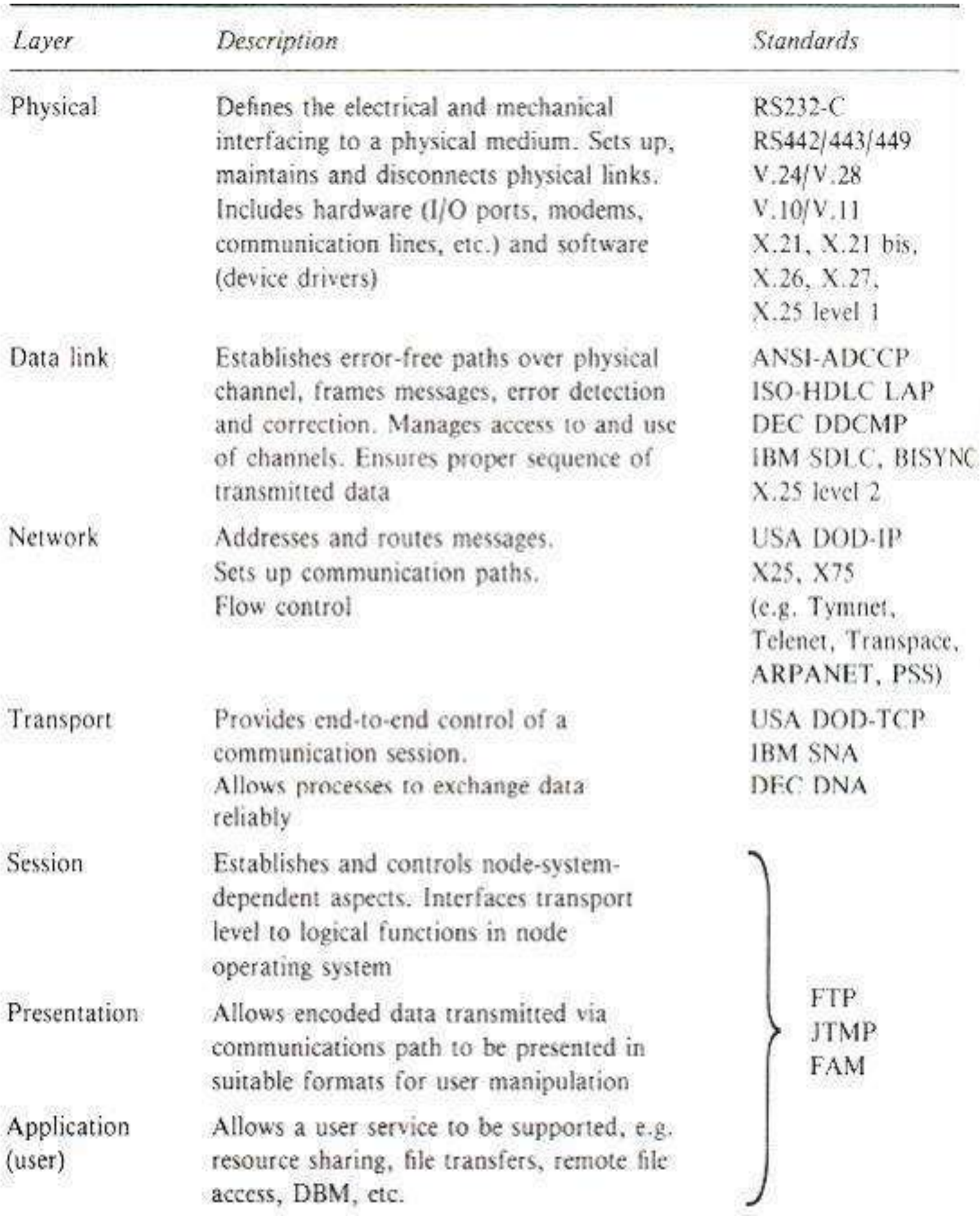


Figure 3.25: ISO seven - layer model.



Layer	*Description*	*Standards*
Physical	Defines the electrical and mechanical interfacing to a physical medium. Sets up, maintains and disconnects physical links. Includes hardware (I/O ports, modems, communication lines, etc.) and software (device drivers)	RS232-C RS442/443/449 V.24/V.28 V.10/V.11 X.21, X.21 bis, X.26, X.27, X.25 level 1
Data link	Establishes error-free paths over physical channel, frames messages, error detection and correction. Manages access to and use of channels. Ensures proper sequence of transmitted data	ANSI-ADCCP ISO-HDLC LAP DEC DDCMP IBM SDLC, BISYNC X.25 level 2
Network	Addresses and routes messages. Sets up communication paths. Flow control	USA DOD-IP X25, X75 (e.g. Tymnet, Telenet, Transpace, ARPANET, PSS)
Transport	Provides end-to-end control of a communication session. Allows processes to exchange data reliably	USA DOD-TCP IBM SNA DEC DNA
Session	Establishes and controls node-system-dependent aspects. Interfaces transport level to logical functions in node operating system	FTP JTMP FAM
Presentation	Allows encoded data transmitted via communications path to be presented in suitable formats for user manipulation	
Application (user)	Allows a user service to be supported, e.g. resource sharing, file transfers, remote file access, DBM, etc.	

Table: ISO seven layer model.

The bus can connect up to a maximum of 15 devices and is only suited to laboratory or small, simple control applications. The ISO (International Organization for Standardization) have promulgated a standard protocol system in the Open Systems Interconnection (OSI) model. This is a layered (hierarchical) model with seven layers running from the basic physical connection to the highest application protocol.

Recommended questions:

1. There are a number of different types of analog-to-digital converters. List them and discuss typical applications for each type (see, for example, Woolvet (1977) or Barney (1985)).
2. The clock on a computer system generates an interrupt every 20 ms. Draw a flowchart for the interrupt service routine. The routine has to keep a 24 hour clock in hours, minutes and seconds.
3. Twenty analog signals from a plant have to be processed (sampled and digitized) every 1s. the analog-to-digital converter and multiplexer which is available can operate in two modes: automatic scan and computer-controlled scan. In the automatic scan mode, on receipt of a 'start' signal the converter cycles through each channel in turn.
4. A turbine flow meter generates pulses proportional to the flow rate of a liquid. What methods can be used to interface the device to a computer?
5. Why is memory protection important in real-time systems?
6. What methods can be used to provide memory protection?

MODULE- 3

Languages For Real –Time Applications

Introduction, Syntax Layout and Readability, Declaration and Initialization of Variables and Constants, Modularity and Variables, Compilation , Data Type, Control Structure, Exception Handling, Low –Level Facilities, Co routines, Interrupts and Device Handling, Concurrency, Real –Time Support, Overview of Real –Time Languages.

Recommended book for reading:

1. **Real –Time Computer control –An Introduction**, Stuart Bennet, 2nd Edn. Pearson Education 2005.
2. **Real-Time Systems Design and Analysis**, Phillip. A. Laplante, Second Edition, PHI, 2005.
3. **Real time Systems Development**, Rob Williams, Elsevier, 2006.

LANGUAGES FOR REAL-TIME APPLICATIONS

4.1 INTRODUCTION

Languages are an important implementation tool for all systems that include embedded computers. To understand fully methods for designing software for such systems one needs to have a sound understanding of the range of implementation languages available and of the facilities which they offer. The range of languages with features for real-time use continues to grow, as do the range and type of features offered. In this chapter we concentrate on the fundamental requirements of a good language for real-time applications and will illustrate these with examples drawn largely from Modula-2 and Ada. Producing safe real-time software places heavy demands on programming languages. Real-time software must be reliable: the failure of a real-time system can be expensive both in terms of lost production, or in some cases, in the loss of human life (for example, through the failure of an aircraft control system).

Real-time systems are frequently large and complex, factors which make development and maintenance costly. Such systems have to respond to external events with a guaranteed response time; they also involve a wide range of interface devices, including non-standard devices. In many applications efficiency in the use of the computer hardware is vital in order to obtain the necessary speed of operation. Early real-time systems were necessarily programmed using assembly level languages, largely because of the need for efficient use of the CPU, and access interface devices and support interrupts. Assembly coding is still widely used for small systems with very high computing speed requirements, or for small systems which will be used in large numbers. In the latter case the high cost of development is offset by the reduction in unit cost through having a small, efficient, program. Dissatisfaction with assemblers (and with high-level languages such as FORTRAN which began to be used as it was recognized that for many applications the advantages of high-level languages outweighed their disadvantages) led to the development of new languages for programming embedded computers.

The limitation of all of them is that they are designed essentially for producing sequential programs and hence rely on operating system support for concurrency. The features that a programmer demands of a real-time language subsume those demanded of a general purpose language and so many of the features described below are also present (or desirable) in languages which do not support real-time operations. Barnes (1976) and Young (1982) divided the requirements that a user looked for in a programming language into six general areas. These are listed below in order of importance for real-time applications:

- Security.
- Readability.
- Flexibility.
- Simplicity.
- Portability.
- Efficiency.

In the following sections we will examine how the basic features of languages meet the requirements of the user as given above. The basic language features examined are:

- Variables and constants: declarations, initialization.
- Data types - including structured types and pointers.
- Control structures and program layout and syntax.
- Scope and visibility rules.
- Modularity and compilation methods.

- Exception handling.

A language for real-time use must support the construction of programs that exhibit concurrency and this requires support for:

- Construction of modules (software components).
- Creation and management of tasks.
- Handling of interrupts and devices.
- Intertask communication.
- Mutual exclusion.
- Exception handling.

4.1.1 SECURITY:

Security of a language is measured in terms of how effective the compiler and the run-time support system is in detecting programming errors automatically. Obviously there are some errors which cannot be detected by the compiler regardless of any features provided by the language: for example, errors in the logical design of the program. The chance of such errors occurring is reduced if the language encourages the programmer to write clear, well-structured, code. Language features that assist in the detection of errors by the compiler include:

- good modularity support;
- enforced declaration of variables;
- good range of data types, including sub-range types;
- typing of variables; and
- unambiguous syntax.

It is not possible to test software exhaustively and yet a fundamental requirement of real-time systems is that they operate reliably. The intrinsic security of a language is therefore of major importance for the production of reliable programs. In real-time system development the compilation is often performed on a different computer than the one used in the actual system, whereas run-time testing has to be done on the actual hardware and, in the later stages, on the hardware connected to plant. Run-time testing is therefore expensive and can interfere with the hardware development program. Economically it is important to detect errors at the compilation stage rather than at run-time since the earlier the error is detected the less it costs to correct it. Also checks done at compilation time have no run-time overheads.

4.1.2 READABILITY:

Readability is a measure of the ease with which the operation of a program can be understood without resort to supplementary documentation such as flowcharts or natural language descriptions. The emphasis is on ease of reading because a particular segment of code will be written only once but will be read many times. The benefits of good readability are:

- Reduction in documentation costs: the code itself provides the bulk of the documentation. This is particularly valuable in projects with a long life expectancy in which inevitably there will be a series of modifications. Obtaining up-to-date documentation and keeping documentation up to date can be very difficult and costly.
- Easy error detection: clear readable code makes errors, for example logical errors, easier to detect and hence increases reliability.
- Easy maintenance: it is frequently the case that when modifications to a program are required the person responsible for making the modifications was not involved in the original design - changes can only be made quickly and safely if the operation of the program is clear.

4.1.3 FLEXIBILITY:

A language must provide all the features necessary for the expression of all the operations required by the application without requiring the use of complicated constructions and tricks, or resort to assembly level code inserts. The flexibility of a language is a measure of this facility. It is particularly important in real-time systems since frequently non-standard I/O devices will have to be controlled. The achievement of high flexibility can conflict with achieving high security. The compromise that is reached in modern languages is to provide high flexibility and, through the *module* or *package* concept, a means by which the low-level (that is, insecure) operations can be hidden in a limited number of self-contained sections of the program.

4.1.4 SIMPLICITY:

In language design, as in other areas of design, the simple is to be preferred to the complex. Simplicity contributes to security. It reduces the cost of training, it reduces the probability of programming errors arising from misinterpretation of the language features, it reduces compiler size and it leads to more efficient object code. Associated with simplicity is consistency: a good language should not impose arbitrary restrictions (or relaxations) on the use of any feature of the language.

4.1.5 PORTABLITILY:

Portability, while desirable as a means of speeding up development, reducing costs and increasing security, is difficult to achieve in practice. Surface portability has improved with the standardization agreements on many languages. It is often possible to transfer a program in source code form from one computer to another and find that it will compile and run on the computer to which it has been transferred. There are, however, still problems when the word lengths of the two machines differ and there may also be problems with the precision with which numbers are represented even on computers with the same word length.

Portability is more difficult for real-time systems as they often make use of specific features of the computer hardware and the operating system. A practical solution is to accept that a real-time system will not be directly portable, and to Restrict the areas of non-portability to specific modules by restricting the use of low level features to a restricted range of modules. Portability can be further enhanced by writing the application software to run on a virtual machine, rather than for a specific operating system.

4.1.6 EFFICIENCY:

In real-time systems, which must provide a guaranteed performance and meet specific time constraints, efficiency is obviously important. In the early computer control systems great emphasis was placed on the efficiency of the coding - both in terms of the size of the object code and in the speed of operation - as computers were both expensive and, by today's standards, very slow. As a consequence programming was carried out using assembly languages and frequently 'tricks' were used to keep the code small and fast. The requirement for generating efficient object code was carried over into the designs of the early real-time languages and in these languages the emphasis was on efficiency rather than security and readability. The falling costs of hardware and the increase in the computational speed of computers have changed the emphasis. Also in a large number of real-time applications the concept of an efficient language has changed to include considerations of the security and the costs of writing and maintaining the program; speed and compactness of the object code have become, for the majority of applications, of secondary importance.

4.1.7 SYNTAX LAYOUT AND READAILITY:

The language syntax and its layout rules have a major impact on the readability of code written in the language. Consider the program fragment given below: BEGIN

```
NST: = TICKS ( ),. ST;
T: =TICKS ()+ST;
LOOP
WHILE TICKS ( )< NST DO (* nothing *) END;
T: =TICKS ();
C C;
NST: = T+ST;
IF KEYPRESSED ( ) THEN EXIT;
END;
END;
END;
```

Without some explanation and comment the meaning is completely obscure. By using long identifiers instead of, for example N S T and ST, it is possible to make the code more readable.

```
BEGIN
NEXTSAMPLETIME: = TICKSO+SAMPLETIME;
TIME: =TICKS () +SAMPLETIME;
LOOP
WHILE TICKSO< NEXTSAMPLETIME DO (* NOTHING
*) END;
```

```
TIME: =TICKSO;
CONTROLCALCULATION;
NEXTSAMPLETIME: =TIME+SAMPLETIME;
IF KEYPRESSEDOTHEN EXIT;
END;
END;
END;
```

The meaning is now a little clearer, although the code is not easy to read because it is entirely in upper case letters. We find it much easier to read lower case text than upper case and hence readability is improved if the language permits the use of lower case text. It also helps if we can use a different case (or some form of distinguishing mark) to identify the reserved words of the language. Reserved words are those used to identify particular language constructs, for example repetition statements, variable declarations, etc. In the next version we use upper case for the reserved words and a mixture of upper and lower case for user-defined entities.

```
BEGIN
NextSampleTime: = Ticks ( ) +Sample Time;
Time: =Ticks ( ) +Sample Time;
LOOP
WHILE Ticks ( ) < NextSampleTime DO (* nothing *)
END;
Time: =Ticks ( );
Control Calculation;
NextSampleTime: = Time + Sample Time;
IF Key Pressed ( ) THEN EXIT;
END;
END;
END;
```

The program is now much easier to read in that we can easily and quickly pick out the reserved words. It can be made even easier to read if the language allows embedded spaces and tab characters to be used to improve the layout.

4.2 DECLARATION AND INTIALIZATION OF VARIABLES AND CONSTANTS.

DECLARATION:

The purpose of declaring an entity used in a program is to provide the compiler with information on the storage requirements and to inform the system explicitly of the names being used. Languages such as Pascal, Modula-2 and Ada require all objects to be specifically declared and a type to be associated with the entity when it is declared. The provision of type information allows the compiler to check that the entity is used only in operations associated with that type. If, for example, an entity is declared as being of type REA L and then it is used as an operand in logical operation, the compiler should detect the type incompatibility and flag the statement as being incorrect. Some older languages, for example BASIC and FORTRAN, do not require explicit declarations; the first use of a name is deemed to be its declaration. In FORTRAN explicit declaration is optional and entities can be associated with a type jf declared. If entities are not declared then implicit typing takes place: names beginning with the letters I-N are assumed to be integer numbers; names beginning with any other letter are assumed to be real numbers.

Optional declarations are dangerous because they can lead to the construction of syntactically correct but functionally erroneous programs. Consider the following program fragment:

```
100 ERROR=0
   …….
200 IF X=Y THEN GOTO 300
250 EROR=1
300...
```

In FORTRAN (or BASIC), ERROR and EROR will be considered as two different variables whereas the programmer's intention was that they should be the same – the variable ER 0 R in line 250 has been mistyped. FORTRAN compilers cannot detect this type of error and it is a *characteristic* error of FORTRAN. Many organizations which use FORTRAN extensively avoid such errors by insisting that all entities are declared and the code is processed by a preprocessor which checks that all names used are mentioned in declaration statements. INTIALIZATION:

It is useful if a variable can be given an initial value when it is declared. It is bad practice to rely on the compiler to initialize variables to zero or some other value.

This is not, of course, strictly necessary as a value can always be assigned to a variable. In terms of the security of a language it is important that the compiler checks that a variable is not used before it has had a value assigned to it. The security of languages such as Modula-2 is enhanced by the compiler checking that all variables have been given an initial value. However, a weakness of Modula-2 is that variables cannot be given an initial value when they are declared but have to be initialized explicitly using an assignment statement. CONSTANTS

Some of the entities referenced in a program will have constant values either because they are physical or mathematical entities such as the speed of light or because they are a parameter which is fixed for that particular implementation of the program ,for example the number of control loops being used or the bus address for an input or output device. It is always possible to provide constants by initializing a variable to the appropriate quantity, but this has the disadvantage that it is in secure in that the compiler cannot detect if a further assignment is made which changes the value of the constant. It is also confusing to the reader since there is no indication which entities are constants and which are variables (unless the initial assignment is carefully documented). Pascal provides a mechanism for declaring constants, but since the constant declarations must precede the type declarations, only constants of the predefined types can be declared. This is a severe restriction on the constant mechanism. For example, it is not possible to do the following: TYPE

```
A Motor State = (OFF, LOW, MEDIUM, HIGH);
CONST
Motor Stop = A Motor State (OFF);
```

A further restriction in the constant declaration mechanism in Pascal is that the value of the constant must be known at compilation time and expressions are not permitted in constant declarations. The restriction on the use of expressions in constant declarations is removed in Modula-2 (experienced assembler programmers will know the usefulness of being able to use expressions in constant declarations).

For example, in Modula-2 the following are valid constant declarations:

```
CONST
message = 'a string of characters';
```

```
length = 1.6;
breadth = 0.5;
area = length * breadth;
```

4.3 MODULARITY AND VARIABLES:

Scope and visibility:

The scope of a variable is defined as the region of a program in which the variation is potentially accessible or modifiable. The regions in which it may actually accessed or modified are the regions in which it is said to be visible. Most languages provide mechanisms for controlling scope and visibility. There are two general approaches: languages such as FORTRAN provide a single level locality whereas the block-structured languages such as Modula-2 provide multilevel locality. In the block-structured languages entities which are declared within a block, only be referenced inside that block. Blocks can be nested and the scope extended throughout any nested blocks. This is illustrated in Example which shows scope for a nested PROCEDURE in Modula-2.

```
MODULE
ScopeExampLe1;
VAR
A, B: INTEGER;
PROCEDURE Level One;
VAR
B, C: INTEGER;
BEGIN
( *
*)
END (* Level one *);
BEGIN
(*
A and B visible here but not Level One and
Level One .C
*)
END ScopeExample1.
```

The *scope* of variables A and B declared in the main module ScopeExample1 extends throughout the program that is they are global variables.

Global and local variables:

Although the compiler can easily handle the reuse of names, it is not as easy for the programmer and the use of deeply nested PRO CEO UR E blocks with the reuse of names can compromise the security of a Pascal or Modula-2 program. As the program shown in Example illustrates the reuse of names can cause confusion as to which entity is being referenced.

```
MODULE ScopeL2;
VAR X. Y, Z: INTEGER;
PROCEDURE L 1;
VAR Y: INTEGER;
PROCEDURE L2;
VAR X: INTEGER;
PROCEDURE L3;
VAR Z: INTEGER;
PROCEDURE L4;
BEGIN
Y: = 25; (* L1.Y NOT
LO.Y*) END L4;
BEGIN
(* L1.Y. L2.X, L3.Z visible
*) END L3;
BEGIN
(* L1.Y, L2.X. LO.Z visible
*) END L2;
BEGIN
(* LO.X, L1.Y. LO.Z visible *)
END L 1 ;
BEGIN
(* ••• *)Scope L2.
```

It is very easy to assume in assigning the value 25 to Y in PROCEDURE L4 that the global variable Y is being referenced, when in fact it is the variable Y declared in PROCEDURE L 1 that is being referenced.

4.4 COMPILATION OF MODULAR PROGRAM:

If we have to use a modular approach in designing software how do we compile the modules to obtain executable object code? There are two basic approaches: either combine at the source code level to form a single unit which is then compiled, or compile the individual modules separately and then in some way link the compiled version of each module to form the executable program code. Using the second approach a special piece of software called a linker has to be provided as part of the compilation support to do the linking of the modules. A reason for the popularity and widespread use of FORTRAN for engineering and scientific work is that subroutines can be compiled independently from the main program, and from each other. The ability to carry out compilation independently arises from the single-level scope rules of FORTRAN; the compiler makes the assumption that any entity which is referenced in a subroutine, but not declared within that subroutine, will be declared externally and hence it simply inserts the necessary external linkage to enable the linker to attach the appropriate code. It must be stressed that the compilation is independent that is when a main program is compiled the compiler has no information available which will enable it to check that the reference to the subroutine is correct.

For example, a subroutine may expect three real variables as parameters, but if the user supplies four integer variables in the call statement the error will not be detected by the compiler. Independent compilation of most block-structured languages is even more difficult and prone to errors in that arbitrary restrictions on the use of variables have to be imposed. Many errors can be detected at the linking stage. However, because linking comes later in the implementation process errors discovered at this stage are more costly to correct. It is preferable to design the language and compilation system in such a way as to be able to detect as many errors as possible during compilation instead of when linking. Both Modula-2 and Ada have introduced the idea of separate compilation units. Separate compilation implies that the compiler is provided with some information about the previously or separately compiled units which are to be incorporated into a program. In the case of Modula-2 the source code of the DEFINITION part of a separately compiled module must be made available to the user, and hence the compiler. This enables the compiler to carry out the normal type checking and procedure parameter matching checks. Thus in Modula-2 type mismatches and procedure parameter errors are detectable by the compiler. It also makes available the scope control features of Modula-2. The provision of independent compilation of the type introduced in FORTRAN represented a major advance in supporting software development because it enabled the development of extensive object code libraries. Languages which support separate compilation represent a further advance in that they add greater security and easy error checking to library use.

4.5 DATA TYPES:

As we have seen above, the allocation of types is closely associated with the declaration of entities. The allocation of a type defines the set of values that can be taken by an entity of that type and the set of operations that can be performed on the entity. The richness of types supported by a language and the degree of rigour with which type compatibility is enforced by the language are important influences on the security of programs written in the language. Languages which rigorously enforce type compatibility are said to be *strongly* typed; languages which do not enforce type compatibility are said to be *weakly* typed. FORTRAN and BASIC are weakly typed languages: they enforce some type checking; for example, the statements A $ = 2 5 or A = X$ + Yare not allowed in BASIC, but they allow mixed integer and real arithmetic and provide implicit type changing in arithmetic statements. Both languages support only a limited number of types.

An example of a language which is strongly typed is Modula-2. In addition to enforcing type checking on standard types, Modula-2 also supports enumerated types. The enumerated type allows programmers to define their own types in addition to using the predefined types. Consider a simple motor speed control system which has four settings 0 F F, LOW, ME DIU M, H I GH and which is controlled from a computer system. Using Modula-2 the programmer could make the declarations:

```
TYPE
AMotorState = (OFF, LOW, MEDIUM, HIGH);
VAR
motor Speed: AMotorState;
```

The variable motor Speed can be assigned only one of the values enumerated in
the T YP E definition statement. An attempt to assign any other value will be trapped

by the compiler, for example the statement will be flagged as an error.

If we contrast this with the way in which the system could be programmed
using FORTRAN we can see some of the protection which strong typing
provides. In ANSI FORTRAN integers must be used to represent the four states
of the motor Control:

INTEGER OFF, LOW, MEDIUM, HIGH

DATA OFF/0/, LOW/1/, MEDIUM/2/, HIGH/3/

If the programmer is disciplined and only uses the defined integers to set MSPEED then the program is clear and readable, but there is no mechanism to prevent direct assignment of any value to MS PEE D.

Hence the statements

MSPEED = 24

MSPEED = 1 SO

would be considered as valid and would not be flagged as errors either by the compiler or by the run-time system. The only way in which they could be detected is if the programmer inserted some code to check the range of values before sending them to the controller. In FORTRAN a programmer-inserted check would be necessary since the output of a value outside the range 0 to 3 may have an unpredictable effect on the motor speed.

4.6 EXCEPTION HANDLING:

One of the most difficult areas of program design and implementation is the handling of errors, unexpected events (in the sense of not being anticipated and hence catered for at the design stage) and exceptions which make the processing of data by the subsequent segments superfluous, or possibly dangerous. The designer has to make decisions on such questions as what errors are to be detected. What sort of mechanism is to be used to do the detection? And what should be done when an error is detected? Most languages provide some sort of automatic error detection mechanisms as part of their run-time support system. Typically they trap errors such as an attempt to divide by zero, arithmetic overflow, array bound violations, and sub-range violations; they may also include traps for input/output errors. For many of the checks the compiler has to add code to the program; hence the checks increase the size of the code and' reduce the speed at which it executes. In most languages the normal response when an error is detected is to halt the program and display an error message on the user's terminal. In a development environment it may be acceptable for a program to halt following an error; in a real-time system halting the program is not acceptable as it may compromise the safety of the system. Every attempt must be made to keep the system running.

4.7 LOW LEVEL FACILITIES:

In programming real-time systems we frequently need to manipulate directly data in specific registers in the computer system, for example in memory registers, CPU registers and registers in an input! output device. In the older, high-level languages, assembly-coded routines are used to do this. Some languages provide extensions to avoid the use of assembly routines and these typically are of the type found in many versions of BASIC. These take the following form: PEEK (address) - returns as INTEGER variable contents of the location address.

POKE (address, value) - puts the INTEGER value in the location address.

It should be noted that on eight-bit computers the integer values must be in the range o to 255 and on 16 bit machines they can be in the range 0 to 65 535. For computer systems in which the input/output devices are not memory mapped, for example Z80 systems, additional functions are usually provided such as INP (address) and OUT (address, value). A slightly different approach has been adopted in BBC BASIC which uses an 'indirection' operator. The indirection operator indicates that the variable which follows it is to be treated as a pointer which contains the address of the operand rather than the operand itself (the term indirection is derived from the indirect addressing mode in assembly languages). Thus in BBC BASIC the following code

```
100 DACAddress=&FE60
120? DACAddress=&34
```

results in the hexadecimal number 34 being loaded into location FE 60 H; the indirection operator is '?'. In some of the so-called Process FORTRAN languages and in CORAL and

RTL/2 additional features which allow manipulation of the bits in an integer variable are provided, for example

SETBITJ (I),

IF BIT J(I) n1 ,n2 (where I refers to the bit In

variable.

Also available are operations such as AND, 0 R, S LA, S RA, etc., which mimic the operations available at assembly level. The weakness of implementing low-level facilities in this way is that all type checking is lost and it is very easy to make mistakes. A much more secure method is to allow the programmer to declare the address of the register or memory location and to be able to associate a type with the declaration, for example

which declares a variable of type CHAR located at memory location 0 FE60 H.

Characters can then be written to this location by simple assignment

Modula-2 provides a low-level support mechanism through a simple set of primitives which have to be encapsulated in a small nucleus coded in the assembly language of the computer on which the system is to run. Access to the primitives is through a module SYS TEM which is known to the compiler. SYST EM can be thought of as the software bus linking the nucleus to the rest of the software modules. SYSTEM makes available three data types, WORD, ADDRESS, PROCESS, and six procedures, ADR, SIZE, TSIZE, NEWPROCESS, TRANSFER, I 0 TRANS FE R. W0 RD is the data type which specifies a variable which maps onto one unit of the specific computer storage. As such the number of bits in a WORD will vary from implementation to implementation; for example, on a PDP·II implementation a WORD is 16 bits, but on a 68000 it would be 32 bits. ADDRESS corresponds to the definition TYPEA DDRES S = POI NTER TOW0 RD, that is objects of type ADDRES S are pointers to memory units and can be used to compute the addresses of memory words. Objects of type PROC ESS have associated with them storage for the volatile environment of the particular computer on which Modula-2 is implemented; they make it possible to create easily process (task) descriptors. Three of the procedures provided by SYSTEM are for address manipulation:

FROM S

AD

EXPOR

ADR (v) returns the ADDRESS of variable v SIZE

(v) returns the SIZE of variable v in WORDs TSIZE

(t) returns the SIZE of any variable of type t

inWORDs.

In addition variables can be mapped onto specific memory locations. This facility can be used for writing device driver modules in Modula-2. A combination of the low-level access facilities and the module concept allows details of the hardware device to be hidden within a module with only the procedures for accessing the module being made available to the end user.

4.8 CO ROUTINES:

In Modula-2 the basic form of concurrency is provided by co routines. The two procedures NEW PRO C E S sand T RAN S FE R exported by S Y S T EM are defined as follows: PROCEDURE NEWPROCESS (ParameterLessProcedure: PROC);

workspace Address: ADDRESS;

workspace Size: CARDINAL;

VAR co routine: ADDRESS (* PROCESS *));

PROCEDURE TRANSFER (VAR source, destination:

ADDRESS (*PROCESS*));

Any parameter less procedure can be declared as a PROCESS. The procedure NEW PRO C E S S S associates with the procedure storage for the process parameters The amount to be allocated depends on the number and size of the variables local to the procedure forming the coroutine, and to the procedures which it calls. Failure to allocate sufficient space will usually result in a stack overflow error at run-time. The variable co routine is initialized to the address which identifies the newly created co routine and is used as a parameter in calls to T RAN S FER. The transfer of control between co routines is made using a standard procedure T RAN SF ER which has two arguments of type ADD RES S (PROCESS) . The first is the calling co routine and the second is the co routine to which control is to be transferred. The mechanism is illustrated in Example 5.13. In this example the two parameter less procedures form the two co routines which pass control to each other so that the message

Co routine one and Co routine two

is printed out 25 times. At the end of the loop, Co routine 2 passes control back to Main Program.

CONCURRENCY:

Wirth (1982) defined a standard module Processes s which provides a higher-level mechanism than co routines for concurrent programming. The module makes no assumption as to how the processes (tasks) will be implemented; in particular it does not assume that the processes will be implemented on a single processor.

4.9 OVERVIEW OF REAL-TIME:

The best way to start an argument among a group of computer scientists, software engineers or systems engineers is to ask them which is the best language to use for writing software. Rational arguments about the merits and demerits of any particular language are likely to be submerged and lost in a sea of prejudice. Since 1970 high-level languages for the programming and construction of real time systems have become widely available. Early languages include: CORAL (Woodward et a/., 1970) and RTL/2 (Barnes, 1976) as well as modifications to FORTRAN and BASIC. More recently the interest in concurrency and multiprocessing has resulted in many languages with the potential for use with real-time systems. These include Ada (see Young, 1982; Burns and Wellings, 1990), ARGUS (Liskovand Scheifler, 1983), CONIC (Kramer et a/., 1983), CSP (Hoare, 1978), CUTLASS (CEGB, see Bennett and Linkens, 1984), FORTH (Brodie, 1986),

A language suitable for programming real-time and distributed systems must have all the characteristics of a good, modern, non-real-time language; that is it should have a clean syntax, a rational procedure for declarations, initialisation and typing of variables, simple and consistent control structures, clear scope and visibility rules, and should provide support for modular construction. The addition required for real-time use includes support for concurrency or multi-tasking and mechanisms to permit access to the basic computer functions (usually referred low-level constructs).

Recommended question:

1. In the computer science literature you will find lots of arguments about 'global' and 'Local' variables. What guidance would you give to somebody who asked for advice on how to decide on the use of global or local variables?
2. Why is it useful to have available a predefined data type BITSET in Modula-2? Give an example to illustrate how, and under what circumstances, BITSET would be used.
3. How does strong data typing contribute to the security of a programming language?
4. Explain simple table-driven approach used for application oriented software.
5. **W**hat are .the major requirements for CUTLASS? Explain in detail, with host-target

Configuration.

6. How do strong data typing contribute to the security of programming language?
7. What are the requirements, which CUTLASS has to meet? With a neat diagram, show the CUTLASS host-target configuration.

Module-4

Operating Systems

Introduction, Real –Time Multi –Tasking OS, Scheduling Strategies, Priority Structures, Task Management, Scheduler and Real –Time Clock Interrupt Handles, Memory Management ,Code Sharing, Resource control, Task Co-operation and Communication, Mutual Exclusion, Data Transfer, Liveness, Minimum OS Kernel, Examples.

Recommended book for reading:

1. **Real –Time Computer control –An Introduction**, Stuart Bennet, 2nd Edn. Pearson Education 2005.
2. **Real-Time Systems Design and Analysis**, Phillip. A. Laplante, Second Edition, PHI, 2005.
3. **Real time Systems Development**, Rob Williams, Elsevier, 2006.

5.1 OPERATING SYSTEMS

INTRODUCTION

Specific computer using a particular language can be hidden from the designer. An operating system for a given computer converts the hardware of the system into a virtual machine with characteristics defined by the operating system. Operating systems were developed, as their name implies, to assist the operator in running a batch processing computer; they then developed to support both real-time systems and multi-access on-line systems. The traditional approach is to incorporate all the requirements inside a general purpose operating system as illustrated in Figure 6.1. Access to the hardware of the system and to the I/O devices is through the operating system. In many real-time and multi-programming systems restriction of access is enforced by hardware and software traps.

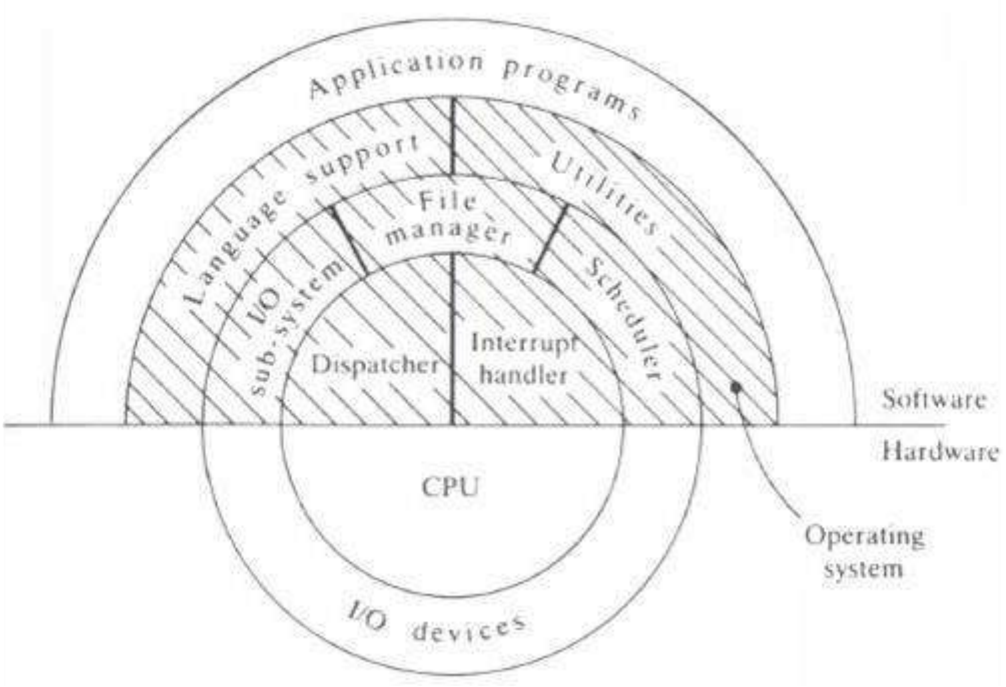


Figure 6.1: General purpose operating system.

The operating system is constructed, in these cases, as a monolithic monitor. In single-job operating systems access through the operating system is not usually enforced; however, it is good programming practice and it facilitates portability since the operating system entry points remain constant across different implementations. In addition to supporting and controlling the basic activities, operating systems provide various utility programs, for example loaders, linkers, assemblers and debuggers, as well as run-time support for high-level languages.

A general purpose operating system will provide some facilities that are not required in a particular application, and to be forced to include them adds unnecessarily to the system overheads. Usually during the installation of an operating system certain features can be selected or omitted. A general purpose operating system can thus be 'tailored' to meet a specific application requirement. Recently operating systems which provide only a minimum kernel or nucleus have become popular; additional features can be added by the applications programmer writing in a high-level language. This structure is shown in Figure 6.2. In this type of operating system the distinction between the operating system and the application software becomes blurred. The approach has many advantages for applications that involve small, embedded systems.

5.2 REAL-TIME MULTI-TASKING OS:

There are many different types of operating systems and until the early 1980s there was a clear distinction between operating systems designed for use in real-time applications and other types of operating system. In recent years the dividing line has become blurred. For example, languages such as Modula-2 enable us to construct multi-tasking real-time applications that run on top of single-user, single· task operating systems. And operating systems such as UNIX and OS/2 support multi-user, multi-tasking applications. Confusion can arise between multi-user or multi-programming operating systems and multi-tasking operating systems. The function of a multi-user operating system is illustrated in Figure 6.4: the operating system ensures that each user can run a single program as if they had the whole of the computer system for their program.

Although at any given instance it is not possible to predict which user will have the use of the CPU, or even if the user's code is in the memory, the operating system ensures that one user program cannot interfere with the operation of another user program. Each user program runs in its own protected environment. A primary concern of the operating system is to prevent one program, either deliberately or through error, corrupting another. In a multi-tasking operating system it is assumed that there is a single user and that the various tasks co-operate to serve the requirements of the user. Co-operation requires that the tasks communicate with each other and share common data. This is illustrated in Figure 6.5. In a good multitasking operating system task communication and data sharing will be regulated so that the operating system is able to prevent inadvertent communication or data access (that is, arising through an error in the coding of one task) and hence protect data which is private to a task (note that deliberate interference cannot be prevented the tasks are assumed to be co-operating).

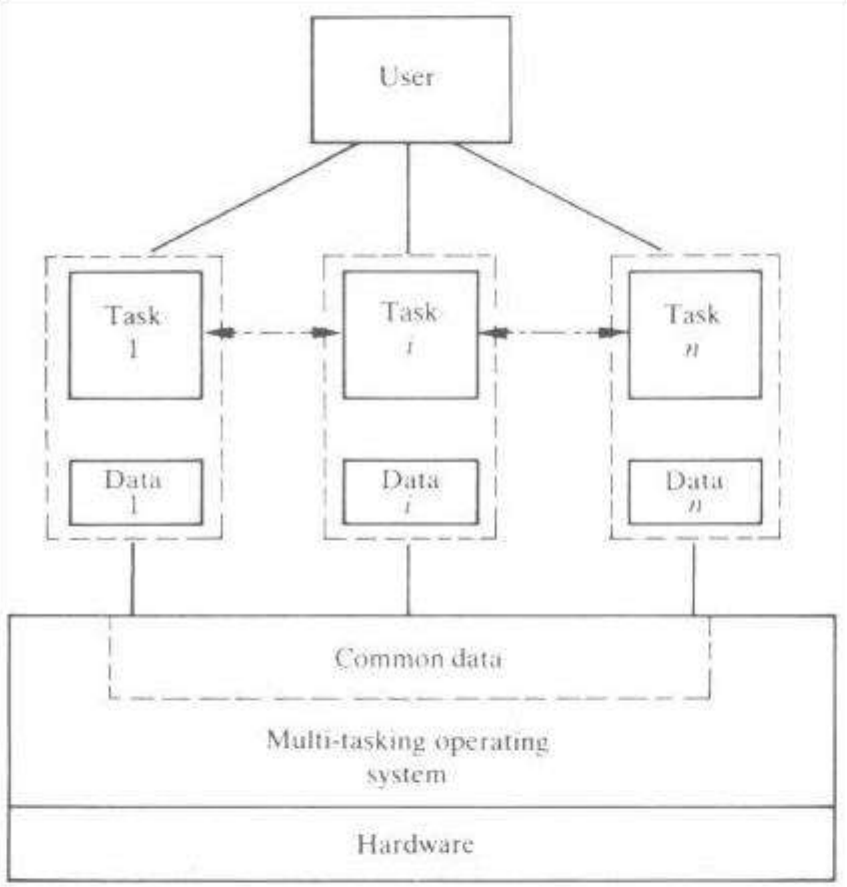


Figure: Multitasking operating system.

A real-time multi-tasking operating system has to support the resource sharing and the timing requirements of the tasks and the functions can be divided as follows:

Task management: the allocation of memory and processor time (scheduling) to tasks.

Memory management: control of memory allocation.

Resource control: control of all shared resources other than memory and CPU time.

Intertask communication and synchronization: provision of support mechanisms to provide safe communication between tasks and to enable tasks to synchronize their activities.

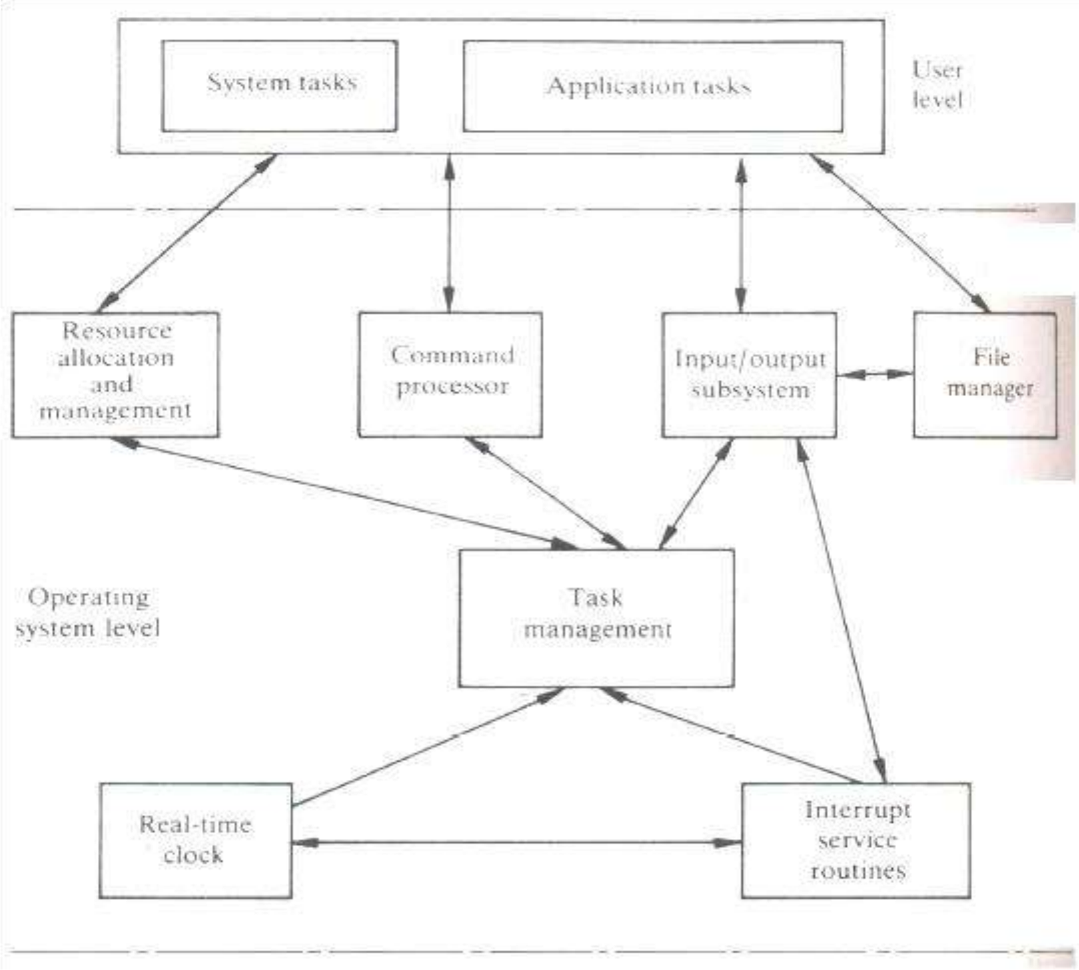


Figure: Typical structure of real-time operating system.

5.3 SCHEDULING STRATEGIES:

If we consider the scheduling of time allocation on a single CPU there are two basic strategies:

1. Cyclic.
2. Pre-emptive.

1. Cyclic

The first of these, cyclic, allocates the CPU to a task in turn. The task uses the CPU for as long as it wishes. When it no longer requires it the scheduler allocates it to the next task in the list. This is a very simple strategy which is highly efficient in that it minimizes the time lost in switching between tasks. It is an effective strategy for small embedded 'systems for which the execution times for each task run are carefully calculated (often by counting the number of machine instruction cycles for. the task) and for which the software is carefully divided into appropriate task segments. In general this approach is too restrictive since it requires that the task units have similar execution times. It is also difficult to deal with random events using this method.

2. Pre-emptive.

There are many pre-emptive strategies. All involve the possibility that a task will be interrupted - hence the term pre-emptive - before it has completed a particular invocation. A consequence of this is that the executive has to make provision to save the volatile environment for each task, since at some later time it will be allocated CPU time and will want to continue from the exact point at which it was interrupted. This process is called context switching and a mechanism for supporting it is described below. The simplest form of pre-emptive scheduling is to use a time slicing approach (sometimes called a round-robin method). Using this strategy each task is allocated a fixed amount of CPU time - a specified number of ticks of the clock – and at the end of this time it is stopped and the next task in the list is run. Thus each task in turn is allocated an equal share of the CPU time. If a task completes before the end of its time slice the next task in the list is run immediately.

The majority of existing RTOSs use a priority scheduling mechanism. Tasks are allocated a priority level and at the end of a predetermined time slice the task with the highest priority of those ready to run is chosen and is given control of the CPU. Note that this may mean that the task which is currently running continues to run. Task priorities may be fixed - a static priority system - or may be changed during system execution - a dynamic priority system. Dynamic priority schemes can increase the flexibility of the system, for example they can be used to increase the priority of particular tasks under alarm conditions. Changing priorities is, however, risky as it makes it much harder to predict the behavior of the system and to test it. There is the risk of locking out certain tasks for long periods of time. If the software is well designed and there is adequate computing power there should be no need to change priorities - all the necessary constraints will be met. If it is badly designed and/or there are inadequate computing resources then dynamic allocation of priorities will not produce a viable, reliable system.

5.4 PRIORITY STRUCTURES:

In a real-time system the designer has to assign priorities to the tasks in the system. The priority will depend on how quickly a task will have to respond to a particular event. An event may be some activity of the process or may be the elapsing of a specified amount of time.

1. Interrupt level: at this level are the service routines for the tasks and devices which require very fast response - measured in milliseconds. One of these tasks will be the real-time clock task and clock level dispatcher.
2. Clock level: at this level are the tasks which require repetitive processing, such as the sampling and control tasks, and tasks which require accurate timing. The lowest-priority task at this level is the base level scheduler.
3. Base level: tasks at this level are of low priority and either have no deadlines to meet or are allowed a wide margin of error in their timing. Tasks at this level may be allocated priorities or may all run at a single priority level - that of the base level scheduler.

Interrupt level:

As we have already seen an interrupt forces a rescheduling of the work of the CPU and the system has no control over the timing of the rescheduling. Because an interrupt-generated rescheduling is outside the control of the system it is necessary to keep the amount of processing to be done by the interrupt handling routine to a minimum. Usually the interrupt handling routine does sufficient processing to preserve the necessary information and to pass this information to a further handling routine which operates at a lower-priority level, either clock level or base level. Interrupt handling routines have to provide a mechanism for task swapping, that is they have to save the volatile environment.

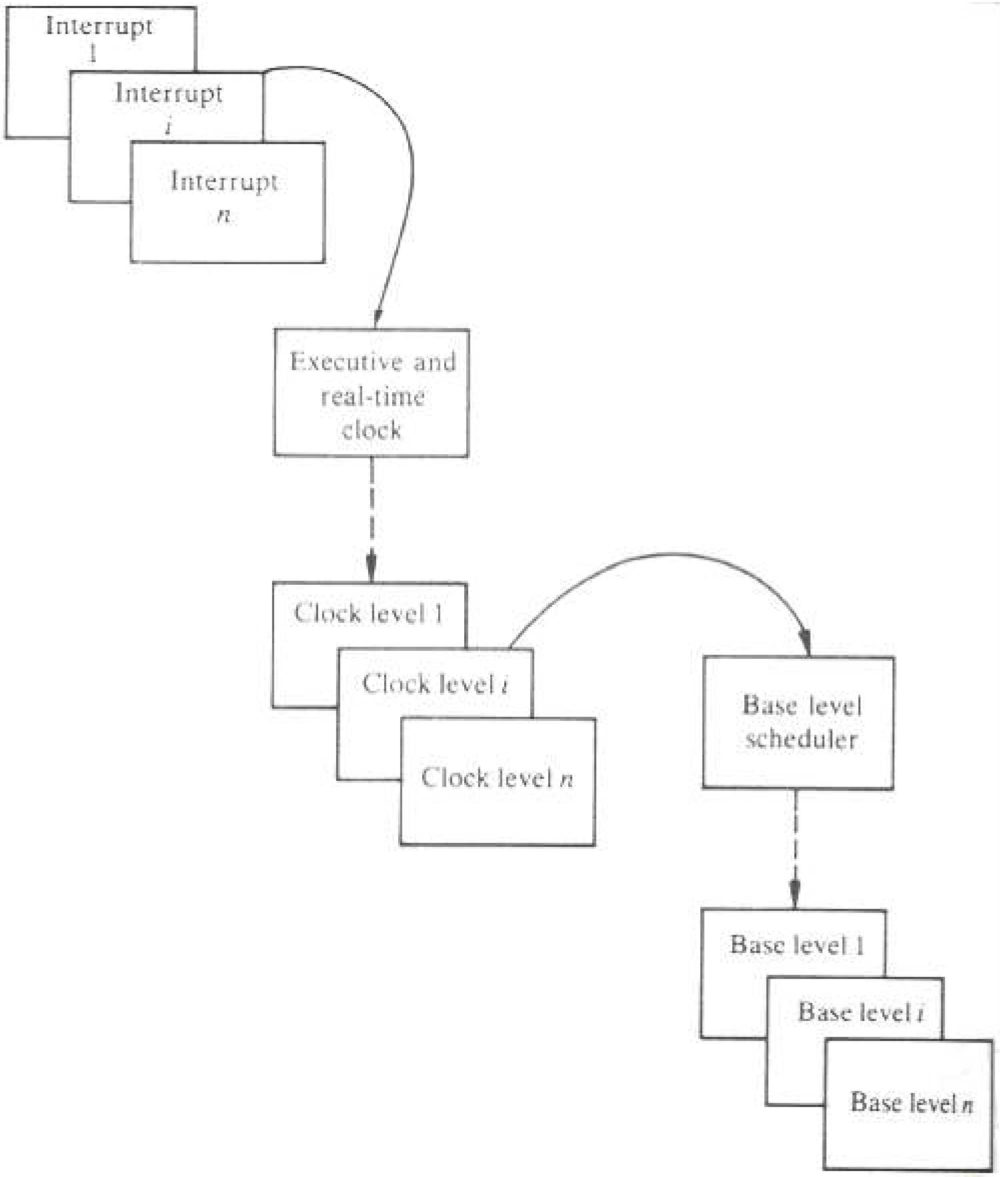


Figure: Priority levels in an RTOS.

Clock level:

One interrupt level task will be the real-time clock handling routine which will be entered at some interval, usually determined by the required activation rate for the most frequently required task. Typical values are I to 200 ms. Each clock interrupt is known as a *tick* and represents the smallest time interval known to the system. The function of the clock interrupt handling routine is to update the time-of-day clock in the system and to transfer control to the dispatcher. The scheduler selects which task is to run at a particular clock tick. Clock level tasks divide into two categories:

1. CYCLIC: these are tasks which require accurate synchronization with the outside world.

2. DELA Y: these tasks simply wish to have a fixed delay between successive repetitions or to delay their activities for a given period of time.

Cyclic tasks:

The *cyclic* tasks are ordered in a priority which reflects the accuracy of timing required for the task, those which require high accuracy being given the highest priority. Tasks of lower priority within the clock level will have some jitter since they will have to await completion of the higher - level tasks.

Delay tasks:

The tasks which wish to delay their activities for a fixed period of time, either to allow some external event to complete (for example, a relay may take 20 ms to close) or because they only need to run at certain intervals (for example, to update the operator display), usually run at the base level. When a task requests a delay its status is changed from runnable to suspended and remains suspended until the delay period has elapsed.

One method of implementing the delay function is to use a queue of task descriptors, say identified by the name DELAYED. This queue is an ordered list of task descriptors, the task at the front of the queue being that whose next running time is nearest to the current time.

Base level:

The tasks at the base level are initiated on demand rather than at some predetermined time interval. The demand may be user input from a terminal, some process event or some particular requirement of the data being processed. The way in which the tasks at the base level are scheduled can vary; one simple way is to use time slicing on a round-robin basis. In this method each task in the runnable queue is selected in turn and allowed to run until either it suspends or the base level scheduler is again entered. For real-time work in which there is usually some element of priority this is not a particularly satisfactory solution. It would not be sensible to hold up a task, which had been delayed waiting for a relay to close but was now ready to run, in order to let the logging task run.

Most real-time systems use a priority strategy even for the base level tasks. This may be either a fixed level of priority or a variable level. The difficulty with a fixed level of priority is in determining the correct priorities for satisfactory operation; the ability to change priorities dynamically allows the system to adapt to particular circumstances. Dynamic allocation of priorities can be carried out using a high-level scheduler or can be done on an *ad hoc* basis from within specific tasks. The high level scheduler is an operating system task which is able to examine the use of the system resources; it may for example check how long tasks have been waiting and increase the priority of the tasks which have been waiting a long time. The difficulty with the high-level scheduler is that the algorithms used can become complicated and hence the overhead in running can become significant.

5.5 TASK MANAGEMENT:

The basic functions of the task management module or executive are:

1. To keep a record of the state of each task;
2. To schedule an allocation of CPU time to each task; and
3. To perform the context switch, that is to save the status of the task that is currently using the CPU and restore the status of the task that is being allocated CPU time.

In most real-time operating systems the executive dealing with the task management functions is split into two parts: a scheduler which determines which task is to run next and which keeps a record of the state of the tasks, and a dispatcher which performs the context switch. Task states:

With one processor only one task can be running at any given time and hence the other tasks must be in some other state. The number of other states, the names given to the states, and the transition paths between the different states vary from operating system to operating system. A typical state diagram is given in Figure6.1 and the various states are as follows (names in parentheses are commonly are. alternatives):

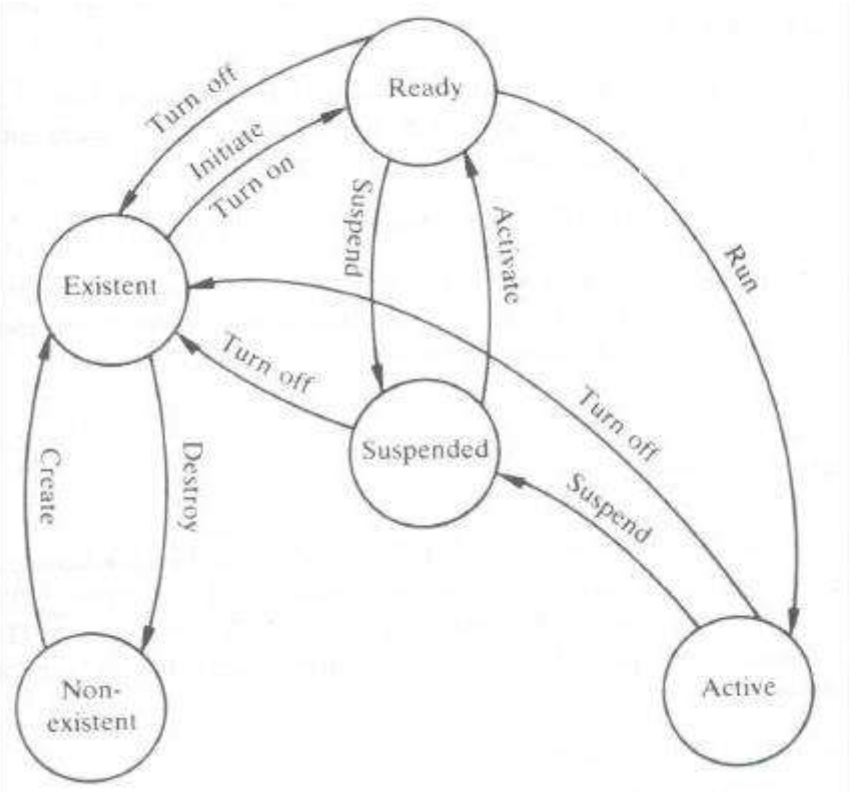


Figure: Example of a typical task state diagram.

• Active (running): this is the task which has control of the CPU. It will normally be the task with the highest priority of the tasks which are ready to run.

• Ready (runnable, on): there may be several tasks in this state. The attribute of the task and the resources required to run the task must be available for the task to be placed in the Ready state.

• Suspended (waiting, locked out, delayed): the execution of tasks placed this state has been suspended because the task requires some resource which is not available or because the task is waiting for some signal from the plant for example input from the analog-to-digital converter, or because the task is waiting for the elapse of time.

• Existent (dormant, off): the operating system is aware of the existence of this task, but the task has not been allocated a priority and has not been made runnable.

• Non-existent (terminated): the operating system has not as yet been made aware of the existence of this task, although it may be resident in the. memory of the computer.

Task descriptor:

Information about the status of each task is held in a block of memory by the RTOS. This block is referred to by various names· task descriptor (TD), process descriptor (PD), task control block (TCB) or task data block (TDB). The information held in the TD will vary from system to system, but will typically consist of the following:

- Task identification (10);
- Task priority (P);
- Current state of task;
- Area to store volatile environment (or a pointer to an area for storing the volatile environment); and
- Pointer to next task in a list.

5.6 SCHEDULER AND REAL-TIME CLOCK INTERRUPT HANDLES:

The real-time clock handler and the scheduler for the clock level tasks must be carefully designed as they run at frequent intervals. Particular attention has to be paid to the method of selecting the tasks to be run at each clock interval. If ached of all tasks were to be carried out then the overheads involved could become significant.

System commands which change task status.

The range of system commands affecting task status varies with the operating system. Typical states and commands are shown in Figure 6.12 and fuller details of the commands are given in Table. Note that this system distinguishes between tasks which are suspended awaiting the passage of time - these tasks are marked as delayed - and those tasks which are waiting for an event or a system resource these are marked as locked out. The system does not explicitly support base level tasks; however, the lowest four priority levels of the clock level tasks can be used to create a base level system A so - called free time executive (FTX) is provided which if used runs at priority level n - 3 where n is the lowest-priority task number. The FTX is used to run tasks at priority levels n - 2, n - I and n; it also provides support for the chaining of tasks. The dispatcher is unaware of the fact that tasks at these three priority levels are being changed; it simply treats whichever tasks are in the lowest three priority levels as low-priority tasks. Tasks run under the FTX do not have access to the system commands (except OFFCO1 that is turn task off).

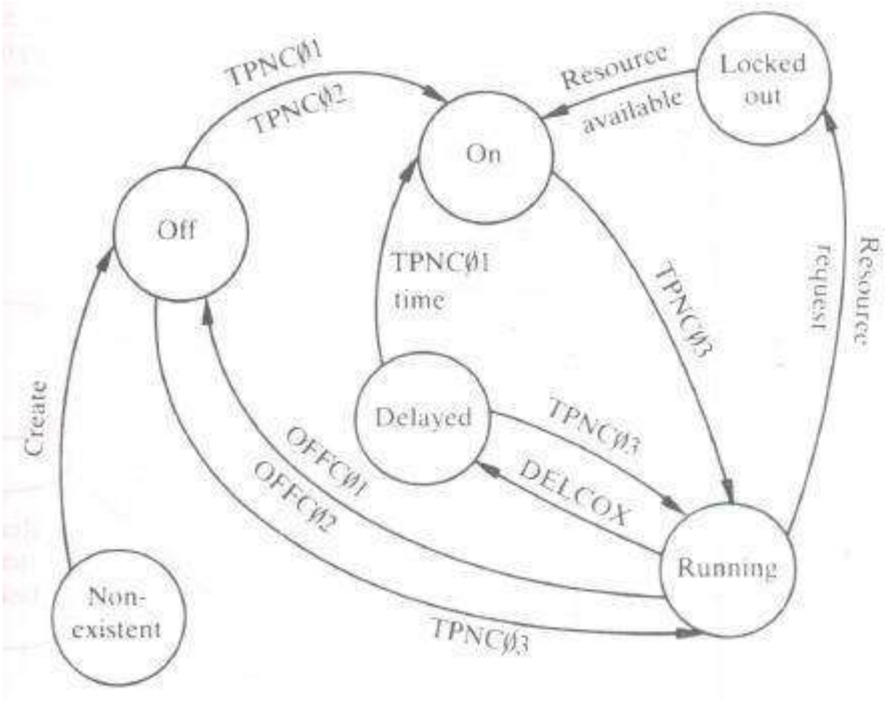


Figure: RTOS state diagram.

Dispatcher- search for work:

The dispatcher/scheduler has two entry conditions:

1. The real-time clock interrupt and any interrupt which signals the completion of an input/output request;
2. A task suspension due to a task delaying, completing or requesting an input/output transfer.

In response to the first condition the scheduler searches for work starting with the highest-priority task and checking each task in priority order (see Figure 6.14). Thus if tasks with a high repetition rate are given a high priority they will be treated as if they were clock level tasks, that is they will be run first during each system clock period. In response to the second condition a search for work is started at the task with the next lowest priority to the task which has just been running. There cannot be another higher-priority task ready to run since a higher-priority task becoming ready always pre-empts a lower-priority-running task. The system commands for task management are issued as calls from the assembly level language and the parameters are passed either in the CPU registers or as a control word immediately following the call statement.

5.7 MEMORY MANAGEMENT:

Since the majority of control application software is static - the software is not dynamically created or eliminated at run-time - the problem of memory management is simpler than for multi-programming, on-line systems. Indeed with the cost of computer hardware, both processors and memory, reducing many control applications use programs which are permanently resident in fast access memory. With permanently resident software the memory can be divided as shown in Figure. The user space is treated as one unit and the software is linked and loaded as a single program into the user area. The information about the various tasks is conveyed to the operating system by means of a create task statement. Such a statement may be of the form the exact form of the statement will depend on the interface between the high-level language and the operating system. An alternative arrangement is shown in Figure. The available memory is divided into predetermined segments and the tasks are loaded individually into the various segments. The load operation would normally be carried out using to command processor. With this type of system the entries in the TD (or the operation system tables) have to be made from the console using a memory examine as change facility.

Divided (partitioned) memory was widely used in many early real-time operating systems and it was frequently extended to allow several tasks to share on:

partition; the tasks were kept on the backing store and loaded into the appropriate partition when required. There was of course a need to keep any tasks in which timing was crucial (hard time constraint tasks) in fast access memory permanent other tasks could be swapped between fast memory and backing store. The difficulty with this method is, of course, in choosing the best mix of partition sizes. The partition size and boundaries have to be determined at system generation.

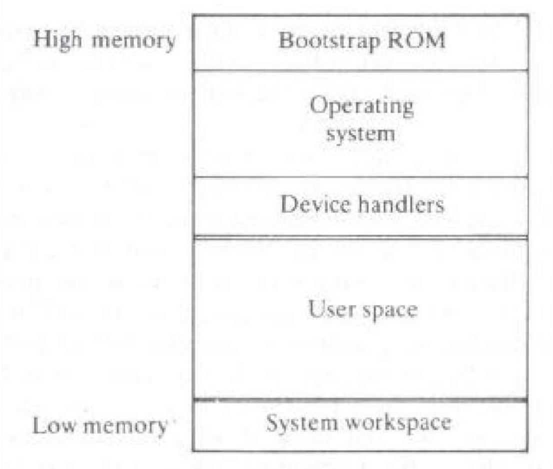


Figure: Non-partitioned memory.

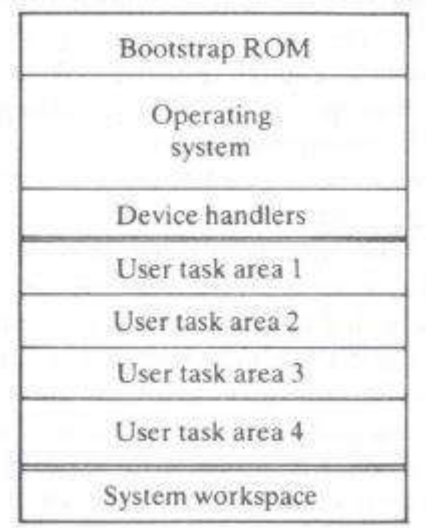


Figure: Partitioned memory.

5.8 CODE SHARING:

In many applications the same actions have to be carried out in several different tasks. In a conventional program the actions would be coded as a subroutine and one copy of the subroutine would be included in the program. In a multi-tasking system each task must have its own copy of the subroutine or some mechanism must be provided to prevent one task interfering with the use of the code by another task. The problems which can arise are illustrated in Figure 6.20. Two tasks share the subroutine S. If task A is using the subroutine but before it finishes some even occurs which causes a rescheduling of the tasks and task B runs and uses the subroutine, then when a return is made to task

A, although it will begin to use subroutine S again at the correct place, the values of locally held data will have been changed and will reflect the information processed within the subroutine by task B. Two methods can be used to overcome this problem:

- serially reusable code; and
- re-entrant code.

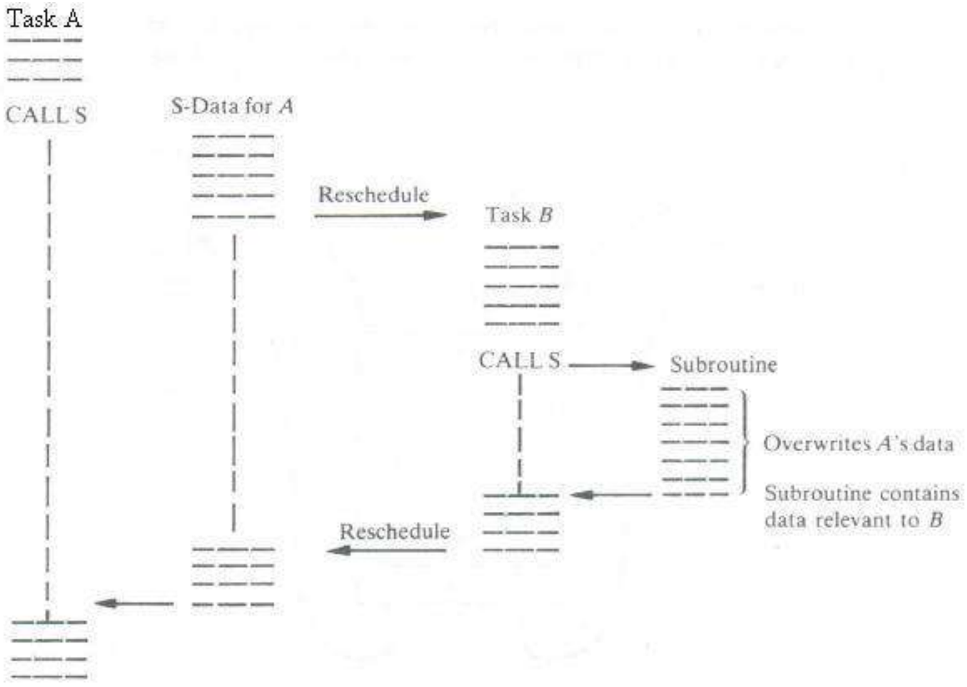


Figure: Sharing a subroutine in multi-tasking system.

Serially reusable code:

As shown in Figure, some form of lock mechanism is placed at the beginning of the routine such that if any task is already using the routine the calling task will not be allowed entry until the task which is using the routine unlocks it. The use of a lock mechanism to protect a subroutine is an example of the need for mechanisms to support mutual exclusion when constructing an operating system.

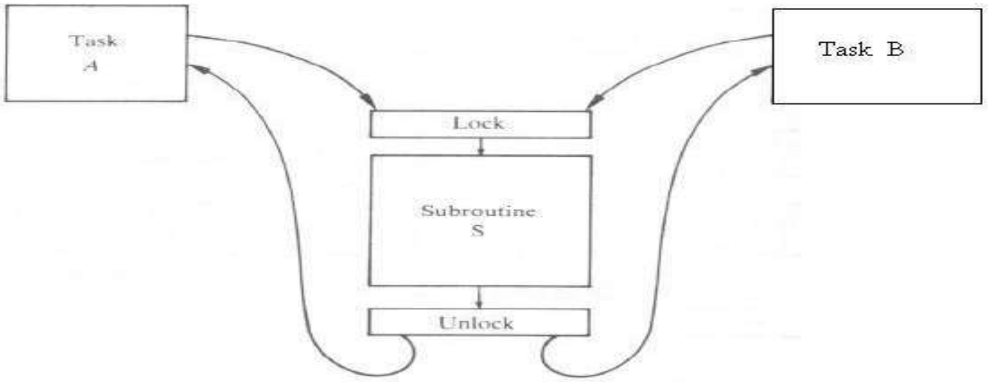


Figure: Serially reusable code.

Re-entrant code:

If the subroutine-can be coded such that it does not hold within it any data that is it is purely code - any intermediate results are stored in the calling task or in a stack associated with the task - then the subroutine is said to be re-entrant. Figure shows an arrangement which can be used: the task descriptor for each task contains a pointer to a data area - usually a stack area - which is used for the storage of all information relevant to that task when using the subroutine. Swapping between tasks while they are using the subroutine will not now cause any problems since the contents of the stack pointer will be saved with the volatile environment of the task and will be restored when the task resumes.

All accesses to data by the subroutine will be through the stack and hence it will automatically manipulate the correct data. Re-entrant routines can be shared between several tasks since they contain no data relevant to a particular task and hence can be stopped and restarted at a different point in the routine without any loss of information. The data held in the working registers of the CPU is stored in the relevant task descriptor when task swapping takes place. Device drivers in conventional operating systems are frequently implemented using re-entrant code. The PID control1er code segment uses the information in the LOOP descriptor and the T ASK to calculate the control value and to send it to the control1er. The actual task is made up of the LOOP descriptor, the TASK segment and the PID control code segment. The addition of another loop to the system requires the provision of new loop descriptors; the actual PID control code remains unchanged.

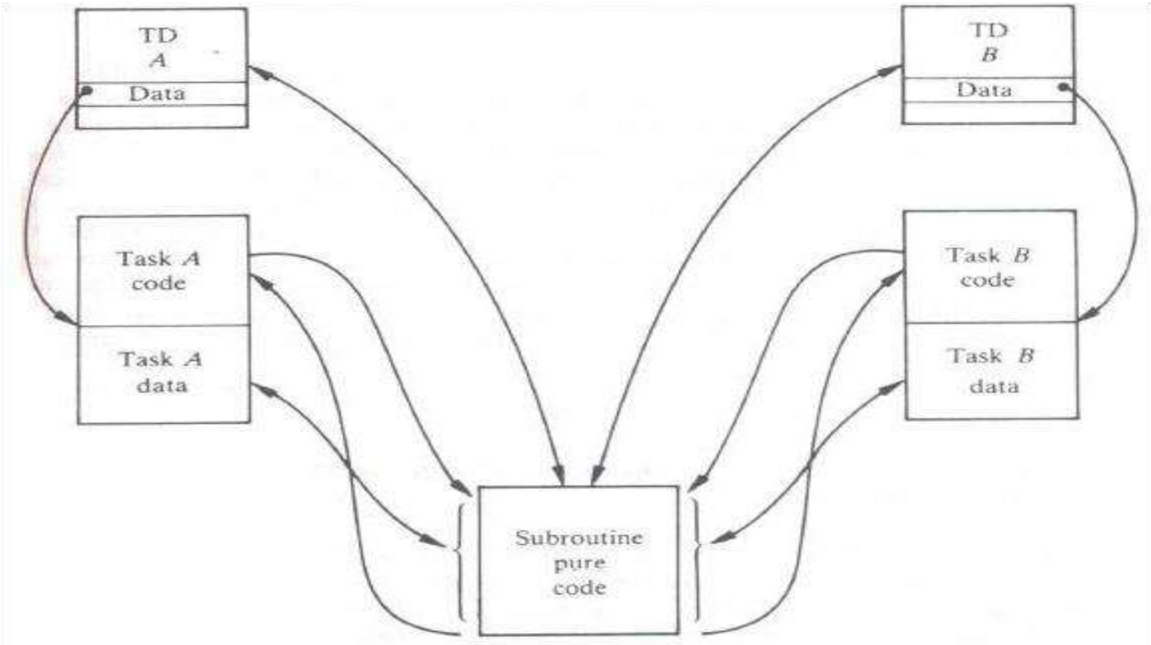


Figure: Use of re-entrant code for sharing.

5.9 RESOURCE CONTROL: AN EXAMPLE OF AN INPUT/OUTPUT SUBSYSTEM (lOSS)

One of the most difficult areas of programming is the transfer of information to and from external devices. The availability of a wel1-designed and implemented input/output subsystem (lOSS) in an operating system is essential for efficient programming. The lOSS handles all the details of the devices. In a multi-tasking system the lOSS should also deal with all the problems of several tasks attempting to access the same device. A typical lOSS will be divided into two levels as shown in Figure. The I/O manager accepts the system calls from the user tasks and transfers the information contained in the calls to the device control block (DCB) for the particular device.

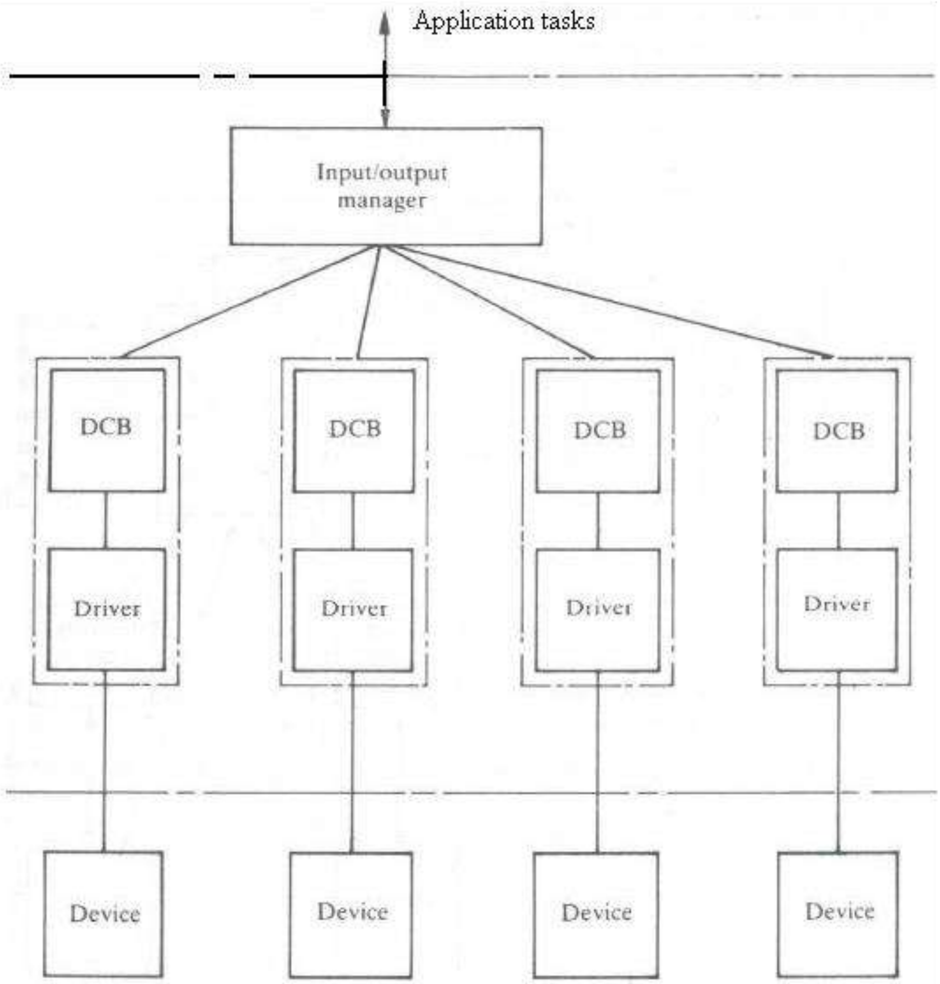


Figure: General structure of IOSS.

5.10 TASK CO-OPERATION AND COMMUNICATION:

In real-time systems tasks are designed to fulfil a common purpose and hence they need to communicate with each other. However, they may also be in competition for the resources of the computer system and this competition must be regulated. Some of the problems which arise have already been met in considering the input/output subsystem and they involve:

- Mutual exclusion;
- Synchronization; and
- Data transfer.

Mutual exclusion:

A multi-tasking, operating system allows the sharing of resources between several concurrently active tasks. This does not imply that the resources can be used simultaneously. The use of some resources is restricted to only one task at a time. For others, for example a re-entrant code module, several tasks can be using them at the same time. The restriction to one task at a time has to be made for resource such as input and output devices; otherwise there is a danger that input intended for one task could get corrupted by input for another task. Similarly problems can arise if two tasks share a data area and both tasks can write to the data area.

Data transfer:

RTOSs typically support two mechanisms for the transfer or sharing of data between tasks: these are the pool and the channel.

Pool is used to hold data common to several tasks, for example tables of values or parameters which tasks periodically consult or update. The write operation on a pool is destructive and the read operation is non-destructive.

Channel supports communication between producers and consumers of data. It can contain one or more items of information. Writing to a channel adds an item without changing items already in it. The read operation is destructive in that it removes an item from the channel. A channel can become empty and also, because in practice its capacity is finite, it can become full.

It is normal to create a large number of pools so as to limit the use of global common data areas. To avoid the problem of two or more tasks accessing a pool simultaneously mutual exclusion on pools is required. The most reliable form of mutual exclusion for a pool is to embed the pool inside a monitor. Given that the read operation does not change the data in a pool there is no need to restrict read access to a pool to one task at a time. Channels provide a direct communication link between tasks, normally on a one-to-one basis. The communication is like a pipe down which successive collections of items of data - messages - can pass. Normally they are implemented so that they can contain several messages and so they act as a buffer between the tasks. One task is seen as the *producer* of information and the other as the *consumer*. Because of the buffer function of the channel the producer and consumer tasks can run asynchronously. There are two basic implementation mechanisms for a channel:

- Queue (linked list); and
- Circular buffer.

The advantage of the queue is that the number of successive messages held in the channel is not fixed. The length of the queue can grow, the only limit being the amount of available memory. The disadvantage of the queue is that as the length of the queue increases the access time that is the time to add and remove items from the queue, increases. For this reason and because it is not good practice to have undefined limits on functions in real-time systems queues are rarely used. The circular buffer uses a fixed amount of memory, the size being defined by the designer of the application. If the producer and consumer tasks run normally they would typically add and remove items from the buffer alternately. If for some reason one or the other is suspended for any length of time the buffer will either fill up or empty. The tasks using the buffer have to check, as appropriate, for buffer full and buffer empty conditions and suspend their operations until the empty or full condition changes.

Synchronization with Data transfer:

There are two main forms of synchronization involving data transfer. The first Involves the producer task simply signaling to say that a message has been produced and is waiting to be collected, and the second is to signal that a message is ready and to wait for the consumer task to reach a point where the two tasks can exchange the data. The first method is simply an extension of the mechanism used in the example in the previous section to signal that a channel was empty or full. Instead of signaling these conditions a signal is sent each time a message is placed in the channel. Either a generalized semaphore or signal that counts the number of sends and waits, or a counter, has to be used.

5.11 LIVENESS:

An important property of a multi-tasking real-time system is Liveness. A system (a set of tasks) is said to possess Iiveness if it is free from livelock, deadlock. and indefinite postponement. Livelock is the condition under which the tasks requiring mutually exclusive access to a set of resources both enter busy wait routines but neither can get out of the busy wait because they are waiting for each other. The CPU appears to be doing useful work and hence the term Livelock.

Deadlock is the condition in which a set of tasks are in a state such that it is impossible for any of them to proceed. The CPU is free but there are no tasks that are ready to run.

5.12 MINIMUM OPERATING SYSTEM KERNEL:

As mentioned in the introduction there has been considerable interest in recent years in the idea of providing a minimum kernel of RTOS support mechanisms and constructing the required additional mechanisms for a particular application or group of applications. One possible set of functions and primitives for *RTGS* is:

Functions:

1. A clock interrupts procedure that decrements a time count for relevant tasks,
2. A basic task handling and context switching mechanism that will support the moving of tasks between queues and the formation of task queues.
3. Primitive device routines (including real-time clock support).

Primitives:

WAIT for some condition (including release of exclusive access rights).

SIGNAL condition and thus release one (or all) tasks waiting on the condition,

ACQUIRE exclusive rights to a resource (option - specify a time-out condition).

RELEASE exclusive rights to a resource.

DELAY task for a specified time.

CYCLE task, that is suspend until the end of its specified cyclic period.

Recommended Questions:

1. Draw up a list of functions that you would expect to find in a real-time operating system. Identify the functions which are essential for a real-time system.
2. Discuss the advantages and disadvantages of using
   (a) Fixed table
   (b) Linked list
   Methods for holding task descriptors in a multi-tasking real-time operating system.

3 A range of real-time operating systems are available with different memory allocation strategies. The strategies range from permanently memory-resident tasks with no task swapping to fully dynamic memory allocation. Discuss the advantages and disadvantages of each type of strategy and give examples of applications for which each is most suited.

4. What are the major differences in requirements between a multi-user operating system and a multi-tasking operating system?

5. What is the difference between static and dynamic priorities? Under what circumstances can the use of dynamic priorities be justified?

1. Choosing the basic clock interval (tick) is an important decision in setting up an RTOS. Why is this decision difficult and what factors need to be considered when choosing the clock interval?

2. List the minimum set of operations that you think a real-time operating system kernel needs to support.

8. What is meant by context switching and why it is required?

Module-5

Design of RTSS General Introduction

Introduction, Specification documentation, Preliminary design, Single –Program Approach, Foreground /Background, Multi- Tasking approach, Mutual Exclusion Monitors.

Recommended book for reading:

1. **Real –Time Computer control –An Introduction**, Stuart Bennet, 2nd Edn. Pearson Education 2005.
2. **Real-Time Systems Design and Analysis**, Phillip. A. Laplante, Second Edition, PHI, 2005.
3. **Real time Systems Development**, Rob Williams, Elsevier, 2006.

7.1 DESIGN O F RTSS- GENERAL INTODUCTION INTRODUCTION

The approach to the design of real-time computer systems is no different in outline from that required for any computer-based system or indeed most engineering systems. The work can be divided into two main sections:

- The planning phase; and
- The development phase.

It is concerned with interpreting use requirements to produce a detailed specification of the system to be developed and an outline plan of the resources - people, time, equipment, costs - required to carry out the development. At this stage preliminary decisions regarding the division of functions between hardware and software will be made. A preliminary assessment of the type of computer structure - a single central computer, a hierarchical system, or a distributed system - will also be made. The outcome of this stage is a specification or requirements document. (The terminology used in books on software engineering can be confusing; some refer to a specification requirement document as well as to specification document and requirements document. It clearer and simpler to consider that documents produced by the user or customer describe requirements, and documents produced by the supplier or designer give the specifications.) It cannot be emphasized too strongly that the specification document for both the hardware and software which results from this phase must be complete, detailed and unambiguous. General experience has shown that a large proportion of errors which appear in the final system can be traced back to unclear, ambiguous or fault) specification documents. There is always a strong temptation to say 'It can be decided later'; deciding it later can result in the need to change parts of the system which have already been designed. Such changes are costly and frequently lead to the introduction of errors. This shows the distribution of errors and cost of rectifying them (the figures are taken from DeMarco, 1978). The detailed design is usually broken down into two stages:

- Decomposition into modules; and
- Module internal design.

For real-time systems additional heuristics are required, one of which is to divide modules into the following categories:

- Real-time, hard constraint;
- Real-time, soft constraint; and
- Interactive.

The arguments given in Chapter 1 regarding the verification and validation of different types of program suggest a rule that aims to minimize the amount of software that falls into the hard constraint category since this type is the most difficult to design and test.

7.2 SPECIFICATION DOCUMENTATION:

To provide an example for the design procedures being described we shall consider a system comprising several of the hot-air blowers described. It is assumed that the planning phase has been completed and a specification document has been prepared. A PID controller with a sampling interval of 40 ms is to be used. The sampling interval may be changed, but will not be less than 40 ms. The controller parameters are to be expressed to the user in standard analog form, that is proportional gain, integral action time and derivative action time. The set point is to be entered from the keyboard. The controller parameters are to be variable and are to be entered from the keyboard.

7.3 PRELIMINARY DESIGN:

Hardware design:

There are many different possibilities for the hardware structure. Obvious arrangements are:

1. Single computer with multi-channel ADC and DAC boards.
2. Separate general purpose computers on each unit.
3. Separate computer-based microcontrollers on each unit linked to a single general. Purpose computer.

Each of these configurations needs to be analyzed and evaluated. Some points to consider are:

Option 1: given that the specification calls for the system to be able to run with a sample interval for the control loop of 40 ms, can this be met with 12units sharing a single processor?

Option 2: is putting a processor that includes a display and keyboard on each unit an expensive solution? Will communication between processors be required? (Almost certainly the answer to this is yes; operators and managers will not want to have to use separate displays and keyboards.)

Option 3: what sort of communication linkage should be used? A shared high speed bus? A local-area network? Where should the microcontrollers be located? At each blower unit or together in a central location? Each option needs careful analysis and evaluation in terms of cost and performance. The analysis must include consideration of development costs, performance operating and maintenance costs. It should also include consideration of reliability and safety. To provide a basis for consideration of the widest range of approaches to software design we will assume that option 1 above is chosen.

Software design:

Examining the specification shows that the software has to perform several
different functions:

- DDC for temperature control;
- Operator display;
- Operator input;
- Provision of management information;
- System start-up and shut-down; and
- clock/calendar function.

The various functions and type of time constraint are shown in Figure. The control module has a hard constraint in that it must run every 40 ms. In practice this constraint may be relaxed a little to, say, 40 ms ± 1 ms with an average value over 1 minute of, say, 40 ms ± 0.5 ms. In general the sampling time can be specified as *Ts* ± *es* with an average value, over time *T*, of *Ts* ± *ea.* The requirement may also be relaxed to allow, for example, one sample in 100 to be missed. These constraints will form part of the test specification. The clock/calendar module must run every 20 ms in order not to miss a clock pulse. The operator display, as specified, has a hard constraint in that an update interval of 5 seconds is given. Common sense suggests that this is unnecessary and an average time of 5 seconds should be adequate; however, a maximum time would also have to be specified, say 10 seconds.. These would have to be decided upon and agreed with the customer. They should form part of the specification in the requirements document. The start-up module does not have to operate in real time and hence can be considered as a standard interactive module. The sub-problems will have to share a certain amount of information and how this is done and how the next stages of the design proceed will depend upon the general approach to the implementation. There are three possibilities:

- Single program;
- Foreground/background system; and
- Multi-tasking.

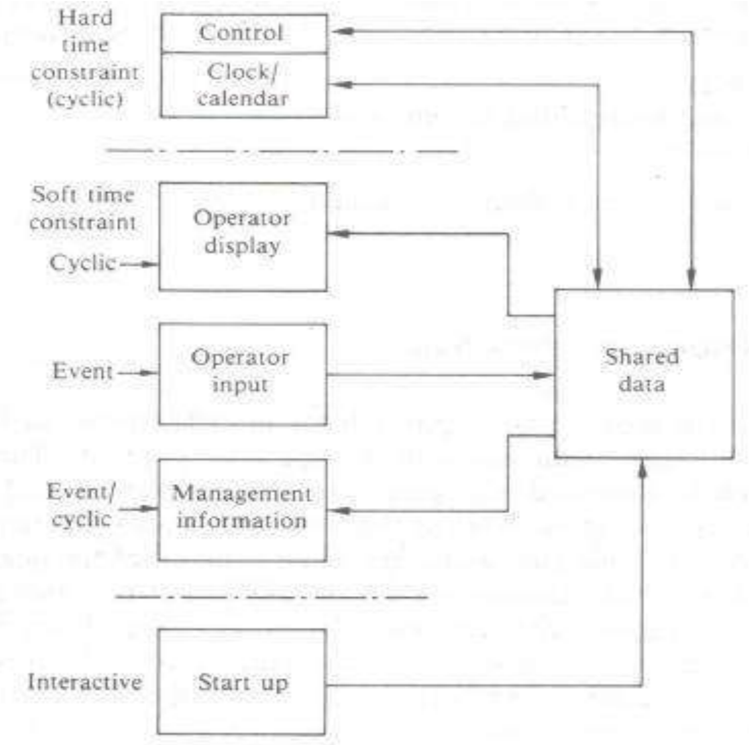


Figure: Basic software modules.

7.4 SINGLE- PROGRAM APPROACH:

Using the standard programming approach the modules shown in Figure are treated as procedures or subroutines of a single main program. The flow chart of such a program is illustrated in Figure. This structure is easy to program; however, it imposes the most severe of the time constraints - the requirement that the clock/calendar module must run every 20 ms - on all of the modules. For the system to work the clock/calendar module and anyone of the other modules must complete their operations within 20 ms. If *fl, fz, f3, f4* and *fs* are the *maximum* computation times for the module's clock/calendar, control, operator display, operator input and management output respectively, then a requirement for the system to work can be expressed as *fl* + *max (tz, f3, f4, fs)* < 20 ms.

The single-program approach can be used for simple, small systems and it lead to a clear and easily understandable design, with a minimum of both hardware and software. Such systems are usually easy to test.. In the above example the management output requirement makes it unsuitable for the single-program approach; if that requirement is removed the approach could be used. It may, however, require the division of the display update module into three modules: display date and time; display process values; and display controller parameters.

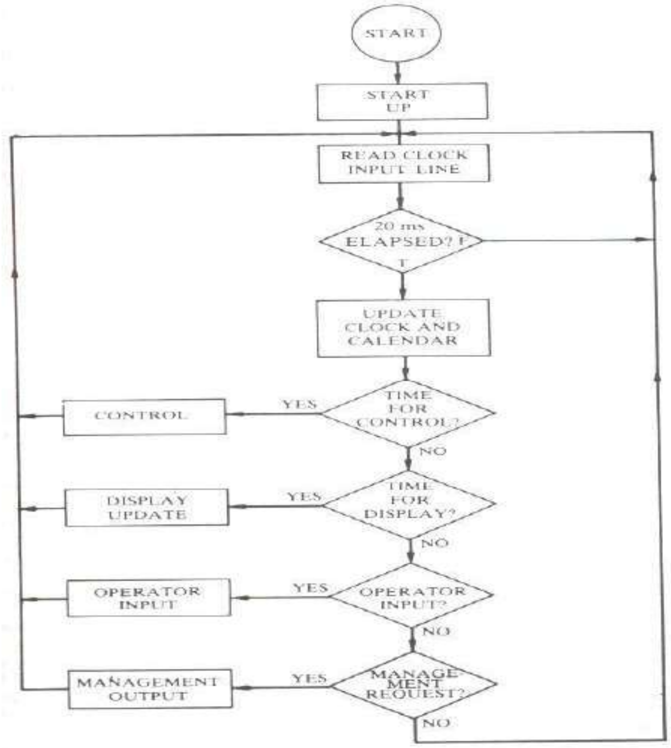


Figure: Single program approach.

7.5 FOREGROUND/BACKGROUND SYSTEMS:

These are obvious advantages - less module interaction, less tight time constraints if the modules with hard time constraints can be separated from, and handled independently of, the modules with soft time constraints or no time constraints. The modules with hard time constraints are run in the so-called 'foreground' and the modules with soft constraints (or no constraints) are run in the 'background'. The foreground modules, or 'tasks' as they are usually termed, have a higher priority than the background tasks and a foreground task must be able to interrupts background task. The partitioning into foreground and background usually requires the support of a real-time operating system, for example the Digital Equipment Corporation's RT/11 system. It is possible, however, to adapt many standard operating systems, for example MS-DOS, to give simple foreground/background operation if the hardware supports interrupts.

The foreground task is written as an interrupt routine and the background task as a standard program. If you use a PC you are in practice using a foreground/background system. The application program that you are using (a word processor, a spreadsheet, graphics package or some program which you have written yourself in a high-level language) is, if we use the terminology given above, running in the background. In the foreground are several interrupt-driven routines - the clock, the keyboard input, the disk controller - and possibly some memory-resident programs which you have installed - a disk caching program or an extended memory manager. The terminology foreground and background can be confusing; literature concerned with non-real-time software uses foreground to refer to the application software and background to refer to interrupt routines that are hidden from the user.

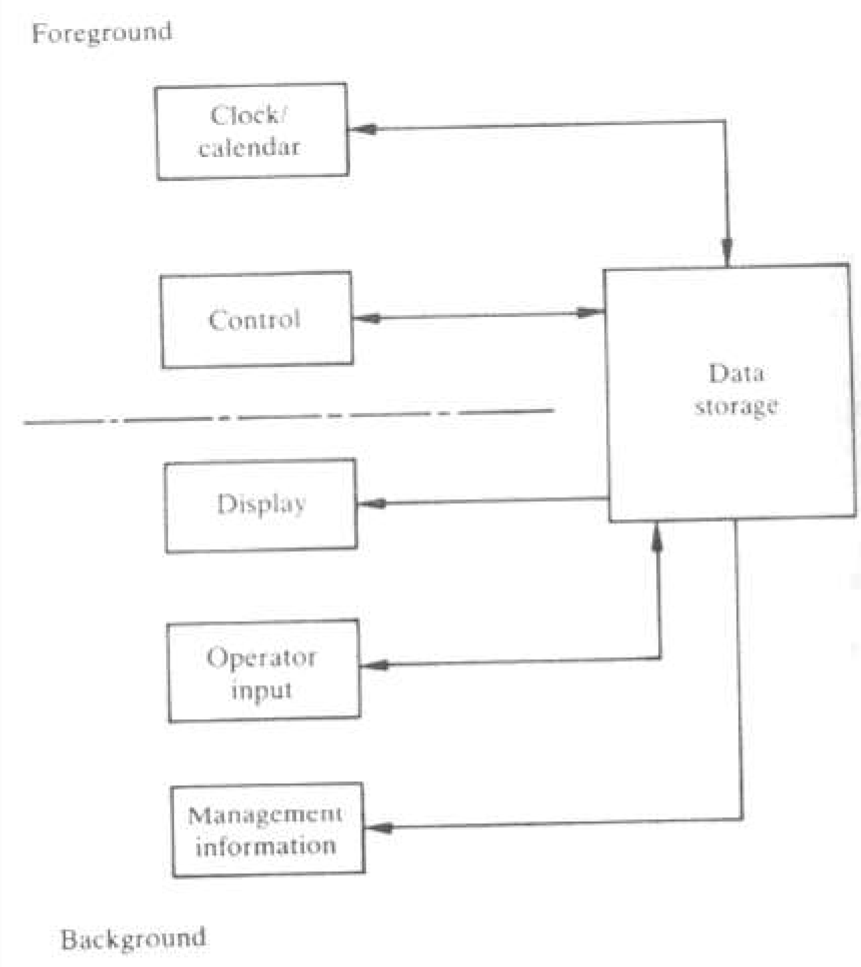


Figure: Software module for foreground/background system.

7.6 MULTI-TASKING APPROACH:

The design and programming of large real-time systems is eased if the foreground/background partitioning can be extended into multiple partitions to allow the concept of many active tasks. At the preliminary design stage each activity is considered to be a separate task. (Computer scientists use the word process rather than task but this usage has not been adopted because of the possible confusion which could arise between internal computer processes and the external processes on the plant.) The implications of this approach are that each task may be carried out in parallel .and there is no assumption made at the preliminary design stage as to how many processors will be used in the system. The implementation of a multi-tasking system requires the ability to:

- Create separate tasks;
- Schedule running of the tasks, usually on a priority basis;

- Share data between tasks;
- Synchronize tasks with each other and with external events;
- Prevent tasks corrupting each other; and
- Control the starting and stopping of tasks.

The facilities to perform the above actions are typically provided by a real-time operating system (RTOS) or a combination of RTOS and a real-time programming language. For simplicity we will assume that we are using only one CPU and that the use of this CPU is time shared between the tasks. We also assume that a number of so-called primitive instructions exist. These are instructions which are part of a programming language or the operating system and their implementation and correctness is guaranteed by the system. All that is of concern to the user is that an accurate description of the syntax and semantics is made available. In practice, with some understanding of the computer system, it should not be difficult to implement the primitive instructions. Underlying the implementation of primitive instructions will be an eventual reliance on the system hardware. For example, in a common memory system some form of arbiter will exist to provide for mutual exclusion in accessing an individual memory location.

7.7 MONITORS:

The basic idea of a monitor is implementation of a monitor in Moouia-2 to protect access to a buffer area is shown. Monitors themselves do not provide a mechanism for synchronizing tasks and hence for this purpose the monitor construct has to be supplemented by allowing, for example, signals to be used within it.

The standard monitor construction outlined above, like the semaphore, does not reflect the priority of the task trying to use a resource; the first task to gain entry can lock out other tasks until it completes. Hence a lower-priority task could hold up a higher-priority. The lack of priority causes difficulties for real-time systems. Traditional operating systems built as monolithic monitors avoided the problem by ensuring that once an operating system call was made (in other words, when a monitor function was invoked) then the call would be completed without interruption from other tasks. The monitor function is treated as a critical section. This does not mean that the whole operation requested was necessarily completed without interruption. For example, a request for access to a printer for output would be accepted and the request queued; once this had been done another task could enter the monitor to request output and either be queued, or receive information from the monitor as to the status of the resource. The return of information is particularly important as it allows the application program to make a decision as to whether to wait for the resource or take some other action.

Preventing lower-priority tasks locking out higher-priority tasks through the monitor access mechanism can be tackled in a number of ways. One solution adopted in some implementations of Modula-2 is to run a monitor with all interrupts locked out; hence a monitor function once invoked runs to completion. In many applications, however, this is too restrictive and some implementations allow the programmer to set a priority level on a monitor such that all lower-priority tasks are locked out - note that this is an interrupt priority level, not a task priority, The monitor has proved to be a popular idea and in practice it provides a good solution to many of the problems of concurrent programming.

The benefits and popularity of the monitor constructs stem from its modularity which means that it can be built and tested separately from other parts of the system, in particular from the tasks which will use it. Once a fully tested monitor is introduced into the system the integrity of the data or resource which it protects is guaranteed and a fault in a task using the monitor cannot corrupt the monitor or the resource which it protects. Although it does rely on the use of signals for inter task synchronization it does have the benefit that the signal operations are hidden within the monitor.

The monitor is an ideal vehicle for creating abstract mechanisms and thus fits in well with the idea of top-down design. However, the nested monitor call problem calling procedures in one monitor from within another monitor can lead to deadlock. Providing that nested monitor calls are prohibited the use of the monitor concept provides a satisfactory solution to many of the problems for a single processor machine or for a multi-processor machine with shared memory, It can also be used on distributed systems.

The monitor's usefulness in some real-time applications is restricted because a task leaving a monitor can only signal and awaken one other task - to do otherwise would breach the requirement that only one task be active within a monitor. This means that a single controlling synchronizer task, for example a clock level scheduler, cannot be built as a monitor. The problem can be avoided by allowing signals to be used outside a monitor but then all the problems associated with signals and semaphores re-emerge.

Recommended Questions:

1. Explain the different phases involved in the design of a RTS.
2. Explain foreground and background system with flowchart.
3. Explain mutual exclusion, using conditional flags.
4. With a neat flow chart, describe the single program approach, with reference to RTS design.
5. Write a note on basic software module, with respect to RTS.
6. Considering a system comprising of several hot air blowers. Prepare specification documents of the same.
7. Explain the data concept of data sharing using common memory.
8. Explain software design for RTS using software module
9. Mention the importance of conditions flag and binary semaphores

RTS Development Methodologies

Introduction, Yourdon Methodology, Requirement definition For Drying Oven, Ward and Mellor Method, Hately and Pirbhai Method.

Recommended book for reading:

1. **Real –Time Computer control –An Introduction**, Stuart Bennet, 2nd Edn. Pearson Education 2005.
2. **Real-Time Systems Design and Analysis**, Phillip. A. Laplante, Second Edition, PHI, 2005.
3. **Real time Systems Development**, Rob Williams, Elsevier, 2006.

8.1 RTS DEVELOPMENT METHODOLOGIES INTRODUCTION

The production of robust, reliable software of high quality for real-time computer control applications is a difficult task which requires the application of engineering methods. During the last ten years increasing emphasis has been placed on formalizing the specification, design and construction of such software, and several methodologies are now extant. All of the methodologies address the problem in three distinct phases. The production of a logical or abstract model - the process of specification; the development of an implementation model for a virtual machine from the logical model - the process of design; and the construction of software for the virtual machine together with the implementation of the virtual machine on a physical system - the process of implementation. These phases, although differently named, correspond to the phases of development generally recognized in software engineering texts. Abstract model: the equivalent of a requirements specification, it is the result of the requirements capture and analysis phase. Implementation model: this is the equivalent of the system design; it is the product of the design stages - architectural design and the detail design

Although there is a logical progression from abstract model to implementation model to implemented software, and although three separate and distinct artifacts abstract model, implementation model, and deliverable system - are produced, the phases overlap in time. The phases overlap because complex systems are best handled by a hierarchical approach: determination of the detail of the lower levels in the hierarchy of the logical model must be based on knowledge of higher - level design decisions, and similarly the lower-level design decisions must be based on the higher-level implementation decisions. Another way of expressing this is to say that the higher-level design decisions determine the requirements specification for the lower levels in the system.

8.2 YOURDON METHODOLOGY:

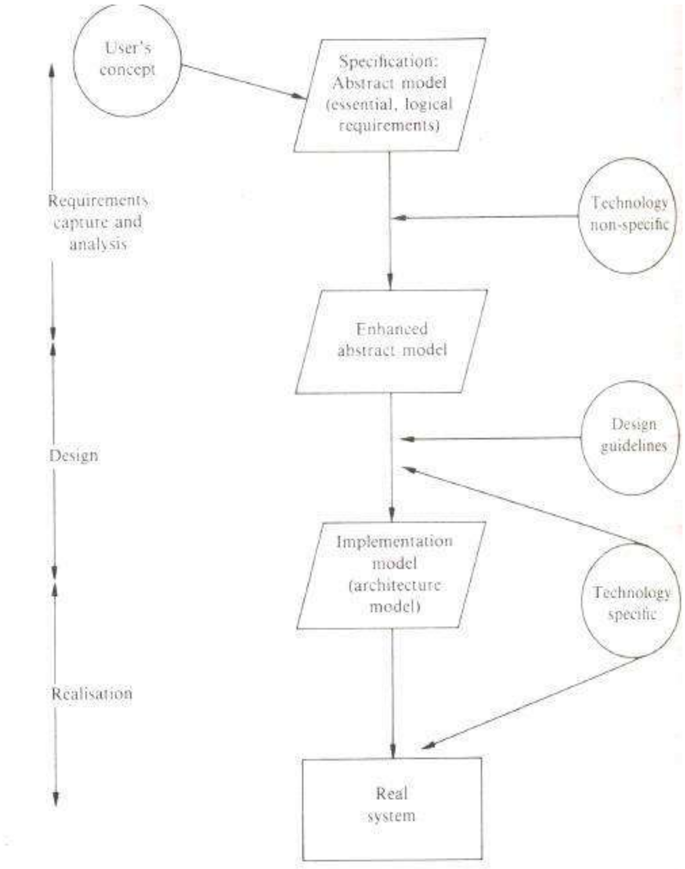

User's concept
Specification: Abstract model (essential, logical requirements)
Requirements capture and analysis
Technology non-specific
Enhanced abstract model
Design
Design guidelines
Implementation model (architecture model)
Technology specific
Realisation
Real system

Figure: Software modeling

The Yourdon methodology has been developed over many years. It is a structured methodology based on using data-flow modeling techniques and junctional decomposition. It supports development from the initial analysis stage through to implementation. Both Ward and Mellor (1986) and Hatley and Pirbhai (1988) have introduced extensions to support the use of the Yourdon approach for the development of real-time systems and the key ideas of their methodologies are:

- Subdivision of system into activities;
- Hierarchical structure;
- Separation of data and control flows;
- No early commitment to a particular technology; and
- Traceability between specification, design and implementation.

8.3 REQUIREMENT DEFINITION FOR DRYING OVEN:

Components are dried by being passed through an oven. The components are placed on a conveyor belt which conveys them slowly through the drying oven. The oven is heated by three gas-fired burners placed at intervals along the oven. The temperature in each of the areas heated by the burners is monitored and controlled. An operator console unit enables the operator to monitor and control the operation of the unit. The system is presently controlled by a hard wired control system. The requirement is to replace this hard wired control system with a computer-based system. The new computer-based system is also to provide links with the management computer over a communication link.

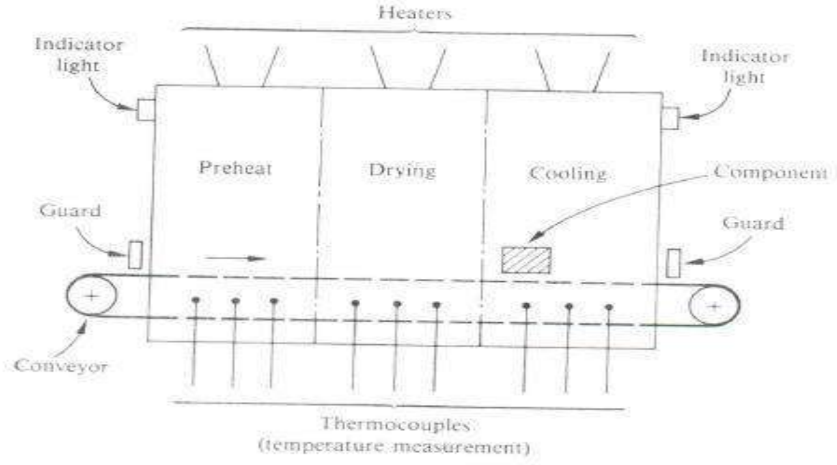


Figure: General arrangement of drying oven.

I
nput/output

The inputs come from a plant interface cubicle and from the operator. There will need to be inputs obtained from the communication interface.

Plant Inputs

A thermocouple is provided in each heater area - the heater areas are pre-heat, drying, and cooling. The inputs are available as voltages in the range 0 to 10 volts at pins 1 to 9 on socket j22 in the interface cubicle.

The conveyor speed is measured by a pulse counting system and is available on pin 3 at socket j23 in the interface cubicle. It is referred to as con-speed..

There are three interlocked safety guards on the conveyor system and these are in-guard, out-guard, and drop-guard. Signals from these guards are provided on pins 4, 5, 6 of socket j23. These signals are set at logic HIGH to indicate that the guards are locked in place.

A conveyor-halted signal is provided on pin I of socket j23. This signal is logic HIGH when the conveyor is running.

Plant Outputs

Heater Control: each of the three heaters has a control unit. The input to the control unit is a voltage in the range 0 to 10 volts which corresponds to no heat output to maximum heat output. Conveyor Start-up: a signal convey-start is output to the conveyor motor control unit.

Guard Locks: asserting the *guard-lock* line, pin 8 on j10 , causes the guards to be locked in position and turns on the red indicator light on the outside of the unit. Operator Inputs

The operator sends the following command inputs: *Start, Stop, Reset, Re-start,* and *Pause.* The operator can also adjust the desired set point for each area of the dryer. Operator Outputs

The operator VDU displays the temperature in each area, the conveyor belt speed, and the alarm status. It should also display the current date and time and the last operator command issued.

8.4 WARD AND MELLOR METHOD:

The outline of the Ward and Mellor method is shown in Figure. The starring point is to build, from the analysis of the requirements, a software model representing the requiremel1ls in terms of the abstract entities. This model is called the essential model. It is in two parts: an environmental model which describes the relationship of the system being modeled with its environment; and the behavioral model which describes the internal structure of the system.

The second stage the design stage - is to derive from the essential model an implementation model which defines how the system is implemented on a particular technology and shows the allocation of parts of the system to processors, the subdivision of activities allocated to each processor into tasks, and the structure of the code for each task. The essential model represents what the system is required to do; the implementation model shows how the system will do what has to be done. The implementation model provides the design from which the implementers of the physical system can work. Correct use of the method results in documentation that provides traceability from the physical system to the abstract speci- fication model.

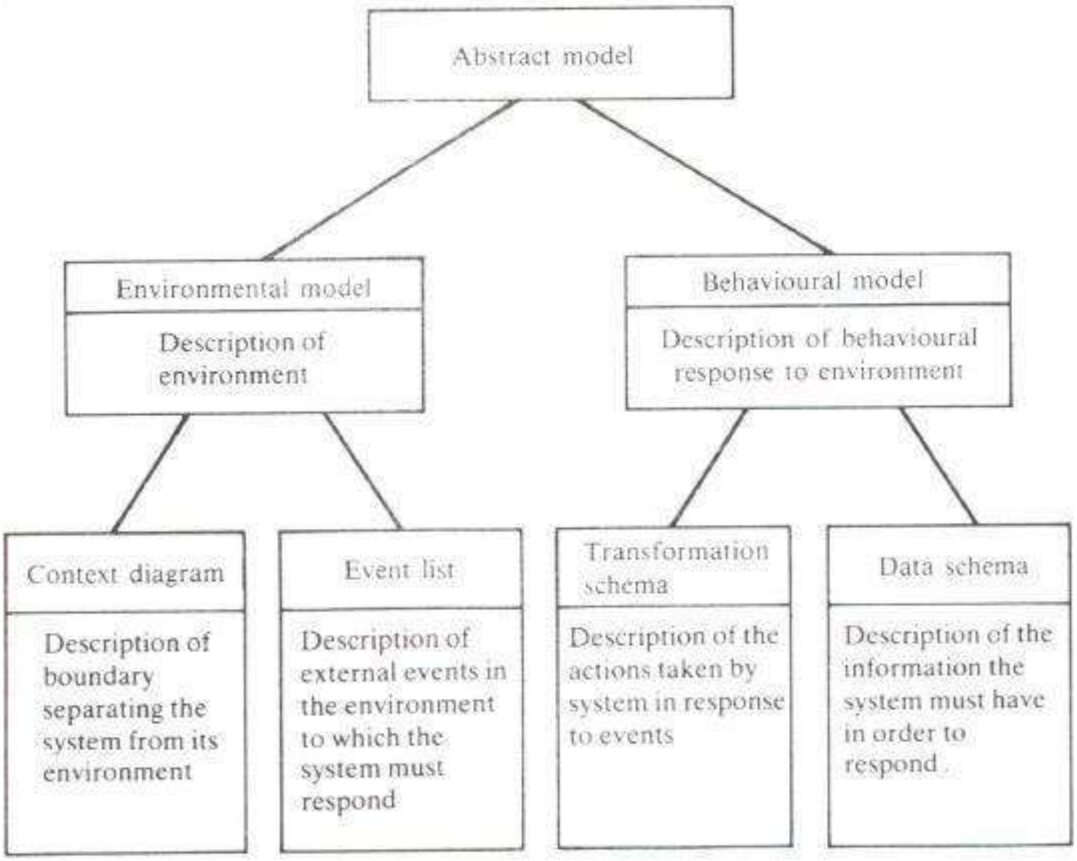
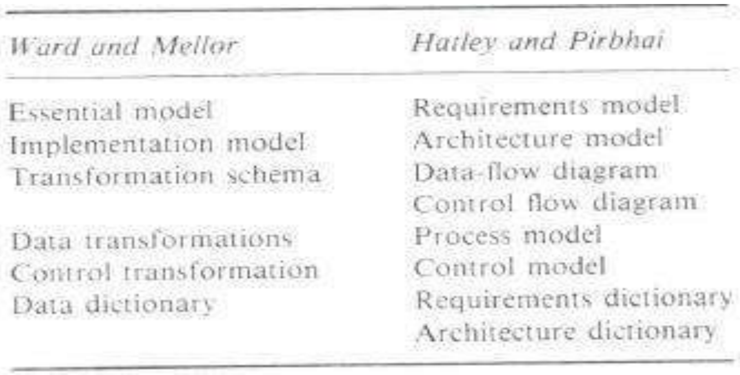

Abstract model
Environmental model
Description of environment
Behavioural model
Description of behavioural response to environment
Context diagram
Description of boundary separating the system from its environment
Event list
Description of external events in the environment to which the system must respond
Transformation schema
Description of the actions taken by system in response to events
Data schema
Description of the information the system must have in order to respond

Figure: Outline of abstract of modeling approach Ward and Mellor.

8.5 HATELY AND PIRBHAI METHOD:

As might be expected the general approach of the Hatley and Pirbhai methodology is very close to that of Ward and Mellor. There are some differences in terminology which are summarized in Table.



Ward and Mellor	*Hatley and Pirbhai*
Essential model	Requirements model
Implementation model	Architecture model
Transformation schema	Data-flow diagram
	Control flow diagram
Data transformations	Process model
Control transformation	Control model
Data dictionary	Requirements dictionary
	Architecture dictionary

Separate diagrams are used for data and control;

- only one CSPEC can appear at any given CFD level; and
- all data f10ws and control f10ws are shown with single arrow heads;

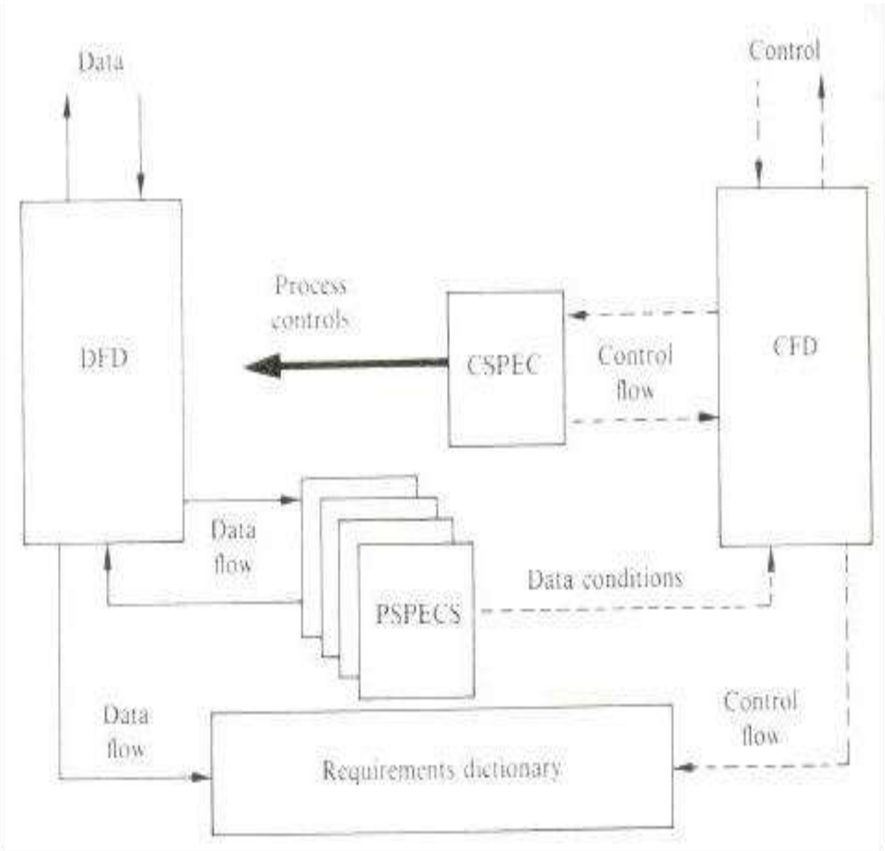
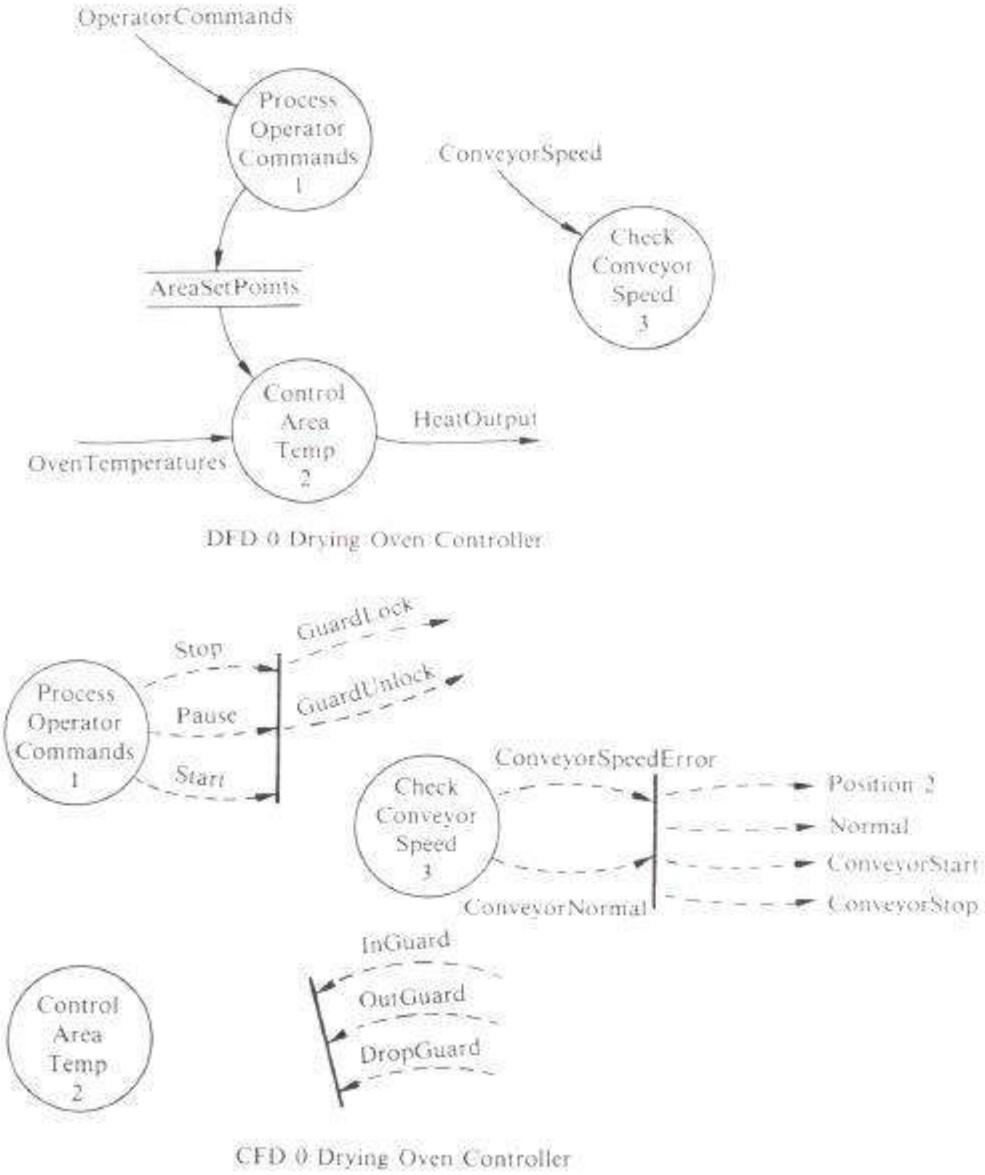


Figure: Structure of requirement model.



DFD 0 Drying Oven Controller

CFD 0 Drying Oven Controller

Figure: Hately and Pirbhai notation.

8.6 COMMENTS ON THE YOURDON METHODOLOGY:

Both methodologies - Ward and Mellor and Hatley and Pirbhai - are simple to learn and have been widely used. They are founded on the well-established structured methods developed by the Yourdon organization and hence over the years a lot of experience in using the techniques has been gained. For serious use on large scale systems they both require the support of CASE tools. The labour involved in checking the models by hand is such that short cuts are likely to be taken and mistakes are bound to occur. It can be argued that the methods are really only a set of procedures for documenting a specification and a design and to some extent this is true.

The analysis procedures are minimal and adequate checking for consistency can be performed only with the support of a CASE tool. However, the methodologies are still useful in that the procedures they recommend provide a sensible way of preparing both a specification and a design in that they encourage the development of hierarchical, modular structures. the two, the Hatley and Pirbhai method is the more structured and formalized in its approach. Its diagrams are less cluttered than those of the Ward and Mellor method and, once the separation is understood, are easier to follow. Many CASE tools provide alternative displays which allow a choice of either separate diagrams or a combined diagram with switching between the two forms. The weakness of both methods lies in the allocation of processors and tasks. The suggestion that one allocates activities to processors and then subdivides the activities into tasks allocated to each processor appears at first sight a sensible way to proceed. However, when it is tried one soon realizes that the information required to do this is not available.

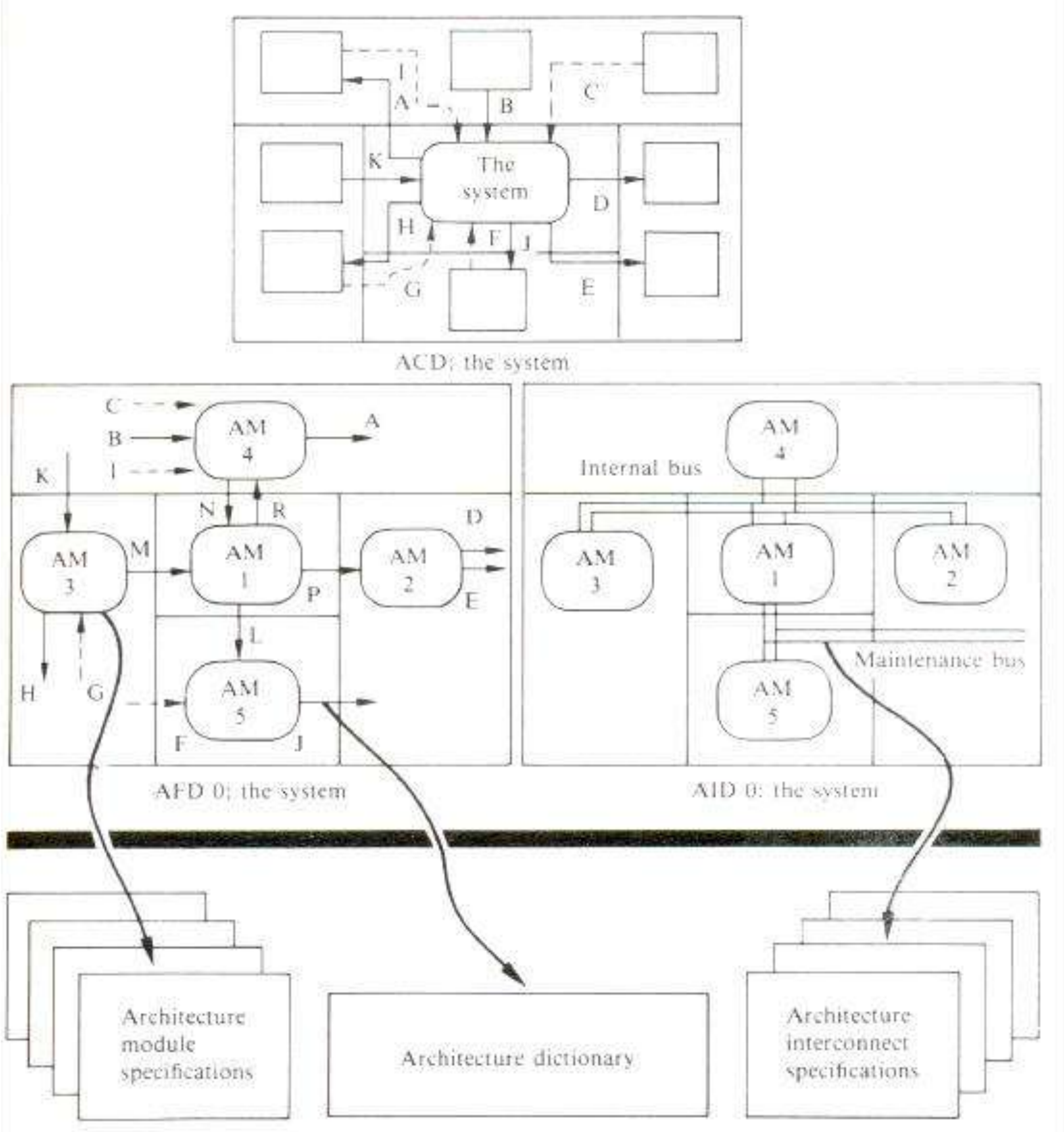


Figure: The structure of architectural model.

Recommended Questions:

1. With a general arrangement for a drying oven, explain its requirements.
2. Write about environmental model, with context diagram for drying oven.

3. Write explanatory notes on the following: A).Hatley and pirbhai methods. B).Ward and millar methods.

4. Show the outline of abstract modeling approach of ward and Mellor and explain.

5. Differentiate between Ward Mellor and hatley and pirbhai mythologies.

4. Explain the CFDO drying over controller using Hatley and pirbhai notation.

6. What do you mean by enhancing the model? Explain with a neat diagram, the relationship between real environment and virtual environment.

7. Write short notes on: i) PSPECs and CSPECs ii) Software modeling iii) YOUR DON methodology