

Module 1:

ARM Cortex-M3 Processor

- The ARM Cortex™-M3 processor, the first of the Cortex generation of processors released by ARM in 2006, was primarily designed to target the 32-bit microcontroller market.
- The Cortex-M3 processor provides excellent performance at low gate count and comes with many new features previously available only in high-end processors.
- The Cortex-M3 addresses the requirements for the 32-bit embedded processor market in the following ways:
 - Greater performance efficiency: allowing more work to be done without increasing the frequency or power requirements
 - **Low power consumption:** enabling longer battery life, especially critical in portable products including wireless networking applications
 - **Enhanced determinism:** guaranteeing that critical tasks and interrupts are serviced as quickly as possible and in a known number of cycles
 - **Improved code density:** ensuring that code fits in even the smallest memory footprints
 - **Ease of use:** providing easier programmability and debugging for the growing number of 8-bit and 16-bit users migrating to 32 bits
 - **Lower cost solutions:** reducing 32-bit-based system costs close to those of legacy 8-bit and 16-bit devices and enabling low-end, 32-bit microcontrollers to be priced at less than US\$1 for the first time
 - **Wide choice of development tools:** from low-cost or free compilers to full-featured development suites from many development tool vendors.

Cortex-M3 processor-based microcontrollers can be easily programmed using the C language and are based on a well-established architecture, application code can be ported and reused easily, reducing development time and testing costs.

Additionally, the Cortex-M3 processor introduces a number of features and technologies that meet the specific requirements of the microcontroller applications, such as nonmaskable interrupts for critical tasks, highly deterministic nested vector interrupts, atomic bit manipulation, and an optional Memory Protection Unit (MPU). These factors make the Cortex-M3 processor attractive to existing ARM processor users as well as many new users considering use of 32-bit MCUs in their products.

Instruction Set Development

- Two different instruction sets are supported on the ARM processor: the ARM instructions that are 32 bits and Thumb instructions that are 16 bits.
- During program execution, the processor can be dynamically switched between the ARM state and the Thumb state to use either one of the instruction sets.
- The Thumb instruction set provides only a subset of the ARM instructions, but it can provide higher code density. It is useful for products with tight memory requirements.

The Thumb-2 Technology and Instruction Set Architecture

The Thumb-2 technology extended the Thumb Instruction Set Architecture (ISA) into a highly efficient and powerful instruction set that delivers significant benefits in terms of ease of use, code size, and performance (see Figure 1.1).

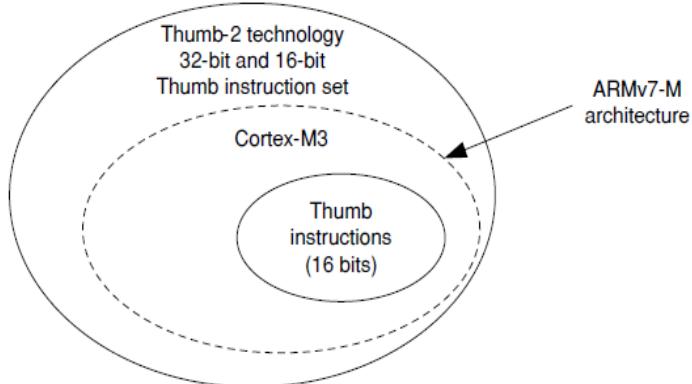


Figure 1.1: The Relationship between the Thumb Instruction Set in Thumb-2 Technology and the Traditional Thumb.

- The extended instruction set in Thumb-2 is a superset of the previous 16-bit Thumb instruction set, with additional 16-bit instructions alongside 32-bit instructions.
- It allows more complex operations to be carried out in the Thumb state, thus allowing higher efficiency by reducing the number of states switching between ARM state and Thumb state. Focused on small memory system devices such as microcontrollers and reducing the size of the processor, the Cortex-M3 supports only the Thumb-2 (and traditional Thumb) instruction set. Instead of using ARM instructions for some operations, as in traditional ARM processors, it uses the Thumb-2 instruction set for all operations. As a result, the Cortex-M3 processor is not backward compatible with traditional ARM processors.
- With support for both 16-bit and 32-bit instructions in the Thumb-2 instruction set, there is no need to switch the processor between Thumb state (16-bit instructions) and ARM state (32-bit instructions).
- For example, in ARM7 or ARM9 family processors, you might need to switch to ARM state if you want to carry out complex calculations or a large number of conditional operations and good performance is needed, whereas in the Cortex-M3 processor, you can mix 32-bit instructions with 16-bit instructions without switching state, getting high code density and high performance with no extra complexity.
- The Thumb-2 instruction set is a very important feature of the ARMv7 architecture. Compared with the instructions supported on ARM7 family processors (ARMv4T architecture), the Cortex-M3 processor instruction set has a large number of new features.
- For the first time, hardware divide instruction is available on an ARM processor, and a number of multiply instructions are also available on the Cortex-M3 processor to improve data-crunching performance. The Cortex-M3 processor also supports unaligned data accesses, a feature previously available only in high-end processors.

Cortex-M3 Processor Applications

- **Low-cost microcontrollers:** The Cortex-M3 processor is ideally suited for low-cost microcontrollers, which are commonly used in consumer products, from toys to electrical appliances. Its lower power, high performance, and ease-of-use advantages enable embedded developers to migrate to 32-bit systems and develop products with the ARM architecture.
- **Automotive:** Another ideal application for the Cortex-M3 processor is in the automotive industry. The Cortex-M3 processor has very high-performance efficiency and low interrupt latency, allowing it to be used in real-time systems. The Cortex-M3 processor supports up to 240 external vectored interrupts, with a built-in interrupt controller with nested interrupt supports and an optional MPU, making it ideal for highly integrated and cost-sensitive automotive applications.
- **Data communications:** The processor's low power and high efficiency, coupled with instructions in Thumb-2 for bit-field manipulation, make the Cortex-M3 ideal for many communications applications, such as Bluetooth and ZigBee.
- **Industrial control:** In industrial control applications, simplicity, fast response, and reliability are key factors. Again, the Cortex-M3 processor's interrupt feature, low interrupt latency, and enhanced fault-handling features make it a strong candidate in this area.
- **Consumer products:** In many consumer products, a high-performance microprocessor (or several of them) is used. The Cortex-M3 processor, being a small processor, is highly efficient and low in power and supports an MPU enabling complex software to execute while providing robust memory protection.

The Memory Map

- The Cortex-M3 has a predefined memory map. This allows the built-in peripherals, such as the interrupt controller and the debug components, to be accessed by simple memory access instructions.
- The predefined memory map also allows the Cortex-M3 processor to be highly optimized for speed and ease of integration in system-on-a-chip (SoC) designs.
- Overall, the 4 GB memory space can be divided into ranges as shown in [Figure 2.6](#).
- The Cortex-M3 design has an internal bus infrastructure optimized for this memory usage. In addition, the design allows these regions to be used differently. For example, data memory can still be put into the CODE region, and program code can be executed from an external Random Access Memory (RAM) region.

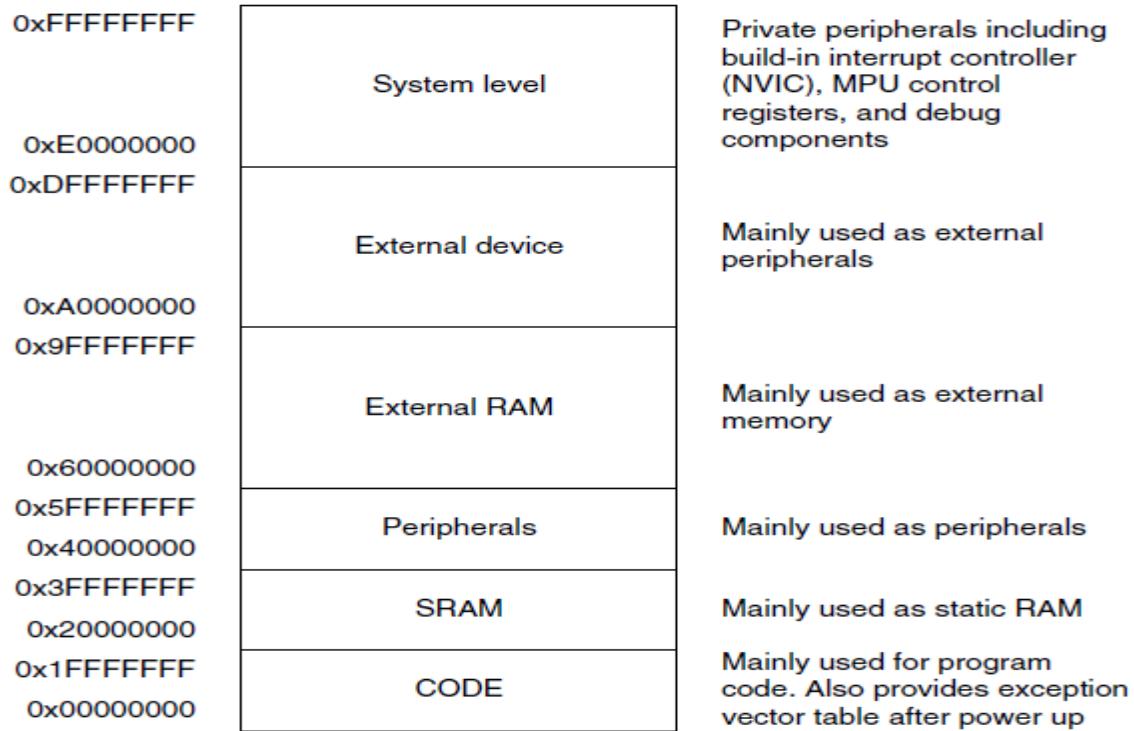


Figure 2.6 The Cortex-M3 Memory Map.

The Bus Interface

There are several bus interfaces on the Cortex-M3 processor. They allow the Cortex-M3 to carry instruction fetches and data accesses at the same time. The main bus interfaces are as follows:

- Code memory buses
- System bus
- Private peripheral bus
 - The code memory region access is carried out on the code memory buses, which physically consist of two buses, one called I-Code and other called D-Code. These are optimized for instruction fetches for best instruction execution speed.
 - The system bus is used to access memory and peripherals. This provides access to the Static Random Access Memory (SRAM), peripherals, external RAM, external devices, and part of the system level memory regions.
 - The private peripheral bus provides access to a part of the system-level memory dedicated to private peripherals, such as debugging components.

The MPU

- The Cortex-M3 has an optional MPU. This unit allows access rules to be set up for privileged access and user program access.
- When an access rule is violated, a fault exception is generated, and the fault exception handler will be able to analyze the problem and correct it, if possible.
- The MPU can be used in various ways. In common scenarios, the OS can set up the MPU to protect data use by the OS kernel and other privileged processes to be protected from untrusted user programs.
- The MPU can also be used to make memory regions read-only, to prevent accidental erasing of data or to isolate memory regions between different tasks in a multitasking system. Overall, it can help make embedded systems more robust and reliable.

Debugging Support

- The Cortex-M3 processor includes a number of debugging features, such as program execution controls, including halting and stepping, instruction breakpoints, data watchpoints, registers and memory accesses, profiling, and traces.
- The debugging hardware of the Cortex-M3 processor is based on the CoreSight™ architecture.
- Unlike traditional ARM processors, the CPU core itself does not have a Joint Test Action Group (JTAG) interface. Instead, a debug interface module is decoupled from the core, and a bus interface called the Debug Access Port (DAP) is provided at the core level. Through this bus interface, external debuggers can access control registers to debug hardware as well as system memory, even when the processor is running. The control of this bus interface is carried out by a Debug Port (DP) device.
- The DPs currently available are the Serial-Wire JTAG Debug Port (SWJ-DP) (supports the traditional JTAG protocol as well as the Serial-Wire protocol) or the SW-DP (supports the Serial-Wire protocol only). A JTAG-DP module from the ARM CoreSight product family can also be used. Chip manufacturers can choose to attach one of these DP modules to provide the debug interface.
- Chip manufacturers can also include an Embedded Trace Macrocell (ETM) to allow instruction trace. Trace information is output via the Trace Port Interface Unit (TPIU),

and the debug host (usually a Personal Computer [PC]) can then collect the executed instruction information via external trace capturing hardware.

- Within the Cortex-M3 processor, a number of events can be used to trigger debug actions. Debug events can be breakpoints, watchpoints, fault conditions, or external debugging request input signals.
- When a debug event takes place, the Cortex-M3 processor can either enter halt mode or execute the debug monitor exception handler.
- The data watchpoint function is provided by a Data Watchpoint and Trace (DWT) unit in the Cortex -M3 processor. This can be used to stop the processor (or trigger the debug monitor exception routine) or to generate data trace information. When data trace is used, the traced data can be output via the TPIU. (In the CoreSight architecture, multiple trace devices can share one single trace port.)
- In addition to these basic debugging features, the Cortex-M3 processor also provides a Flash Patch and Breakpoint (FPB) unit that can provide a simple breakpoint function or remap an instruction access from Flash to a different location in SRAM.

Registers

The Cortex™-M3 processor has registers R0 through R15 and a number of special registers. R0 through R12 are general purpose, but some of the 16-bit Thumb® instructions can only access R0 through R7 (low registers), whereas 32-bit Thumb-2 instructions can access all these registers.

Special registers have predefined functions and can only be accessed by special register access instructions.

General Purpose Registers R0 through R7

The R0 through R7 general purpose registers are also called *low registers*. They can be accessed by all 16-bit Thumb instructions and all 32-bit Thumb-2 instructions. They are all 32 bits; the reset value is unpredictable.

General Purpose Registers R8 through R12

The R8 through R12 registers are also called *high registers*. They are accessible by all Thumb-2 instructions but not by all 16-bit Thumb instructions. These registers are all 32 bits; the reset value is unpredictable

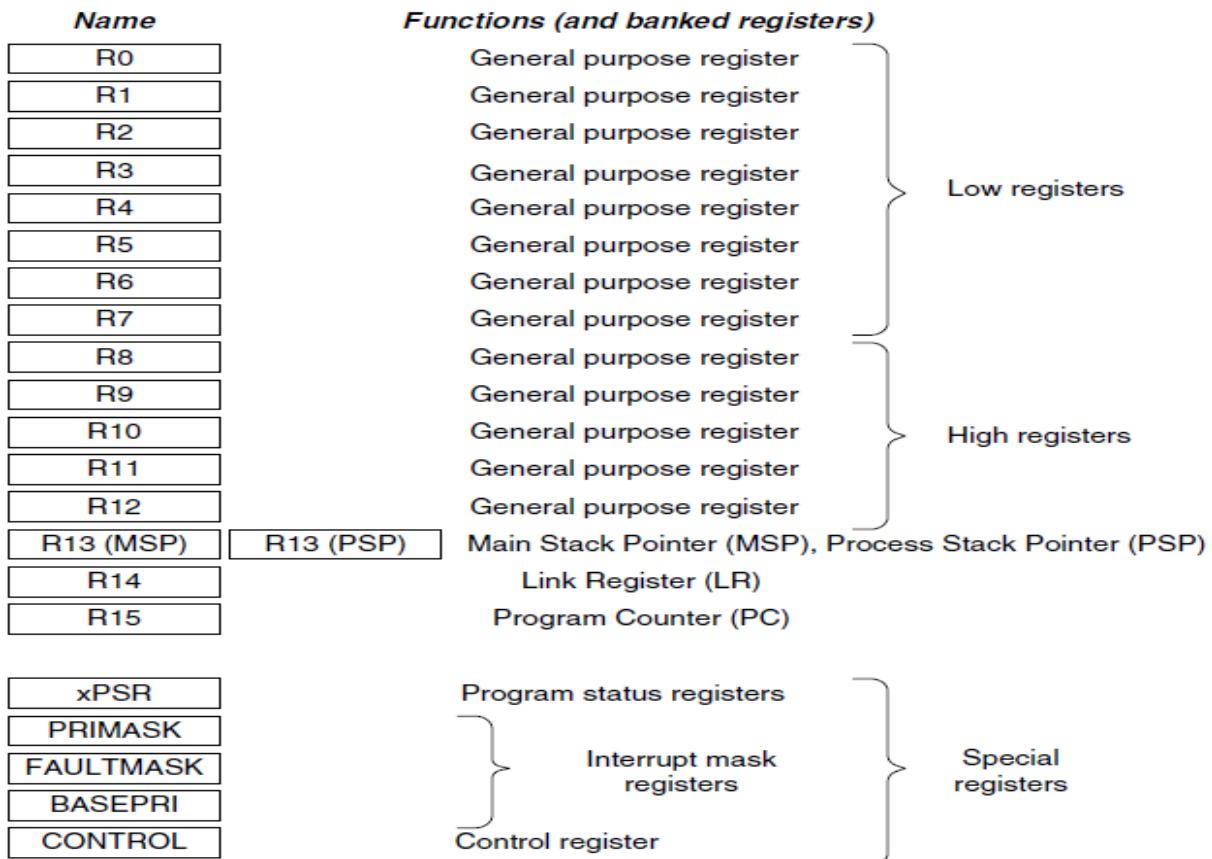


Figure: Registers in the Cortex-M3.

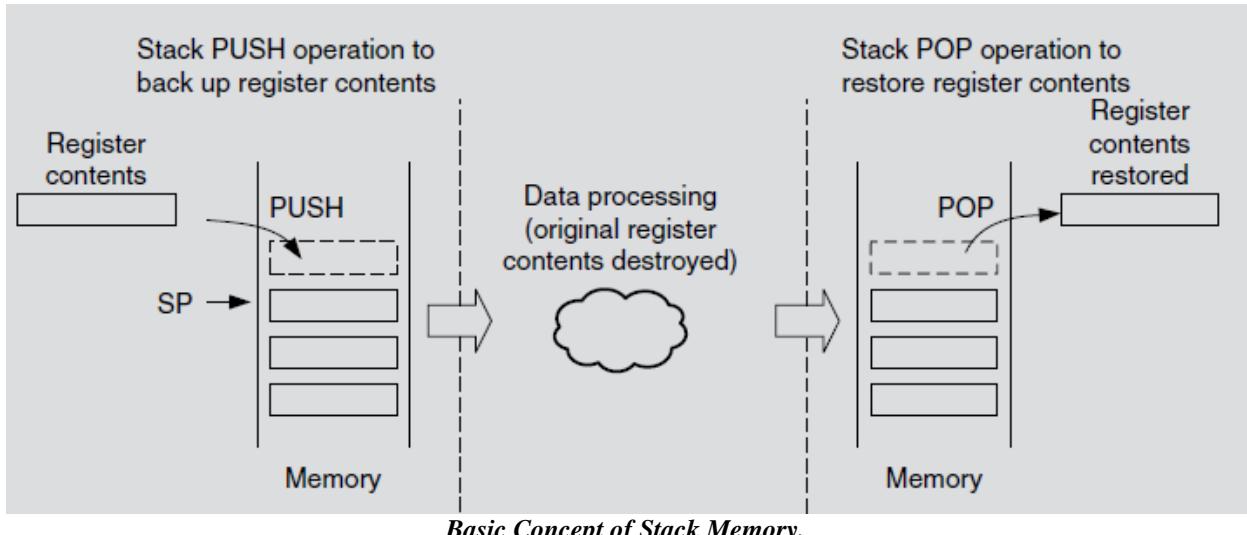
Stack Pointer R13

R13 is the stack pointer (SP). In the Cortex-M3 processor, there are two SPs. This duality allows two separate stack memories to be set up. When using the register name R13, you can only access the current SP; the other one is inaccessible unless you use special instructions to move to special register from general-purpose register (MSR) and move special register to general-purpose register (MRS). The two SPs are as follows:

- *Main Stack Pointer (MSP) or SP main in ARM documentation:* This is the default SP; it is used by the operating system (OS) kernel, exception handlers, and all application codes that require privileged access.
- *Process Stack Pointer (PSP) or SP process in ARM documentation:* This is used by the base-level application code (when not running an exception handler).

Stack PUSH and POP

Stack is a memory usage model. It is simply part of the system memory, and a pointer register (inside the processor) is used to make it work as a first-in/last-out buffer. The common use of a stack is to save register contents before some data processing and then restore those contents from the stack after the processing task is done.



It is not necessary to use both SPs. Simple applications can rely purely on the MSP. The SPs are used for accessing stack memory processes such as PUSH and POP.

In the Cortex-M3, the instructions for accessing stack memory are PUSH and POP. The assembly language syntax is as follows (text after each semicolon [;] is a comment):

```
PUSH {R0}; R13=R13-4, then Memory[R13] = R0
POP {R0}; R0 = Memory[R13], then R13 = R13 + 4
```

The Cortex-M3 uses a full-descending stack arrangement. Therefore, the SP decrements when new data is stored in the stack. PUSH and POP are usually used to save register contents to stack memory at the start of a subroutine and then restore the registers from stack at the end of the subroutine. You can PUSH or POP multiple registers in one instruction:

```
subroutine_1
PUSH {R0-R7, R12, R14} ; Save registers
... ; Do your processing
POP {R0-R7, R12, R14} ; Restore registers
BX R14 ; Return to calling function
```

The MSP, also called *SP_main* in ARM documentation, is the default SP after power-up; it is used by kernel code and exception handlers. The PSP, or *SP_process* in ARM documentation, is typically used by thread processes in system with embedded OS running.

Link Register R14

R14 is the link register (LR). Inside an assembly program, you can write it as either *R14* or *LR*. LR is used to store the return program counter (PC) when a subroutine or function is called—for example, when you're using the branch and link (BL) instruction:

```
main ; Main program
...
BL function1 ; Call function1 using Branch with Link instruction.
; PC = function1 and
; LR = the next instruction in main
```

```

...
function1
... ; Program code for function 1
BX LR ; Return

```

Program Counter R15

R15 is the PC. You can access it in assembler code by either R15 or PC. Because of the pipelined nature of the Cortex-M3 processor, when you read this register, you will find that the value is different than the location of the executing instruction, normally by 4. For example:

```
0x1000 : MOV R0, PC ; R0 = 0x1004
```

Special Registers

The special registers in the Cortex-M3 processor include the following

- Program Status registers (PSRs)
- Interrupt Mask registers (PRIMASK, FAULTMASK, and BASEPRI)
- Control register (CONTROL)

Special registers can only be accessed via MSR and MRS instructions; they do not have memory addresses:

```
MRS <reg>, <special_reg>; Read special register
```

```
MSR <special_reg>, <reg>; write to special register
```

Program Status Registers

The PSRs are subdivided into three status registers:

- Application Program Status register (APSR)
- Interrupt Program Status register (IPSR)
- Execution Program Status register (EPSR)

The three PSRs can be accessed together or separately using the special register access instructions MSR and MRS. When they are accessed as a collective item, the name *xPSR* is used.

You can read the PSRs using the MRS instruction. You can also change the APSR using the MSR instruction, but EPSR and IPSR are read-only. For example:

```

MRS r0, APSR ; Read Flag state into R0
MRS r0, IPSR ; Read Exception/Interrupt state
MRS r0, EPSR ; Read Execution state
MSR APSR, r0 ; Write Flag state

```

	31	30	29	28	27	26:25	24	23:20	19:16	15:10	9	8	7	6	5	4:0
APSR	N	Z	C	V	Q											
IPSR													Exception number			
EPSR					ICI/IT	T				ICI/IT						

Program Status Registers (PSRs) in the Cortex-M3.

	31	30	29	28	27	26:25	24	23:20	19:16	15:10	9	8	7	6	5	4:0
xPSR	N	Z	C	V	Q	ICI/IT	T			ICI/IT						Exception number

Combined Program Status Registers (xPSR) in the Cortex-M3.

Bit Fields in Cortex-M3 Program Status Registers

Bit	Description
N	Negative
Z	Zero
C	Carry/borrow
V	Overflow
Q	Sticky saturation flag
ICI/IT	Interrupt-Continuable Instruction (ICI) bits, IF-THEN instruction status bit
T	Thumb state, always 1; trying to clear this bit will cause a fault exception
Exception number	Indicates which exception the processor is handling

In ARM assembler, when accessing xPSR (all three PSRs as one), the symbol *PSR* is used:

MRS r0, PSR ; Read the combined program status word

MSR PSR, r0 ; Write combined program state word

If you compare this with the Current Program Status register (CPSR) in ARM7, you might find that some bit fields that were used in ARM7 are gone. The Mode (M) bit field is gone because the Cortex-M3 does not have the operation mode as defined in ARM7. Thumb-bit (T) is moved to bit 24. Interrupt status (I and F) bits are replaced by the new interrupt mask registers (PRIMASKs), which are separated from PSR. For comparison, the CPSR in traditional ARM processors is shown below.

	31	30	29	28	27	26:25	24	23:20	19:16	15:10	9	8	7	6	5	4:0
ARM (general)	N	Z	C	V	Q	IT	J	Reserved	GE[3:0]	IT	E	A	I	F	T	M[4:0]
ARM7 TDMI	N	Z	C	V					Reserved				I	F	T	M[4:0]

Figure: Current Program Status Registers in Traditional ARM Processors.

In ARM assembler, when accessing xPSR (all three PSRs as one), the symbol *PSR* is used:

MRS r0, PSR ; Read the combined program status word

MSR PSR, r0 ; Write combined program state word

If you compare this with the Current Program Status register (CPSR) in ARM7, you might find that some bit fields that were used in ARM7 are gone. The Mode (M) bit field is gone because the Cortex-M3 does not have the operation mode as defined in ARM7. Thumb-bit (T) is moved to bit 24. Interrupt status (I and F) bits are replaced by the new interrupt mask registers (PRIMASKs), which are separated from PSR. For comparison, the CPSR in traditional ARM processors is shown in above figure.

PRIMASK, FAULTMASK, and BASEPRI Registers

The PRIMASK and BASEPRI registers are useful for temporarily disabling interrupts in timing-critical tasks. An OS could use FAULTMASK to temporarily disable fault handling when a task has crashed. In this scenario, a number of different faults might be taking place when a task crashes. Once the core starts cleaning up, it might not want to be interrupted by other faults caused by the crashed process. Therefore, the FAULTMASK gives the OS kernel time to deal with fault conditions.

Table 3.2 Cortex-M3 Interrupt Mask Registers

Register Name	Description
PRIMASK	A 1-bit register, when this is set, it allows nonmaskable interrupt (NMI) and the hard fault exception; all other interrupts and exceptions are masked. The default value is 0, which means that no masking is set.
FAULTMASK	A 1-bit register, when this is set, it allows only the NMI, and all interrupts and fault handling exceptions are disabled. The default value is 0, which means that no masking is set.
BASEPRI	A register of up to 8 bits (depending on the bit width implemented for priority level). It defines the masking priority level. When this is set, it disables all interrupts of the same or lower level (larger priority value). Higher priority interrupts can still be allowed. If this is set to 0, the masking function is disabled (this is the default).

To access the PRIMASK, FAULTMASK, and BASEPRI registers, a number of functions are available in the device driver libraries provided by the microcontroller vendors. For example, the following:

```
x = __get_BASEPRI(); // Read BASEPRI register  
x = __get_PRIMARK(); // Read PRIMASK register  
x = __get_FAULTMASK(); // Read FAULTMASK register  
__set_BASEPRI(x); // Set new value for BASEPRI  
__set_PRIMASK(x); // Set new value for PRIMASK  
__set_FAULTMASK(x); // Set new value for FAULTMASK  
__disable_irq(); // Clear PRIMASK, enable IRQ  
__enable_irq(); // Set PRIMASK, disable IRQ
```

In assembly language, the MRS and MSR instructions are used. For example:

```
MRS r0, BASEPRI ; Read BASEPRI register into R0  
MRS r0, PRIMASK ; Read PRIMASK register into R0  
MRS r0, FAULTMASK ; Read FAULTMASK register into R0  
MSR BASEPRI, r0 ; Write R0 into BASEPRI register  
MSR PRIMASK, r0 ; Write R0 into PRIMASK register  
MSR FAULTMASK, r0 ; Write R0 into FAULTMASK register
```

The PRIMASK, FAULTMASK, and BASEPRI registers cannot be set in the user access level.

The Control Register

The control register is used to define the privilege level and the SP selection. This register has 2 bits.

CONTROL[1]

In the Cortex-M3, the CONTROL[1] bit is always 0 in handler mode. However, in the thread or base level, it can be either 0 or 1.

CONTROL[0]

The CONTROL[0] bit is writable only in a privileged state. Once it enters the user state, the only way to switch back to privileged is to trigger an interrupt and change this in the exception handler.

To access the control register in C, the following CMSIS functions are available in CMSIS compliant device driver libraries:

```
x = __get_CONTROL(); // Read the current value of CONTROL  
__set_CONTROL(x); // Set the CONTROL value to x
```

To access the control register in assembly, the MRS and MSR instructions are used:

MRS r0, CONTROL ; Read CONTROL register into R0

MSR CONTROL, r0 ; Write R0 into CONTROL register

Table 3.3 Cortex-M3 Control Register

Bit	Function
CONTROL[1]	Stack status: 1 = Alternate stack is used 0 = Default stack (MSP) is used If it is in the thread or base level, the alternate stack is the PSP. There is no alternate stack for handler mode, so this bit must be 0 when the processor is in handler mode.
CONTROL[0]	0 = Privileged in thread mode 1 = User state in thread mode If in handler mode (not thread mode), the processor operates in privileged mode.

Exceptions and Interrupts

- The Cortex-M3 supports a number of exceptions, including a fixed number of system exceptions and a number of interrupts, commonly called *IRQ*.
- The number of interrupt inputs on a Cortex-M3 microcontroller depends on the individual design. Interrupts generated by peripherals, except System Tick Timer, are also connected to the interrupt input signals.
- The typical number of interrupt inputs is 16 or 32. However, you might find some microcontroller designs with more (or fewer) interrupt inputs.
- Besides the interrupt inputs, there is also a nonmaskable interrupt (NMI) input signal. The actual use of NMI depends on the design of the microcontroller or system-on-chip (SoC) product you use.
- In most cases, the NMI could be connected to a watchdog timer or a voltage-monitoring block that warns the processor when the voltage drops below a certain level.

- The NMI exception can be activated any time, even right after the core exits reset.
- The list of exceptions found in the Cortex-M3 is given below. A number of the system exceptions are fault-handling exceptions that can be triggered by various error conditions.
- The NVIC also provides a number of fault status registers so that error handlers can determine the cause of the exceptions.

Table 3.4 Exception Types in Cortex-M3

Exception Number	Exception Type	Priority	Function
1	Reset	-3 (Highest)	Reset
2	NMI	-2	Nonmaskable interrupt
3	Hard fault	-1	All classes of fault, when the corresponding fault handler cannot be activated because it is currently disabled or masked by exception masking
4	MemManage	Settable	Memory management fault; caused by MPU violation or invalid accesses (such as an instruction fetch from a nonexecutable region)
5	Bus fault	Settable	Error response received from the bus system; caused by an instruction prefetch abort or data access error
6	Usage fault	Settable	Usage fault; typical causes are invalid instructions or invalid state transition attempts (such as trying to switch to ARM state in the Cortex-M3)
7–10	—	—	Reserved
11	SVC	Settable	Supervisor call via SVC instruction
12	Debug monitor	Settable	Debug monitor
13	—	—	Reserved
14	PendSV	Settable	Pendable request for system service
15	SYSTICK	Settable	System tick timer
16–255	IRQ	Settable	IRQ input #0–239

Table 3.5 Vector Table Definition after Reset

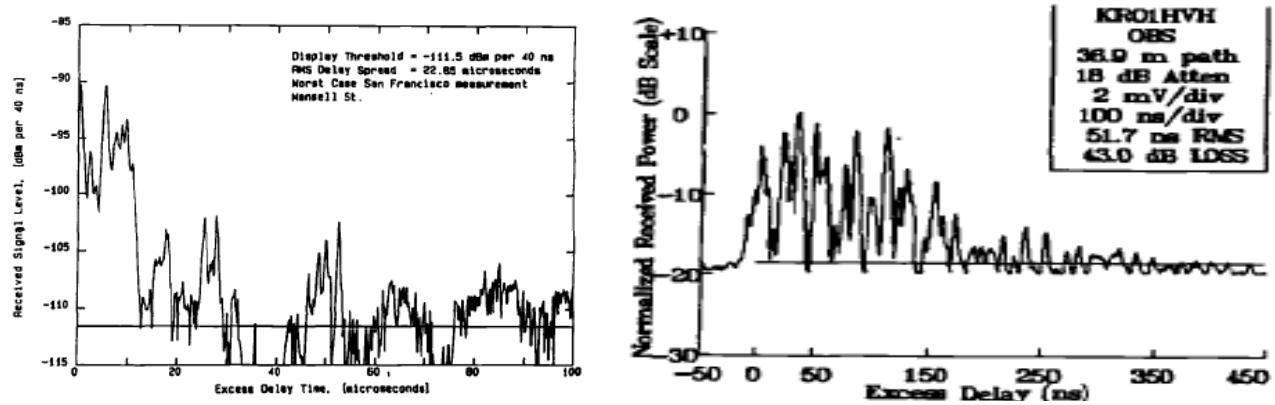
Exception Type	Address Offset	Exception Vector
18–255	0x48–0x3FF	IRQ #2–239
17	0x44	IRQ #1
16	0x40	IRQ #0
15	0x3C	SYSTICK
14	0x38	PendSV
13	0x34	Reserved
12	0x30	Debug monitor
11	0x2C	SVC
7–10	0x1C–0x28	Reserved
6	0x18	Usage fault
5	0x14	Bus fault
4	0x10	MemManage fault
3	0x0C	Hard fault
2	0x08	NMI
1	0x04	Reset
0	0x00	Starting value of the MSP

Stack Memory Operations

In the Cortex-M3, besides normal software-controlled stack PUSH and POP, the stack PUSH and POP operations are also carried out automatically when entering or exiting an exception/interrupt handler.

Parameters of Mobile Multipath Channels:

Many multipath channel parameters are derived from the power delay profile. Depending on the time resolution of the probing pulse and the type of multipath channels studied, researchers often choose to sample at spatial separations of a quarter of a wavelength and over receiver movements no greater than 6 m in outdoor channels and no greater than 2 m in indoor channels in the 450 MHz - 6 GHz range. This small-scale sampling avoids averaging bias in the resulting small-scale statistics.



Measured multipath power delay profile a) From a 900 MHz cellular system in San Francisco b) Inside a grocery store at 4 GHz

1. Time Dispersion Parameters:

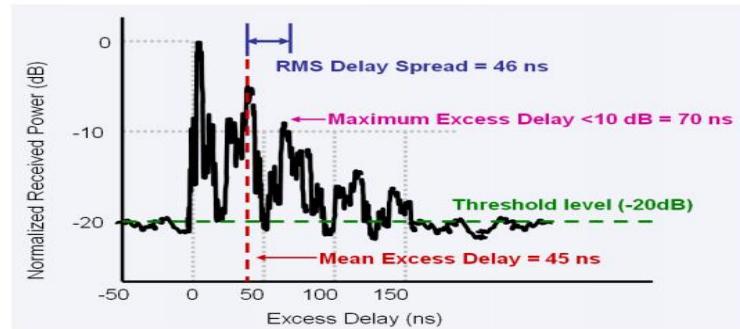
The mean excess delay, rms delay spread, and excess delay spread (X dB) are multipath channel parameters that can be determined from a power delay profile. The time dispersive properties of wide band multipath channels are most commonly quantified by their mean excess delay ($\bar{\tau}$) and rms delay spread (σ_{τ}). The mean excess delay is the first moment of the power delay profile and is defined to be

$$\bar{\tau} = \frac{\sum_k a_k^2 \tau_k}{\sum_k a_k^2} = \frac{\sum_k P(\tau_k) \tau_k}{\sum_k P(\tau_k)}$$

The rms delay spread σ_{τ} : The rms delay spread is the square root of the second central moment of the power delay profile and is defined to be

$$\sigma_{\tau} = \sqrt{\bar{\tau}^2 - (\bar{\tau})^2} \quad \text{and} \quad \bar{\tau}^2 = \frac{\sum_k a_k^2 \tau_k^2}{\sum_k a_k^2} = \frac{\sum_k P(\tau_k) \tau_k^2}{\sum_k P(\tau_k)}$$

Maximum excess delay (X dB): the time delay during which multipath energy falls to X dB below the maximum, $\tau_X - \tau_0$ and **Excess delay spread (X dB), τ_X**



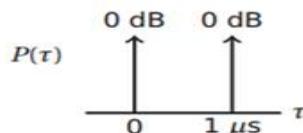
Multipath channel parameters can be determined from power delay profile. The τ , τ^2 and $\sigma\tau$ depend on the choice of noise threshold. These delays are measured relatively to the first detectable signal arriving at $\tau_0 = 0$. Delays do not rely on the absolute power level of $P(\tau)$, but only the relative amplitudes of the multipath components within $P(\tau)$. Analogous to the delay spread parameters in the time domain, coherence bandwidth is used to characterize the channel in frequency domain.

Table 4.1 Typical Measured Values of RMS Delay Spread

Environment	Frequency (MHz)	RMS Delay Spread (σ_τ)	Notes	Reference
Urban	910	1300 ns avg. 600 ns st. dev. 3500 ns max.	New York City	[Cox75]
Urban	892	10-25 μ s	Worst case San Francisco	[Rap90]
Suburban	910	200-310 ns	Averaged typical case	[Cox72]
Suburban	910	1960-2110 ns	Averaged extreme case	[Cox72]
Indoor	1500	10-50 ns 25 ns median	Office building	[Sal87]
Indoor	850	270 ns max.	Office building	[Dev90a]
Indoor	1900	70-94 ns avg. 1470 ns max.	Three San Francisco buildings	[Sei92a]

Example 6

- (a) Compute the RMS delay spread for the following power delay profile. (b) If BPSK modulation is used, what is the maximum bit rate that can be sent through the channel without needing an equalizer?



Solution:

$$(a) \bar{\tau} = \frac{(1)(0)+(1)(1)}{1+1} = \frac{1}{2} = 0.5 \mu s.$$

$$\tau^2 = \frac{(1)(0)^2+(1)(1)^2}{1+1} = \frac{1}{2} = 0.5 \mu s^2.$$

$$\sigma_\tau = \sqrt{\tau^2 - \bar{\tau}^2} = \sqrt{0.5 - (0.5)^2} = \sqrt{0.25} = 0.5 \mu s.$$

$$(b) \frac{\sigma_\tau}{T_s} \leq 0.1 \Rightarrow T_s \geq \frac{\sigma_\tau}{0.1} = \frac{0.5}{0.1} = 5 \mu s$$

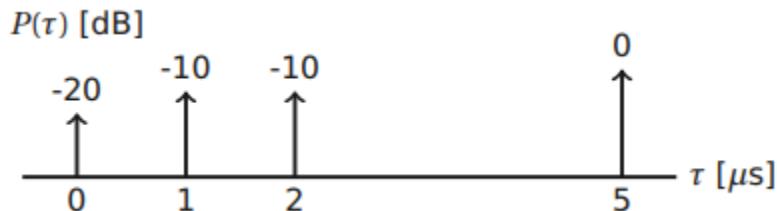
$$R_s = \frac{1}{T_s} = 0.2 \times 10^6 \text{ sps} = 200 \text{ ksp} \Rightarrow R_b = 200 \text{ kbps}$$

2. Coherence Bandwidth: (Time domain: delay spread \Leftrightarrow Frequency domain: coherence bandwidth)

While the delay spread is a natural phenomenon caused by reflected and scattered propagation paths in the radio channel, the coherence bandwidth, is a defined relation derived from the rms delay spread. Coherence bandwidth is a statistical measure of the range of frequencies over which the channel can be considered "flat" (i.e., a channel which passes all spectral components with approximately equal gain and linear phase);

The RMS delay spread and coherence bandwidth are inversely proportional to each other. $\sigma_\tau \propto 1 / B_c$. B_c is a statistical measure of the range of frequencies over which the channel can be considered flat (i.e. a channel which passes all spectral components with approximately equal gain and linear phase). The B_c is the range of frequencies over which two frequency components have a strong potential for amplitude correlation. The bandwidth over which the frequency correlation is above 0.9: $B_c \approx 1 / 50\sigma_\tau$. The bandwidth over which the frequency correlation is above 0.5: $B_c \approx 1 / 5\sigma_\tau$

Problem: Calculate the mean excess delay, RMS delay spread, and the maximum excess delay (10 dB) for the multipath profile below. Estimate the 50% coherence bandwidth of the channel. Would the channel be suitable for AMPS or GSM service without the use of an equalizer?



Solution

$$\tau_{10dB} = 5\mu s, \bar{\tau} = \frac{(1)(5)+(0.1)(1)+(0.1)(2)+(0.01)(0)}{0.01+0.1+0.1+1} = 4.38\mu s$$

$$\overline{\tau^2} = \frac{(1)(5)^2+(0.1)(1)^2+(0.1)(2)^2+(0.01)(0)^2}{0.01+0.1+0.1+1} = 21.07\mu s^2$$

$$\sigma_\tau = \sqrt{21.07 - (4.38)^2} = 1.37\mu s, B_c \approx \frac{1}{5\sigma_\tau} = \frac{1}{5(1.37\mu s)} = 146 \text{ kHz}$$

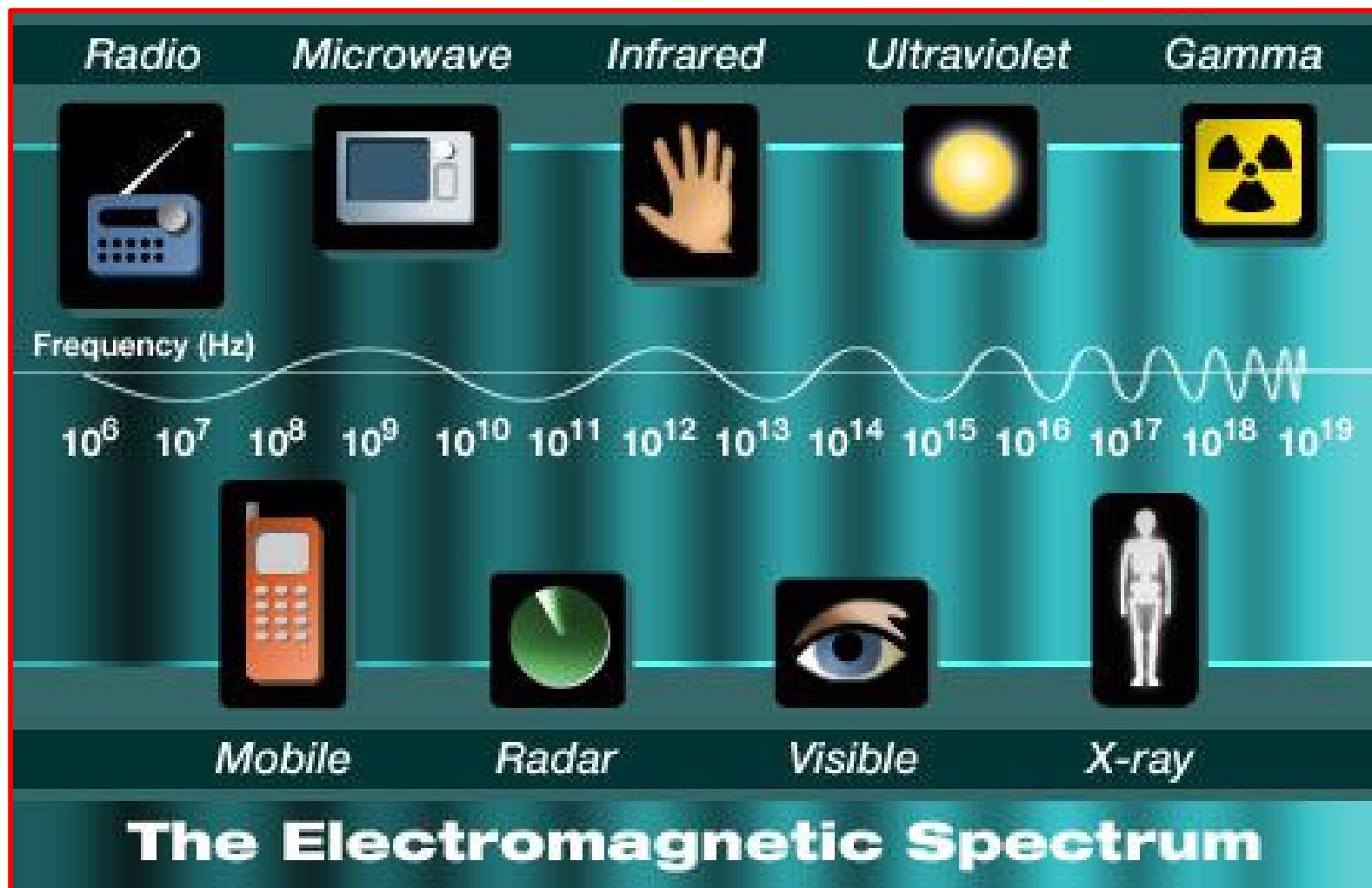
$\because B_c > 30 \text{ kHz} \therefore$ AMPS will work without an equalizer;

$\because B_c < 200 \text{ kHz} \therefore$ GSM will not work without an equalizer.

UNIT I MICROWAVE NETWORK THEORY

Introduction –Microwave frequency range, applications of microwaves– Scattering matrix representation of multi port network properties of S-parameters – S matrix of a two port network with mismatched load – Z and ABCD parameters-Comparison between [S] - [Z] and [Y] matrices

Electromagnetic spectrum



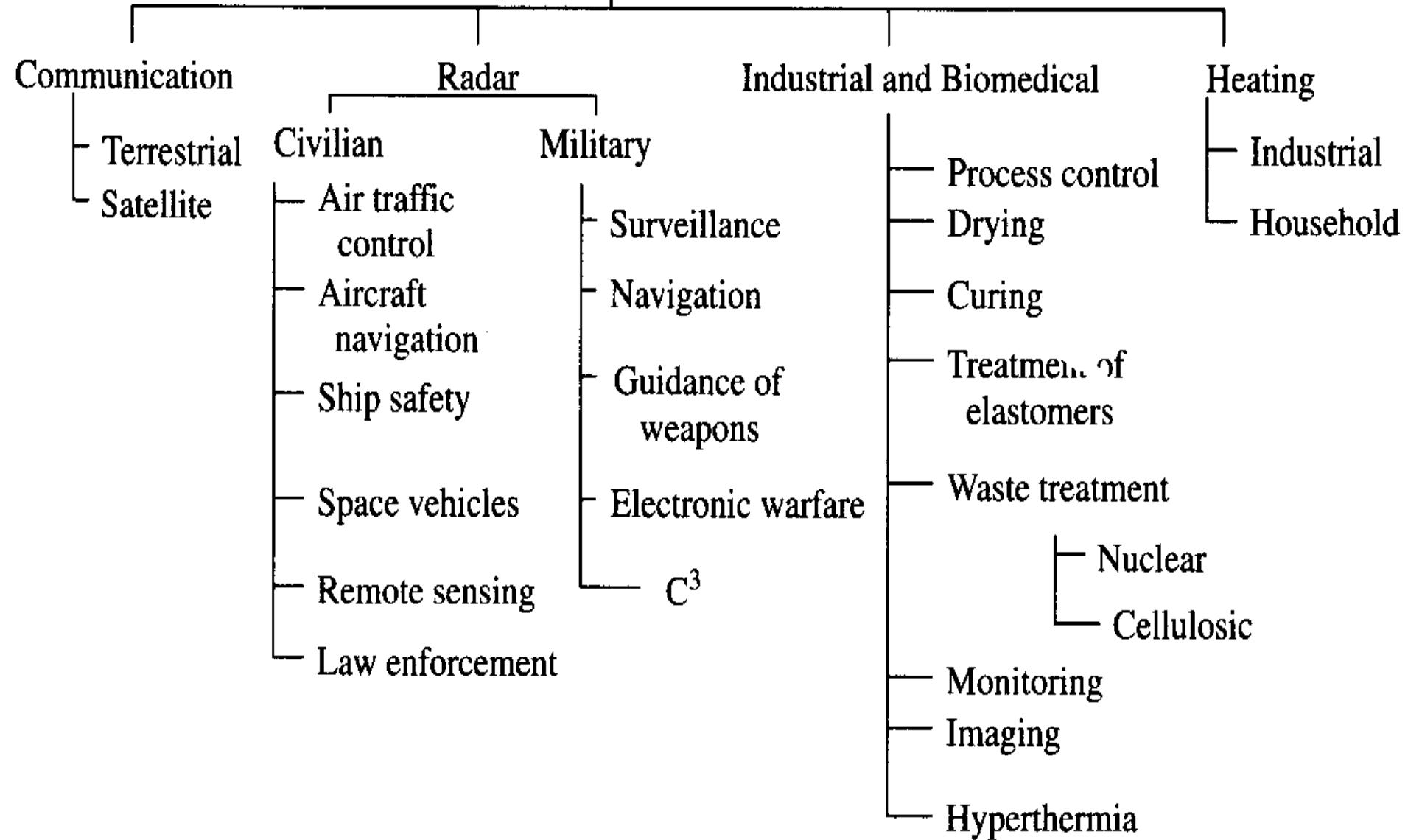
Microwaves

- Microwaves are electromagnetic waves whose frequencies range from about 300 MHz – 300 GHz (1 MHz = 10^6 Hz and 1 GHz = 10^9 Hz) or wavelengths in air ranging from 100 cm – 1 mm.
- The word *Microwave* means *very short wave*, which is the shortest wavelength region of the radio spectrum and a part of the electromagnetic spectrum.

Properties of Microwaves

1. Microwave is an electromagnetic radiation of short wavelength.
2. They can reflect by conducting surfaces just like optical waves since they travel in straight line.
3. Microwave currents flow through a thin outer layer of an ordinary cable.
4. Microwaves are easily attenuated within short distances.
5. They are not reflected by ionosphere

Microwave Applications



Applications

- Microwaves have a wide range of applications in modern technology, which are listed below
1. **Telecommunication:** Intercontinental Telephone and TV, space communication (Earth – to – space and space – to – Earth), telemetry communication link for railways etc.
 2. **Radars:** detect aircraft, track / guide supersonic missiles, observe and track weather patterns, air traffic control (ATC), burglar alarms, garage door openers, police speed detectors etc.

Commercial and industrial applications

- Microwave oven
- Drying machines – textile, food and paper industry for drying clothes, potato chips, printed matters etc.
- Food process industry – Precooling / cooking, pasteurization / sterility, heat frozen / refrigerated precooled meats, roasting of food grains / beans.
- Rubber industry / plastics / chemical / forest product industries
- Mining / public works, breaking rocks, tunnel boring, drying / breaking up concrete, breaking up coal seams, curing of cement.
- Drying inks / drying textiles, drying / sterilizing grains, drying / sterilizing pharmaceuticals, leather, tobacco, power transmission.
- Biomedical Applications (diagnostic / therapeutic) – diathermy for localized superficial heating, deep electromagnetic heating for treatment of cancer, hyperthermia (local, regional or whole body for cancer therapy).

THE SCATTERING MATRIX

- Usually we use Y, Z, H or ABCD parameters to describe a linear two port network.
- These parameters require us to open or short a network to find the parameters.
- At radio frequencies it is difficult to have a proper short or open circuit, there are parasitic inductance and capacitance in most instances.
- Open/short condition leads to standing wave, can cause oscillation and destruction of device.
- For non-TEM propagation mode, it is not possible to measure voltage and current. We can only measure power from E and H fields.

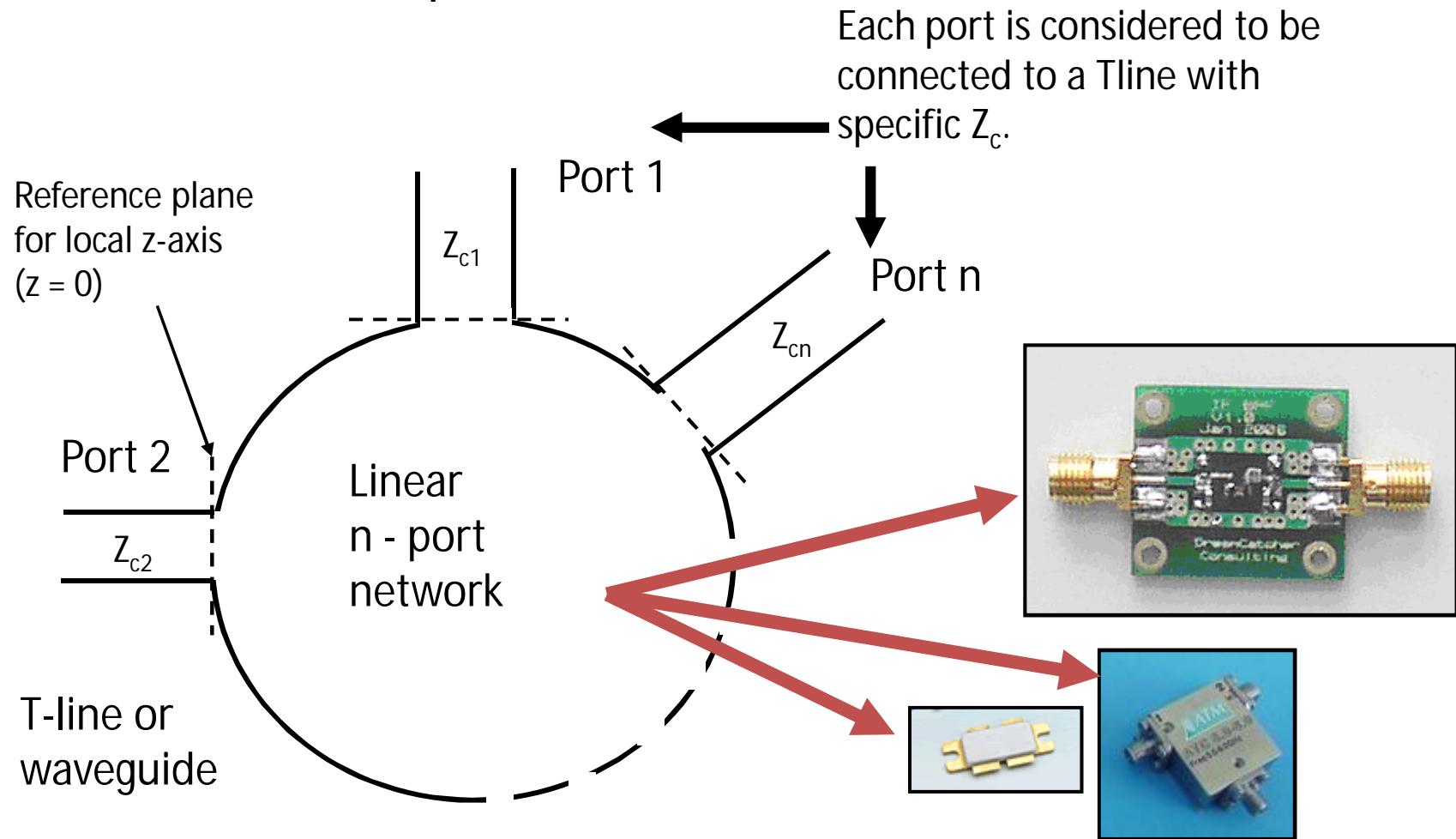
THE SCATTERING MATRIX

- Hence a new set of parameters (S) is needed which
 - Do not need open/short condition.
 - Do not cause standing wave.
 - Relates to incident and reflected power waves, instead of voltage and current.

- As oppose to V and I , S -parameters relate the reflected and incident voltage waves.
- S -parameters have the following advantages:
 1. Relates to familiar measurement such as reflection coefficient, gain, loss etc.
 2. Can cascade S -parameters of multiple devices to predict system performance (similar to $ABCD$ parameters).
 3. Can compute Z , Y or H parameters from S -parameters if needed.

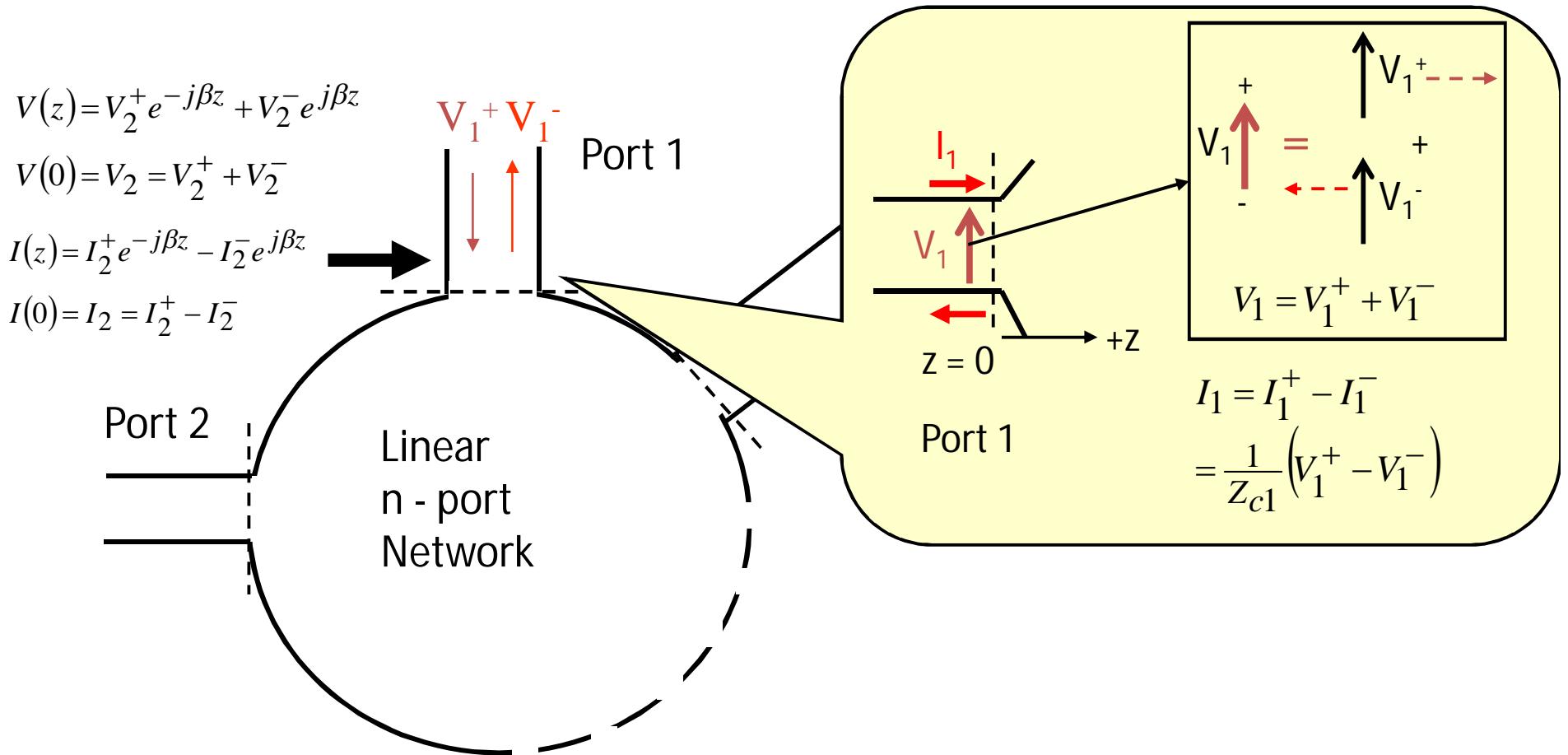
THE SCATTERING MATRIX

- Consider an n – port network:



THE SCATTERING MATRIX

- There is a voltage and current on each port.
- This voltage (or current) can be decomposed into the incident (+) and reflected component (-).



THE SCATTERING MATRIX

- The port voltage and current can be normalized with respect to the impedance connected to it.
- It is customary to define normalized voltage waves at each port as:

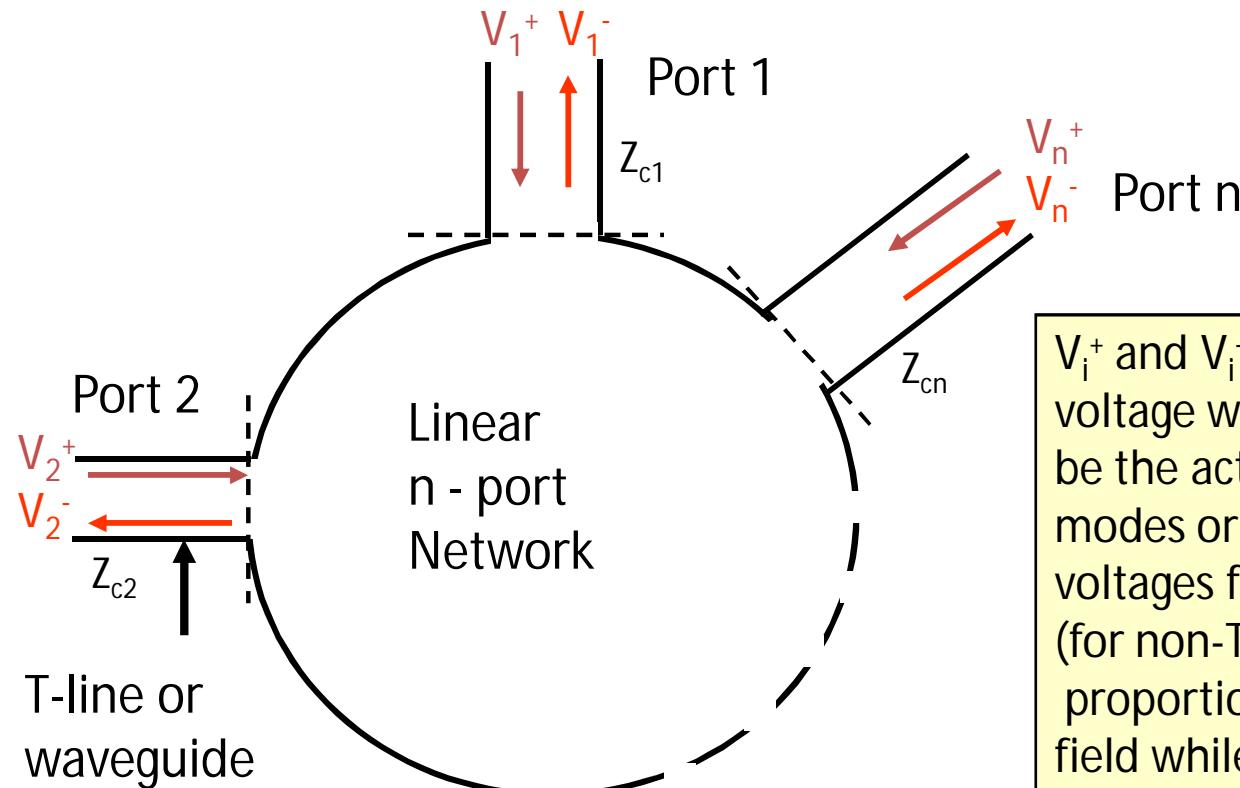
Normalized incident waves

$$a_i = \frac{V_i^+}{\sqrt{Z_{ci}}} \quad (4.3a)$$
$$a_i = I_i^+ \sqrt{Z_{ci}}$$

$$b_i = \frac{V_i^-}{\sqrt{Z_{ci}}} \quad \text{Normalized reflected waves}$$
$$b_i = I_i^- \sqrt{Z_{ci}} \quad (4.3b)$$

THE SCATTERING MATRIX

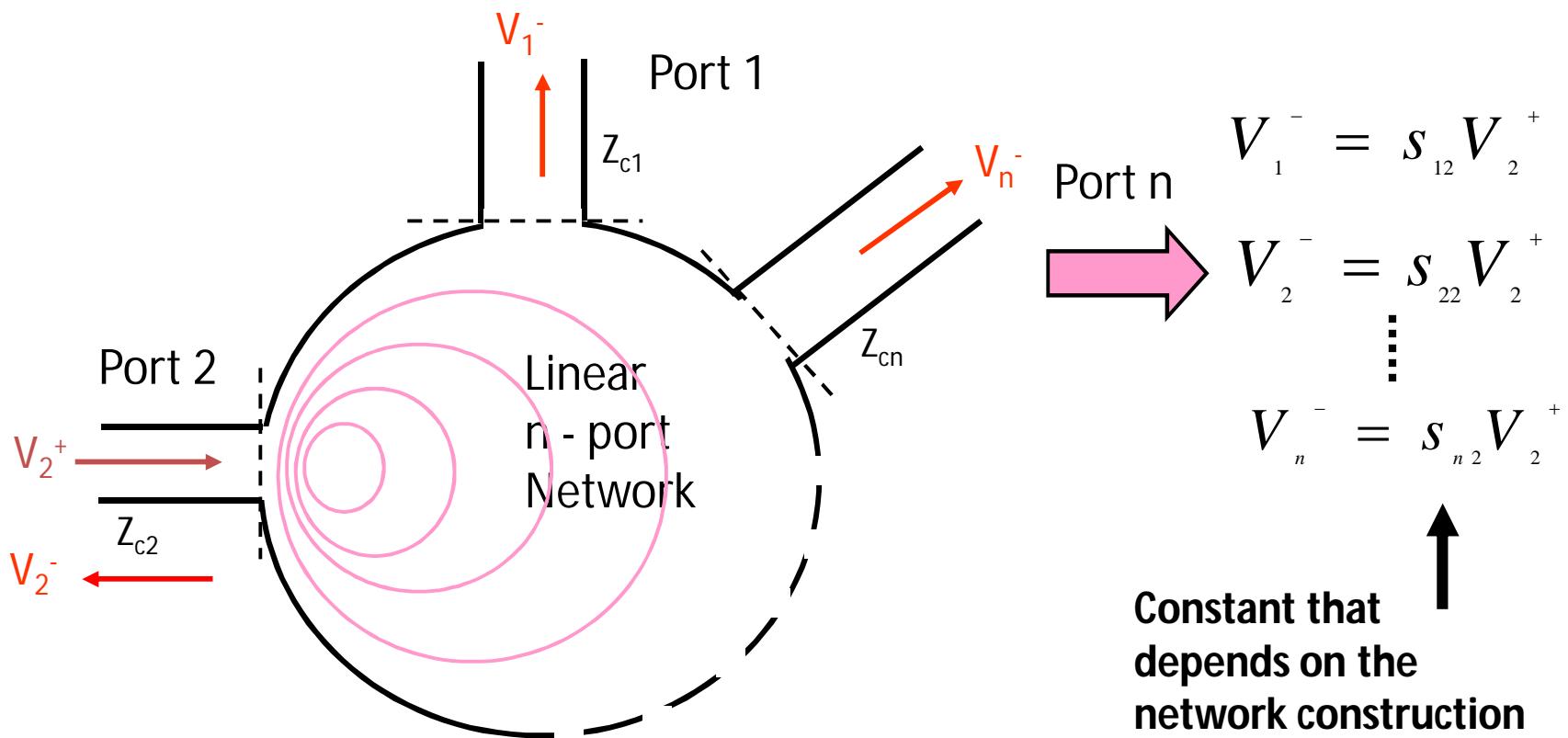
- Thus in general:



V_i^+ and V_i^- are propagating voltage waves, which can be the actual voltage for TEM modes or the equivalent voltages for non-TEM modes. (for non-TEM, V is defined proportional to transverse E field while I is defined proportional to transverse H field, see [1] for details).

THE SCATTERING MATRIX

- If the n – port network is linear (make sure you know what this means!), there is a linear relationship between the normalized waves.
- For instance if we energize port 2:



THE SCATTERING MATRIX

- Considering that we can send energy into all ports, this can be generalized to:

$$\begin{aligned}
 V_1^- &= s_{11} V_1^+ + s_{12} V_2^+ + s_{13} V_3^+ + \cdots + s_{1n} V_n^+ \\
 V_2^- &= s_{21} V_1^+ + s_{22} V_2^+ + s_{23} V_3^+ + \cdots + s_{2n} V_n^+ \\
 &\vdots \\
 V_n^- &= s_{n1} V_1^+ + s_{n2} V_2^+ + s_{n3} V_3^+ + \cdots + s_{nn} V_n^+
 \end{aligned} \tag{4.4a}$$

- Or written in Matrix equation:

$$\overline{V^-} = \overline{\overline{S} \overline{V^+}} \quad \text{or} \quad \begin{bmatrix} V_1^- \\ V_2^- \\ \vdots \\ V_n^- \end{bmatrix} = \begin{bmatrix} s_{11} & s_{12} & \cdots & s_{1n} \\ s_{21} & s_{22} & \cdots & s_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ s_{n1} & s_{n2} & \cdots & s_{nn} \end{bmatrix} \begin{bmatrix} V_1^+ \\ V_2^+ \\ \vdots \\ V_n^+ \end{bmatrix} \tag{4.4b}$$

- Where s_{ij} is known as the i -th j -th element of the scattering matrix, or just S-parameters for short. From (4.3), each port i can have different characteristic impedance Z_{ci}

THE SCATTERING MATRIX

- Consider the N -port network shown in figure 4.1.

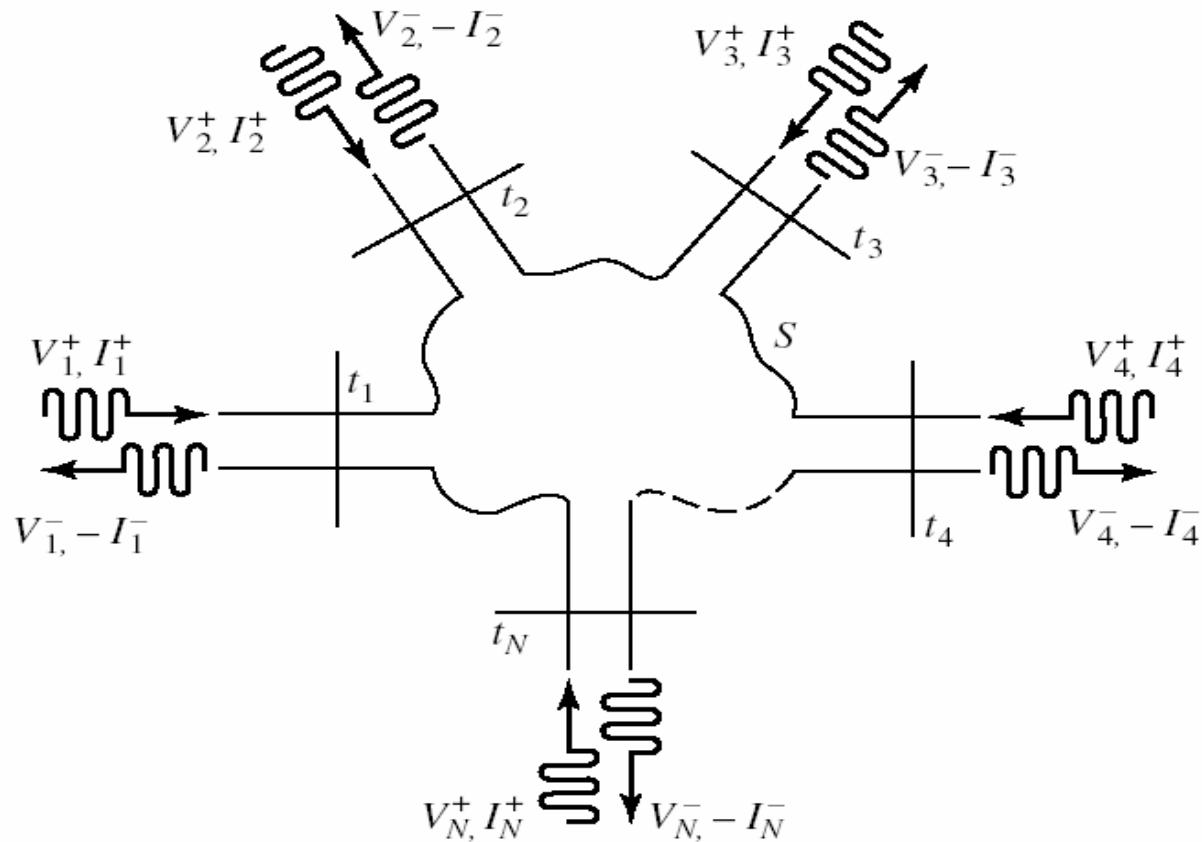


Figure 4.1: An arbitrary N -port microwave network

THE SCATTERING MATRIX

- V_n^+ is the amplitude of the voltage wave incident on port n .
- V_n^- is the amplitude of the voltage wave reflected from port n .
- The scattering matrix or [S] matrix, is defined in relation to these incident and reflected voltage wave as:

$$\begin{bmatrix} V_1^- \\ V_2^- \\ \vdots \\ V_n^- \end{bmatrix} = \begin{bmatrix} S_{11} & S_{12} & \cdot & \cdot & \cdot & S_{1N} \\ S_{21} & \cdot & \cdot & \cdot & \cdot & \cdot \\ \vdots & \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & \ddots & \vdots \\ V_1^+ \\ V_2^+ \\ \vdots \\ V_n^+ \end{bmatrix} \quad [4.1a]$$

THE SCATTERING MATRIX

or $[V^-] = [S][V^+]$ [4.1b]

A specific element of the [S] matrix can be determined as:

$$S_{ij} = \left. \frac{V_i^-}{V_j^+} \right|_{V_k^+ = 0, \text{ for } k \neq j} \quad [4.2]$$

S_{ij} is found by driving port j with an incident wave V_j^+ , and measuring the reflected wave amplitude, V_i^- , coming out of port i .

The incident waves on all ports except j -th port are set to zero (which means that all ports should be terminated in matched load to avoid reflections).

Thus, S_{ii} is the reflection coefficient seen looking into port i when all other ports are terminated in matched loads, and S_{ij} is the transmission coefficient from port j to port i when all other ports are terminated in matched loads.

THE SCATTERING MATRIX

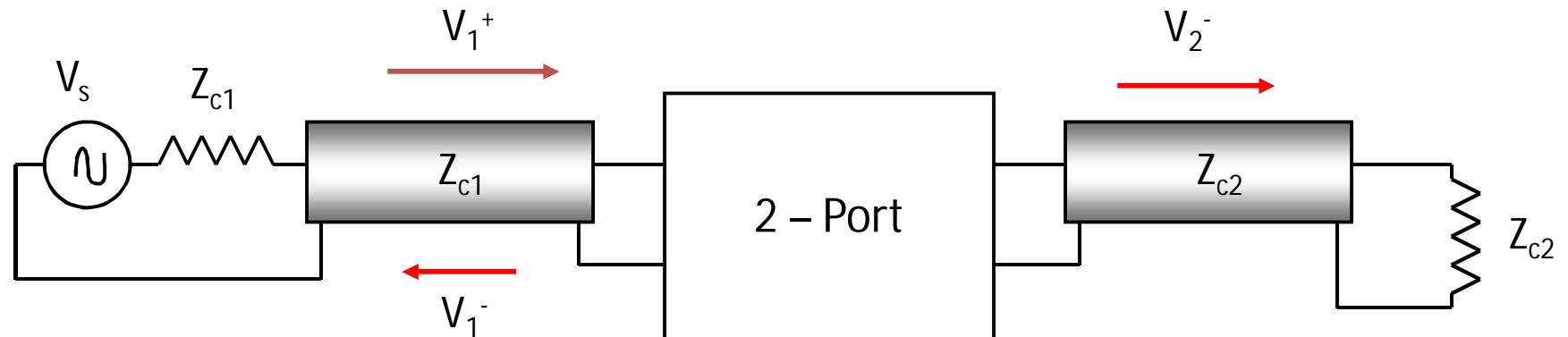
- For 2-port networks, (4.4) reduces to:

$$\begin{bmatrix} V_1^- \\ V_2^- \end{bmatrix} = \begin{bmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{bmatrix} \begin{bmatrix} V_1^+ \\ V_2^+ \end{bmatrix} = \bar{S} \begin{bmatrix} V_1^+ \\ V_2^+ \end{bmatrix} \quad (4.5a)$$

$$S_{11} = \left. \frac{V_1^-}{V_1^+} \right|_{V_2^+ = 0} \quad S_{21} = \left. \frac{V_2^-}{V_1^+} \right|_{V_2^+ = 0} \quad S_{22} = \left. \frac{V_2^-}{V_2^+} \right|_{V_1^+ = 0} \quad S_{12} = \left. \frac{V_1^-}{V_2^+} \right|_{V_1^+ = 0} \quad (4.5b)$$

- Note that $V_i^+ = 0$ implies that we terminate i th port with its characteristic impedance.
- Thus zero reflection eliminates standing wave.

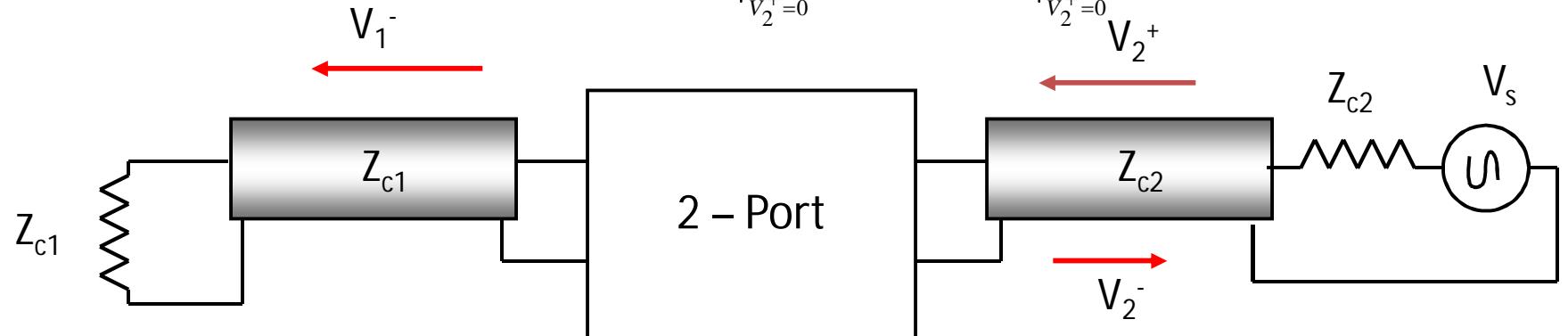
THE SCATTERING MATRIX



Measurement of s_{11} and s_{21} :

$$s_{11} = \left. \frac{V_1^-}{V_1^+} \right|_{V_2^+=0}$$

$$s_{21} = \left. \frac{V_2^-}{V_1^+} \right|_{V_2^+=0}$$



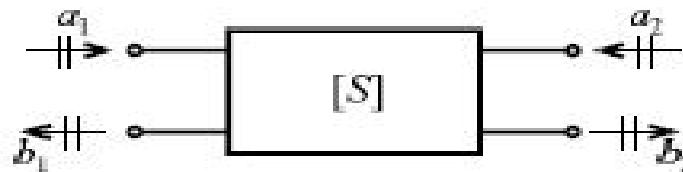
Measurement of s_{22} and s_{12} :

$$s_{22} = \left. \frac{V_2^-}{V_2^+} \right|_{V_1^+=0}$$

$$s_{12} = \left. \frac{V_1^-}{V_2^+} \right|_{V_1^+=0}$$

THE SCATTERING MATRIX

- Input-output behavior of network is defined in terms of normalized power waves
- S-parameters are measured based on properly terminated transmission lines (and not open/short circuit conditions)



$$S_{11} = \frac{b_1}{a_1} \Big|_{a_2=0}$$

$$S_{22} = \frac{b_2}{a_2} \Big|_{a_1=0}$$

$$S_{21} = \frac{b_2}{a_1} \Big|_{a_2=0}$$

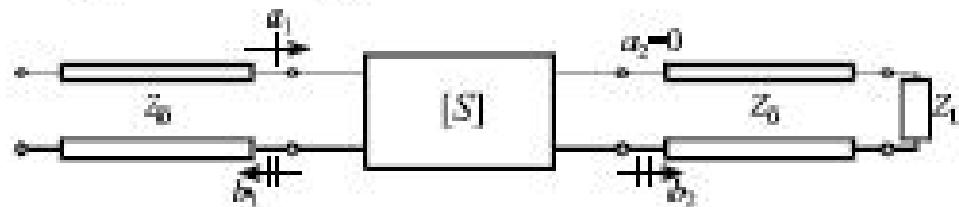
$$S_{12} = \frac{b_1}{a_2} \Big|_{a_1=0}$$

Require proper termination
on port 2

Require proper termination
on port 1

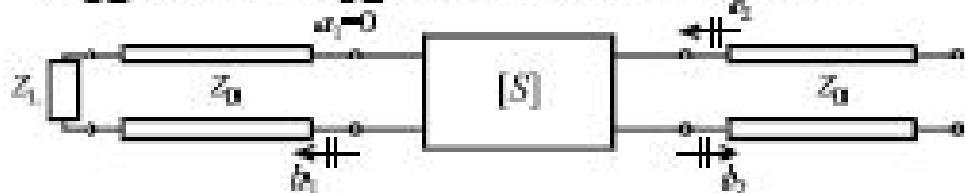
THE SCATTERING MATRIX

Properly terminated port 2 in order to make
 S_{11} and S_{21} measurements



Load impedance =
line impedance

Properly terminated port 1 in order to make
 S_{22} and S_{12} measurements



input impedance =
line impedance

Properties of S-parameter

Reciprocal Networks and *S* Matrices

In the case of reciprocal networks, it can be shown that

$$[S] = [S]^T \quad (4.48), (1)$$

where $[S]^T$ indicates the transpose of $[S]$. In other words, (1) is a statement that $[S]$ is symmetric about the main diagonal, which is what we also observed for the Z and Y matrices.

Lossless Networks and *S* Matrices

The condition for a lossless network is a bit more obtuse for *S* matrices. As derived in your text, if a network is lossless then

Properties of S-parameter

$$[S]^* = \{[S]'\}^{-1} \quad (4.51), (2)$$

which, as it turns out, is a statement that $[S]$ is a **unitary matrix**.

This result can be put into a different, and possibly more useful, form by pre-multiplying (2) by $[S]'$

$$[S]' \cdot [S]^* = [S]' \cdot \{[S]'\}^{-1} = [I] \quad (3)$$

$[I]$ is the unit matrix defined as

$$[I] = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

Expanding (3) we obtain

$$\underbrace{\begin{bmatrix} S_{11} & S_{21} & \cdots & S_{N1} \\ S_{12} & S_{22} & \cdots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ S_{1N} & \cdots & \cdots & S_{NN} \end{bmatrix}}_{-[S]'} \cdot \begin{bmatrix} S_{11}^* & S_{12}^* & \cdots & S_{1N}^* \\ S_{21}^* & S_{22}^* & \cdots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ S_{N1}^* & \cdots & \cdots & S_{NN}^* \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \quad (4)$$

Properties of S-parameter

Shifting Reference Planes

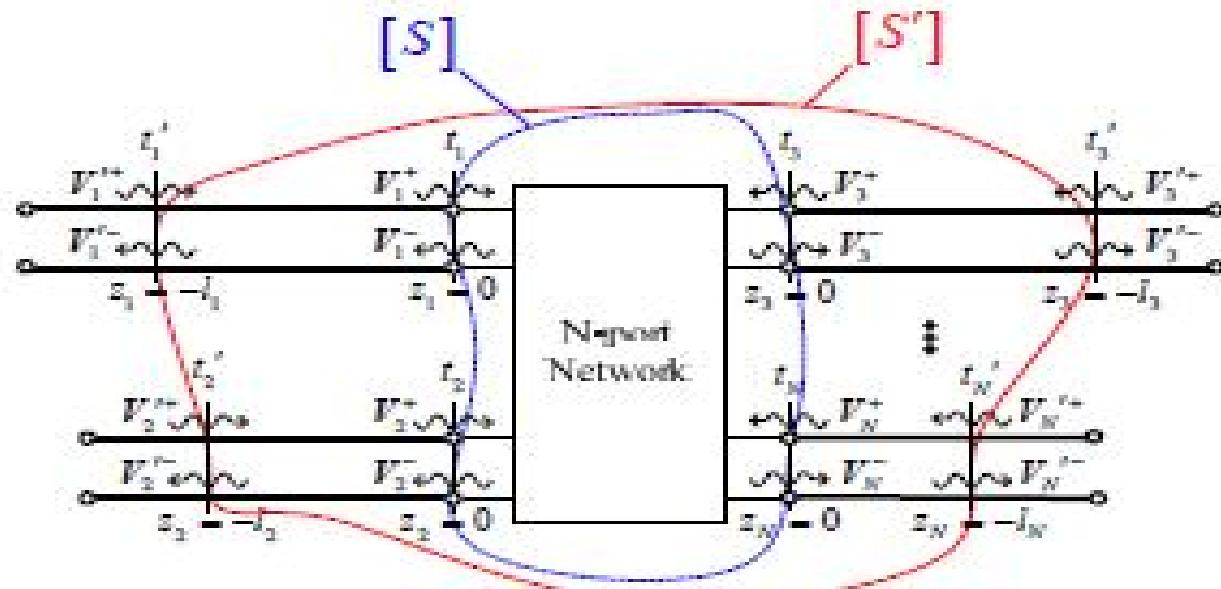
Recall that when we defined S parameters for a network, terminal planes were defined for all ports. These are arbitrarily chosen phase = 0° locations on TLs connected to the ports.

It turns out that S parameters change very simply and predictably as the reference planes are varied along lossless TLs. This fact can prove handy, especially in the lab.

To be specific, let $[S]$ be the scattering matrix of a network with reference planes (i.e., ports) at t_n , while $[S']$ is the scattering matrix of the network with the reference planes shifted to t'_n .

Applying the definition of the scattering matrix in these two situations yields

Properties of S-parameter



Many times you'll find that your measured S parameters differ from simulation by a phase angle, even though the magnitude is in good agreement. This likely occurred because your **terminal planes were defined differently** in your simulations than was set during measurement.

S matrix of a two port network

11. Two-port device with its S-matrix

$$S_{11} = \left. \frac{b_1}{a_1} \right|_{a_2=0} : \text{reflection coefficient at port 1 with port 2 matched}$$

$$S_{21} = \left. \frac{b_2}{a_1} \right|_{a_2=0} : \text{forward transmission coefficient with port 2 matched}$$

$$S_{12} = \left. \frac{b_1}{a_2} \right|_{a_1=0} : \text{reverse transmission coefficient with port 1 matched}$$

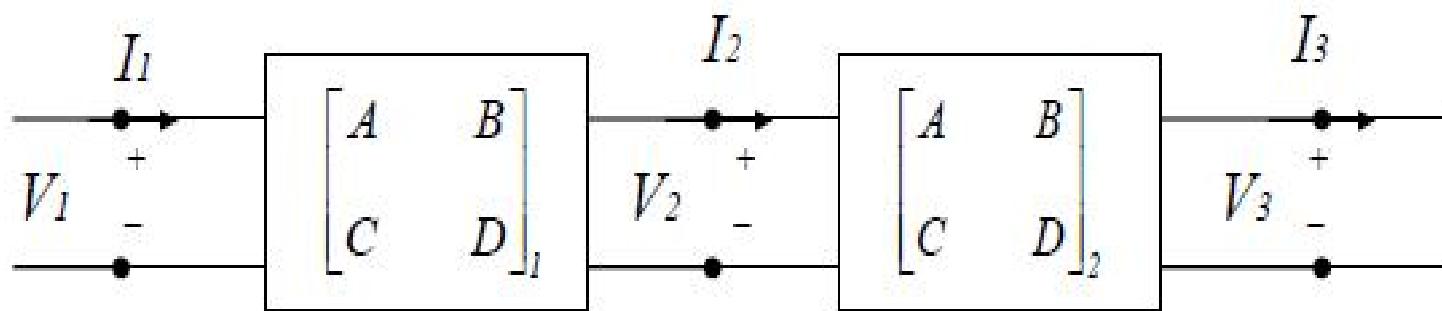
$$S_{22} = \left. \frac{b_2}{a_2} \right|_{a_1=0} : \text{reflection coefficient at port 2 with port 1 matched}$$

$$\begin{bmatrix} b_1 \\ b_2 \end{bmatrix} = \begin{bmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \end{bmatrix} \quad \begin{aligned} b_1 &= a_1 S_{11} + a_2 S_{12} \\ b_2 &= a_1 S_{21} + a_2 S_{22} \end{aligned}$$

ABCD MATRIX

4.4 The transmission (ABCD) matrix

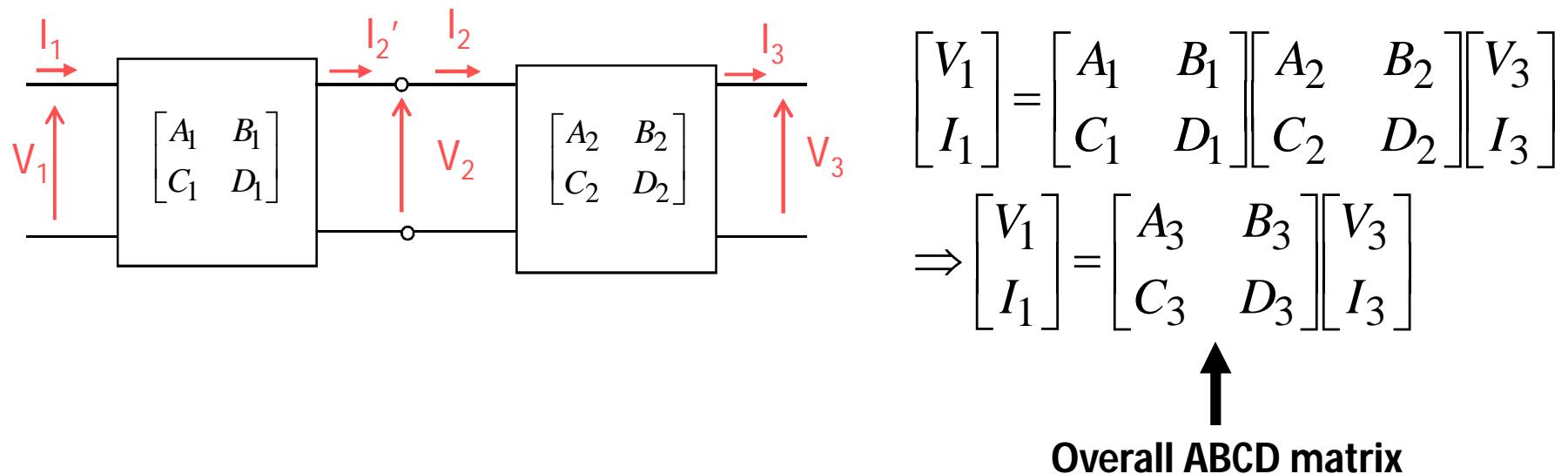
- Cascade network



$$\begin{bmatrix} V_1 \\ I_1 \end{bmatrix} = \begin{bmatrix} A & B \\ C & D \end{bmatrix}_1 \begin{bmatrix} V_2 \\ I_2 \end{bmatrix} = \begin{bmatrix} A & B \\ C & D \end{bmatrix}_1 \begin{bmatrix} A & B \\ C & D \end{bmatrix}_2 \begin{bmatrix} V_3 \\ I_3 \end{bmatrix}$$

ABCD MATRIX

- The ABCD matrix is useful for characterizing the overall response of 2-port networks that are cascaded to each other.



UNIT II MICROWAVE PASSIVE DEVICES

Coaxial cables-connectors and adapters –
Wave guides- Matched terminations –
Rectangular to circular wave guide transition–
Wave guide corners – Bends and twists –
Windows –Attenuators – Phase shifters – Wave
guide tees– E plane tee – H plane tee – Magic tee
– Isolators – Circulators –Directional couplers –
scattering matrix derivation for all components .

Microwave coaxial connectors

For high-frequency operation, the average circumference of a coaxial cable must be limited to about one wavelength in order to reduce multimodal propagation and eliminate erratic reflection coefficients, power losses, and signal distortion. Except for the sexless APC-7 connector, all other connectors are identified as either male (plugs) which have a center conductor that is a probe or female (jacks) which have a center conductor that is a receptacle. Sometimes it is hard to distinguish them as some female jacks may have a hollow center "pin" which appears to be male, yet accepts a smaller male contact. An adapter is an \approx zero loss interface between two connectors and is called a barrel when both connectors are identical. Twelve types of coaxial connectors are described below, however other special purpose connectors exist, including blind mate connectors where spring fingers are used in place of threads to obtain shielding (desired connector shielding should be at least 90 dB). Figure 1 shows the frequency range of several connectors and Figure 2 shows most of these connectors pictorially (\approx actual size).

Microwave coaxial connectors

1. **APC-2.4** (2.4mm) - The $50\ \Omega$ APC-2.4 (Amphenol Precision Connector-2.4 mm) is also known as an OS-50 connector. It was designed to operate at extremely high microwave frequencies (up to 50 GHz).
2. **APC-3.5** (3.5mm) - The APC-3.5 was originally developed by Hewlett-Packard (HP), but is now manufactured by Amphenol. The connector provides repeatable connections and has a very low VSWR. Either the male or female end of this $50\ \Omega$ connector can mate with the opposite type of SMA connector. The APC-3.5 connector can work at frequencies up to 34 GHz.
3. **APC-7** (7mm) - The APC-7 was also developed by HP, but has been improved and is now manufactured by Amphenol. The connector provides a coupling mechanism without male or female distinction and is the most repeatable connecting device used for very accurate $50\ \Omega$ measurement applications. Its VSWR is extremely low up to 18 GHz. Other companies have 7mm series available.
4. **BNC (OSB)** - The BNC (Bayonet Navy Connector) was originally designed for military system applications during World War II. The connector operates best at frequencies up to about 4 GHz; beyond that it tends to radiate electromagnetic energy. The BNC can accept flexible cables with diameters of up to 6.35 mm (0.25 in.) and characteristic impedance of 50 to $75\ \Omega$. It is now the most commonly used connector for frequencies under 1 GHz.

Microwave coaxial connectors

5. **SC (OSSC)** - The SC coaxial connector is a medium size, older type constant $50\ \Omega$ impedance. It is larger than the BNC, but about the same as Type N. It has a frequency range of 0-11 GHz.
6. **C** - The C is a bayonet (twist and lock) version of the SC connector.
7. **SMA (OSM/3mm)** - The SMA (Sub-Miniature A) connector was originally designed by Bendix Scintilla Corporation, but it has been manufactured by the Omni-Spectra division of M/A-COM (as the OSM connector) and many other electronic companies. It is a $50\ \Omega$ threaded connector. The main application of SMA connectors is on components for microwave systems. The connector normally has a frequency range to 18 GHz, but high performance varieties can be used to 26.5 GHz.
8. **SSMA (OSSM)** - The SSMA is a microminiature version of the SMA. It is also $50\ \Omega$ and operates to 26.5 GHz with flexible cable or 40 GHz with semi-rigid cable.
9. **SMC (OSMC)** - The SMC (Sub-Miniature C) is a $50\ \Omega$ or $75\ \Omega$ connector that is smaller than the SMA. The connector can accept flexible cables with diameters of up to 3.17 mm (0.125 in.) for a frequency range of up to 7-10 GHz.

Microwave coaxial connectors



APC 2.4 Jack - APC 3.5 Jack



SC Jack - Type N Jack



Type N Jack - TNC Jack



SMA Plug - TNC Plug



SSMA Jack - BNC Jack



Type N Plug - TNC Jack

Figure 2. . Microwave Coaxial Connectors (Connector Orientation Corresponds to Name Below It)

Microwave coaxial connectors



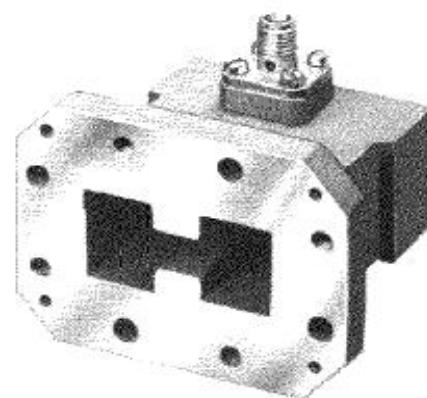
Standard
Waveguide - 7mm



SMC Plug - SMA Jack



7mm - 3.5mm Plug



Double ridge
Waveguide - SMA Jack

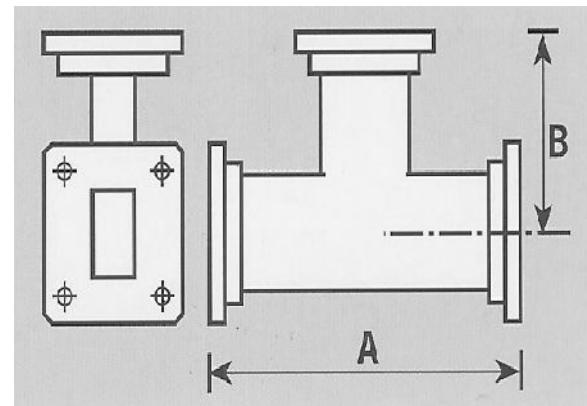
Figure 2. Microwave Coaxial connectors (Continued)

Waveguide tees:

- Waveguide junctions are used in microwave technologies when power in a waveguide needs to be split or some extracted.
- There are a number of different types of waveguide junction that can be used.
- Each type having different properties - the different types of waveguide junction affect the energy contained within the waveguide in different ways.
- When selecting a waveguide junction balances between performance and cost need to be made and therefore an understanding of the different types of waveguide junction is useful.

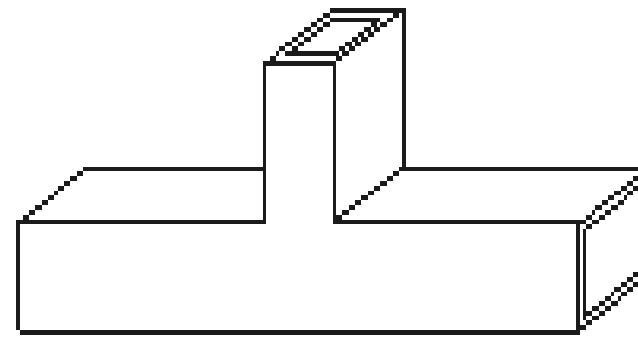
Types of Waveguide Tee Junctions:

- There are a number of different types of waveguide junction. The major types are listed below:
 1. H-type T Junction
 2. E-Type T Junction
 3. Magic T waveguide junction
 4. Hybrid Ring Waveguide Junction



E-Type Waveguide Junction

- It is called an E-type T junction because the junction arm, i.e. the top of the "T" extends from the main waveguide in the same direction as the E field.
- It is characterized by the fact that the outputs of this form of waveguide junction are 180° out of phase with each other.



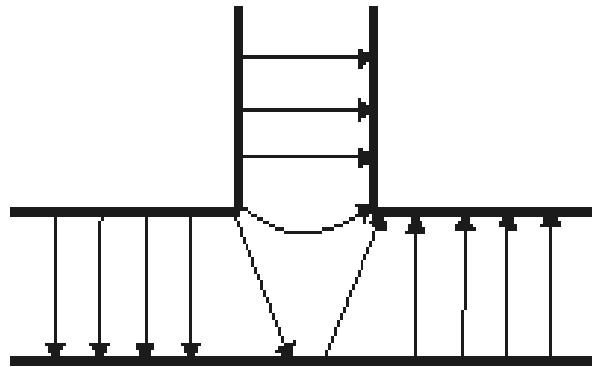
Waveguide E-type junction

E-Type Waveguide Junction:

- The basic construction of the waveguide junction shows the three port waveguide device.
- Although it may be assumed that the input is the single port and the two outputs are those on the top section of the "T", actually any port can be used as the input, the other two being outputs.
- WORKNG:
 - To see how the waveguide junction operates, and how the 180° phase shift occurs, it is necessary to look at the electric field. The magnetic field is omitted from the diagram for simplicity.

Working:

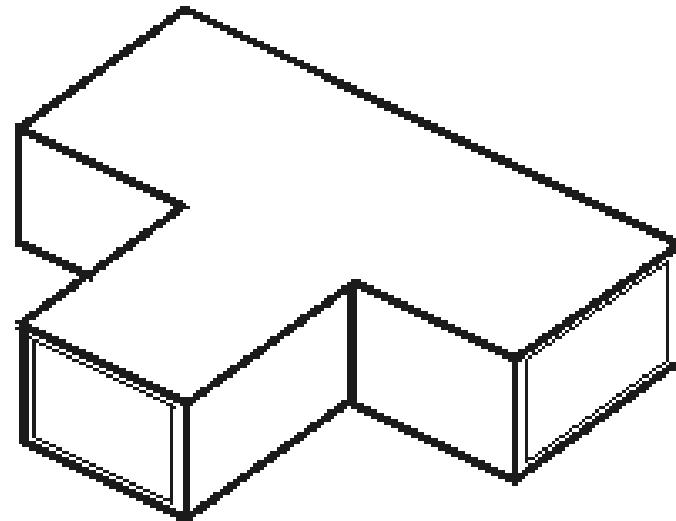
- It can be seen from the electric field that when it approaches the T junction itself, the electric field lines become distorted and bend.
- They split so that the "positive" end of the line remains with the top side of the right hand section in the diagram, but the "negative" end of the field lines remain with the top side of the left hand section. In this way the signals appearing at either section of the "T" are out of phase.
- These phase relationships are preserved if signals enter from either of the other ports.



Waveguide E-type junction E fields

H-type waveguide junction

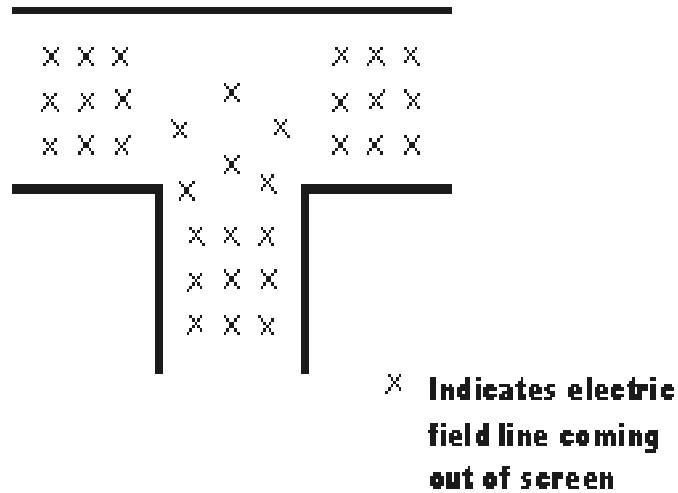
- This type of waveguide junction is called an H-type T junction because the long axis of the main top of the "T" arm is parallel to the plane of the magnetic lines of force in the waveguide.
- It is characterized by the fact that the two outputs from the top of the "T" section in the waveguide are in phase with each other.



Waveguide H-type junction

Working:

- To see how the waveguide junction operates, the diagram below shows the electric field lines.



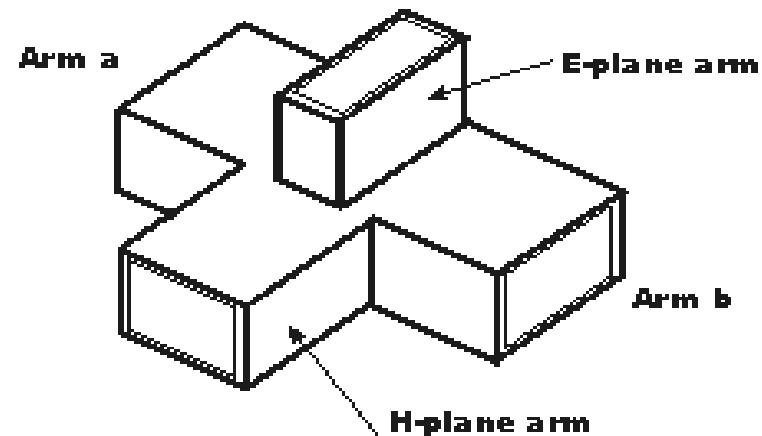
Waveguide H-type junction electric fields

Working

- The electric field lines are shown using the traditional notation - a cross indicates a line coming out of the screen, whereas a dot indicates an electric field line going into the screen.
- It can be seen from the diagram that the signals at all ports are in phase.
- Although it is easiest to consider signals entering from the lower section of the "T", any port can actually be used - the phase relationships are preserved whatever entry port is used.

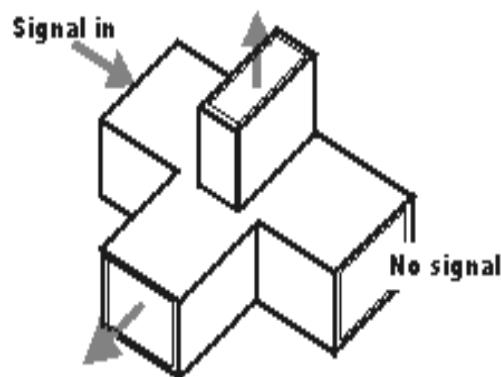
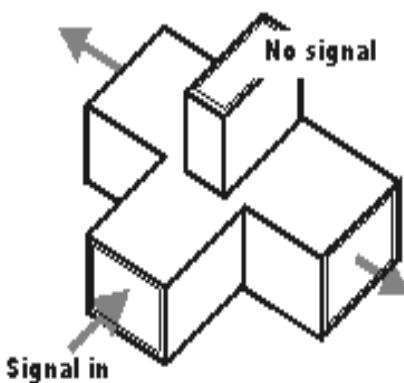
Magic T hybrid waveguide junction

- The magic-T is a combination of the H-type and E-type T junctions. The most common application of this type of junction is as the mixer section for microwave radar receivers.
- The diagram besides gives simplified version of the Magic T waveguide junction with its four ports.



Magic T waveguide junction

Working:



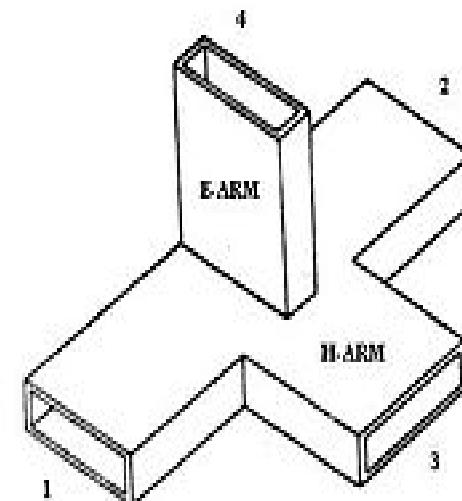
Working:

- **E-plane:**

- To look at the operation of the Magic T waveguide junction, take the example of when a signal is applied into the "E plane" arm.
- A signal injected into the E-plane port will also be divided equally between ports 1 and 2, but will be 180 degrees out of phase.

- **H-plane:**

- A signal injected into the H-plane port will be divided equally between ports 1 and 2, and will be in phase.



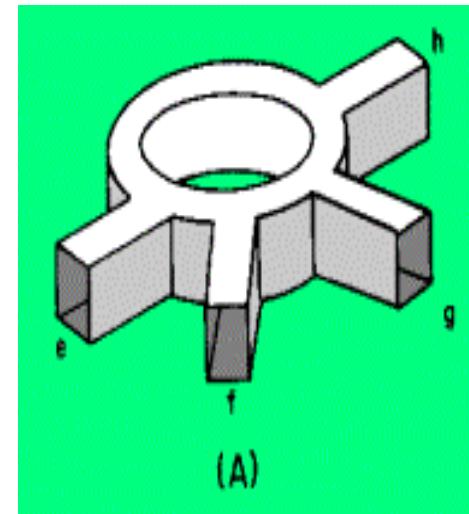
Disadvantage:

- One of the disadvantages of the Magic-T waveguide junction are that reflections arise from the impedance mismatches that naturally occur within it.
- These reflections not only give rise to power loss. The reflections can be reduced by using matching techniques.



Hybrid ring waveguide junction:

- This form of waveguide junction overcomes the power limitation of the magic-T waveguide junction.
- A hybrid ring waveguide junction is a further development of the magic T.
- It is constructed from a circular ring of rectangular waveguide.
- The ports are then joined to the holes at the required points. Again, if signal enters one port, it does not appear at all the others.



Practical Use:

- The hybrid ring is used primarily in high-power radar and communications systems where it acts as a duplexer - allowing the same antenna to be used for transmit and receive functions.
- During the transmit period, the hybrid ring waveguide junction couples microwave energy from the transmitter to the antenna while blocking energy from the receiver input.
- Then as the receive cycle starts, the hybrid ring waveguide junction couples energy from the antenna to the receiver.
- During this period it prevents energy from reaching the transmitter.

Rectangular to circular waveguide transition

FEATURES:

- ❖ Minimum VSWR
- ❖ Minimum Insertion Loss
- ❖ Optional Pressurized Models Available
- ❖ Efficient Conversion from TE₁₁ Mode
Rectangular Waveguide to TE₁₁ or TE₀₁ Mode
Circular Waveguide

APPLICATIONS:

- ❖ Radar Systems
- ❖ Test Setup



DESCRIPTION:

CRC series

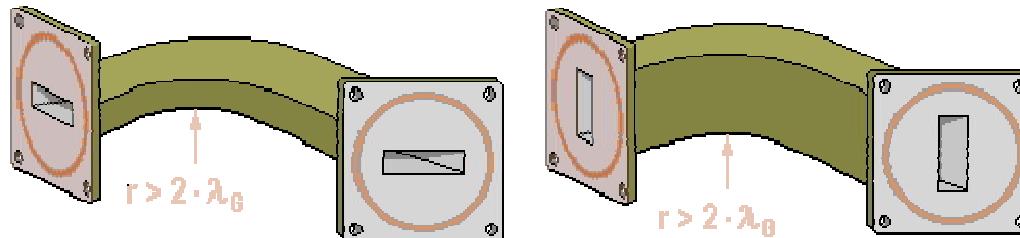
Cernex's CRC series TE₀₁ mode transitions are available for operation from 18.0 to 170.0 GHz. These reciprocal devices have a standard rectangular TE₁₁ mode waveguide input and a circular TE₁₁ or TE₀₁ mode output. Because of the different frequency ranges of circular TE₁₁ or TE₀₁ mode waveguide, it is possible for a standard sized rectangular waveguide input to have one of several different circular waveguide size outputs. The CRC series circular mode waveguide features low VSWR and insertion loss. For maximum mode purity, filtering is recommended for all TE₀₁ propagation (please refer to Table 1-1).

Waveguides Bends and Twists

- The size, shape, and dielectric material of a waveguide must be constant throughout its length for energy to move from one end to the other without reflections. Any abrupt change in its size or shape can cause reflections and a loss in overall efficiency. When such a change is necessary, the bends, twists, and joints of the waveguides must meet certain conditions to prevent reflections.
- **Bends**
- Waveguides may be bent in several ways that do not cause reflections. One way is the gradual bend shown in the right part of the following figure. This gradual bend is known as an E bend because it distorts the E fields. The E bend must have a radius greater than two wavelengths to prevent reflections.

Waveguides Bends

- H - bend
- E - bend
- *Figure 1: Waveguide bends*
- Another common bend is the gradual H bend shown in the leftt part of the figure. It is called an H bend because the H fields are distorted when a waveguide is bent in this manner. Again, the radius of the bend must be greater than two wavelengths to prevent reflections.
- A sharp bend in either dimension may be used if it meets certain requirements. Notice the two 45-degree bends in figure; the bends are $1/4\cdot\lambda$ apart. The reflections that occur at the 45-degree bends cancel each other, leaving the fields as though no reflections have occurred.



Waveguide **Twists**

- Sometimes the electromagnetic fields must be rotated so that they are in the proper phase to match the phase of the load. This may be accomplished by twisting the waveguide as shown in the figure. The twist must be gradual and greater than two wavelengths ($2\cdot\lambda$).



- *Figure 3: Waveguide twist*
- The flexible waveguide allows special bends which some equipment applications might require. It consists of a specially wound ribbon of conductive material, most commonly brass, with the inner surface plated with chromium. Power losses are greater in the flexible waveguide because the inner surfaces are not perfectly smooth. Therefore, it is only used in short sections where no other reasonable solution is available.

Phase shifters

Phase Shifters are devices, in which the phase of an electromagnetic wave of a given frequency can be shifted when propagating through a transmission line.

In many fields of electronics, it is often necessary to change the phase of signals.

RF and microwave Phase Shifters have many applications in various equipments such as phase discriminators, beam forming networks, power dividers, linearization of power amplifiers, and phase array antennas.

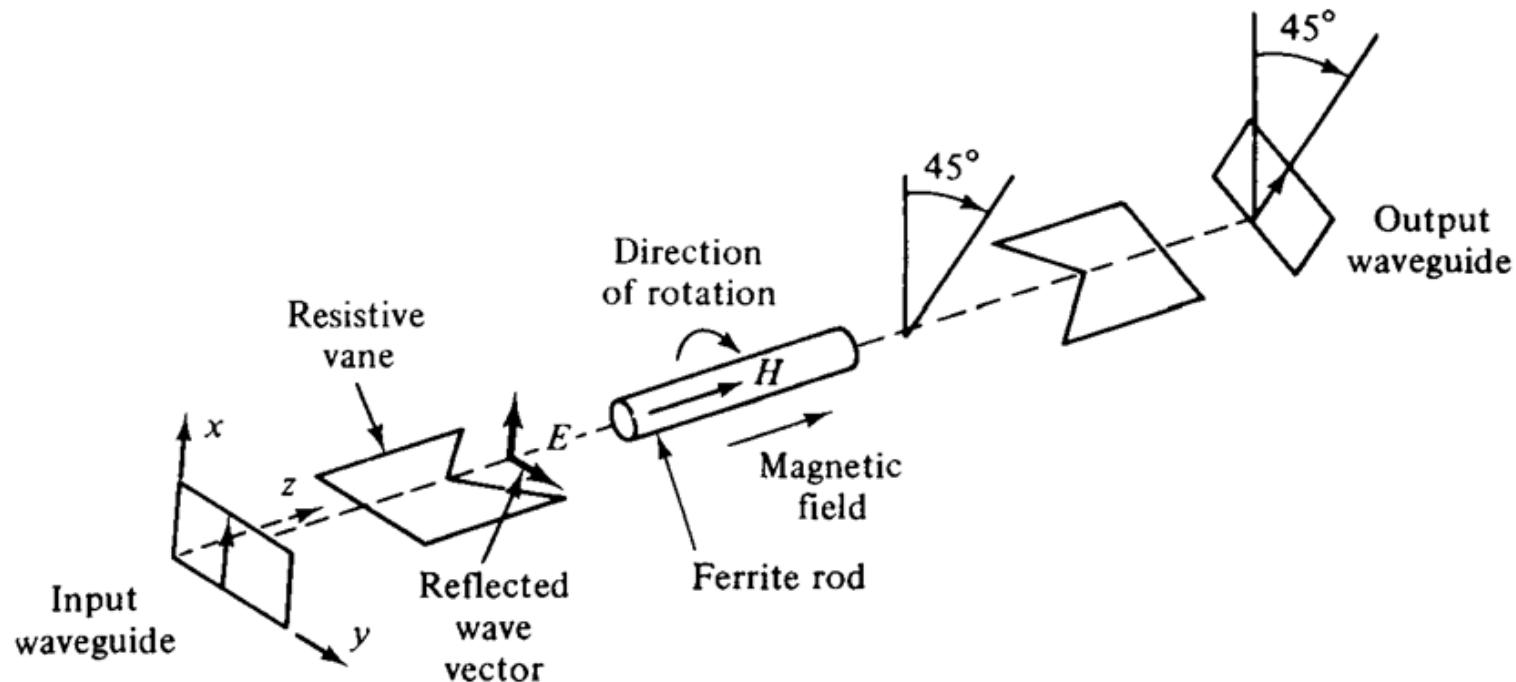
The major parameters which define the RF and microwave Phase Shifters are:

- frequency range,
- bandwidth (BW),
- total phase variance ($\Delta\varphi$),
- insertion loss (IL),
- switching speed,
- power handling (P),
- accuracy and resolution,
- input/output matching ($VSWR$) or return loss (RL),
- harmonics level.

Isolator

- An isolator is a non reciprocal transmission device that is used to isolate one component from reflections of the other in a transmission line.
- An ideal isolator completely absorbs the power from propagation in one direction and provide loss less transmission in opposite direction
- It is also known as UNILINE
- It is used to improve the frequency stability.
- One type of isolator is Faraday rotation Isolator, the input resistive card is in y-z plane, the output resistive card is displaced 45° with respect to the input card.
- The magnetic field which is applied longitudinally to the ferrite rod rotates the wave plane by 45° .
- This is normal to the output resistive card
- As the result of rotation the wave arrives at the out put end without attenuation at all.
- On the other end a reflected wave from the output end is similarly rotated clockwise 45° by the ferrite rod, since the reflected wave is parallel to the input resistive card the wave is absorbed by the input card.

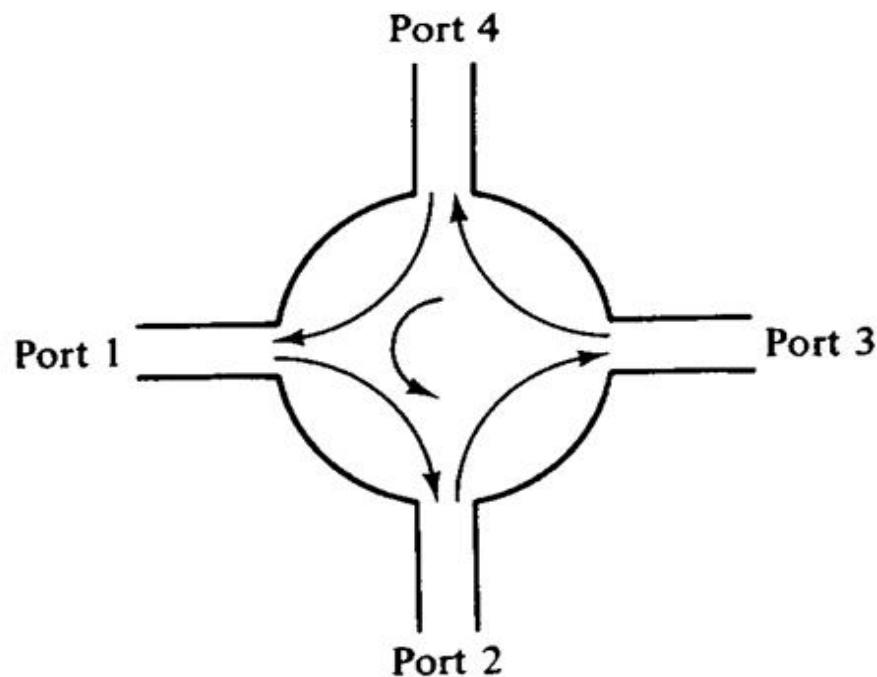
Isolator



Isolator

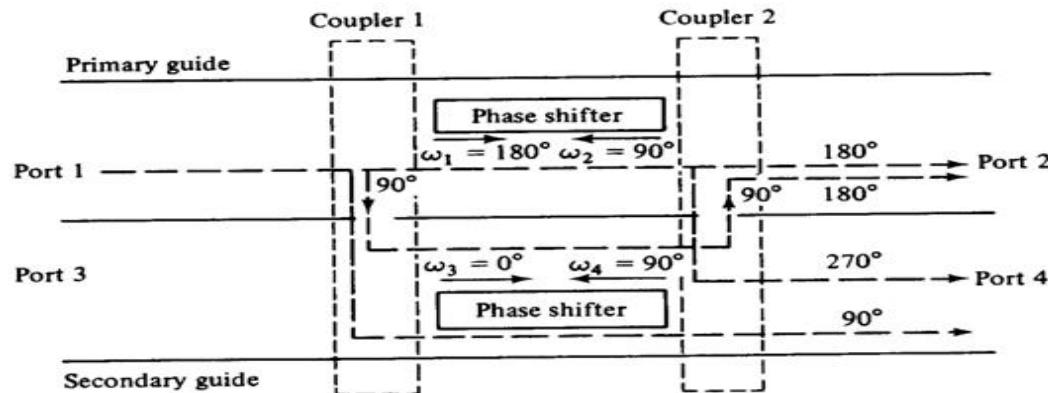
circulator

- A circulator is a multiport wave guide junction in which wave can flow only from nth port to n+1th port in one direction.
- There is no restriction on number of ports.



Circulator

- The operating principle of a microwave circulator can be analyzed with the help of figure below.



- Each of the two 3dB couplers in circulator introduces a phase shift of 90° and each of the two phase shifters produce a certain phase change, the wave is split in to two components by the coupler 1.
- The wave in primary guide arrives at port 2 with a relative phase change of 180° . The second wave propagates through the two couplers & secondary guide arrives at port 2 with relative phase shift of 180° , since the two waves reaching port 2 are in phase , the power transmitted is obtained from port 1 to port 2.
- The waves propagating through primary guide , phase shifter, & coupler 2 arrives at port 4 with a 270° phase change.

Circulator

- The wave travelling through coupler I & secondary guide arrives at port 4 with a phase shift of 90° .
- Since the two waves reaching port 4 are opposite in phase the power transmission from 1-4 is zero.
- A perfectly matched lossless nonreciprocal four port circulator has an S matrix of the form.

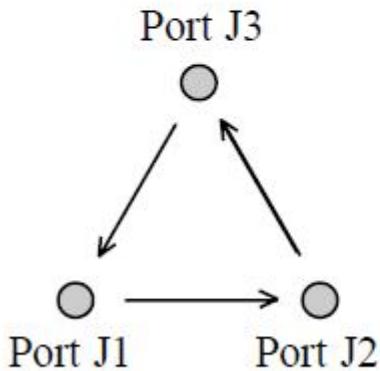
$$\mathbf{S} = \begin{bmatrix} 0 & S_{12} & S_{13} & S_{14} \\ S_{21} & 0 & S_{23} & S_{24} \\ S_{31} & S_{32} & 0 & S_{34} \\ S_{41} & S_{42} & S_{43} & 0 \end{bmatrix}$$

- Using the parameters of S parameters the above matrix is simplifies as

$$\mathbf{S} = \begin{bmatrix} 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix}$$

Isolators and circulators

INTRODUCTION: Isolators and Circulators are usually three port devices, and they are used to force the microwave energy into one direction only. The typical junction Circulator consists of a stripline circuit, sandwiched between two ferrite discs or triangles, an upper and a lower ground plane, magnetically biased by permanent magnets located outside the ground planes. In a Circulator, the magnetic field, applied through the vertical axis of the assembly, results into a circulation of the microwave energy from one port to the other, depending on where the energy is coming from.



Microwave energy entering the device from port J1 is directed to port J2. Energy entering from port J2, is directed to port J3. Signals entering from port J3, are directed to port J1, etc. If one of the ports is terminated into a 50 Ohms load, the device becomes an Isolator. Signals then only can pass the unit with low loss in one direction, and only with high loss in the reverse direction. If e.g. port J3 is terminated into a 50 Ohms line, microwave energy only can pass the device with low loss from port J1 to port J2. An Isolator is used to "isolate" microwave components from each other, or to protect units from receiving damages when working into an open or short circuit. The output of an oscillator is usually protected by an isolator.

Isolators and circulators

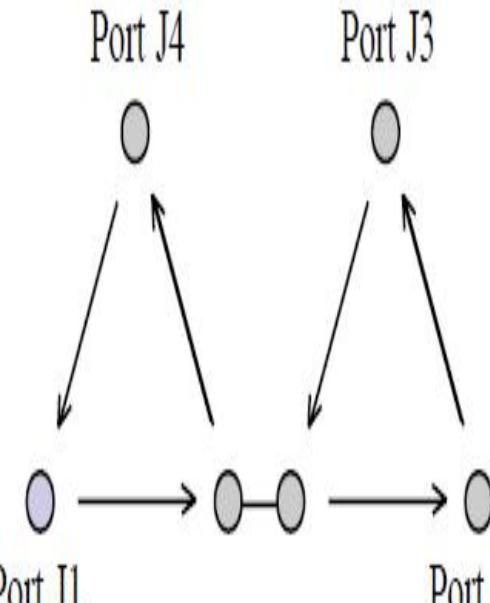
Frequency and Bandwidth: Coaxial and microstrip circulators and isolators operate either in the bias region above resonance or below resonance. Above-resonance circuits are usually used for smaller bandwidths and higher power designs, while below-resonance circuits achieve wider bandwidths. Theoretically, the above-resonance circuits have no lower frequency limit.

Operating Temperature: The performance depends on the magnetic field, applied to saturate the ferrite material. Temperature compensated magnets and ferrites need to be used where wide temperature ranges are required. Internal heaters can be installed, where temperature range and ferrite material do not allow other compensation.

Input VSWR: The input VSWR is a function of the VSWR of the other ports. At an isolator the higher output VSWR will cause reflected energy towards the terminated port, where it will be attenuated by the value of the isolation, and the balance is reflected back to the input, increasing the input VSWR.

Isolators and circulators

Four Port Devices: Four Port Circulators and Isolators are used where higher directivity is needed. An Isolator would have the ports J3 and J4 terminated. In the schematic to the left, microwave energy is forced from port J1 to J2, or from port J3 to port J4, when crossing two ferrite junctions. The high isolation only applies when two ferrite junctions have been crossed, here between ports J2 and J1 with ports J3 and J4 terminated with matched loads.

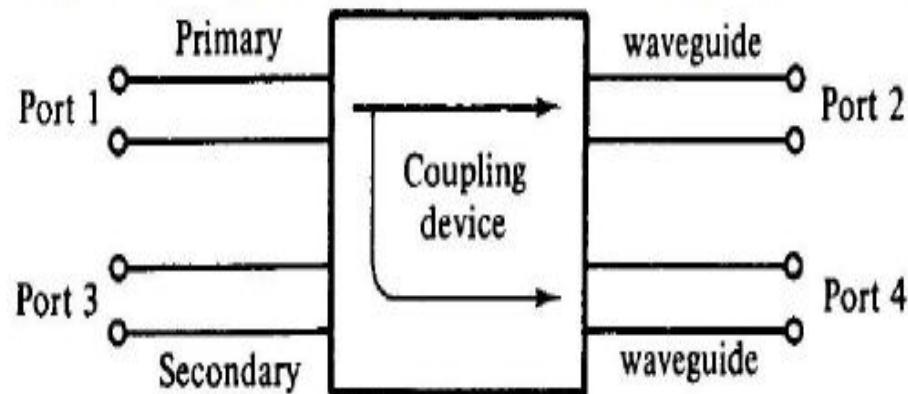


The schematic diagram illustrates a four-port device. It shows four circular ports labeled Port J1, Port J2, Port J3, and Port J4. Port J1 is at the bottom left, Port J2 is at the bottom right, Port J3 is at the top right, and Port J4 is at the top left. Arrows indicate the flow of microwave energy: one arrow points from Port J1 to Port J2, and another arrow points from Port J3 to Port J4. Both of these arrows pass through a series of two small circles representing ferrite junctions. The path from Port J1 to Port J2 passes through the first junction, then the second junction, and finally reaches Port J2. The path from Port J3 to Port J4 also passes through both junctions and reaches Port J4. The other two ports, J1 and J3, are shown with a single circle, indicating they are terminated with matched loads.



Directional Coupler

A directional coupler is a four-port waveguide junction as shown below. It consists of a primary waveguide 1-2 and a secondary waveguide 3-4. When all ports are terminated in their characteristic impedances, there is free transmission of the waves without reflection, between port 1 and port 2, and there is no transmission of power between port 1 and port 3 or between port 2 and port 4 because no coupling exists between these two pairs of ports. The degree of coupling between port 1 and port 4 and between port 2 and port 3 depends on the structure of the coupler. The characteristics of a directional coupler can be expressed in terms of its Coupling factor and its directivity. Assuming that the wave is propagating from port to port 2 in the primary line, the coupling factor and the directivity are defined,



Directional Coupler

Where PI = power input to port I

P3 = power output from port 3

P4 = power output from port 4

$$\text{Coupling factor (dB)} = 10 \log_{10} \frac{P_1}{P_4}$$

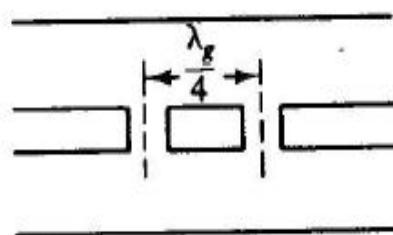
$$\text{Directivity (dB)} = 10 \log_{10} \frac{P_4}{P_3}$$

It should be noted that port 2, port 3, and port 4 are terminated in their characteristic impedances. The coupling factor is a measure of the ratio of power levels in the primary and secondary lines. Hence if the coupling factor is known, a fraction of power measured at port 4 may be used to determine the power input at port 1.

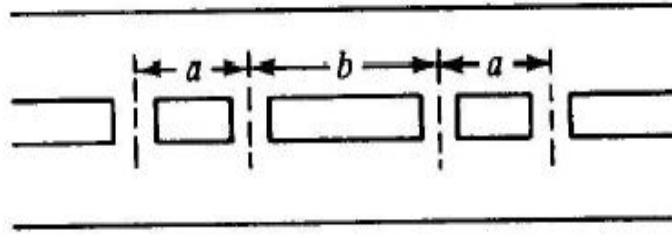
This significance is desirable for microwave power measurements because no disturbance, which may be caused by the power measurements, occurs in the primary line. The directivity is a measure of how well the forward traveling wave in the primary waveguide couples only to a specific port of the secondary waveguide ideal directional coupler should have infinite directivity. In other words, the power at port 3 must be zero because port 2 and port A are perfectly matched. Actually well-designed directional couplers have a directivity of only 30 to 35 dB.

Directional Coupler

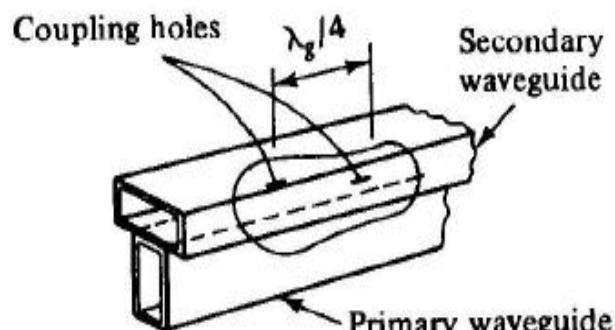
Several types of directional couplers exist, such as a two-hole direct coupler, four-hole directional coupler, reverse-coupling directional coupler , and Bethe-hole directional coupler the very commonly used two-hole directional coupler is described here.



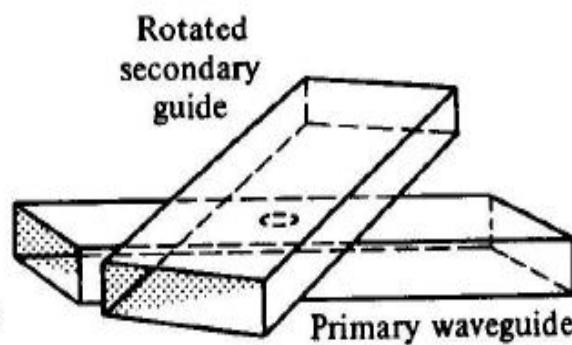
(a)



(b)



(c)

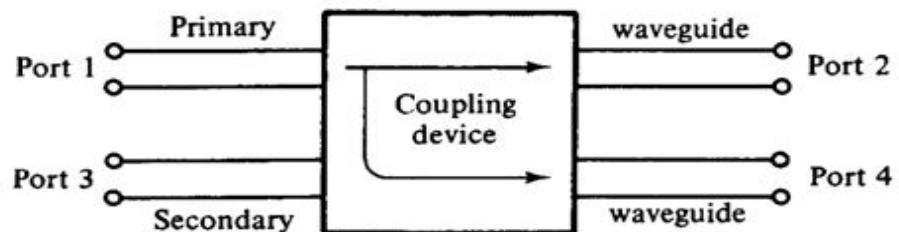


(d)

Directional Coupler

➤ A Directional coupler is a four port wave guide junction as shown in figure.

➤ The primary waveguide is 1-2
➤ Secondary waveguide is 3-4



➤ When all ports are terminated there is a free transmission of power without reflection between ports 1 & 2.

➤ There is no transmission between 1 - 3 & 2 – 4 because of no coupling.

➤ The characteristics of directional coupler can be expressed in terms of C

$$\text{Coupling factor (dB)} = 10 \log_{10} \frac{P_1}{P_4}$$

where P_1 = power input to port 1

$$\text{Directivity (dB)} = 10 \log_{10} \frac{P_4}{P_3}$$

P_3 = power output from port 3

P_4 = power output from port 4

Directional Coupler

- A two hole directional coupler with travelling wave propagation in it is illustrated in the figure given.
- The spacing between the centre of two holes should be

$$L = (2n + 1) \frac{\lambda_g}{4}$$

- In directional Coupler all four ports are completely matched. So

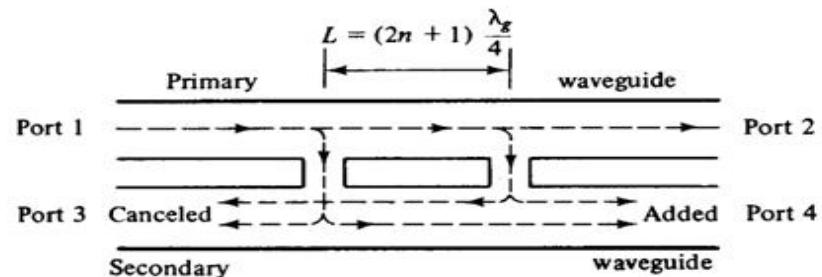
$$S_{11} = S_{22} = S_{33} = S_{44} = 0$$

- There is no coupling between port 2 & 4, thus

$$S_{13} = S_{31} = S_{24} = S_{42} = 0$$

- Consequently, the S matrix of Directional Coupler is

$$\mathbf{S} = \begin{bmatrix} 0 & S_{12} & 0 & S_{14} \\ S_{21} & 0 & S_{23} & 0 \\ 0 & S_{32} & 0 & S_{34} \\ S_{41} & 0 & S_{43} & 0 \end{bmatrix}$$



This equation can be reduced using zero property

$$S_{12} S_{14}^* + S_{32} S_{34}^* = 0$$

$$S_{21} S_{23}^* + S_{41} S_{43}^* = 0$$

Unitary Property

$$S_{12} S_{12}^* + S_{14} S_{14}^* = 1$$

We have

$$|S_{12}| |S_{14}| = |S_{32}| |S_{34}|$$

$$|S_{21}| |S_{23}| = |S_{41}| |S_{43}|$$

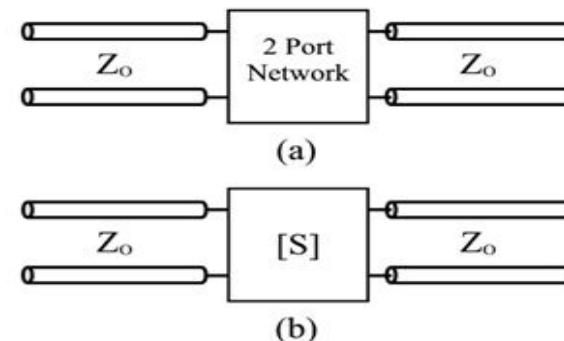
scattering matrix derivation for all components .

Scattering Parameters

Consider a circuit or device inserted into a T-Line as shown in the Figure. We can refer to this circuit or device as a two-port network.

The behavior of the network can be completely characterized by its scattering parameters (S-parameters), or its scattering matrix, [S].

Scattering matrices are frequently used to characterize multiport networks, especially at high frequencies. They are used to represent microwave devices, such as amplifiers and circulators, and are easily related to concepts of gain, loss and reflection.



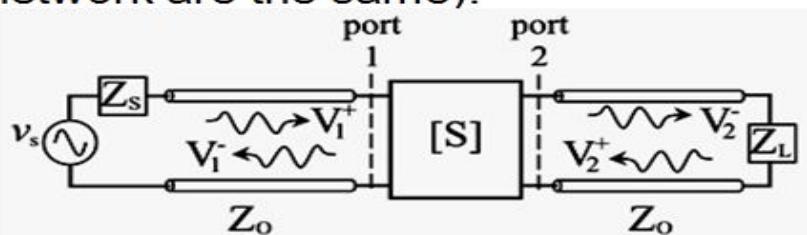
Scattering matrix

$$[S] = \begin{bmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{bmatrix}$$

scattering matrix derivation for all components .

Scattering Parameters (S-Parameters)

The scattering parameters represent ratios of voltage waves entering and leaving the ports (If the same characteristic impedance, Z_0 , at all ports in the network are the same).

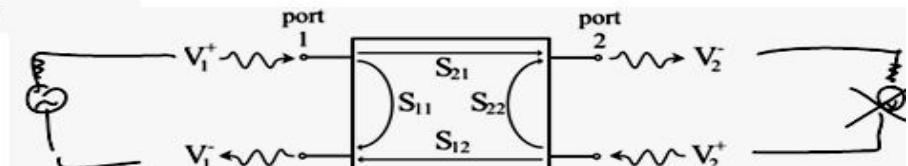


$$V_1^- = S_{11}V_1^+ + S_{12}V_2^+.$$

$$V_2^- = S_{21}V_1^+ + S_{22}V_2^+.$$

In matrix form this is written

$$\begin{bmatrix} V_1^- \\ V_2^- \end{bmatrix} = \begin{bmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{bmatrix} \begin{bmatrix} V_1^+ \\ V_2^+ \end{bmatrix}, [V]^- = [S][V]^+$$



Where,

$$S_{11} = \left. \frac{V_1^-}{V_1^+} \right|_{V_2^+=0}$$

Reflection
Coefficient
at port 1

$$S_{12} = \left. \frac{V_1^-}{V_2^+} \right|_{V_1^+=0}$$

Transmission
coefficient from
port 2 to port 1

$$S_{21} = \left. \frac{V_2^-}{V_1^+} \right|_{V_2^+=0}$$

Transmission
coefficient from
port 1 to port 2

$$S_{22} = \left. \frac{V_2^-}{V_2^+} \right|_{V_1^+=0}$$

Reflection
Coefficient
at port 2

UNIT III MICROWAVE VACCUM TUBE DEVICES

Introduction – Two cavity klystron amplifier – Mechanism and mode of operation –Power output and efficiency -Applications – Reflex klystron oscillator – Mechanism and mode of operation-Power output – Efficiency – Mode curve –Applications – TWT amplifier – Principle of operation-gain and applications – Magnetron oscillator – Hull cut-off voltage mechanism of operation– Power output and efficiency – Applications – Numerical problems.

WHY MICROWAVE VACUUM TUBES?

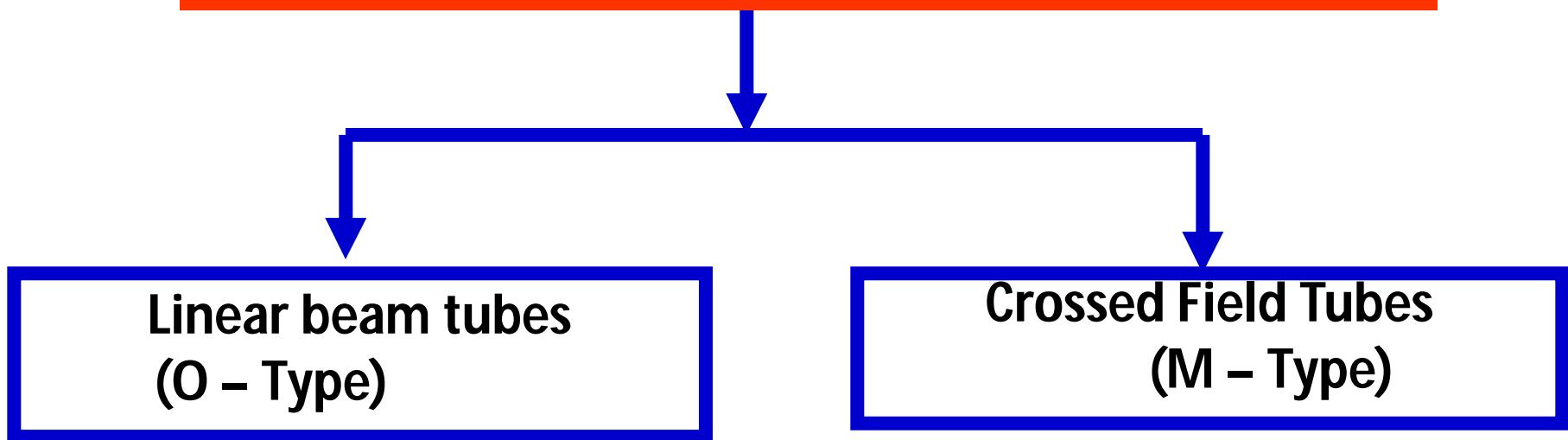
- Because of

- The electron transit from the cathode to the grid become comparable to time period of the sinusoidal signal.
- Appearance of stray reactance's due to the lead wire inductances and the inter-electrode capacitances.

MICROWAVE OSCILLATIONS OR AMPLIFICATION

- The principles uses an electron beam on which space-charge waves
 - interact with EM fields in the microwave cavities to transfer energy to the output circuit of the cavity or
 - interact with EM fields in a slow-wave structure to give amplification through transfer of energy.

Types of Microwave Tubes



Eg:

Klystron

Reflex klystron

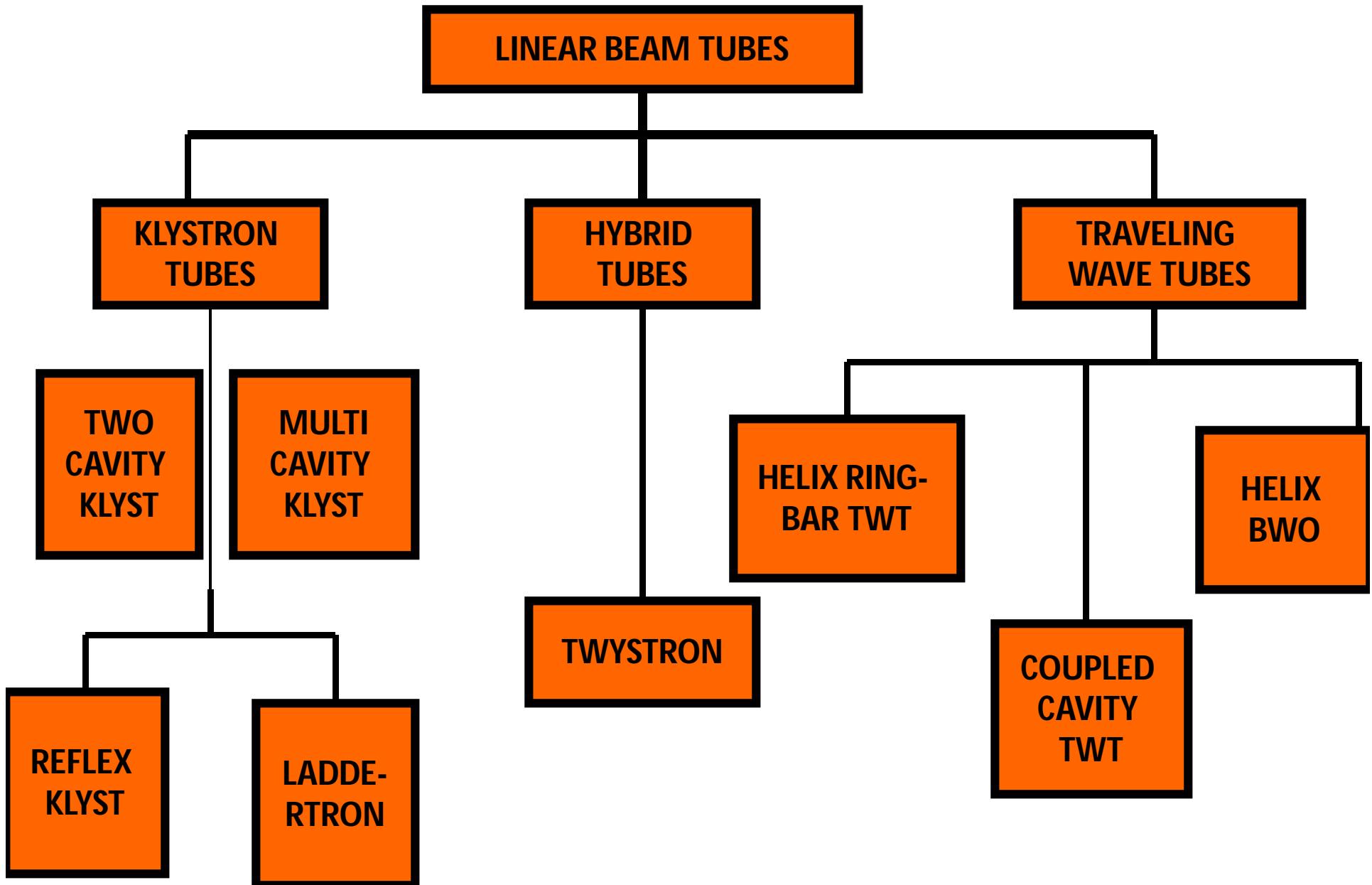
TWT

Eg:

Magnetron

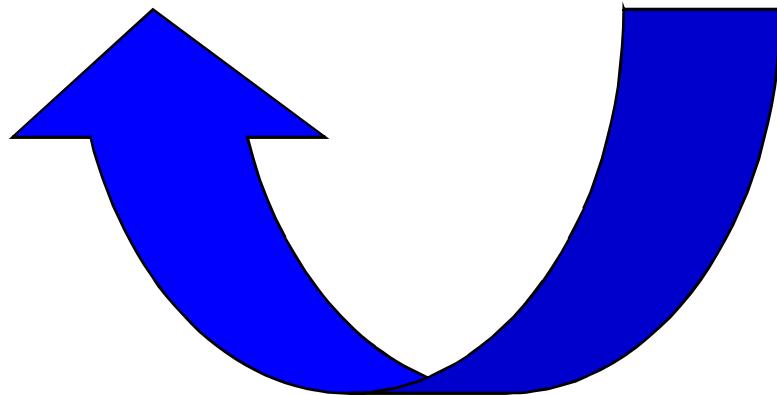
Linear beam devices	Crossed field devices
(I) Straight path taken by the electron beam	A principle feature of such tubes is that electrons travel in a curved path
(i) DC magnetic field is in parallel with DC electric field to focus the electron beam	DC magnetic field is perpendicular to DC electric field

Types of Linear Beam Tubes



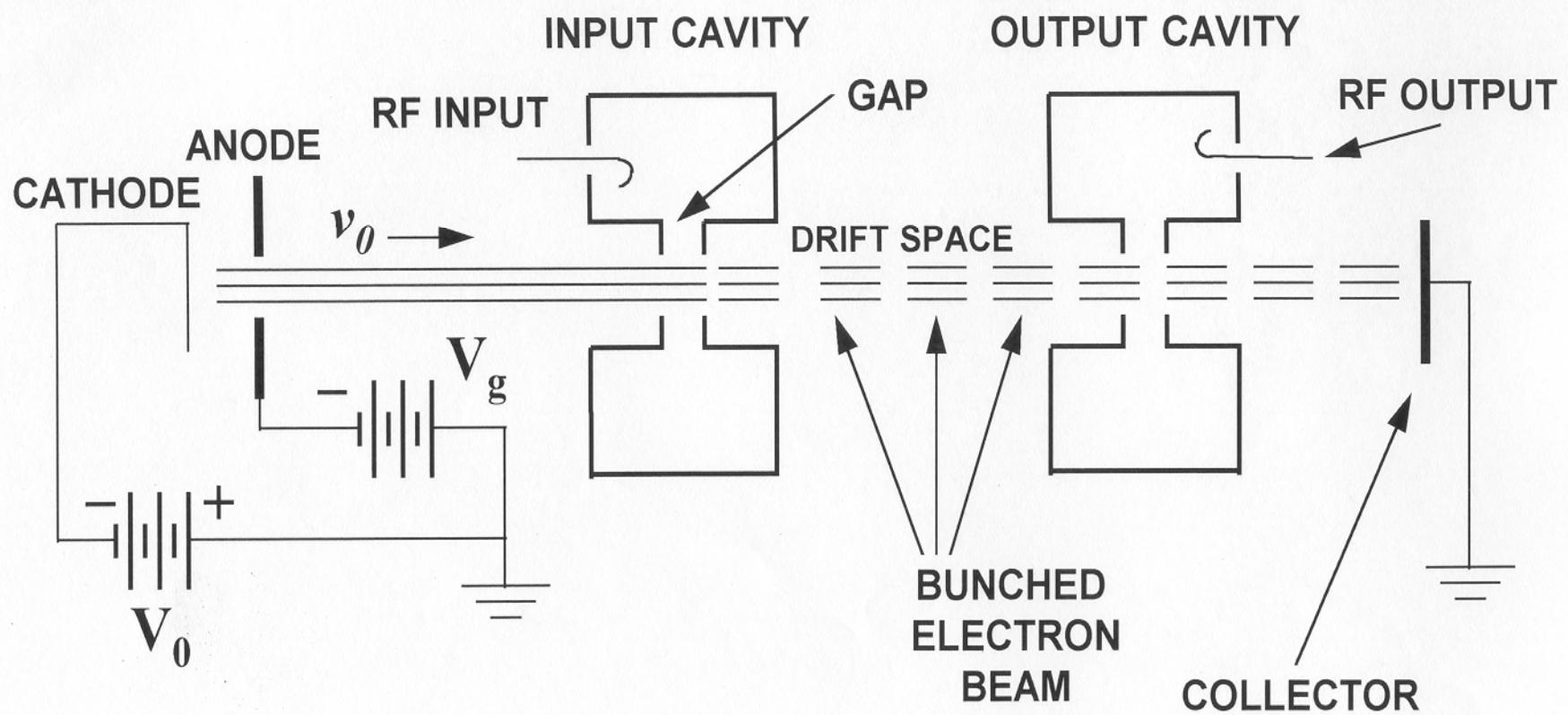
TWYSTRON

- KLYSTRON + TWT = TWYSTRON



- It is hybrid amplifier that uses the combinations of klystron and TWT components

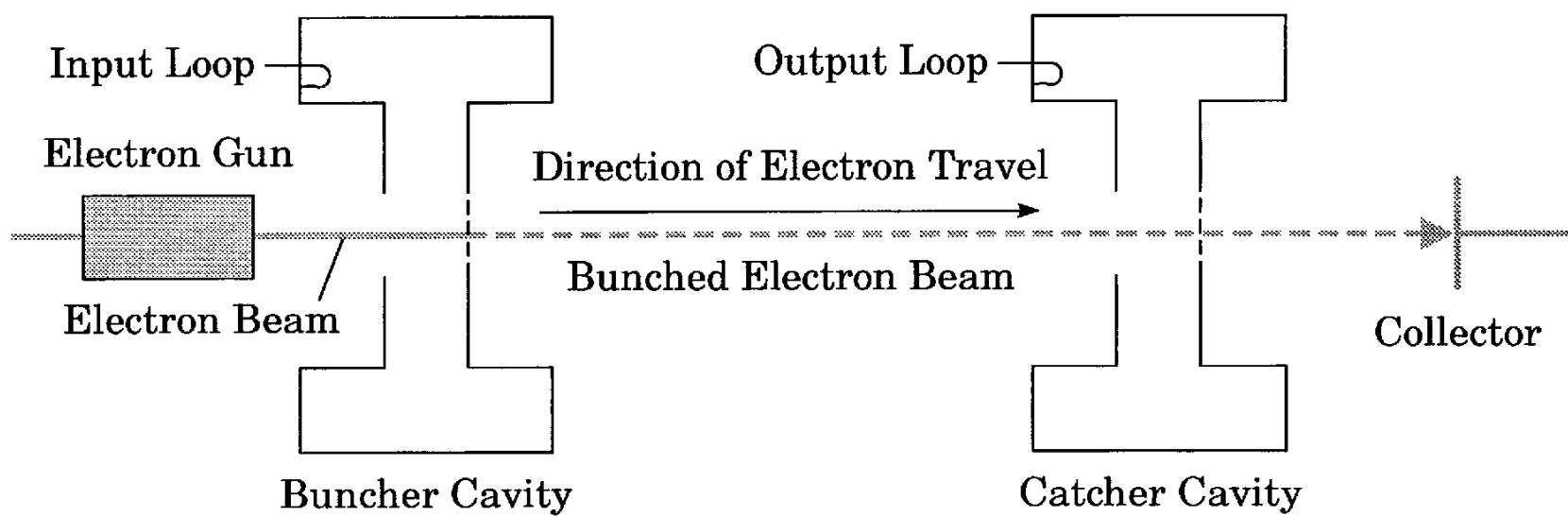
KLYSTRON STRUCTURE



Klystron

- Used in high-power amplifiers
- Electron beam moves down tube past several cavities.
- Input cavity is the *buncher*, output cavity is the *catcher*.
- *Buncher* modulates the velocity of the electron beam

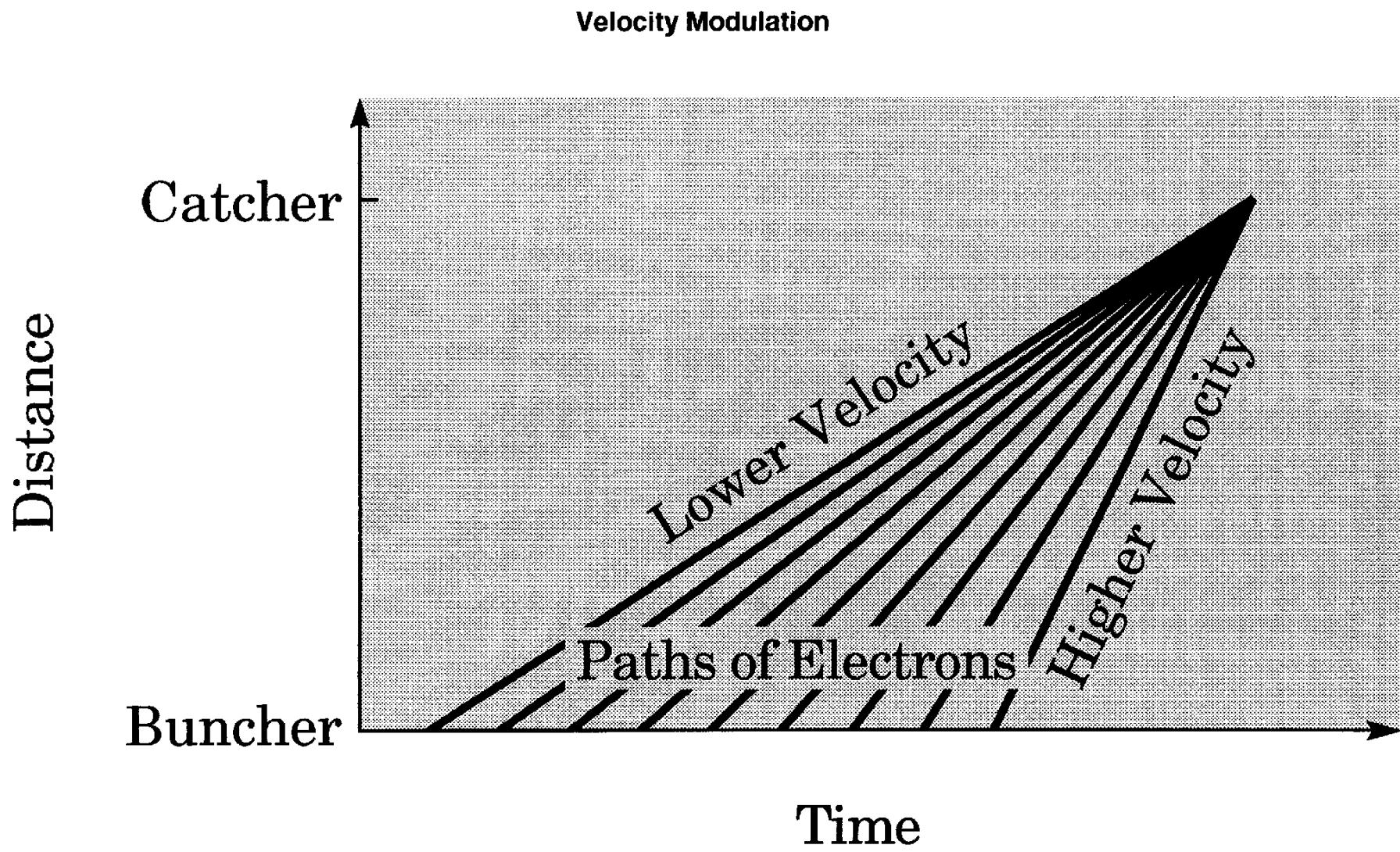
Klystron Cross Section



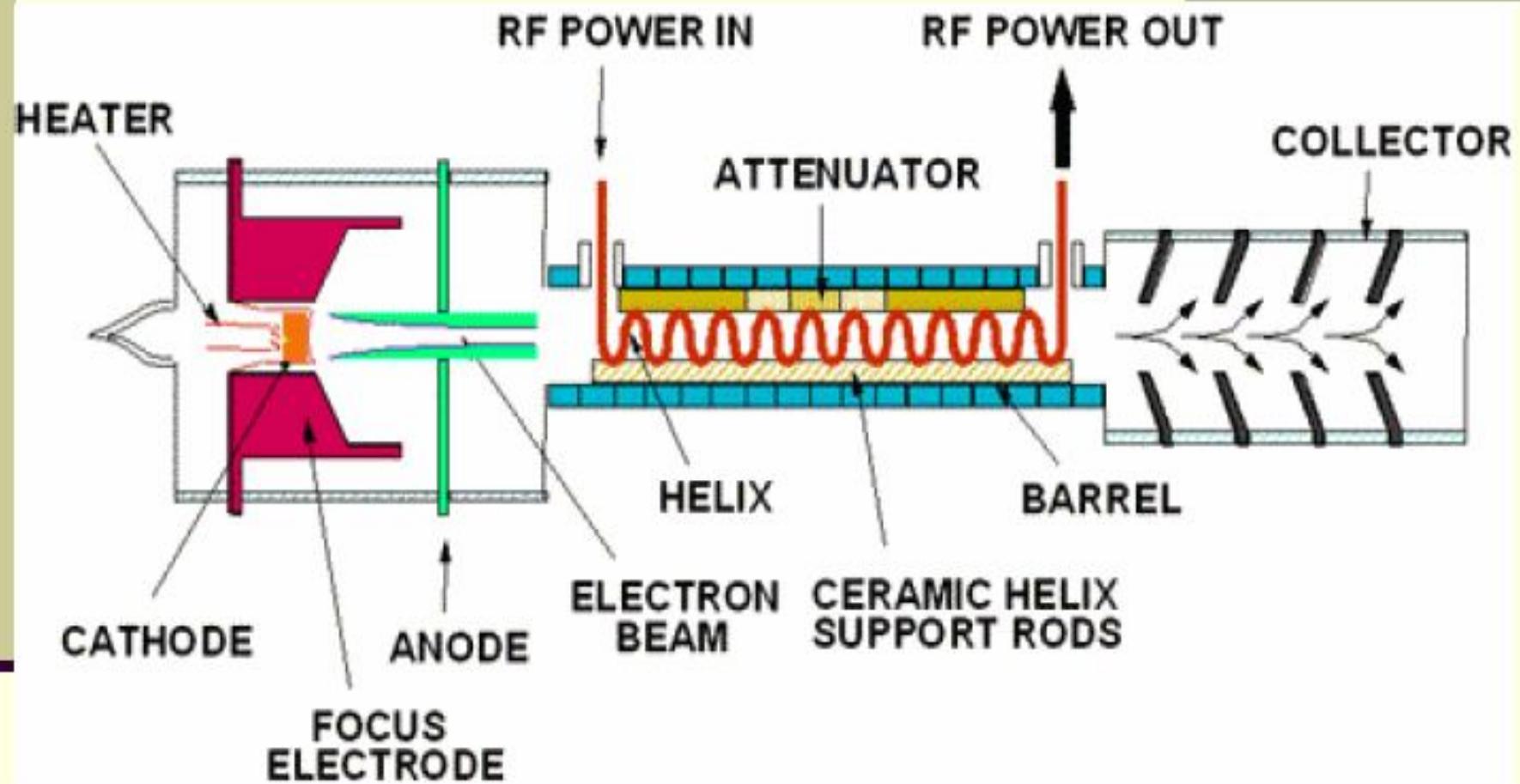
PRINCIPLE

Velocity Modulation

- Electric field from microwaves at buncher alternately speeds up and slows electron beam .
- This causes electrons to bunch up Electron bunches at catcher induce microwaves with more energy.
- The cavities form a slow-wave structure



BASICS of Traveling Wave Tube (TWT) Amplifier

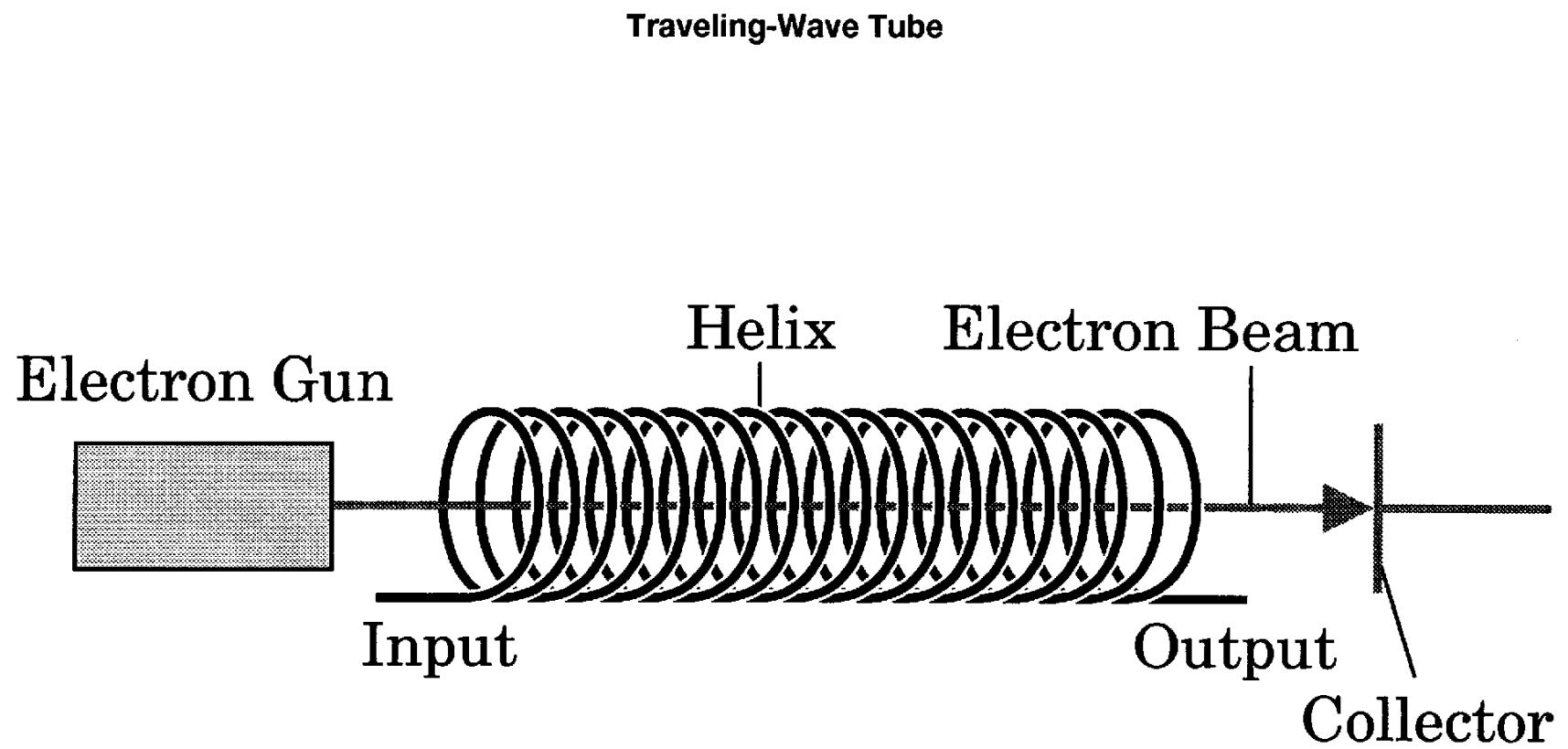


Traveling-Wave Tube (TWT)

- Uses a helix as a slow-wave structure
- Microwaves input at cathode end of helix, output at anode end
- Energy is transferred from electron beam to microwaves

Traveling-Wave Tube (TWT)

- Heater/Filament is closest to Cathode Voltage.
- Heater and Cathode act as electron gun, and they are on the side RF Input.
- Collectors sits on RF output.
- Electrons are fired from Cathode and received from Collectors.
- RF signal is amplified through bunching effect after traveling along the path of Helix coil.*
- Higher Cathode voltage → Higher RF Power *
- Advantage of TWTA (over solid state amplification) is the linearity and output power*
- TWTA Efficiency: 50% to 60% vs. Solid State: 25% to 30%
- Ranges of Frequency for TWTA: 1Ghz – 40 Ghz



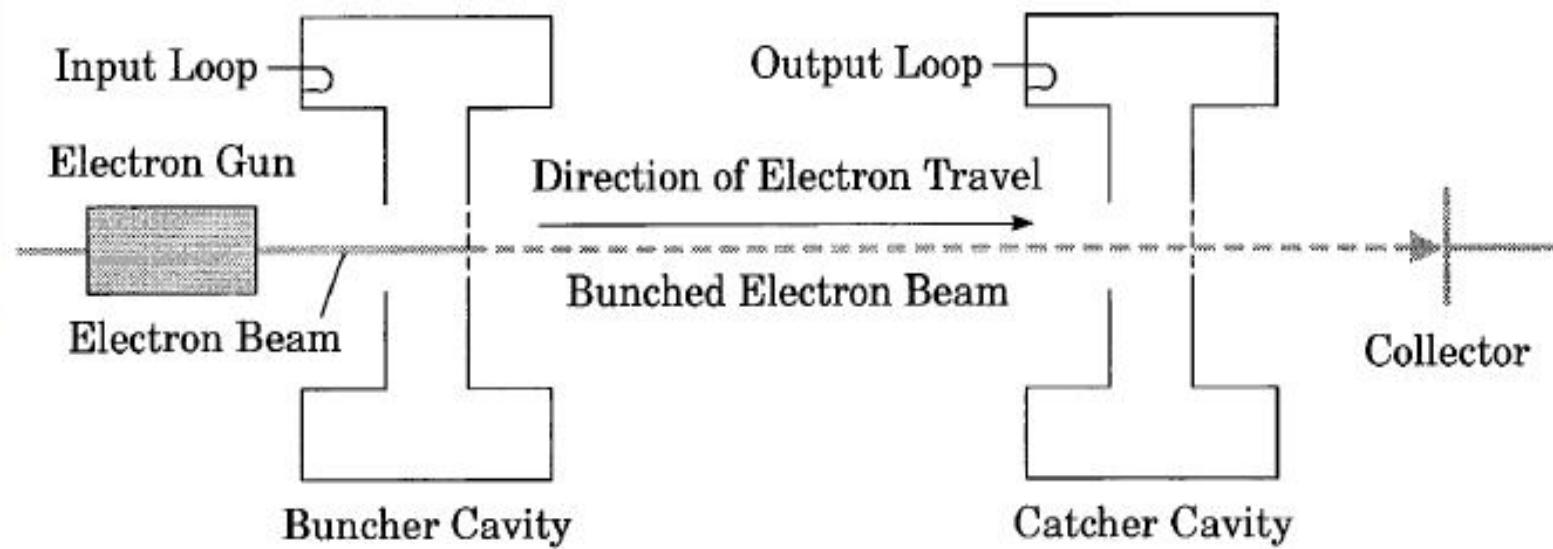
Application of TWT

- Point to Point Communication
- Satellite communication and Rader Appz
- Missile tracking application for military
- Television live broadcasting
 - LIVE news vans with satellite dishes on the roof carry TWTA inside

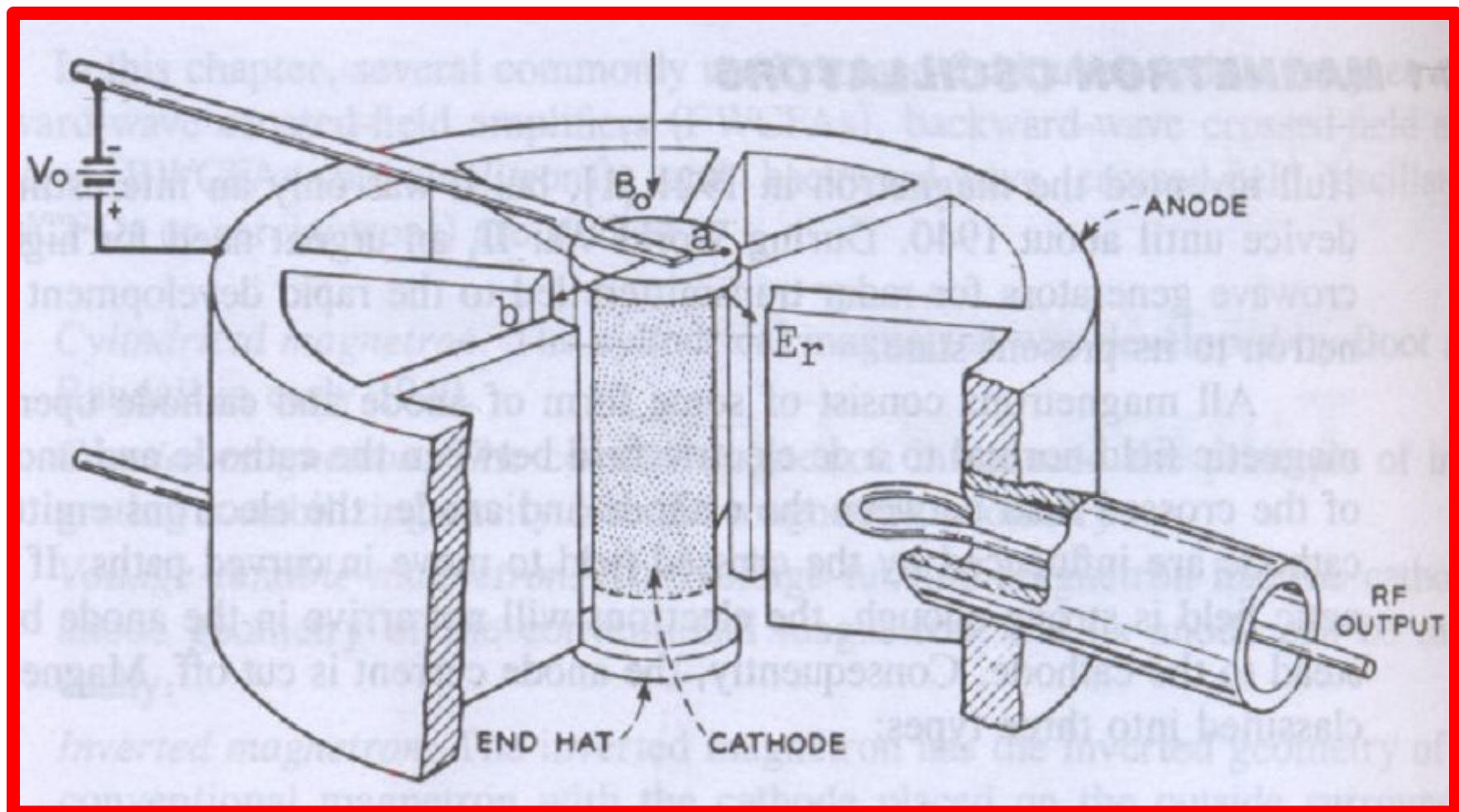


MULTI-CAVITY KLYSTRON

- Electron beam moves down tube past several cavities.
- Input cavity is the *buncher*, output cavity is the *catcher*.
- *Buncher* modulates the velocity of the electron beam



Magnetron Oscillator

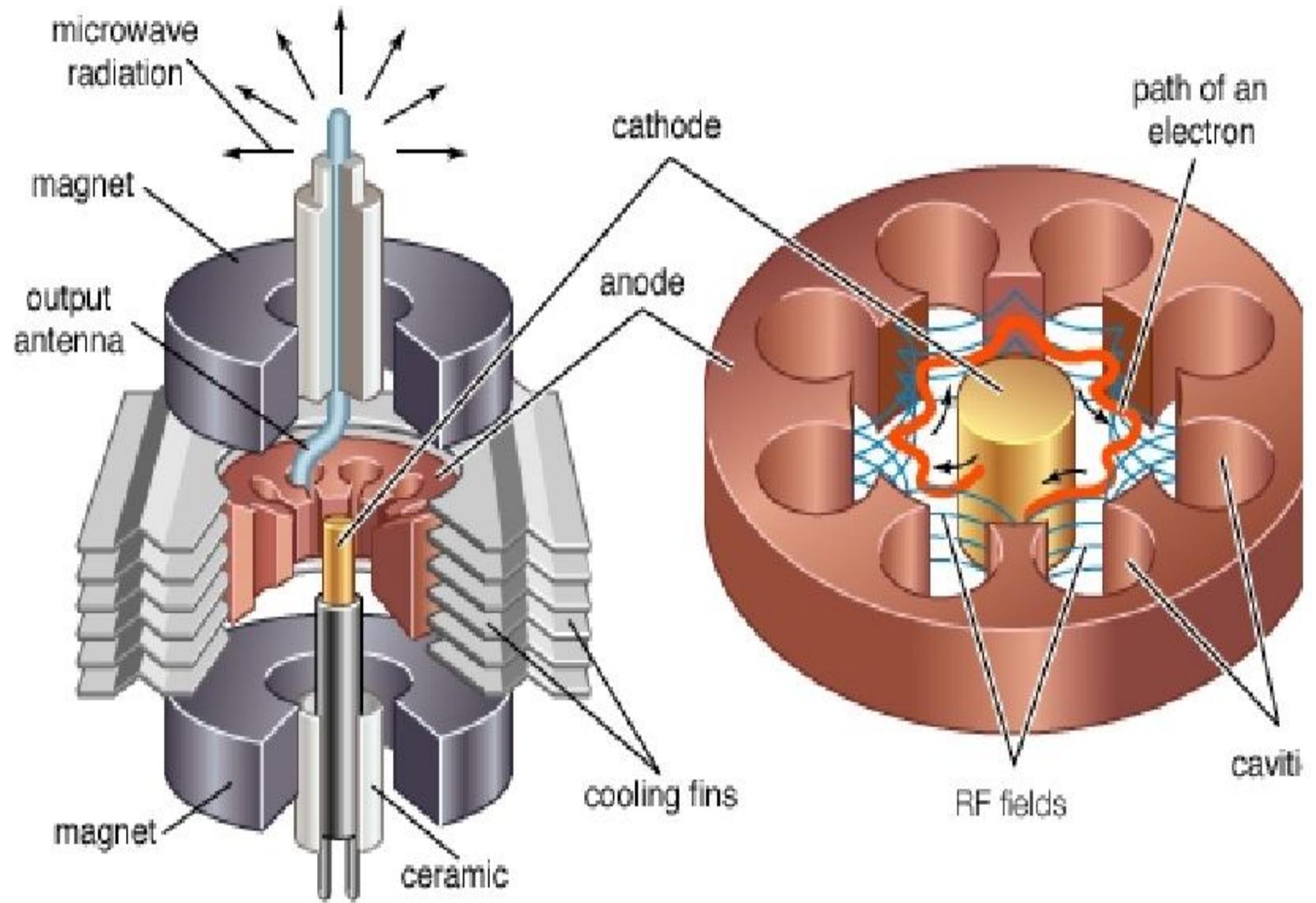


MAGNETRON

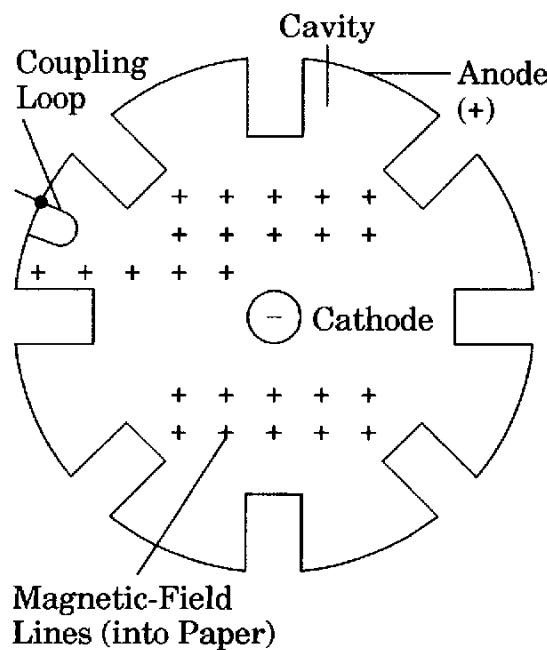
- High-power oscillator
- Common in radar and microwave ovens
- Cathode in center, anode around outside
- Strong dc magnetic field around tube causes electrons from cathode to spiral as they move toward anode
- Current of electrons generates microwaves in cavities around outside

SLOW-WAVE STRUCTURE

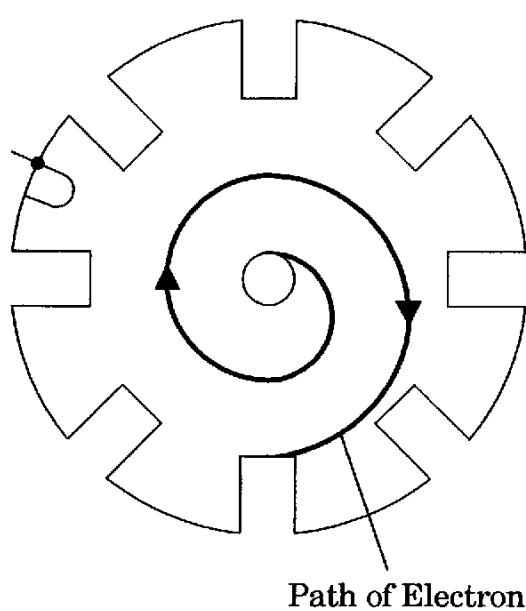
- Magnetron has cavities all around the outside
- Wave circulates from one cavity to the next around the outside
- Each cavity represents one-half period
- Wave moves around tube at a velocity much less than that of light
- Wave velocity approximately equals electron velocity



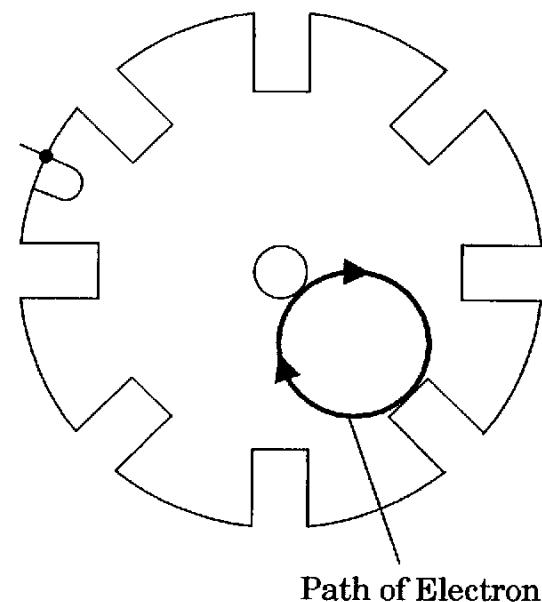
Cavity Magnetron



(a) Cross Section



(b) Electron Paths in Normal Operation



(c) Electron Paths at Cutoff

UNIT IV MICROWAVE SEMICONDUCTOR DEVICES AND CIRCUITS

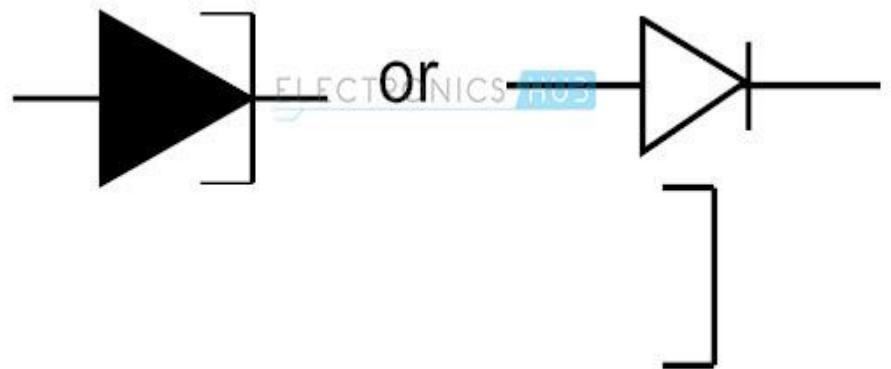
Principles of tunnel diodes - Varactor and Step recovery diodes – Transferred Electron Devices -Gunn diode- Avalanche Transit time devices- IMPATT and TRAPATT Devices- Parametric Amplifiers – Introduction to Micro strip Lines, & Monolithic Microwave Integrated circuits-Materials, MMIC Fabrication Techniques.

Tunnel Diode

- It is used as high speed switch, of order nano-seconds. Due to tunneling effect it has very fast operation in microwave frequency region. It is a two terminal device in which concentration of dopants is too high.
- The transient response is being limited by junction capacitance plus stray wiring capacitance. Mostly used in microwave oscillators and amplifiers. It acts as most negative conductance device. Tunnel diodes can be tuned in both mechanically and electrically. The symbol of tunnel diode is as shown below.

Tunnel Diode Applications

- Oscillatory circuits.
- Microwave circuits.
- Resistant to nuclear radiation.



Varactor Diode

- These are also known as Varicap diodes. It acts like the variable capacitor. Operations are performed mainly at reverse bias state only. These diodes are very famous due to its capability of changing the capacitance ranges within the circuit in the presence of constant voltage flow.
- They can able to vary capacitance up to high values. In varactor diode by changing the reverse bias voltage we can decrease or increase the depletion layer. These diodes have many applications as voltage controlled oscillator for cell phones, satellite pre-filters etc. The symbol of varactor diode is given below.

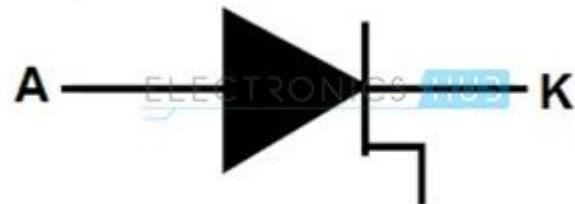
Varactor Diode Applications

- Voltage-controlled capacitors.
- Voltage-controlled oscillators.
- Parametric amplifiers.
- Frequency multipliers.
- FM transmitters and Phase locked loops in radio, television sets and cellular telephone.



Step Recovery Diodes

- It is also called as snap-off diode or charge-storage diode. These are the special type of diodes which stores the charge from positive pulse and uses it in the negative pulse of the sinusoidal signals. The rise time of the current pulse is equal to the snap time. Due to this phenomenon it has speed recovery pulses.
- The applications of these diodes are in higher order multipliers and in pulse shaper circuits. The cut-off frequency of these diodes is very high which are nearly at Giga hertz order.
- As multiplier this diode has the cut-off frequency range of 200 to 300 GHz. In the operations which are performing at 10 GHz range these diodes play a vital role. The efficiency is high for lower order multipliers. The symbol for this diode is as shown below.



Transferred Electron Devices

Gunn diodes are also known as transferred electron devices, TED, are widely used in microwave RF applications for frequencies between 1 and 100 GHz.

The Gunn diode is most commonly used for generating microwave RF signals - these circuits may also be called a transferred electron oscillator or TEO. The Gunn diode may also be used for an amplifier in what may be known as a transferred electron amplifier or TEA.

As Gunn diodes are easy to use, they form a relatively low cost method for generating microwave RF signals.

Gunn diode basics

The Gunn diode is a unique component - even though it is called a diode, it does not contain a PN diode junction. The Gunn diode or transferred electron device can be termed a diode because it does have two electrodes. It depends upon the bulk material properties rather than that of a PN junction. The Gunn diode operation depends on the fact that it has a voltage controlled negative resistance.

The mechanism behind the transferred electron effect was first published by Ridley and Watkins in a paper in 1961. Further work was published by Hilsum in 1962, and then in 1963 John Battiscombe (J. B.) Gunn independently observed the first transferred electron oscillation using Gallium Arsenide, GaAs semiconductor.

Gunn Diode

Gunn diode symbol for circuit diagrams

The Gunn diode symbol used in circuit diagrams varies. Often a standard diode is seen in the diagram, however this form of Gunn diode symbol does not indicate the fact that the Gunn diode is not a PN junction. Instead another symbol showing two filled in triangles with points touching is used as shown below.

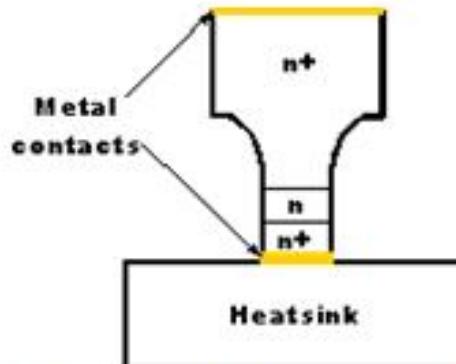


Gunn diode construction

Gunn diodes are fabricated from a single piece of n-type semiconductor. The most common materials are gallium Arsenide, GaAs and Indium Phosphide, InP. However other materials including Ge, CdTe, InAs, InSb, ZnSe and others have been used. The device is simply an n-type bar with n+ contacts. It is necessary to use n-type material because the transferred electron effect is only applicable to electrons and not holes found in a p-type material.

Within the device there are three main areas, which can be roughly termed the top, middle and bottom areas.

Gunn Diode



A discrete Gunn diode with the active layer mounted onto a heatsink for efficient heat transfer

The most common method of manufacturing a Gunn diode is to grow an epitaxial layer on a degenerate n+ substrate. The active region is between a few microns and a few hundred microns thick. This active layer has a doping level between 10^{14}cm^{-3} and 10^{16}cm^{-3} - this is considerably less than that used for the top and bottom areas of the device. The thickness will vary according to the frequency required.

The top n+ layer can be deposited epitaxially or doped using ion implantation. Both top and bottom areas of the device are heavily doped to give n+ material. This provides the required high conductivity areas that are needed for the connections to the device.

Devices are normally mounted on a conducting base to which a wire connection is made. The base also acts as a heat sink which is critical for the removal of heat. The connection to the other terminal of the diode is made via a gold connection deposited onto the top surface. Gold is required because of its relative stability and high conductivity.

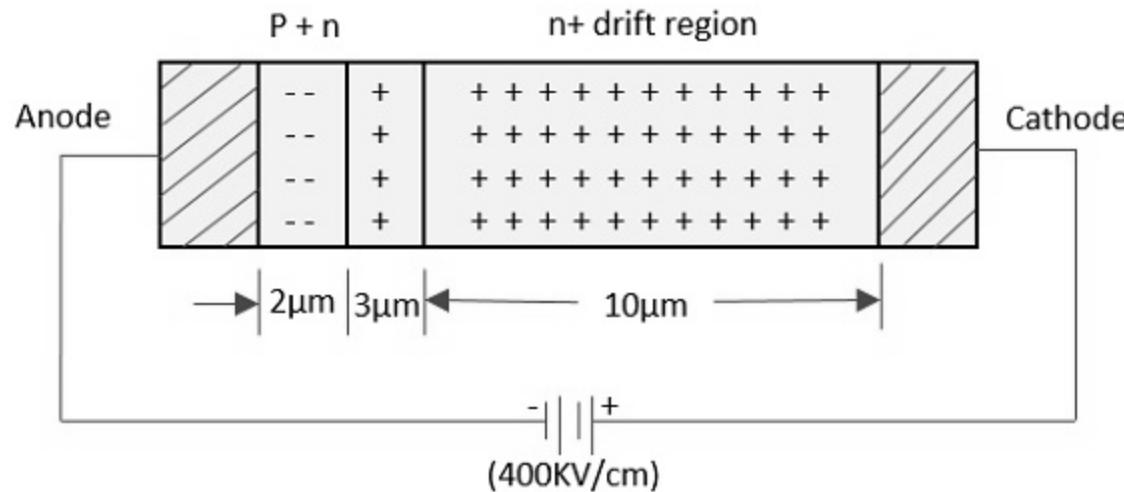
Avalanche transit time devices

- The process of having a delay between voltage and current, in avalanche together with transit time, through the material is said to be Negative resistance. The devices that helps to make a diode exhibit this property are called as **Avalanche transit time devices**.
- The examples of the devices that come under this category are IMPATT, TRAPATT and BARITT diodes. Let us take a look at each of them, in detail.

IMPATT Diode

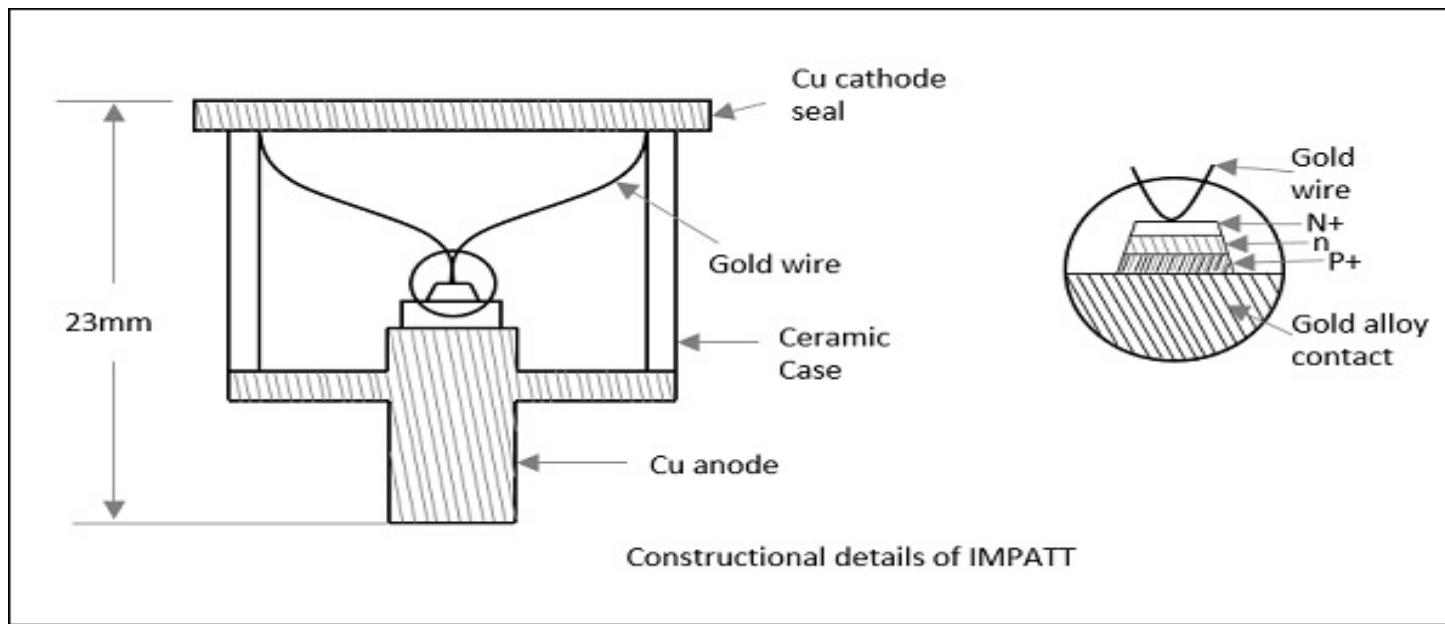
- This is a high-power semiconductor diode, used in high frequency microwave applications. The full form IMPATT is **IMPact ionization Avalanche Transit Time diode**.
- A voltage gradient when applied to the IMPATT diode, results in a high current. A normal diode will eventually breakdown by this. However, IMPATT diode is developed to withstand all this. A high potential gradient is applied to back bias the diode and hence minority carriers flow across the junction.
- Application of a RF AC voltage if superimposed on a high DC voltage, the increased velocity of holes and electrons results in additional holes and electrons by thrashing them out of the crystal structure by Impact ionization. If the original DC field applied was at the threshold of developing this situation, then it leads to the avalanche current multiplication and this process continues. This can be understood by the following figure.

IMPATT Diode



- Due to this effect, the current pulse takes a phase shift of 90° . However, instead of being there, it moves towards cathode due to the reverse bias applied. The time taken for the pulse to reach cathode depends upon the thickness of **n+** layer, which is adjusted to make it 90° phase shift. Now, a dynamic RF negative resistance is proved to exist. Hence, IMPATT diode acts both as an oscillator and an amplifier.

IMPATT Diode



The efficiency of IMPATT diode is represented as

$$\eta = \frac{P_{dc}}{P_{ac}} = \frac{V_a V_d}{I_a I_d} \quad \eta = \frac{P_{dc}}{P_{ac}} = \frac{V_a V_d}{I_a I_d}$$

Where,

• P_{ac} = AC power V_d & I_d = DC voltage & current

• P_{dc} = DC power V_a & I_a = AC voltage & current

IMPATT Diode

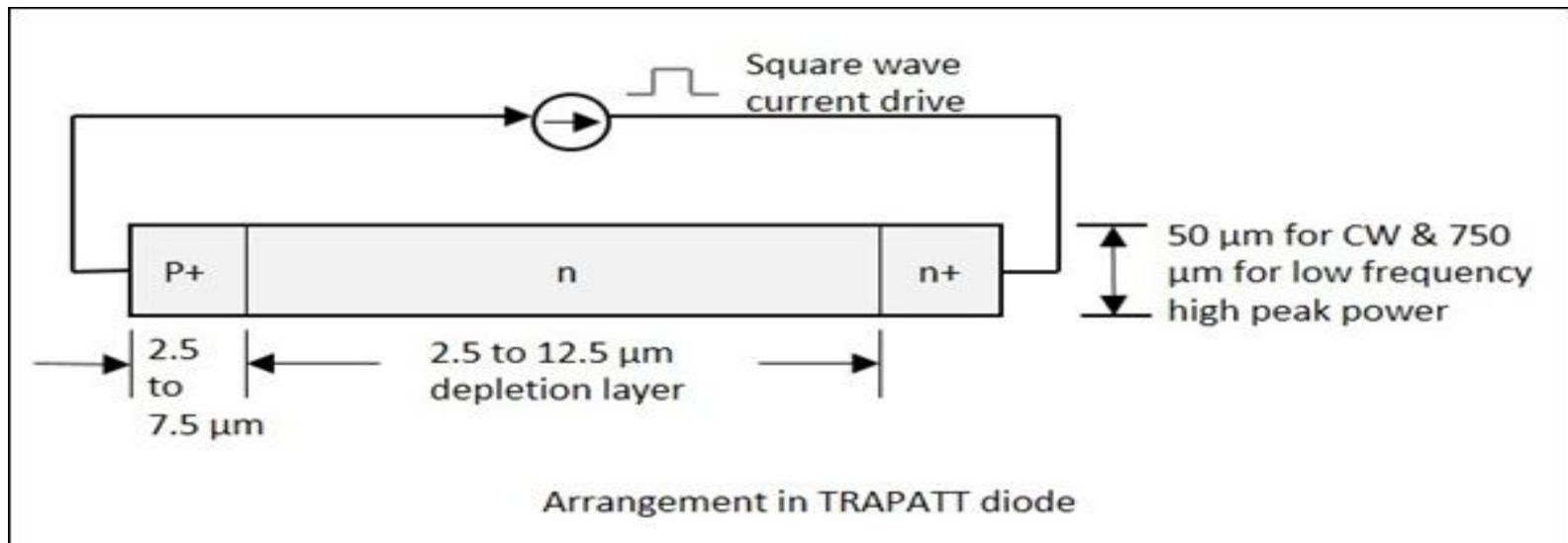
Disadvantages

- Following are the disadvantages of IMPATT diode.
- It is noisy as avalanche is a noisy process
- Tuning range is not as good as in Gunn diodes

Applications

- Following are the applications of IMPATT diode.
- Microwave oscillator
- Microwave generators
- Modulated output oscillator
- Receiver local oscillator
- Negative resistance amplifications
- Intrusion alarm networks (high Q IMPATT)
- Police radar (high Q IMPATT)
- Low power microwave transmitter (high Q IMPATT)
- FM telecom transmitter (low Q IMPATT)
- CW Doppler radar transmitter (low Q IMPATT)

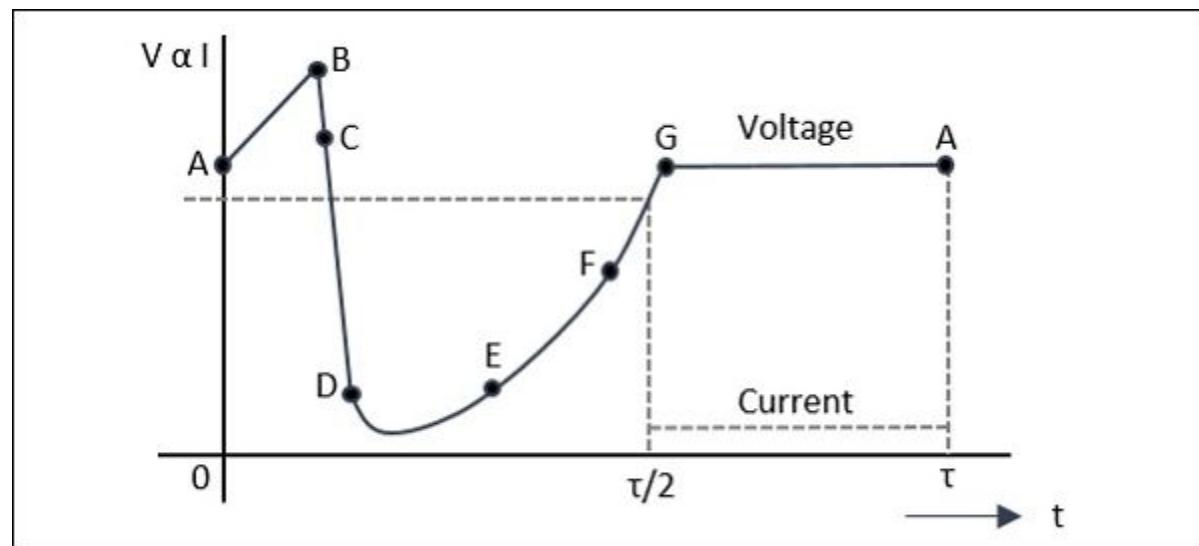
TRAPATT Diode



The full form of TRAPATT diode is **TRApped Plasma Avalanche Triggered Transit diode**. A microwave generator which operates between hundreds of MHz to GHz. These are high peak power diodes usually **n+- p-p+** or **p+-n-n+** structures with n-type depletion region, width varying from 2.5 to 1.25 μm . The following figure depicts this.

TRAPATT Diode

- The electrons and holes trapped in low field region behind the zone, are made to fill the depletion region in the diode. This is done by a high field avalanche region which propagates through the diode.
- The following figure shows a graph in which AB shows charging, BC shows plasma formation, DE shows plasma extraction, EF shows residual extraction, and FG shows charging. Let us see what happens at each of the points.



TRAPATT Diode

- **A:** The voltage at point A is not sufficient for the avalanche breakdown to occur. At A, charge carriers due to thermal generation results in charging of the diode like a linear capacitance.
- **A-B:** At this point, the magnitude of the electric field increases. When a sufficient number of carriers are generated, the electric field is depressed throughout the depletion region causing the voltage to decrease from B to C.
- **C:** This charge helps the avalanche to continue and a dense plasma of electrons and holes is created. The field is further depressed so as not to let the electrons or holes out of the depletion layer, and traps the remaining plasma.
- **D:** The voltage decreases at point D. A long time is required to clear the plasma as the total plasma charge is large compared to the charge per unit time in the external current.
- **E:** At point E, the plasma is removed. Residual charges of holes and electrons remain each at one end of the deflection layer.

TRAPATT Diode

- **E to F:** The voltage increases as the residual charge is removed.
- **F:** At point F, all the charge generated internally is removed.
- **F to G:** The diode charges like a capacitor.
- **G:** At point G, the diode current comes to zero for half a period. The voltage remains constant as shown in the graph above. This state continues until the current comes back on and the cycle repeats.

The avalanche zone velocity V_s is represented as

$$V_s = \frac{dx}{dt} = JqNA$$

Where

J = Current density

q = Electron charge 1.6×10^{-19}

N = Doping concentration

TRAPATT Diode

The avalanche zone will quickly sweep across most of the diode and the transit time of the carriers is represented as

$$\tau_s = L/V_s$$

Where

V_s = Saturated carrier drift velocity

L = Length of the specimen

The transit time calculated here is the time between the injection and the collection. The repeated action increases the output to make it an amplifier, whereas a microwave low pass filter connected in shunt with the circuit can make it work as an oscillator.

Applications

- There are many applications of this diode.
- Low power Doppler radars
- Local oscillator for radars
- Microwave beacon landing system
- Radio altimeter
- Phased array radar, etc.

Parametric Amplifier

Parametric amplification is a process of RF-RF power conversion that operates by pumping a nonlinear reactance with a large-signal RF pumping source to either produce mixing products with gain or to generate a negative resistance. Parametric amplifiers (paramps) were traditionally grouped into two types: the phase-incoherent upconverting parametric amplifier and the negative-resistance parametric amplifier. With phase-incoherent upconverting parametric amplifiers, a fixed-frequency phase-incoherent incommensurate pump, at frequency f_p , mixes with an RF small-signal source input, at frequency f_s , to produce an upconverted output with gain that can be predicted by the Manley-Rowe relations [27, 28]. Negative-resistance parametric amplifiers are also mixers, but differ from phase-incoherent upconverting paramps in that the frequency relationship $f_i = f_p - f_s$ must be satisfied, where f_i is the so-called “idler” frequency [20]. The Manley-Rowe relations show that negative-resistance parametric amplifiers present a regenerative condition with the possibility of oscillation at both the source and idler frequencies.

Parametric Amplifier

2.1 The Manley-Rowe Relations

In 1956, J. M. Manley and H. E. Rowe published a manuscript that analyzed the power flow into and out of a nonlinear reactive element under excitation at its different harmonic frequencies [27]. The results of this analysis were two simple mathematical expressions quantifying how the total outgoing power flow would distribute itself among the harmonic terms. These two mathematical relationships, which will now be referred to as the Manley-Rowe relations, have the following important properties:

1. They are independent of the particular shape of the capacitance-voltage or inductance-current curve for a nonlinear capacitance or nonlinear inductance, respectively.
2. The power levels of the various driving sources are irrelevant.
3. The external circuitry connected to the nonlinear reactance will not affect how the power is distributed to the harmonic frequencies.

Microstrip lines

Microstrip Structure

The diagram illustrates the cross-section of a microstrip line. It consists of a central 'Conducting strip' of width W and thickness t , situated above a 'Ground plane'. This assembly rests on a 'Dielectric substrate' of height h and dielectric constant ϵ_r . The entire structure is shown on a grid background.

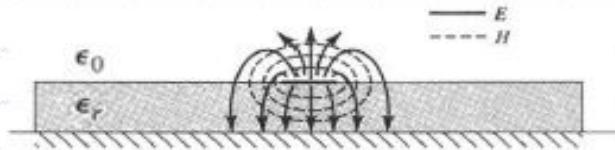
On the right side of the diagram, a separate section shows the internal electric (E) and magnetic (H) field distributions. The electric field E is represented by solid arrows pointing upwards from the ground plane, while the magnetic field H is represented by dashed arrows forming concentric loops around the conductor.

- Inhomogeneous structure:
Due to the fields within two guided-wave media, the microstrip does not support a pure TEM wave.
- When the longitudinal components of the fields for the dominant mode of a microstrip line is much smaller than the transverse components, the **quasi-TEM approximation** is applicable to facilitate design.

Microstrip lines

- Transmission Line Parameters

Effective Dielectric Constant (ϵ_{re}) and Characteristic Impedance (Z_c)



➤ For thin conductors (i.e., $t \rightarrow 0$), closed-form expression (error $\leq 1\%$):

◆ $W/h \leq 1$:

$$\epsilon_{re} = \frac{\epsilon_r + 1}{2} + \frac{\epsilon_r - 1}{2} \left[\left(1 + 12 \frac{h}{W} \right)^{-0.5} + 0.04 \left(1 - \frac{W}{h} \right)^2 \right]$$

$$Z_c = \frac{\eta}{2\pi\sqrt{\epsilon_{re}}} \ln \left(\frac{8h}{W} + 0.25 \frac{W}{h} \right)$$

◆ $W/h \geq 1$:

$$\epsilon_{re} = \frac{\epsilon_r + 1}{2} + \frac{\epsilon_r - 1}{2} \left(1 + 12 \frac{h}{W} \right)^{-0.5}$$

$$Z_c = \frac{\eta}{\sqrt{\epsilon_{re}}} \left[\frac{W}{h} + 1.393 + 0.677 \ln \left(\frac{W}{h} + 1.444 \right) \right]^{-1}$$

➤ For thin conductors (i.e., $t \rightarrow 0$), more accurate expressions:

◆ Effective dielectric constant (error $\leq 0.2\%$):

$$\epsilon_{re} = \frac{\epsilon_r + 1}{2} + \frac{\epsilon_r - 1}{2} \left(1 + \frac{10}{u} \right)^{-ab}$$

$$a = 1 + \frac{1}{49} \ln \left(\frac{u^4 + \left(\frac{u}{52} \right)^2}{u^4 + 0.432} \right) + \frac{1}{18.7} \ln \left[1 + \left(\frac{u}{18.1} \right)^3 \right]$$

$$b = 0.564 \left(\frac{\epsilon_r - 0.9}{\epsilon_r + 3} \right)^{0.053}$$

◆ Characteristic impedance (error $\leq 0.03\%$):

$$Z_c = \frac{\eta}{2\pi\sqrt{\epsilon_{re}}} \ln \left[\frac{F}{u} + \sqrt{1 + \left(\frac{2}{u} \right)^2} \right]$$

$$F = 6 + (2\pi - 6) \exp \left[- \left(\frac{30.666}{u} \right)^{0.7528} \right]$$

Microstrip lines

- Transmission Line Parameters

- Guided wavelength

$$\lambda_g = \frac{\lambda_0}{\sqrt{\epsilon_{re}}} \quad \text{or} \quad \lambda_g = \frac{300}{f(\text{GHz})\sqrt{\epsilon_{re}}} \text{ mm}$$

- Propagation constant

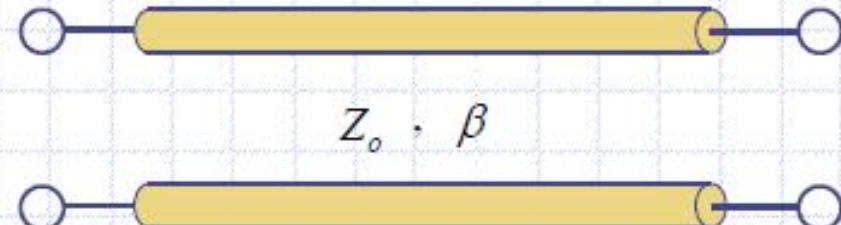
$$\beta = \frac{2\pi}{\lambda_g}$$

- Phase velocity

$$v_p = \frac{\omega}{\beta} = \frac{c}{\sqrt{\epsilon_{re}}}$$

- Electrical length

$$\theta = \beta l$$



Microstrip lines

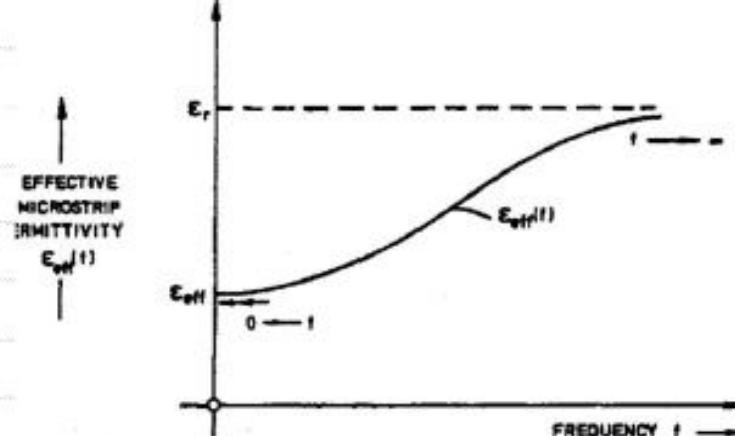
- Transmission Line Parameters

- Losses

- ◆ Conductor loss
- ◆ Dielectric loss
- ◆ Radiation loss

- Dispersion

- ◆ $\epsilon_{re}(f)$
- ◆ $Z_o(f)$



- Surface Waves and higher-order modes

- ◆ Coupling between the quasi-TEM mode and surface wave mode become significant when the frequency is above f_s

$$f_s = \frac{c \tan^{-1} \epsilon_r}{\sqrt{2\pi h} \sqrt{\epsilon_r - 1}}$$

- ◆ Cutoff frequency f_c of first higher-order modes in a microstrip

$$f_c = \frac{c}{\sqrt{\epsilon_r} (2W + 0.8h)}$$

- ◆ The operating frequency of a microstrip line $< \text{Min } (f_s, f_c)$

Monolithic Microwave Integrated Circuit (MMIC)

- Microwave ICs are the best alternative to conventional waveguide or coaxial circuits, as they are low in weight, small in size, highly reliable and reproducible. The basic materials used for monolithic microwave integrated circuits are –
- Substrate material
- Conductor material
- Dielectric films
- Resistive films
- These are so chosen to have ideal characteristics and high efficiency. The substrate on which circuit elements are fabricated is important as the dielectric constant of the material should be high with low dissipation factor, along with other ideal characteristics. The substrate materials used are GaAs, Ferrite/garnet, Aluminum, beryllium, glass and rutile.

Monolithic Microwave Integrated Circuit (MMIC)

- The conductor material is so chosen to have high conductivity, low temperature coefficient of resistance, good adhesion to substrate and etching, etc. Aluminum, copper, gold, and silver are mainly used as conductor materials. The dielectric materials and resistive materials are so chosen to have low loss and good stability.
- Fabrication Technology
- In hybrid integrated circuits, the semiconductor devices and passive circuit elements are formed on a dielectric substrate. The passive circuits are either distributed or lumped elements, or a combination of both.
- Hybrid integrated circuits are of two types.
- Hybrid IC
- Miniature Hybrid IC
- In both the above processes, Hybrid IC uses the distributed circuit elements that are fabricated on IC using a single layer metallization technique, whereas Miniature hybrid IC uses multi-level elements.
- Most analog circuits use meso-isolation technology to isolate active n-type areas used for FETs and diodes. Planar circuits are fabricated by implanting ions into semi-insulating substrate, and to provide isolation the areas are masked off.

UNIT V MICROWAVE MEASUREMENTS

Introduction – Slotted line carriage --- Spectrum analyzer – Network analyzer – Power measurements – Schottky barrier diode sensor –Bolometer sensor – Power sensor – High power measurement – Insertion loss and attenuation measurement – VSWR measurement – Low and high VSWR – Impedance measurement – Frequency measurement – Measurement of cavity Q – Dielectric measurement of a solid by wave-guide method – Antenna measurement – Radiation pattern – Phase and gain.

Slotted Line carriage

A slotted line to measure voltage standing wave ratio. You might turn up such an instrument if you work in a lab that is more than 25 years old. Basically it is a coax line with a slot down one side where a probe can be moved longitudinally to measure varying electric field strength. The probe has a detector that converts RF energy to DC voltage, so you can measure peaks and valleys using an voltmeter. For circuits that were extremely mismatched (or open or short circuited), the peaks and valleys are the most noticeable. The ratio of the peak voltage to the valley voltage was the most directly calculated piece of data you can get with a slotted line... hence "voltage standing wave ratio".

Using a slotted line, you could also measure an unknown frequency by measuring the distance between the voltage peaks and noting that the distance is $1/2$ wavelength.



Spectrum Analyzer

- An RF spectrum analyzer is a device used to examine the spectral composition of some electrical waveform. It may also measure the power spectrum.
Wikipedia
- Translation: It is a fast-sweeping tuned radio receiver that displays signal amplitudes at various frequencies.



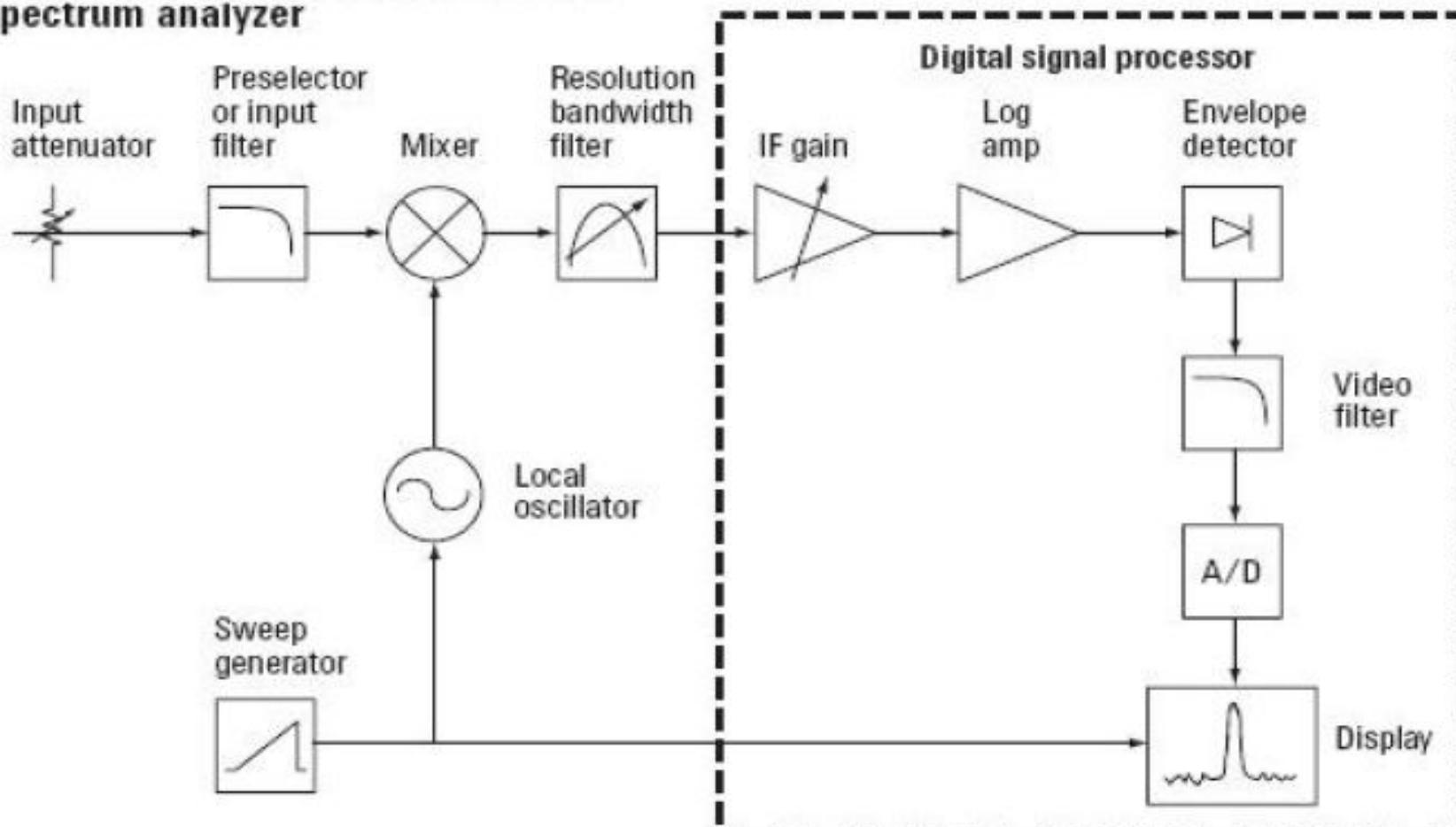
HP 494AP SA



HP 141T SA

Spectrum Analyzer

Block diagram of a superheterodyne spectrum analyzer



Buying a Spectrum Analyzer?

- Prices vary from ~ hundreds\$ to thousands\$.
- Good for general purpose experimenting and test. For example, required for cavity duplexer adjustment.
- CRT Trace Storage capability is a must.
- Addition of Tracking generator is a wise move.
- It is nice to be able to capture the plots. Need 1980's grade (GP-IB bus). Otherwise take screen pictures.
- The HP-141T mainframe SA is good entry point.
- Lab-Grade units keep their value.

Two Types of Spectrum Analyzers

- **Swept**
 - Traditional Heterodyne design.
 - The most popular and least costly
 - Has wider frequency coverage (GHz...)
 - Has limitations in capturing bursty or complex events
 - Provides only amplitude information
- **Real-Time (Fourier Transform)**
 - RF samples are taken by ADC in the time domain
 - Fourier Transform and other post-processing (math) is applied to the samples at various frequency bins.
 - Is much better in capturing complex or fast changing signals
 - Provides both amplitude and phase info, thanks to FFT
 - Has frequency range limited by ADC.
 - More costly

Network Analyzer - Definition

- An instrument used to analyze the properties of electrical networks, especially those properties associated with the reflection and transmission of electrical signals known as scattering parameters (S-parameters).
Wikipedia
- Translation: It is a fast-sweeping tuned or wideband radio receiver that displays relative signal amplitudes (and optionally phases) when compared to a reference at various frequencies.



Network Analyzer – Two Types

- Scalar Network Analyzer (SNA)
 - Measures amplitude properties only. Simpler design (\$)
 - Usually requires an external sweeping RF source
 - May have external RF detectors
- Vector Network Analyzer (VNA)
 - Measures both amplitude **and phase** properties with greater dynamic range and accuracy. Complex unit (\$\$\$)
 - Has built-in sweeping RF source (generally)
 - Has built-in Tuned RF receiver

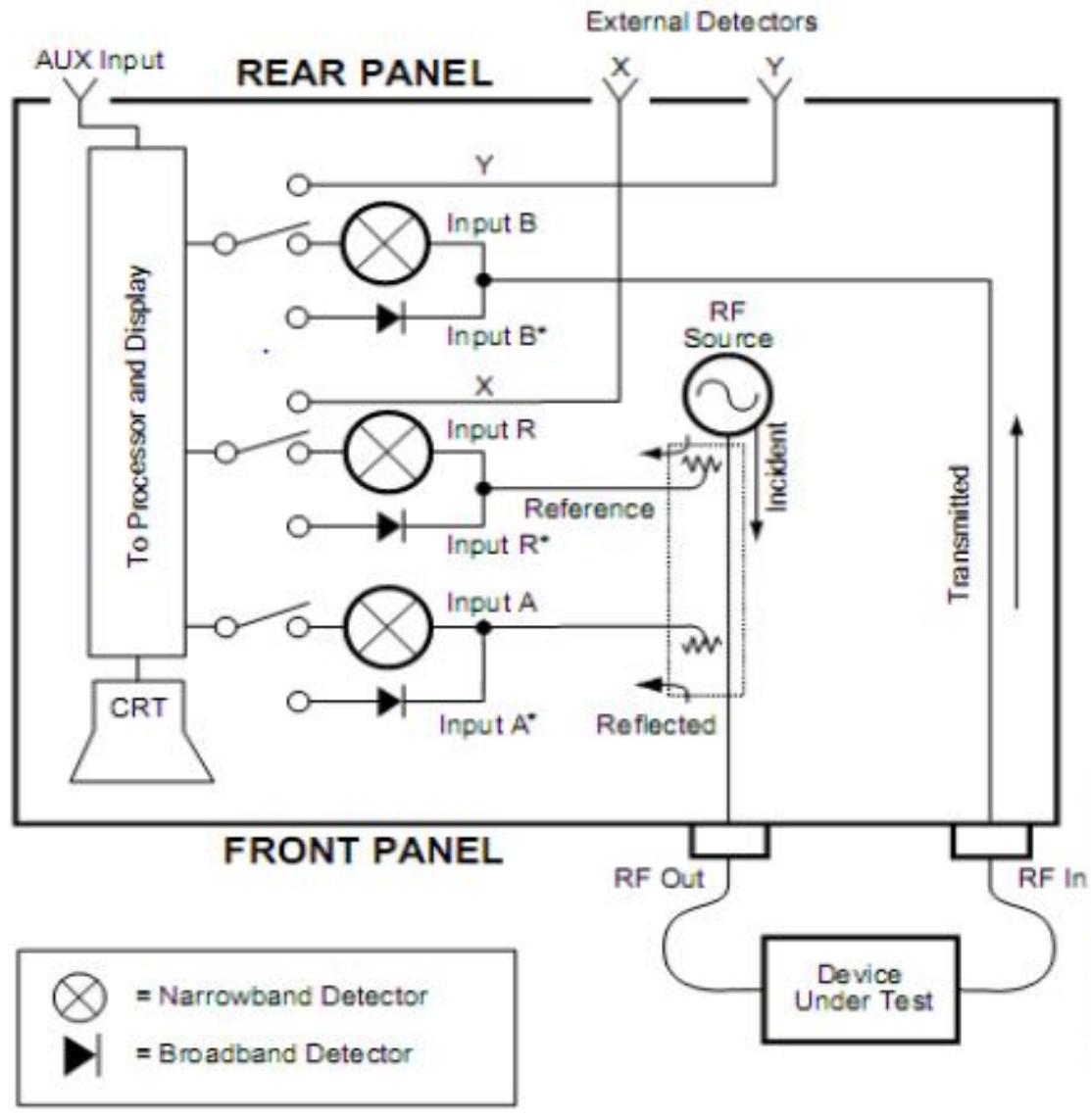


Wiltron 560 SNA



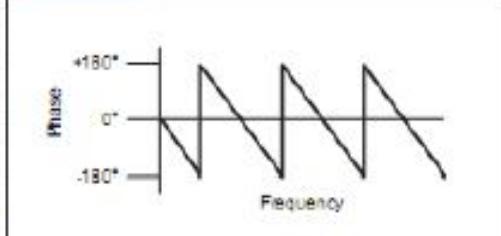
HP 8753C VNA

Network Analyzer – Block Diagram

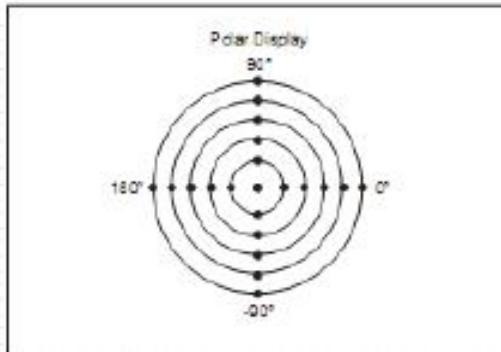


Vector Network Analyzers

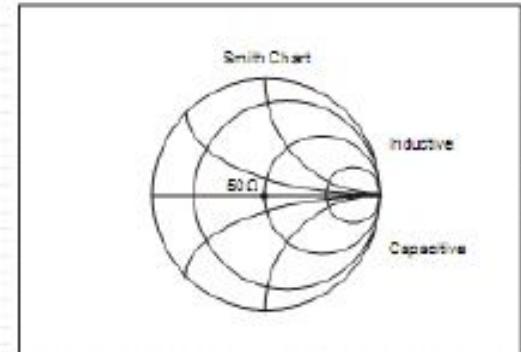
- Can display data in various forms



Linear Phase with Frequency



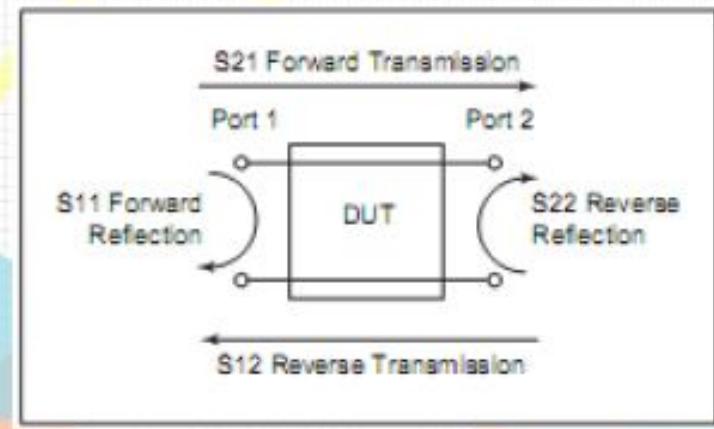
Polar Plot



Smith Chart

- Can usually express results in the form of S-parameters directly.

- Completely characterize a one-port or two-port linear or passive device



Buying a Network Analyzer?

- Classic VNAs are expensive (min. 1000\$). Keep their value.
- SNA....Better off with SA and Tracking Generator.
- Cheaper newer models available (use the computer for display/control). Not as broadband, not as accurate as classic lab-grade VNAs. Require a PC.
 - MiniVNA, max. 180MHz
 - N2PK Vector Network Analyzer, max. 60MHz, a kit.
 - VNA 2180, max. 180MHz
- Antenna Analyzers, a possibility...Limited in frequency and measurement range, accuracy, but are small and can be connected up at the antenna feedpoint.
 - MFJ-269, HF, VHF, UHF
 - AEA CIA-HF, HF
 - Autek RF-1, HF only



POWER MEASUREMENT

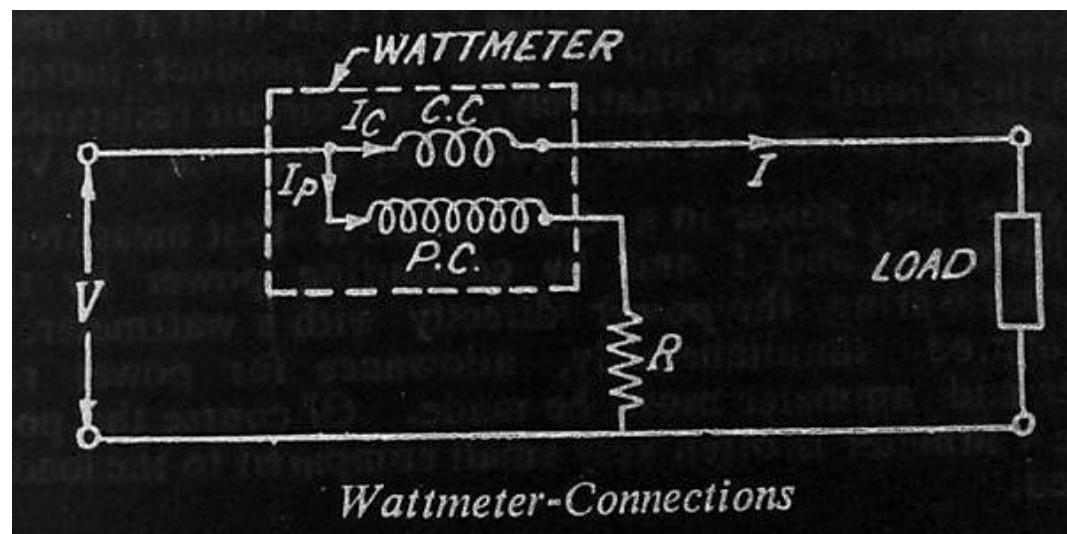
- Power is defined as the quantity of energy dissipated or stored per unit time.
- Methods of measurement of power depend on the frequency of operation, levels of power and whether the power is continuous or pulsed.
- The range of microwave power is divided into three categories :-
 - i. Low power ($< 10\text{mW} @ 0\text{dBm}$)
 - ii. Medium power (from $10 \text{ mW} - 10 \text{ W} @ 0 - 40 \text{ dBm}$)
 - iii. High power ($> 10 \text{ W} @ 40 \text{ dBm}$)
- The microwave power meter consists of a power sensor, which converts the microwave power to heat energy.
- The sensors used for power measurements are the Schottky barrier diode, bolometer and the thermocouple.

Power Measurement

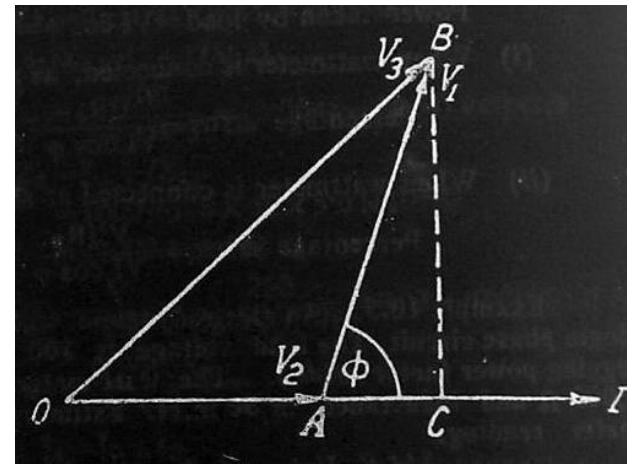
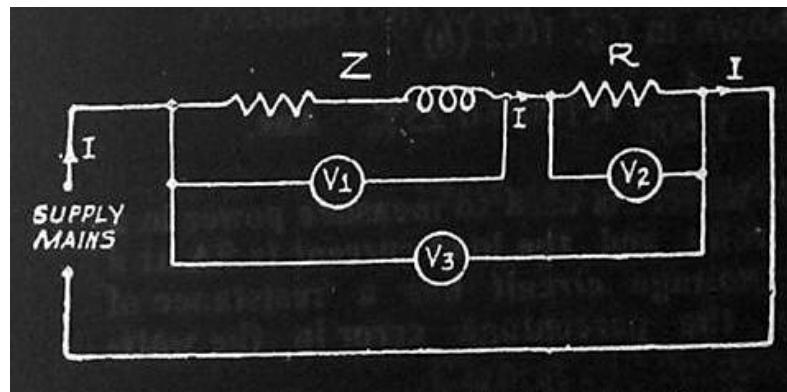
- Power may be defined as the rate at which energy is transformed or made available
- In almost all cases the power in a d.c. circuit is best measured by separately measuring quantities, V and I and by computing $P=VI$
- In case of a.c. circuits the instantaneous power varies continuously as the current and voltage go through a cycle of values
- The fact that the power factor is involved in the expression for the power means that a wattmeter must be used instead of merely an ammeter and voltmeter.

Wattmeter

- A wattmeter is essentially an inherent combination of an ammeter and a voltmeter and, therefore , consists of two coils known as *current coil* and *pressure coil*.
- Wattmeter connection:



Measurement of Power in Single Phase A.C. Circuit

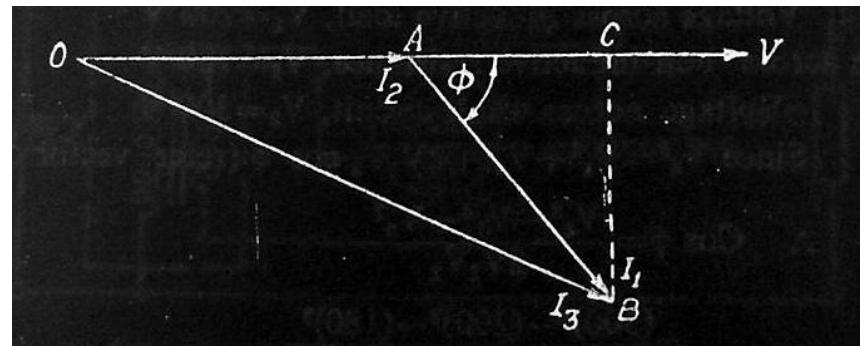
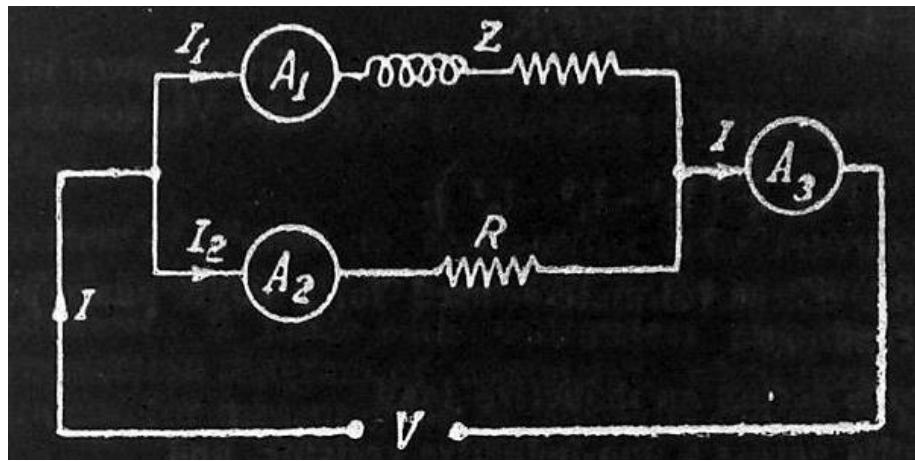


- 3-voltmeter method

$$P = \frac{V_3^2 - V_1^2 - V_2^2}{2R}$$

$$\cos \phi = \frac{V_3^2 - V_1^2 - V_2^2}{2V_1 V_2}$$

- Disadvantages : (i) Even small errors in measurement of voltages may cause serious errors in the value of power, (ii) Supply voltage higher than normal voltage is required



- 3-Ammeter method

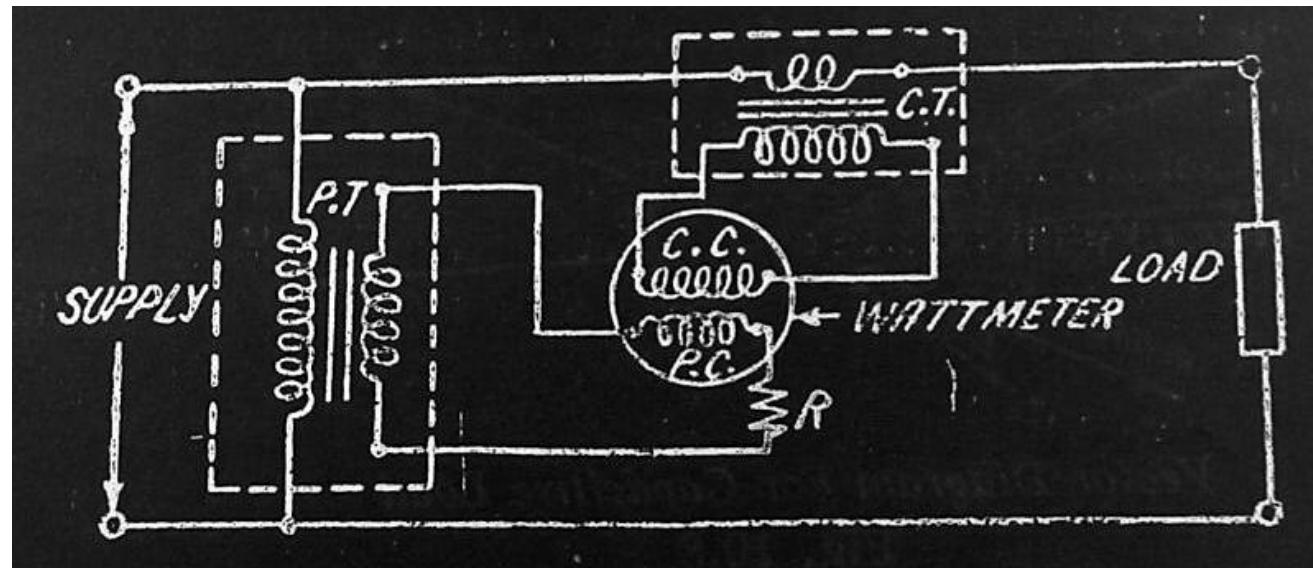
$$P = \frac{R}{2} (I_3^2 - I_1^2 - I_2^2)$$

$$\cos \varphi = \frac{I_3^2 - I_1^2 - I_2^2}{2I_1 I_2}$$

- The disadvantages of measurement of power by 3 voltmeters are overcome in this method

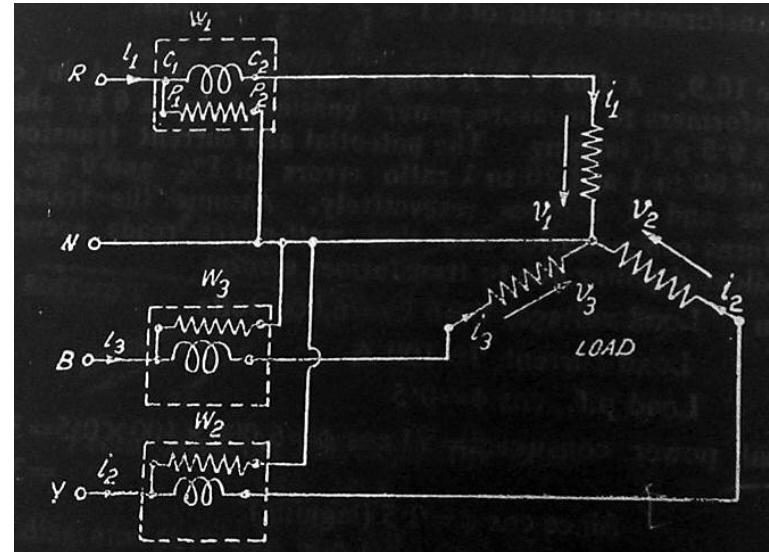
Measurement of power in conjunction with instrument transformers

- This method is used when the currents and voltages of the circuits to be measured are high
- Figure below shows a measurement of power with wattmeter in conjunction with instrument transformers in single phase A.C. circuits



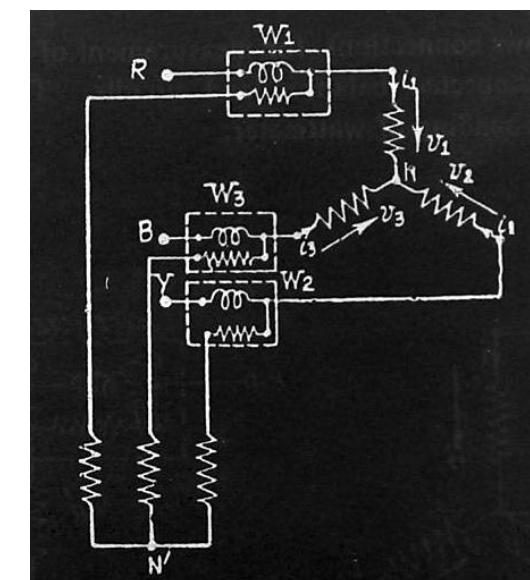
Measurement of Power in 3-Phase Circuit

- Measurement of power in 3-phase, 4-wire circuits-----→



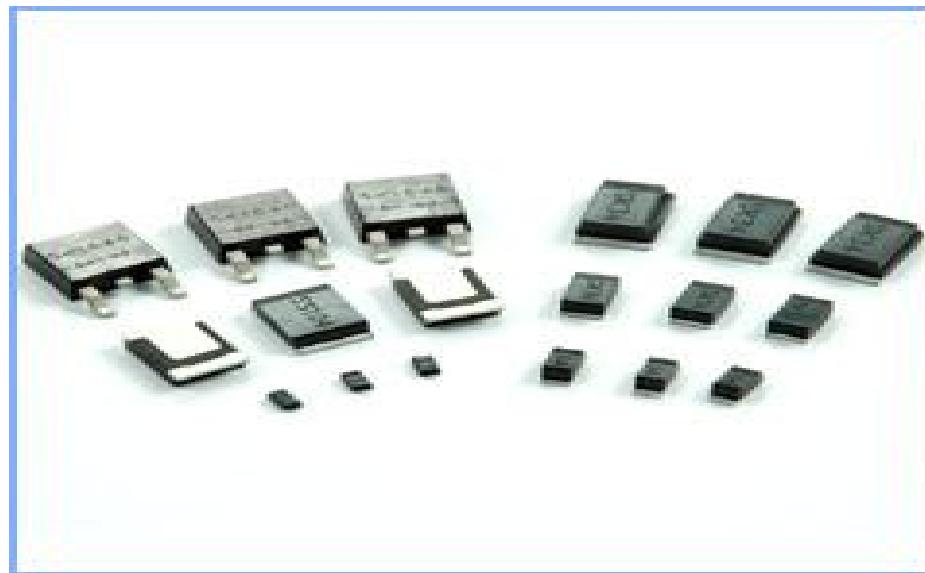
- Measurement of power in 3-phase, 3-wire circuits-----→

- $P=W_1+W_2+W_3$



SCHOTTKY BARRIER DIODE

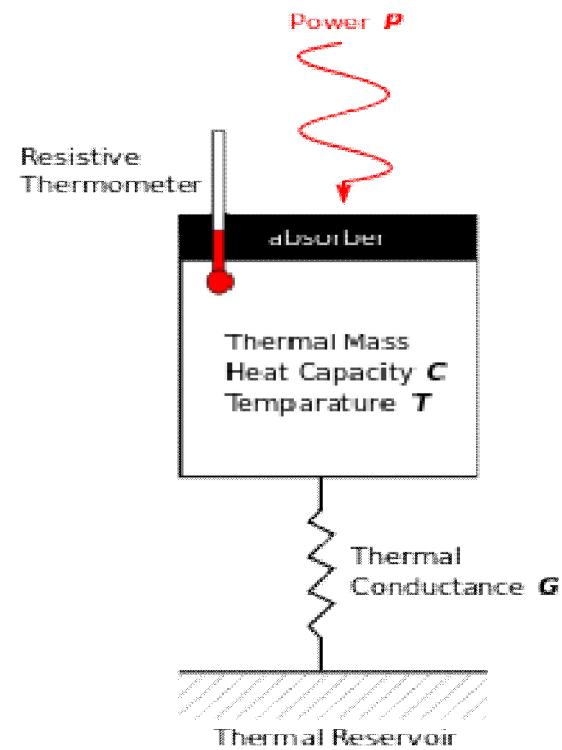
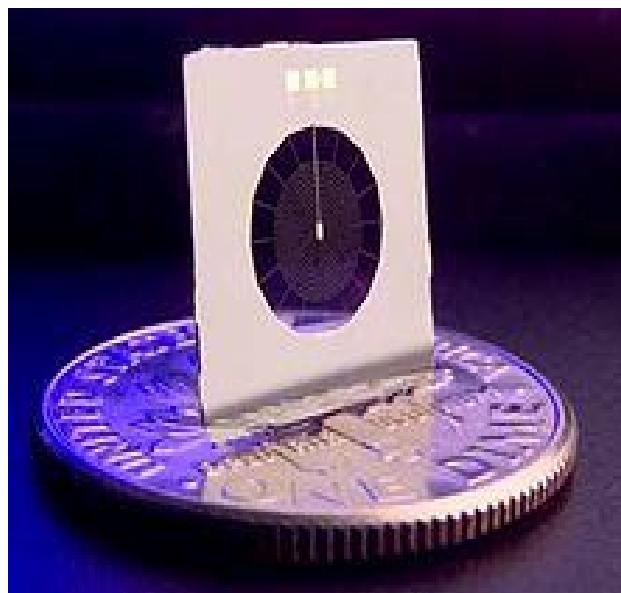
- A zero-biased Schottky Barrier Diode is used as a square-law detector whose output is proportional to the input power.
- The diode detectors can be used to measure power levels as low as 70dBm.



BOLOMETERS

- A Bolometer is a power sensor whose resistance changes with temperature as it absorbs microwave power.
- Are power detectors that operate on thermal principles. Since the temperature of the resistance is dependent on the signal power absorbed, the resistance must also be in proportion to the signal power.
- The two most common types of bolometer are, the barretter and the thermistor. Both are sensitive power detectors and is used to indicate microwatts of power. They are used with bridge circuits to convert resistance to power using a meter or other indicating devices.

BOLOMETER

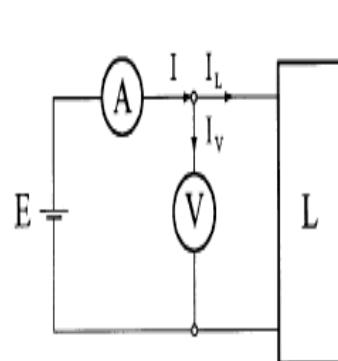


Power in DC circuits

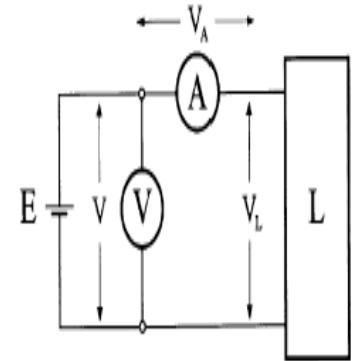
- Power
- Can be carried ~~Puting~~ using a voltmeter and an ammeter (generally)
- Two measurement arrangements
- Wattmeters:
 - Dynamometer
 - Digital wattmeter
 - Thermal wattmeter
 - Hall-power meter

DC circuits

- a) Ammeter measures current which flow into the voltmeter and load
- b) Voltmeter measures voltage drop across the ammeter in addition to that dropping across the load



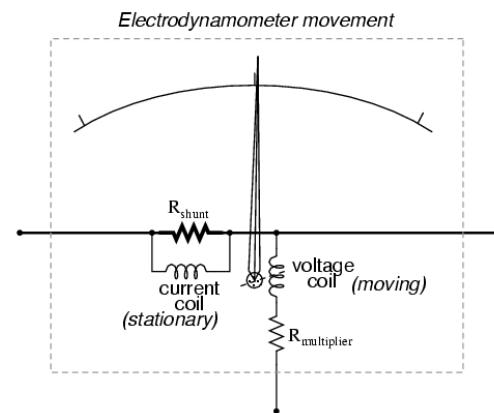
(a)



(b)

Dynamometer

- Power (direct) measurement device for DC and AC systems
- Accuracy better than 0,25 %
- Two coils: static and movable
- Torque is proportional product of current in current coil and current in voltage coil



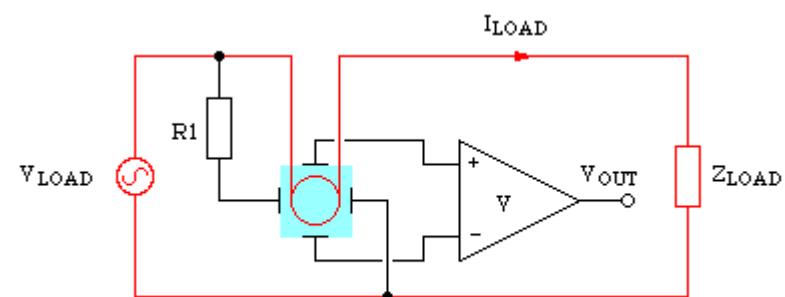
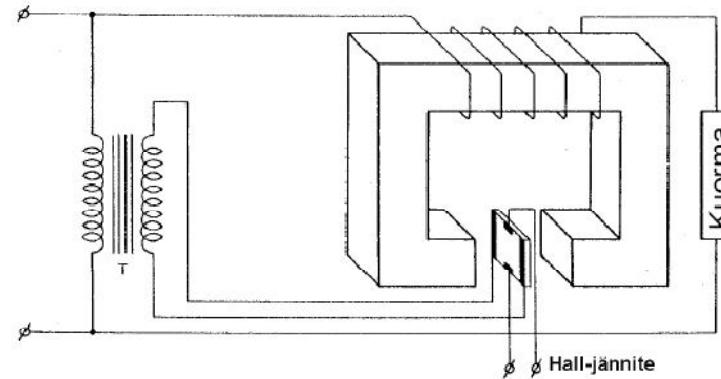
Digital wattmeter (up to 100 kHz)

- Advantages:
 - High-resolution
 - Accuracy
- Several techniques (multiplication of signals)
- Electronic multiplier is an analog system which gives as its output a voltage proportional to the power indication required
→ A/D conversion



Hall-power meter

- Coil generates magnetic field which is proportional to load current
- The sensor excitation current passes through R1 and is proportional to the load voltage
→ Hall voltage is proportional to load power
- Problems: offset and linearity



Circuit 9. Schematic wattmeter based on Hall effect sensor

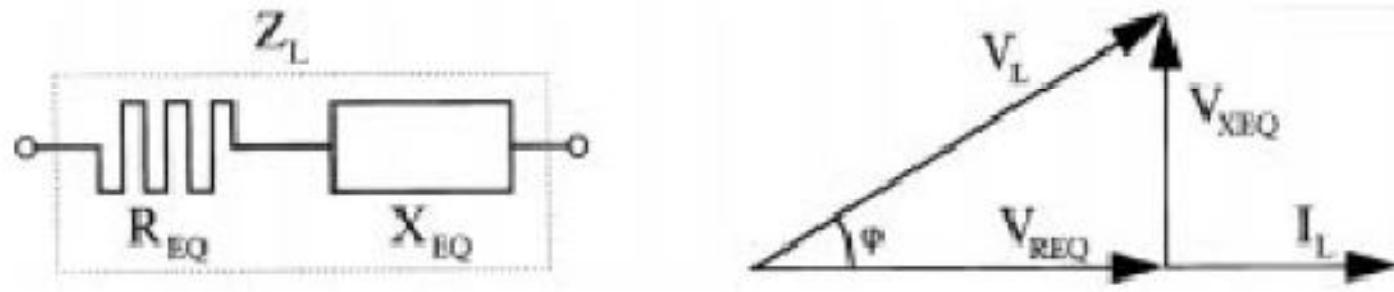
Power in AC circuits

- Instantaneous power (time dependence)
- Mean power (usually the most interesting)
- Real power (active work), reactive power, apparent power
- Measures can be done same way as DC circuit (single-phase)

$$p(t) = v(t)i(t)$$

$$P = \frac{1}{T} \int_0^T p(t) dt$$

AC circuits



$$P = V_L I_L \cos \varphi$$

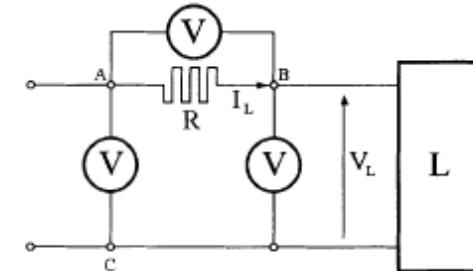
$$Q = V_L I_L \sin \varphi$$

$$S = \sqrt{P^2 + Q^2}$$

Low- and Medium-Frequency Power Measurements

- Three-Voltmeter Method
 - Single-phase arrangements
 - Power in load can be measured using a non-inductive resistor and measuring the three voltage
 - Also in DC circuits

$$P_L = \frac{V_{AC}^2 - V_{AB}^2 - V_{BC}^2}{2R}$$



Line-Frequency Power Measurements

- Polyphase Power Measurements
 - Three-phase systems are most commonly used in industrial applications
 - Energy and power generation and distribution
 - “Real power for consumer”
 - Reactive power also important (loading)
 - Power can measured several ways
 - Power factor

Line-Frequency Power Measurements

(2)

- Four (main) different cases which affects to the measurement arrangements:
 1. Symmetrical load with neutral conductor
 2. Symmetrical load without neutral conductor
 3. Unsymmetrical load with neutral conductor
 4. Unsymmetrical load without neutral conductor

Insertion Loss

Insertion loss measures the energy absorbed by the transmission line in the direction of the signal path in dB/meter or dB/feet. Transmission line losses are dependent on cable type, operating frequency and the length of the cable run. Insertion loss of a cable varies with frequency; the higher the frequency, the greater the loss.

Insertion loss measurements help troubleshoot the network by verifying the cable installation and cable performance. High insertion loss in the feedline or jumpers can contribute to poor system performance and loss of coverage. Measuring insertion loss using Site Master assures accurate and repeatable measurements.

Insertion Loss Measurement Setup

The insertion loss measurement set up for a typical transmission feed line system is shown in Figure 2. Remove the antenna and connect an enclosed precision "short" at the end of the transmission line.

If a Tower Mounted Amplifier (TMA) is in the transmission feed line system, remove the TMA and antenna and connect an enclosed short at the end of the transmission line. Insertion loss measurement for a transmission feed line system with a tower mounted amplifier is shown in Figure 3.

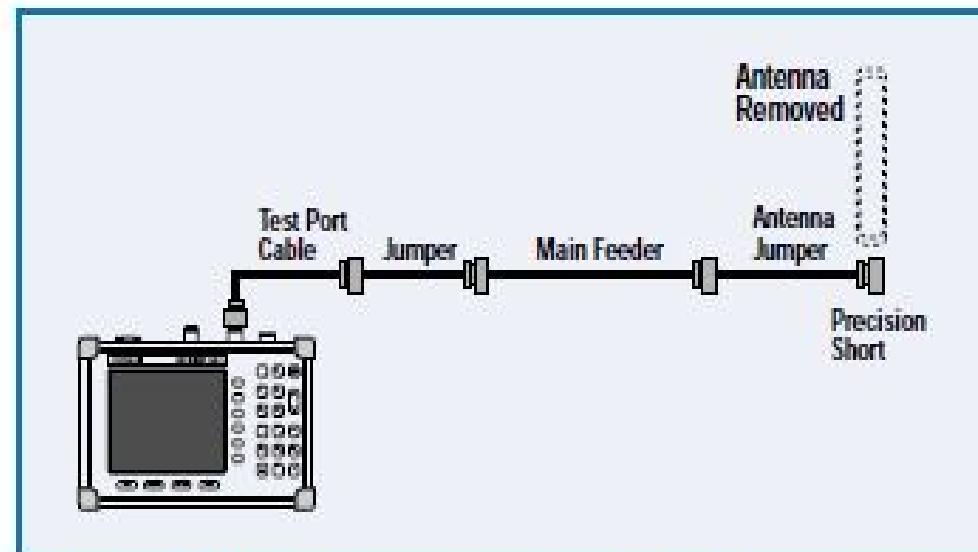
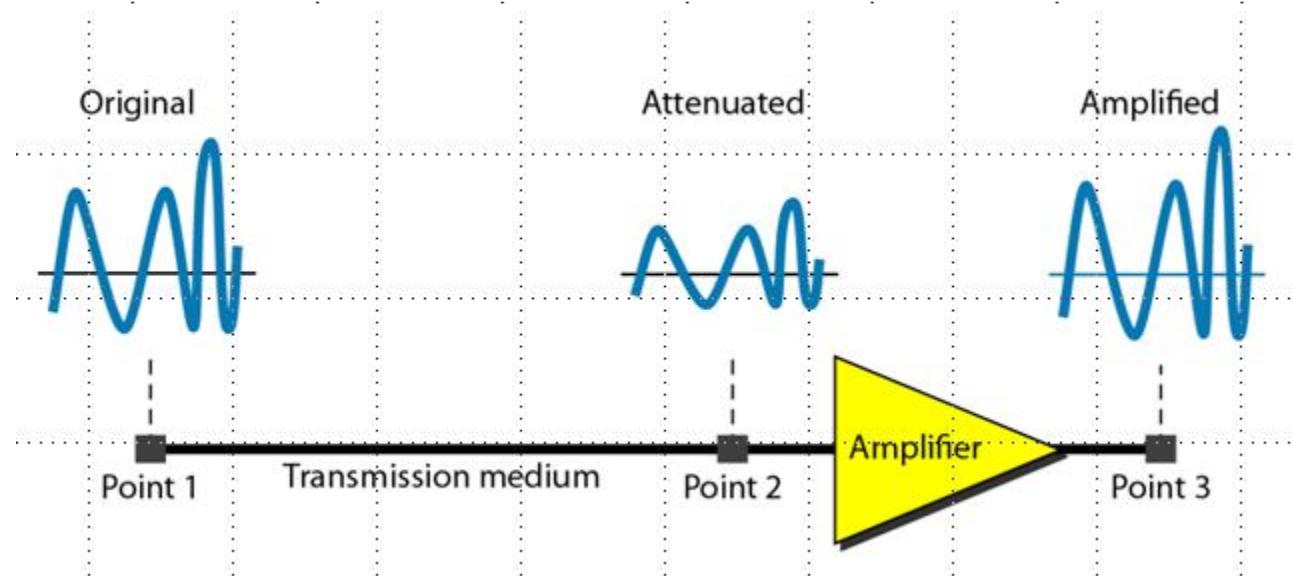


Figure 2. An insertion loss measurement setup after antenna is removed.

Attenuation Measurement

- Means loss of energy -> weaker signal
- When a signal travels through a medium it loses energy overcoming the resistance of the medium
- Amplifiers are used to compensate for this loss of energy by amplifying the signal.



Measurement of Attenuation

- To show the loss or gain of energy the unit "decibel" is used.

$$dB = 10 \log_{10} P_2/P_1$$

P_1 - input signal

P_2 - output signal

VSWR Measurement

- VSWR is defined as the ratio of the maximum voltage to the minimum voltage in standing wave pattern along the length of a transmission line structure. It varies from 1 to (plus) infinity and is always positive. Unless you have a piece of slotted line-test equipment this is a hard definition to use, especially since the concept of voltage in a microwave structure has many interpretations.
- Sometimes VSWR is called SWR to avoid using the term voltage and to instead use the concept of power waves. This in turn leads to a mathematical definition of VSWR in terms of a reflection coefficient. A reflection coefficient is defined as the ratio of reflected wave to incident wave at a reference plane. This value varies from -1 (for a shorted load) to +1 (for an open load), and becomes 0 for matched impedance load. It is a complex number. This helps us because we can actually measure power.
-

VSWR Measurement

- The reflection coefficient, commonly denoted by the Greek letter gamma (Γ), can be calculated from the values of the complex load impedance and the transmission line characteristic impedance which in principle could also be a complex number.
- $\Gamma = (Z_L - Z_0) / (Z_L + Z_0)$
- The square of $| \Gamma |$ is then the power of the reflected wave, the square hinting at a historical reference to voltage waves.
- Now we can define VSWR (SWR) as a scalar value:
- $VSWR = (1 + | \Gamma |) / (1 - | \Gamma |)$ or in terms of s-parameters: $VSWR = (1 + | S_{11} |) / (1 - | S_{11} |)$
- This is fine but what has it to do with common usage in ads and specifications. Generally, VSWR is sometimes used as a stand-in for a figure of merit for impedance matching. Sometimes this simplification of a scalar quantity and its restricted definition can lead to confusion in the matter of a source to load match. Most of the time there is no problem but, technically, VSWR derives from the ratio using the load impedance and the characteristic impedance of the transmission line in which the standing waves reside and not specifically to a source to load match. I prefer to think of VSWR as a figure of merit and to use the reflection coefficient whenever I am trying to solve problems.
- By the way, if you think you have never experienced a standing wave personally, it's very unlikely. Standing waves in a microwave oven are the reason that food is cooked unevenly (the turntable is a partial solution to that problem). The wavelength of the 2.45 GHz signal is about 12 centimeters, or about five inches. Nulls in the radiation (and heating) will be separated at a distance similar to wavelength.

FREQUENCY MEASUREMENT

- The frequency meter used has a cavity which is coupled to the waveguide by a small coupling hole which is used to absorb only a tiny fraction of energy passing along the waveguide.
- Adjusting the micrometer of the Frequency Meter will vary the plunger into the cavity. This will alters the cavity size and hence the resonance frequency.
- The readings on the micrometer scales are calibrated against frequency. As the plunger enters the cavity, its size is reduced and the frequency increases.

- The wavemeter is adjusted for maximum or minimum power meter readings depending on whether the cavity is a transmission or absorption type device. With the transmission-type device, the power meter will be adjusted for a maximum. It only allows frequency close to resonance to be transmitted through them. Other frequencies are reflected down the waveguide. The wavemeter acts as a short circuit for all other frequencies.
- For the absorption-type wavemeter, the power meter will be adjusted for a minimum. Its absorb power from the line around resonant frequency and act as a short to other frequencies.
- The absorbing material used is to absorb any unwanted signal that will cause disturbance to the system.

VSWR (VOLTAGE STANDING WAVE RATIO)

MEASUREMENT

- Used to determine the degree of mismatch between the source and load when the value $\text{VSWR} \neq 1$.
- Can be measured by using a slotted line. **Direct Method Measurement** is used for VSWR values upto about 10. Its value can be read directly using a standing wave detector .
- The measurement consists simply of adjusting attenuator to give an adequate reading, making sure that the frequency is correct and then using the dc voltmeter to measure the detector output at a maximum on the slotted section and then at the nearest minimum.

The ratio of the voltage maximum to the minimum gives the VSWR i.e

$$\mathbf{VSWR} = \mathbf{V}_{\max} / \mathbf{V}_{\min}$$

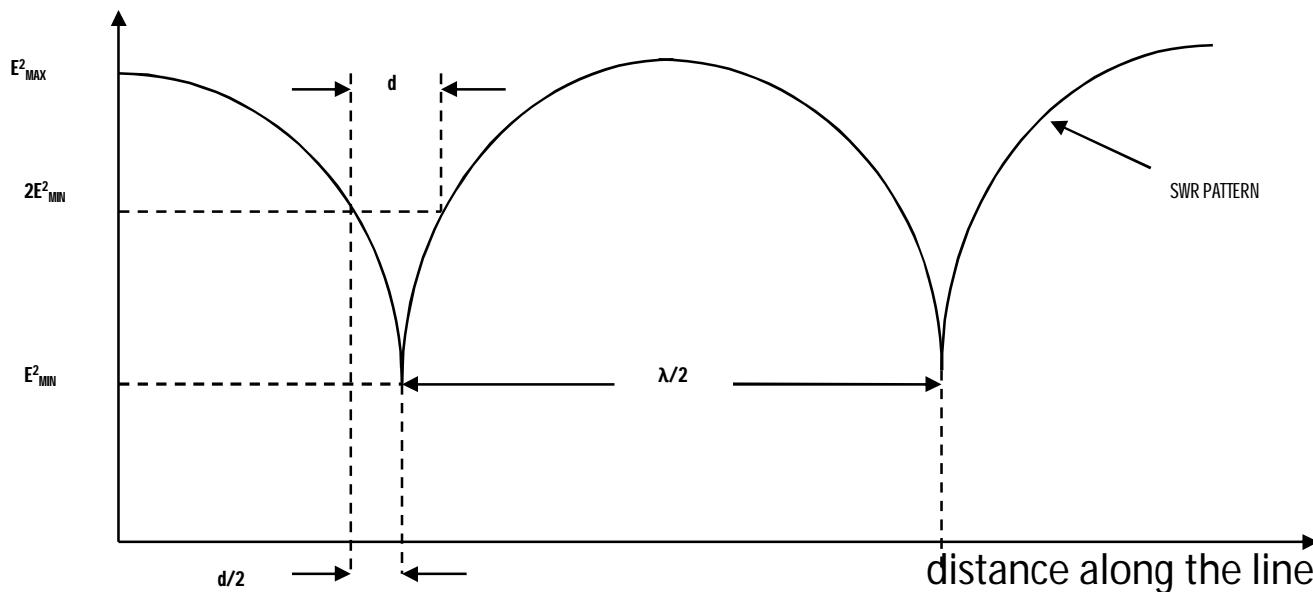
$$\begin{aligned}\mathbf{ISWR} &= \mathbf{I}_{\max} / \mathbf{I}_{\min} \\&= k (V_{\max})^2 / k (V_{\min})^2 \\&= (V_{\max} / V_{\min})^2 \\&= \mathbf{VSWR}^2\end{aligned}$$

$$\boxed{\mathbf{VSWR} = \sqrt{(\mathbf{I}_{\max} / \mathbf{I}_{\min})} = \sqrt{\mathbf{ISWR}}}$$

- Methods used depends on the value of VSWR whether it is high or low. If the load is not exactly matched to the line, standing wave pattern is produced.
- Reflections can be measured in terms of voltage, current or power. Measurement using voltage is preferred because it is simplicity.
- When reflection occurred, the incident and the reflected waves will reinforce each other in some places, and in others they will tend to cancel each other out.

DOUBLE MINIMUM METHOD MEASUREMENT (VSWR > 10)

- 'Double Minimum' method is usually employed for VSWR values greater than about 10.



- The detector output (proportional to field strength squared) is plotted against position. The probe is moved along the line to find the minimum value of signal.
- It is then moved either side to determine 2 positions at which twice as much detector signal is obtained. The distance d between these two positions then gives the VSWR according to the formula :

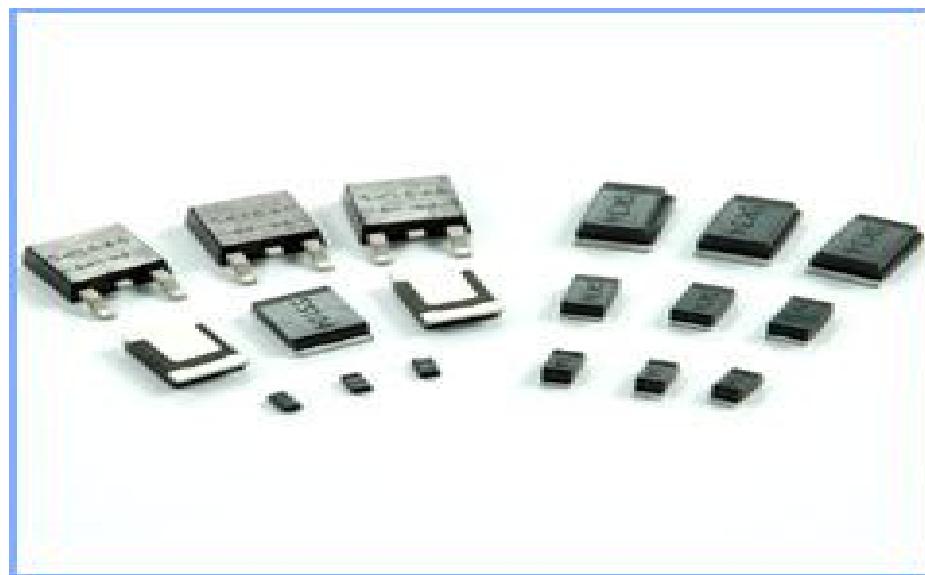
$$S = \sqrt{1 + 1/\sin^2(\pi d/\lambda)}$$

POWER MEASUREMENT

- Power is defined as the quantity of energy dissipated or stored per unit time.
- Methods of measurement of power depend on the frequency of operation, levels of power and whether the power is continuous or pulsed.
- The range of microwave power is divided into three categories :-
 - i. Low power ($< 10\text{mW}$ @ 0dBm)
 - ii. Medium power (from $10\text{ mW} - 10\text{ W}$ @ $0 - 40\text{ dBm}$)
 - iii. High power ($> 10\text{ W}$ @ 40 dBm)
- The microwave power meter consists of a power sensor, which converts the microwave power to heat energy.
- The sensors used for power measurements are the Schottky barrier diode, bolometer and the thermocouple.

SCHOTTKY BARRIER DIODE

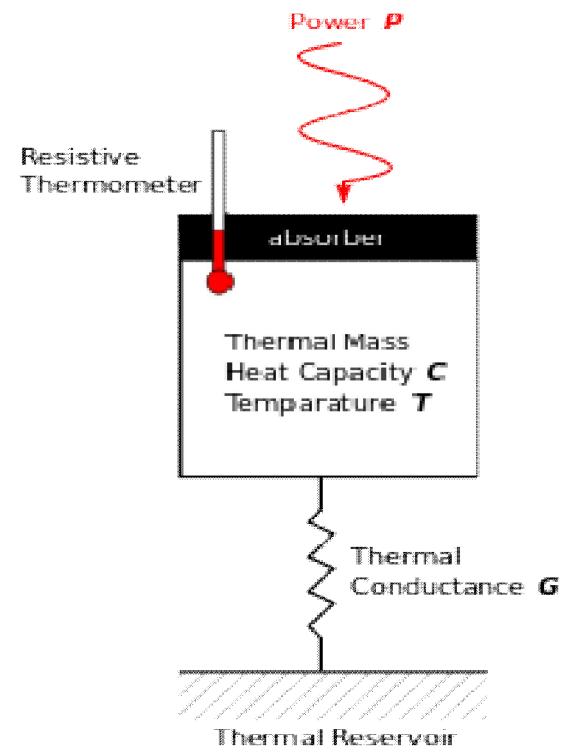
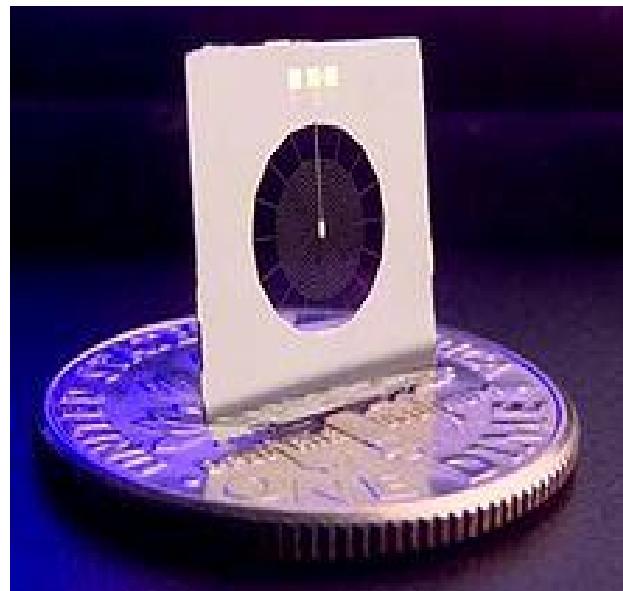
- A zero-biased Schottky Barrier Diode is used as a square-law detector whose output is proportional to the input power.
- The diode detectors can be used to measure power levels as low as 70dBm.



BOLOMETERS

- A Bolometer is a power sensor whose resistance changes with temperature as it absorbs microwave power.
- Are power detectors that operate on thermal principles. Since the temperature of the resistance is dependent on the signal power absorbed, the resistance must also be in proportion to the signal power.
- The two most common types of bolometer are, the barretter and the thermistor. Both are sensitive power detectors and is used to indicate microwatts of power. They are used with bridge circuits to convert resistance to power using a meter or other indicating devices.

BOLOMETER



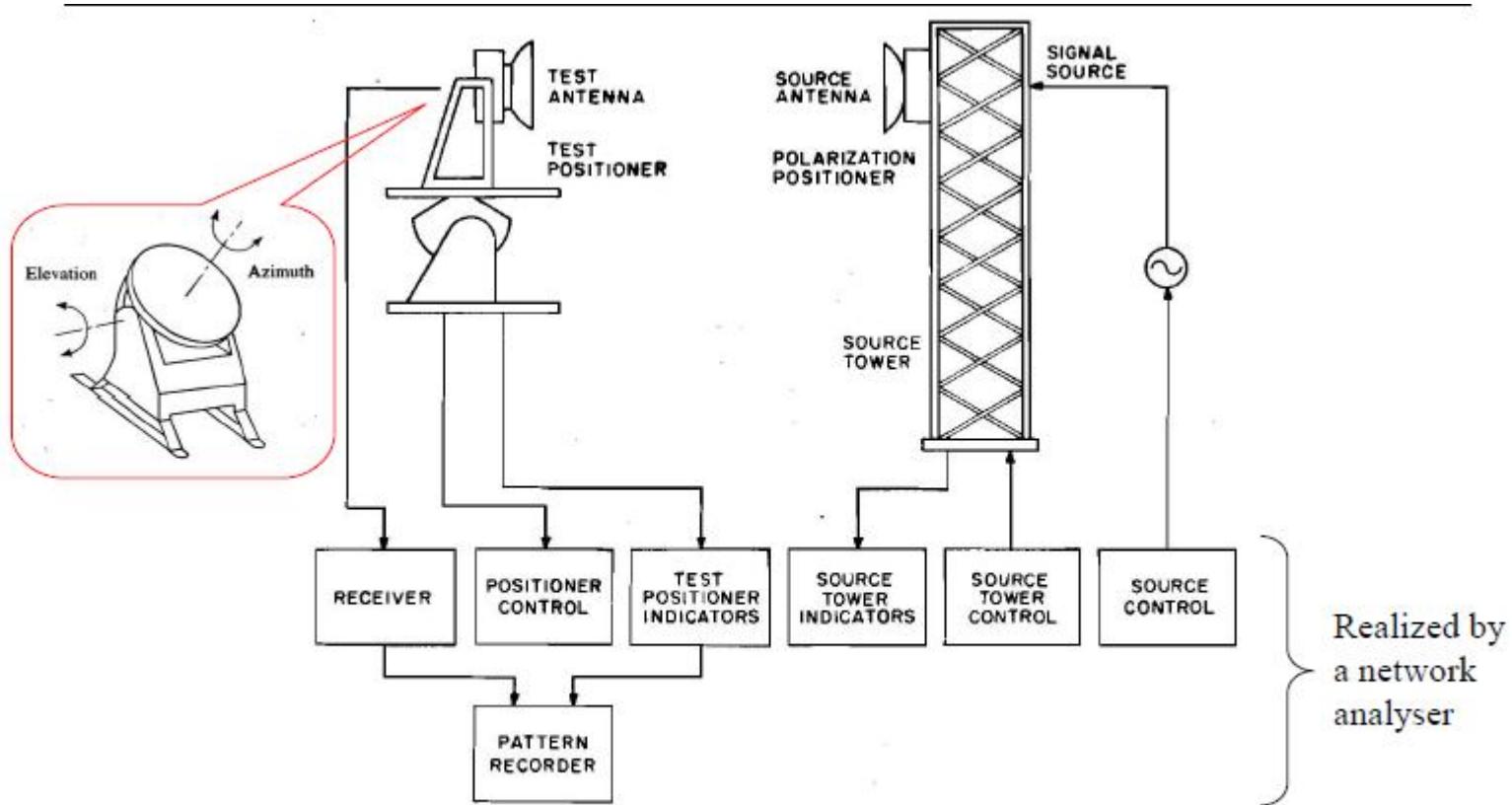
Antenna Measurement

1 Antenna Ranges

An **antenna range** is a facility where antenna radiation characteristics are measured. An antenna range includes the following typical components:

1. A substantial space for hosting the test antenna and the source antenna
2. A source antenna
3. An antenna positioner
4. A transmitter and receiver system (e.g. a Network Analyser)

Antenna Measurement



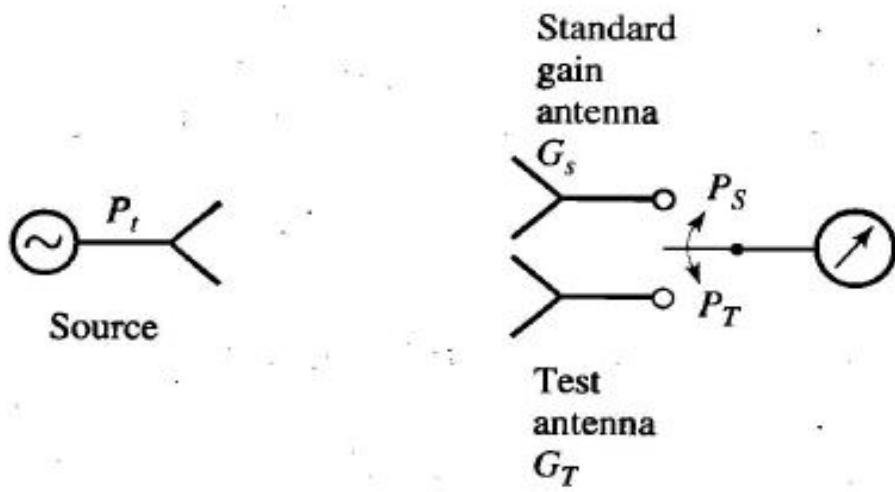
Block diagram of a typical antenna-measurement system

Gain Measurement

Gain Measurement

Comparison Method

The gain of an antenna can be measured by the comparison method using a **standard gain antenna** whose **gain** and **reflection coefficient** are known accurately. The power received by the standard gain antenna and the test antenna are measured, respectively, under the same conditions.



Radiation Pattern

The radiation pattern of an antenna is, generally, its most basic requirement since it determines the spatial distribution of the radiated energy. This is usually the first property of an antenna that is specified, once the operating frequency has been stated. An *antenna radiation pattern* or *antenna pattern* is defined as a graphical representation of the radiation properties of the antenna as a function of space coordinates. Since antennas are commonly used as parts of wireless telecommunication systems, the radiation pattern is determined in the far-field region where no change in pattern with distance occurs. Using a spherical coordinate system, shown in Fig. 1, where the antenna is at the origin, the radiation properties of the antenna depend only on the angles ϕ and θ along a path or surface of constant radius. A trace of the radiated or received power at a constant radius is called a *power pattern*, while the spatial variation of the electric or magnetic field along a constant radius is called an *amplitude field pattern*. In practice, the necessary information from the complete three-dimensional pattern of an antenna can be received by taking a few two-dimensional patterns, according to the complexity of radiation pattern of the specific antenna. Usually, for most applications, a number of plots of the pattern as a function of θ for some particular values of ϕ , plus a few plots as a function of ϕ for some particular values of θ , give the needed information.

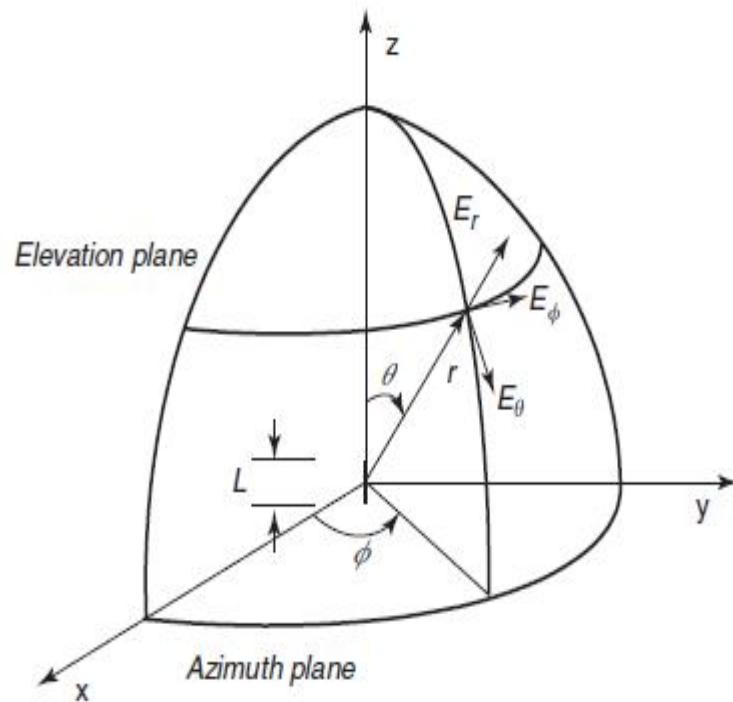
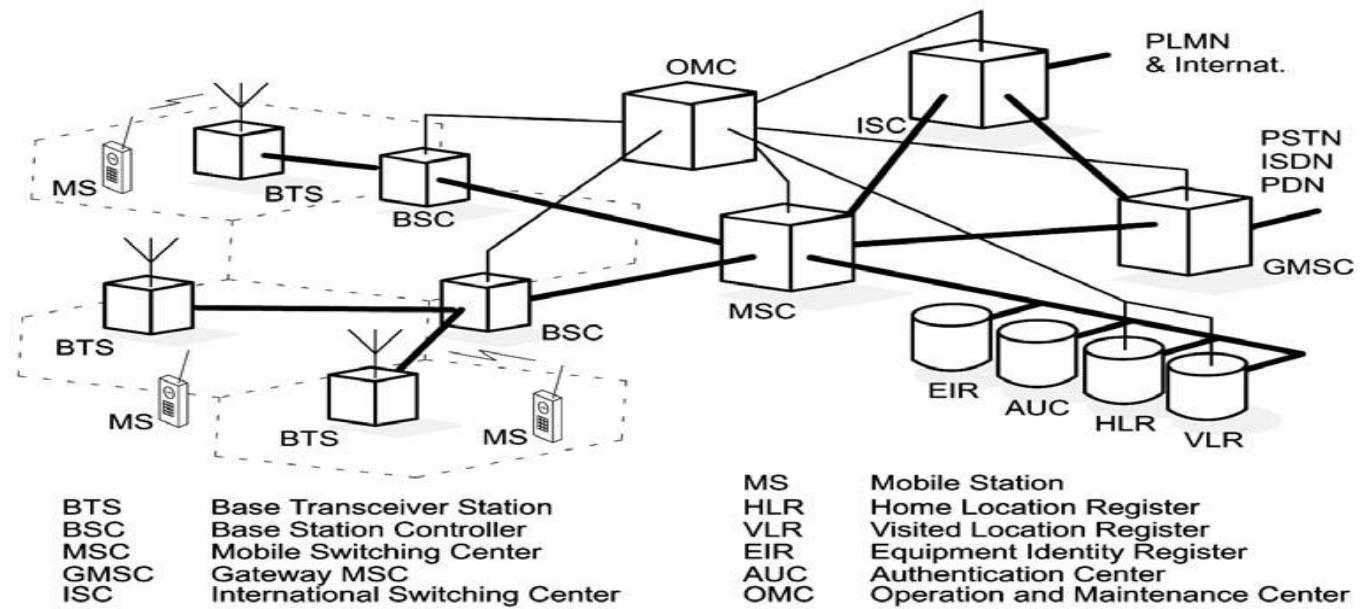


Figure 1. Spherical coordinate system for antenna analysis purposes. A very short dipole is shown with its non-zero field component directions.

Module 3

PART A => System architecture and addressing

Introduction: GSM (Global System for Mobile Communications, originally *Groupe Spécial Mobile*) is a standard developed by the European Telecommunications Standards Institute (ETSI) to describe the protocols for second-generation digital cellular networks used by mobile devices such as tablets, first deployed in Finland in December 1991.^[2] As of 2014, it has become the global standard for mobile communications – with over 90% market share, operating in over 193 countries and territories. 2G networks developed as a replacement for first generation (1G) analog cellular networks, and the GSM standard originally described as a digital, circuit-switched network optimized for full duplex voice telephony. This expanded over time to include data communications, first by circuit switched transport, then by packet data transport via GPRS (General Packet Radio Services) and EDGE (Enhanced Data rates for GSM Evolution, or EGPRS). Subsequently, the 3GPP developed third-generation (3G) UMTS standards, followed by fourth-generation (4G) LTE Advanced standards, which do not form part of the ETSI GSM standard.



1.1 GSM system architecture

The fundamental components of a GSM network are shown above. A user carries a Mobile Station (MS), which can communicate over the air with a base station, called Base Transceiver Station (BTS) in GSM. The BTS contains transmitter and receiver equipment, such as antennas and amplifiers, as well as a few components for signal and protocol processing. For example, error protection coding is performed in the BTS, and the link-level protocol for signaling on the radio path is terminated here.

In order to keep the base stations small, the essential control and protocol intelligence resides in the Base Station Controller (BSC). It contains, for example, protocol functions for radio channel allocation, channel setup and management of handovers. Typically, several BTSs are controlled by one BSC. In practice, the BTS and BSC are connected by fixed lines or point-to-point radio links. BTS and BSC together form the radio access network. The combined traffic of the users is routed through a switch, called the Mobile Switching Center (MSC). It performs all of the switching functions of a switching node in a fixed telephone network, e.g., in an Integrated Services Digital Network (ISDN). This includes path search, data forwarding and service feature processing.

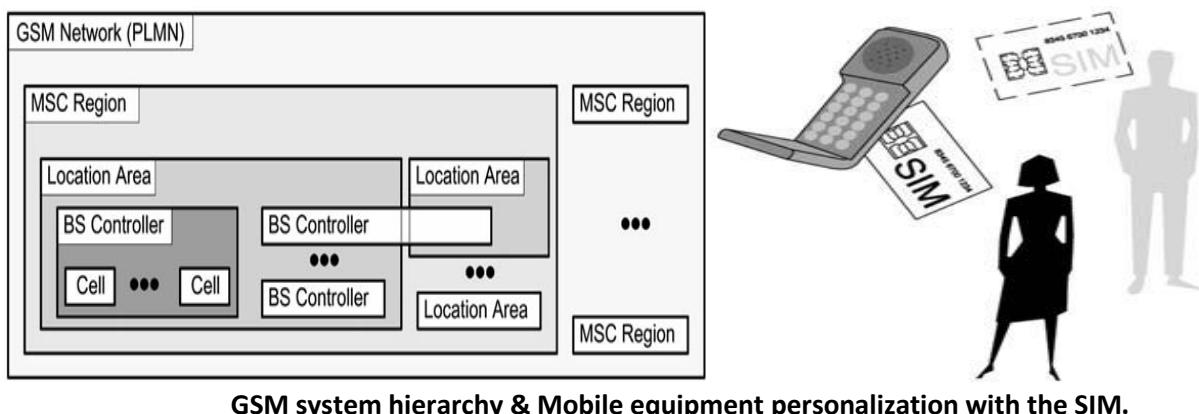
The main difference between an ISDN switch and an MSC is that the MSC also has to consider the allocation and administration of radio resources and the mobility of the users. The MSC therefore has to provide additional functions for location registration of users and for the handover of a connection in the case of changing from cell to cell. A cellular network can have several MSCs with each being responsible for a part of the network (e.g., a city or metropolitan area). Calls originating from or terminating in the fixed network are handled by a dedicated

Gateway MSC (GMSC). The interworking of a cellular network and a fixed network (e.g., PSTN, ISDN) is performed by the Interworking Function (IWF). It is needed to map the protocols of the cellular network onto those of the respective fixed network. Connections to other mobile or international networks are typically routed over the International Switching Center (ISC) of the respective country.

A GSM network also contains several types of databases. The Home Location Register (HLR) and the Visited Location Register (VLR) store the current location of a mobile user. This is needed since the network must know the current cell of a user to establish a call to the correct base station. In addition, these registers store the profiles of users, which are required for charging and billing and other administrative issues. Two further databases perform security functions: the Authentication Center (AUC) stores security-related data such as keys used for authentication and encryption; the Equipment Identity Register (EIR) registers equipment data rather than subscriber data. The network management is organized from a central place, the Operation and Maintenance Center (OMC). Its functions include the administration of subscribers, terminals, charging data, network configuration, operation, and performance monitoring and network maintenance. The operation and maintenance functions are based on the concept of the Telecommunication Management Network (TMN) which is standardized in the ITU-T series M.30.

A GSM network can be divided into three subnetworks: the radio access network, the core network and the management network. These subnetworks are called subsystems in the GSM standard. The respective three subsystems are called the Base Station Subsystem (BSS), the Network Switching Subsystem (NSS) and the Operation and Maintenance Subsystem (OMSS).

Figure Below summarizes the hierarchical relationship between the network components MSC, BSC and BTS. The entire network is divided into MSC regions. Each of these is composed of at least one Location Area (LA), which in turn consists of several cell groups. Each cell group is assigned to a BSC. For each LA there exists at least one BSC, but cells of one BSC may belong to different LAs. The exact partitioning of the network area with respect to LAs, BSCs and MSCs is not, however, uniquely determined and is left to the network operator who thus has many possibilities for optimization.



1.2 The SIM concept

Each GSM user owns a personal chip card, the **Subscriber Identity Module (SIM)**. As illustrated in previous right side Figure, it can be plugged into a piece of mobile equipment. In fact, only the SIM of a subscriber turns a piece of mobile equipment into a complete mobile station with network usage privileges, which can be used to make calls or receive calls.

This concept allows us to distinguish between equipment mobility and subscriber mobility. The subscriber can register to the locally available network with their SIM card on different mobile stations, or the SIM card could be used as a normal telephone card in the fixed telephone network. This enables international roaming independent of mobile equipment and network technology, provided that the interface between SIM and end terminal is standardized.

Beyond that, the SIM can store short messages and charging information, and it has a telephone book function and short list of call numbers storing names and telephone numbers for efficient and fast number selection. These functions, in particular, contribute to a genuine personalization of a mobile terminal, since the subscriber can use their normal 'environment' plus telephone list and short message archive with any piece of mobile equipment. In addition to subscriber-specific data, the SIM can also store network-specific data, e.g., lists of carrier frequencies used by the network to broadcast system information periodically.

Use of the SIM and thus of the whole MS can be protected with a Personal Identification Number (PIN) against unauthorized access. The SIM also takes over security functions: all of the cryptographic algorithms to be kept confidential are realized on the SIM, which implements important functions for the authentication and user data encryption based on the subscriber identity and secret keys.

1.3 Addressing

As in each communication network, the entities of a GSM network must be assigned certain addresses or identities. These serve to identify, authenticate and localize the network entities. The most commonly known GSM address is the telephone number of a user. In addition to telephone numbers, several other identifiers have been defined; they are needed for the management of user mobility and for addressing all remaining network elements. GSM distinguishes explicitly between a user and their equipment. Hence, there are specific address types for users and specific address types for MSs. The user identities are stored on the SIM; the equipment identities on the mobile equipment. In addition, GSM distinguishes between user identity and their telephone number. This leaves some scope for development of services when each subscriber may be called personally, independent of reachability or type of connection (mobile or fixed). In addition to the personal identifier, each GSM subscriber is assigned one or several ISDN numbers.

The following sections explain the most important addresses and identifiers used in GSM.

1.3.1 International mobile station equipment identity: The International Mobile Station Equipment Identity (IMEI) uniquely identifies a mobile station internationally and gives clues about its manufacturer and the date of manufacturing. It is a kind of serial number. The IMEI is allocated by the equipment manufacturer and registered by the network operator, who stores it in the EIR. By means of the IMEI one recognizes obsolete, stolen or nonfunctional equipment and can deny service if required.

For this purpose, the IMEI is assigned to one or more of the following three categories within the EIR.

- The white list is a register of all equipment.
- The black list contains all suspended equipment. This list is periodically exchanged among network operators.
- Optionally, an operator may maintain a gray list, in which malfunctioning equipment or equipment with obsolete software versions is registered. Such equipment has network access, but its use is reported to the operating personnel.

The IMEI is usually requested by the network at registration, but it can be requested repeatedly. It is a hierarchical address, containing the following parts:

- Type Approval Code (TAC), six digits, centrally assigned;
- Final Assembly Code (FAC), six digits, assigned by the manufacturer;
- Serial Number, six digits, assigned by the manufacturer;
- Spare, one digit.

1.3. 2. International mobile subscriber identity

When registering for service with a mobile network operator, each subscriber receives a unique identifier, the International Mobile Subscriber Identity (IMSI). This IMSI is stored in the SIM. A mobile station can only be operated if a SIM with a valid IMSI is inserted into equipment with a valid IMEI, since this is the only way to correctly bill the associated subscriber.

The IMSI uses a maximum of 15 decimal digits and consists of three parts:

- Mobile Country Code (MCC), three digits, internationally standardized;
- Mobile Network Code (MNC), two digits, for unique identification of mobile networks within a country;
- Mobile Subscriber Identification Number (MSIN), maximum of 10 digits, identification number of the subscriber in their mobile home network.

The IMSI is a GSM-specific addressing concept and is different from the ISDN numbering plan. A three-digit MCC has been assigned to each of the GSM countries, and two-digit MNCs have been assigned within countries (e.g., 262 as MCC for Germany; and MNC 01, 02 and 07 for the networks of T-Mobile, Vodafone, and O2, respectively). Whereas the MCC is defined internationally, the National Mobile Subscriber Identity (NMSI = MNC + MSIN) is assigned by the operator of the home network.

1.3.3. Mobile subscriber ISDN number

The ‘real telephone number’ of a mobile user is called the Mobile Subscriber ISDN Number (MSISDN). It is assigned to the subscriber (their SIM), such that a mobile station can have several MSISDNs depending on the SIM. With this concept, GSM was the first mobile system to distinguish between subscriber identity and the number to call. The separation of call number (MSISDN) and subscriber identity (IMSI) primarily serves to protect the confidentiality of the IMSI. In contrast to the MSISDN, the IMSI need not be made public. With this separation, one cannot derive the subscriber identity from the MSISDN, unless the association of IMSI and MSISDN as stored in the HLR has been made public. It is the rule that the IMSI used for subscriber identification is not known, and thus the faking of a false identity is significantly more difficult.

In addition, a subscriber can hold several MSISDNs for selection of different services. Each MSISDN of a subscriber is reserved for specific service (voice, data, fax, etc.). In order to realize this service, service-specific resources have to be activated in the MS as well as in the network. The service desired and the resources needed for the specific call can be derived from the MSISDN. Thus, an automatic activation of service-specific resources is already possible during the setup of a connection. The MSISDN categories follow the international ISDN numbering plan, having the following structure:

- Country Code (CC), up to three digits;
- National Destination Code (NDC), typically two or three digits;
- Subscriber Number (SN), a maximum of 10 digits.

The CCs are internationally standardized, complying with the ITU-T recommendation E.164. There are country codes with one, two, or three digits, e.g. the country code for the USA is 1, for the UK it is 44 and for Finland it is 358. The national operator or regulatory administration assigns the NDC as well as the SN, which may have variable length. The NDC of the mobile networks in Germany have three digits (e.g., 170, 171, and 172). The MSISDN is stored centrally in the HLR.

1.3.4. Mobile station roaming number

The Mobile Station Roaming Number (MSRN) is a temporary location-dependent ISDN number. It is assigned by the locally responsible VLR to each MS in its area. Calls are routed to the MS by using the MSRN. On request, the MSRN is passed from the HLR to the GMSC.

The MSRN has the same structure as the MSISDN:

- CC of the visited network;
- NDC of the visited network;
- SN in the current mobile network.

The components CC and NDC are determined by the network visited and depend on the current location. The SN is assigned by the current VLR and is unique within the mobile network. An MSRN is assigned in such a way that the currently responsible switching node MSC in the visited network can be determined from the subscriber number, which allows routing decisions to be made. The MSRN can be assigned in two different ways by the VLR: either at each registration when the MS enters a new LA or each time when the HLR requests it for setting up a connection for incoming calls to the MS.

In the first case, the MSRN is also passed on from the VLR to the HLR, where it is stored for routing. In the case of an incoming call, the MSRN is first requested from the HLR of this MS. In this way the currently responsible MSC can be determined, and the call can be routed to this switching node. Additional localization information can be obtained there from the responsible VLR.

In the second case, the MSRN cannot be stored in the HLR, since it is only assigned at the time of call setup. Therefore, the address of the current VLR must be stored in the tables of the HLR. Once routing information is requested from the HLR, the HLR itself goes to the current VLR and uses a unique subscriber identification (IMSI and MSISDN) to request a valid roaming number MSRN. This allows further routing of the call.

1.3.5. Location area identity

Each LA of a cellular network has its own identifier. The Location Area Identifier (LAI) is also structured hierarchically and internationally unique, with LAI again consisting of an internationally standardized part and an operator-dependent part:

- CC, three digits;
- MNC, two digits;
- Location Area Code (LAC), a maximum of five digits or a maximum of 2×8 bits, coded in hexadecimal.

This LAI is broadcast regularly by the base station on the Broadcast Control Channel (BCCH). Thus, each cell is identified uniquely on the radio channel as belonging to an LA, and each MS can determine its current location through the LAI. If the LAI that is 'heard' by the MS changes, the MS notices this LA change and requests an update to its location information in the VLR and HLR (location update).

The significance for GSM networks is that the MS itself rather than the network is responsible for monitoring the local conditions of signal reception, to select the base station that can be received best, and to register with the VLR of that LA which the current base station belongs to. The LAI is requested from the VLR if the connection for an incoming call has been routed to the current MSC using the MSRN. This determines the precise location of the MS where the mobile can be subsequently paged.

When the MS answers, the exact cell and therefore also the base station become known; this information can then be used to switch the call through.

1.3.6. Temporary mobile subscriber identity

The VLR responsible for the current location of a subscriber can assign a Temporary Mobile Subscriber Identity (TMSI), which has only local significance in the area handled by the VLR. It is used in place of the IMSI for the definite identification and addressing of the MS. In this way nobody can determine the identity of the subscriber by listening to the radio channel, since this TMSI is only assigned during the presence of the MS in the area of one VLR, and can even be changed during this period (ID hopping).

The MS stores the TMSI on the SIM card. The TMSI is stored on the network side only in the VLR and is not passed to the HLR. A TMSI may therefore be assigned in an operator-specific way; it can consist of up to 4×8 bits, but the HEX value FFFF FFFF is excluded, because the SIM marks empty fields internally with logical 1. Together with the current location area, a TMSI allows a subscriber to be identified uniquely, i.e., for the ongoing communication the IMSI is replaced by the 2-tuple (TMSI, LAI).

1.3.7. Other identifiers

The VLR can assign an additional searching key to each MS within its area to accelerate database access; this is the Local Mobile Station Identity (LMSI). The LMSI is assigned when the MS registers with the VLR and is also sent to the HLR. The LMSI is no longer used by the HLR, but each time messages are sent to the VLR concerning a MS, the LMSI is added, so the VLR can use the short searching key for transactions concerning this MS. This kind of additional identification is only used when the MSRN is newly assigned with each call. In this case, fast processing is very important to achieve short times for call setup. Like the TMSI, an LMSI is also assigned in an operator-specific way, and it is only unique within the administrative area of a VLR. An LMSI consists of four octets (4×8 bits).

Within an LA, the individual cells are uniquely identified with a Cell Identifier (CI), a maximum of 2×8 bits. Together with the global CI cells are thus also internationally defined in a unique way. In order to distinguish neighboring base stations, these receive a Base Transceiver Station Identity Code (BSIC) which consists of two components:

- Network Color Code (NCC), a color code within a mobile network (3 bits);
- Base Transceiver Station Color Code (BCC), a BTS color code (3 bits).

The BSIC is broadcast periodically by the base station. Directly adjacent mobile networks must have different NCCs, and neighboring base stations of a mobile network must have different BCCs. MSCs and location registers (HLR, VLR) are addressed with ISDN numbers. In addition, they may have a Signaling Point Code (SPC)

within a mobile network, which can be used to address them uniquely within the Signaling System Number 7 network (SS#7). The number of the VLR in whose area a MS is currently roaming must be stored in the HLR data for this MS, if the MSRN distribution is on a call-by-call basis; thus the MSRN can be requested for incoming calls and the call can be switched through to the MS.

1.4 Registers and subscriber data

1.4.1. Location registers (HLR and VLR)

The GSM standard defines two database types for the management of user data and location: the HLR and the VLR. These databases are queried by the network for user registration and localization. The HLR has a record for all subscribers registered with a network operator. It stores, for example, each user's telephone number, service subscriptions, permissions and authentication data. In addition to this permanent administrative data, it also contains temporary data, such as the current location of a subscriber. In the case of incoming traffic to a mobile user, the HLR is queried to determine the user's current location. This enables the gateway to route the traffic to the appropriate MSC.

The MS must inform the network about its current location area; to do so it sends a location update message to the network whenever it changes its location area. The full list of subscriber data stored in the HLR is given in Table below. A VLR is responsible for a group of location areas and stores the data of all users that are currently located in this area. The data includes part of the permanent user data, which is copied from the HLR to the VLR for fast access.

In addition, the VLR may also assign and store local data, such as temporary identifiers. A user may either be registered with a VLR of their home network or a foreign network. Upon a location update, the MSC forwards the user's identity and current location to the VLR, which subsequently updates its database. If the user has not been registered with this VLR before, the HLR is informed about the current VLR of the user. This process enables incoming calls to be routed to this MS. Table 2 summarizes the subscriber data stored in the VLR. Typically, there is one central HLR per network and one VLR for each MSC. This organization depends on the number of subscribers, the processing and storage capacity of the switches and the structure of the network.

Table 3.1 Mobile subscriber data in the HLR.

Subscriber and subscription data:	<ul style="list-style-type: none"> - IMSI - MSISDN - Service subscriptions - Service restrictions (e.g., roaming restrictions) - Information on the subscriber's equipment (if available) - Authentication data (subject to implementation)
Tracking and routing information:	<ul style="list-style-type: none"> - Mobile Station Roaming Number (MSRN) - Current VLR address (if available) - Current MSC address (if available) - Local Mobile Subscriber Identity (LMSI) (if available)

Table 3.2 Mobile subscriber data in the VLR.

Subscriber and subscription data:	<ul style="list-style-type: none"> - IMSI - MSISDN - Parameters for supplementary services - Information on the subscriber's equipment (if available) - Authentication data (subject to implementation)
Tracking and routing information:	<ul style="list-style-type: none"> - MSRN - TMSI - LMSI (if available) - LAI of LA where the MS was registered (used for paging and call setup)

1.4.2. Security-related registers (AUC and EIR)

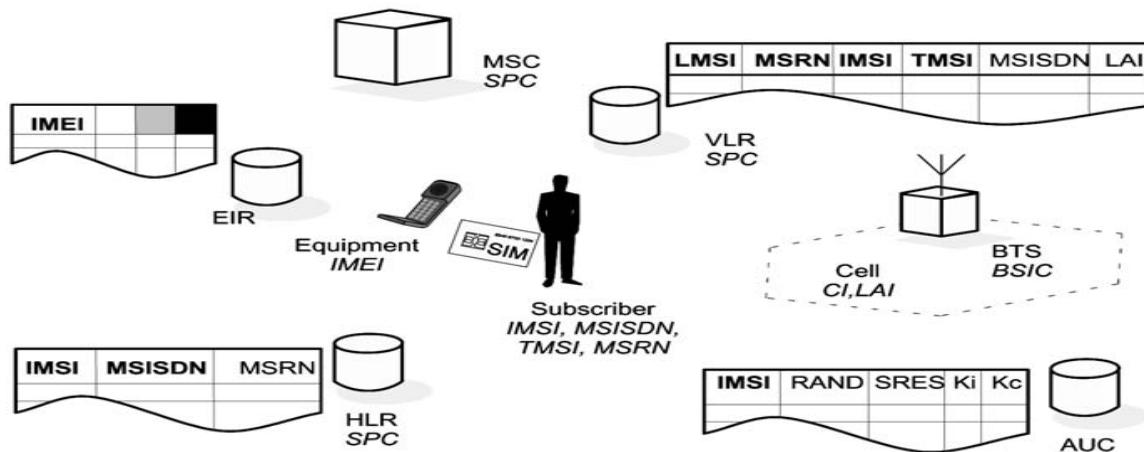
Two additional databases are responsible for various aspects of system security. System security of GSM networks is based primarily on the verification of equipment and subscriber identity; therefore, the databases serve for subscriber identification and authentication and for equipment registration. Confidential data and keys are stored or generated in the AUC. The keys serve for user authentication and authorize the respective service access. The EIR stores the serial numbers (supplied by the manufacturer) of the terminals (IMEI), which makes it possible to check for MSs with obsolete software or to block service access for MSs reported as stolen.

1.4.3. Subscriber data

Service-specific data are used to parameterize and personalize supplementary services. Finally, contracts with subscribers can define different service levels, e.g., booking of special supplementary services or subscriptions to data or teleservices. The contents of such contracts are stored in appropriate data structures in order to enable correct realization or provision of these services. The association of the most important identifiers and their storage locations is summarized in Figure below. Subscriber-related addresses are stored on the SIM and in the HLR and

VLR as well. These data (IMSI, MSISDN, TMSI, and MSRN) serve to address, identify and localize a subscriber or a MS. Whereas IMSI and MSISDN are permanent data items, TMSI and MSRN are temporary values, which change according to the current location of the subscriber. Of the other data items defined for user or network equipment elements (such as IMEI, LAI or SPCs), only some are used (LAI, SPC) for localizing or routing. IMEI and BSIC/CI hold a special position by being used only for identification of network elements.

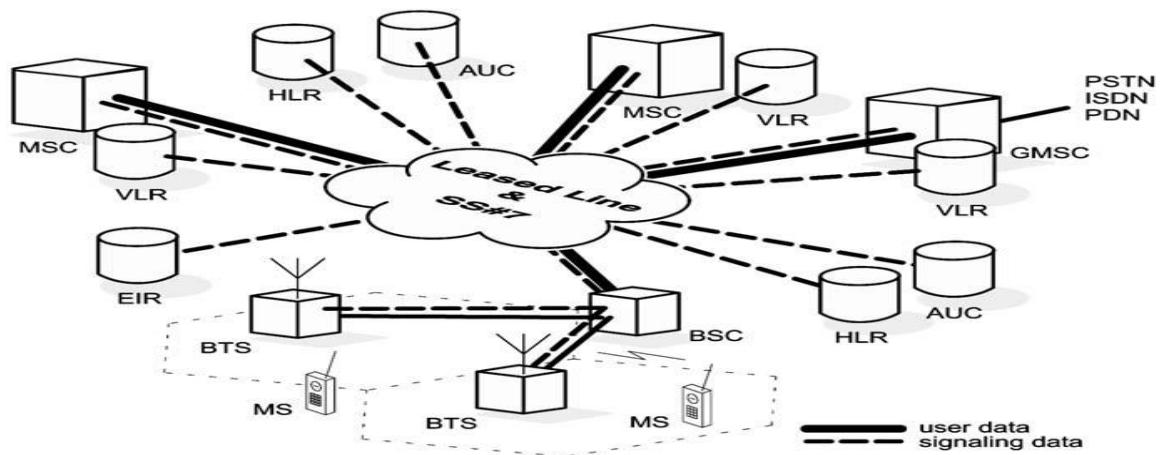
Security-relevant subscriber data are stored in the AUC, which also calculates identifiers and keys for cryptographic processing functions. Each set of data in the AUC contains the IMSI of the subscriber as a search key. For identification and authentication of a subscriber, the AUC stores the subscriber's secret key K_i from which a pair of keys RAND/SRES are pre calculated and stored. Once an authentication request occurs, this pair of keys is queried by the VLR to conduct the identification/authentication process properly. The key K_c for user data encryption on the radio channel is also calculated in advance in the AUC from the secret key K_i and is requested by the VLR at connection setup.



GSM databases and addresses

Above all, the HLR contains the permanent data about the subscriber's contractual relationship, e.g., information about subscribed bearer and teleservices (data, fax, etc.), service restrictions and parameters for supplementary services. Beyond that, the registers also contain information about equipment used by the subscriber (IMEI). Depending on the implementation of the authentication center and the security mechanisms, data and keys used for subscriber authentication and encryption can also be stored there. The search keys used for retrieving subscriber information (such as IMSI, MSISDN, MSRN, TMSI and LMSI) from a register are indicated in boldface (Figure above).

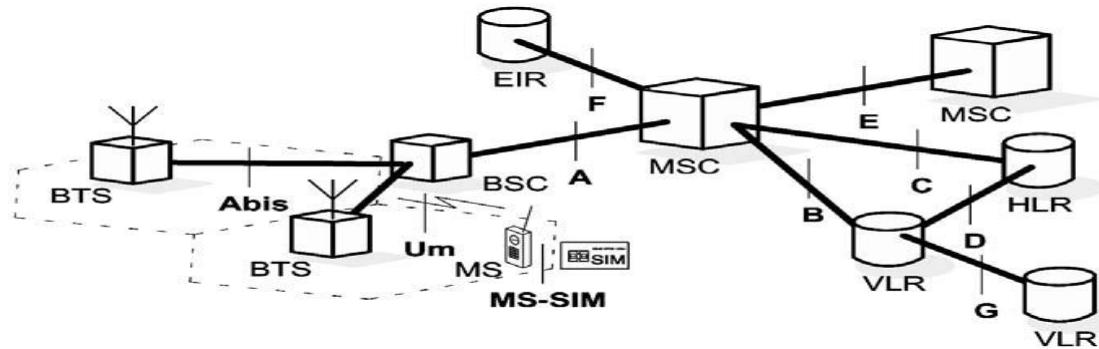
1.5 Network interfaces and configurations



User data transport and signaling in a GSM network.

The GSM system architecture with corresponding user data and signaling links between the network components is shown in below figure. Signaling has two fundamentally different parts: The core network employs the SS#7, which is well known from fixed networks. In order to setup, manage and release calls, the SS#7 protocol called ISDN User Part (ISUP) is used. In order to perform signaling that is specific to mobile networking, an extension to SS#7 has been developed, the so-called Mobile Application Part (MAP). It is implemented in the MSC, HLR and VLR. The radio access network (including the air interface) does not employ the SS#7 protocol, but uses a GSM-specific protocol. Signaling between the radio access network and the MSC uses the Base Station System Application Part (BSSAP).

1.5.1. Interfaces



Interfaces in a GSM network.

The communication relationships between the GSM networks components are formally described by a number of standardized interfaces (Figure above). The A interface between BSS and MSC is used for the transfer of data for BSS management, for connection control and for mobility management. Within the BSS, the Abis interface between BTS and BSC and the air interface um have been defined.

An MSC which needs to obtain data about an MS staying in its administrative area, requests the data from the VLR responsible for this area over the B interface. Conversely, the MSC forwards to this VLR any data generated at location updates by MSs. If the subscriber reconfigures special service features or activates supplementary services, the VLR is also informed first, which then updates the HLR.

This updating of the HLR occurs through the D interface. The D interface is used for the exchange of location-dependent subscriber data and for subscriber management. The VLR informs the HLR about the current location of the mobile subscriber and reports the current MSRN. The HLR transfers all of the subscriber data to the VLR that is needed to give the subscriber their usual customized service access. The HLR is also responsible for giving a cancellation request for the subscriber data to the old VLR once the acknowledgement for the location update arrives from the new VLR. If, during location updating, the new VLR needs data from the old VLR, it is directly requested over the G interface. Furthermore, the identity of subscriber or equipment can be verified during a location update; for requesting and checking the equipment identity, the MSC has an interface F to the EIR.

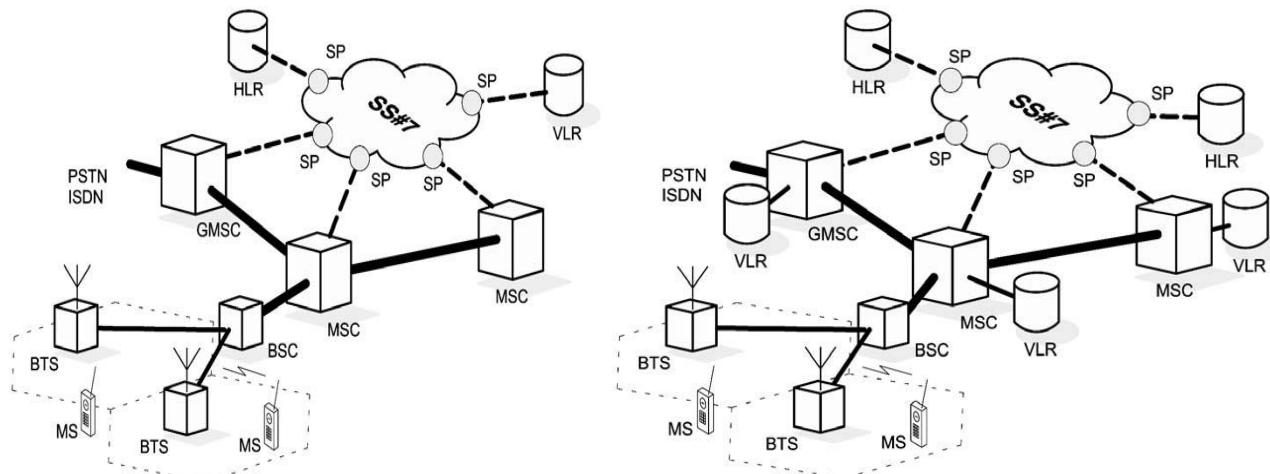
An MSC has two more interfaces in addition to the A and B interfaces, namely the C and E interfaces. Charging information can be sent over the C interface to the HLR. In addition to this, the MSC must be able to request routing information from the HLR during call setup, for calls from the mobile network as well as for calls from the fixed network. In the case of a call from the fixed network, if the fixed network's switch cannot interrogate the HLR directly, initially it routes the call to a GMSC, which then interrogates the HLR. If the mobile subscriber changes during a conversation from one MSC area to another, a handover needs to be performed between these two MSCs, which occurs across the E interface.

1.5.2. Configurations

As mentioned above, the configuration of a mobile network is largely left to the network operator. Figure on next page shows a basic configuration of a GSM network. This configuration contains a central HLR and a central VLR. All database transactions (updates, inquiries, etc.) and handover transactions between the MSC are performed with the help of the MAP over the SS#7 network. For this purpose, each MSC and register is known as a Signaling Point (SP) and is identified by its Signaling Point Code (SPC) within the SS#7 network. Whenever an MS changes its

location area, the location information in the VLR must be updated. Furthermore, the VLR has to be interrogated: the MSC needs subscriber parameters in addition to location data for successful connection setup, such as service restrictions and supplementary services to be activated. Thus, there is a significant message traffic between MSC and VLR, which constitutes an ensuing load on the signaling network. Hence, these two functional units can be combined into one physical unit, i.e. the entire VLR is implemented in distributed form and a VLR is associated with each MSC (see 2nd Figure). The traffic between MSC and VLR no longer needs to be transported through the SS#7 network.

We can go one step further and also distribute the database of the HLR, thus introducing several HLRs into a mobile network. This is especially interesting for a growing pool of subscribers, since a centralized database leads to a high traffic load for this database. If there are several HLRs in a network, the network operator has to define an association rule between MSISDNs and HLRs, such that for incoming calls the routing information to an MSISDN can be derived from the associated HLR. One possible association is geographic partitioning of the whole subscriber identification space (SN field in the MSISDN), where, for example, the first two digits of the SN indicate the region and the associated HLR. In extreme cases, the HLR can be realized with the VLR in a single physical unit. In this case, an HLR would also be associated with each MSC.



Basic configuration of a GSM network. & Configuration of a GSM network with a VLR for each MSC.

PART B => Air Interface – GSM Physical Layer

The GSM physical layer, which resides on the first of the seven layers of the Open Systems Interconnection (OSI) Reference Model (Tanenbaum, 1996), contains very complex functions. The physical channels are defined here by a TDMA scheme. On top of the physical channels, a series of logical channels are defined, which are transmitted in the time slots of the physical channels. Logical channels perform a multiplicity of functions, such as payload transport, signaling, and broadcast of general system information, synchronization and channel assignment.

The logical channels serve as a foundation for understanding the signaling procedures at the air interface. The realization of the physical channels, including GSM modulation, multiple access, duplexing and frequency hopping. Synchronization. The mapping of logical onto physical channels follows where the higher-level multiplexing of logical channels into multiframe is also covered. A discussion of the most important control mechanisms for the air interface (channel measurement, power control, disconnection and cell selection). The conclusion is that a power-up scenario with the sequence of events occurring, from when a MS is turned on to when it is in a synchronized state ready to transmit.

1. Logical channels

On Layer 1 of the OSI Reference Model, GSM defines a series of logical channels, which are made available either in an unassigned random access mode or in a dedicated mode assigned to a specific user. Logical channels are divided into two categories (Table below): **traffic channels and signaling (control) channels.**

Table 4.1 Classification of logical channels in GSM.

Group	Channel	Function	Direction
Traffic channel	TCH	TCH/F, Bm	MS ↔ BSS
		TCH/H, Lm	MS ↔ BSS
Signaling channels (Dm)	BCH	BCCH	MS ← BSS
		FCCH	MS ← BSS
		SCH	MS ← BSS
	CCCH	RACH	MS → BSS
		AGCH	MS ← BSS
		PCH	MS ← BSS
DCCH	NCH	Notification	MS ← BSS
	SDCCH	Stand-alone dedicated control	MS ↔ BSS
		SACCH	MS ↔ BSS
	FACCH	Fast associated control	MS ↔ BSS

1.1. Traffic channels

The Traffic Channels (TCHs) are used for the transmission of user payload data (speech, data). They do not carry any control information of Layer 3. Communication over a TCH can be circuit-switched or packet-switched. In the circuit-switched case, the TCH provides a transparent data connection or a connection that is specially treated according to the carried service (e.g. telephony). For the packet-switched mode, the TCH carries user data of OSI Layers 2 and 3 according to the recommendations of the X.25 standard or similar standard packet protocols.

A TCH may either be fully used (full-rate TCH, TCH/F) or be split into two half-rate channels (half-rate TCH, TCH/H), which can be allocated to different subscribers. Following ISDN terminology, the GSM traffic channels are also designated as Bm channel (mobile B channel) or Lm channel (lower-rate mobile channel, with half the bit rate). A Bm channel is a TCH for the transmission of bit streams of either 13 kbit/s of digitally coded speech or of data streams at 14.5, 12, 6 or 3.6 kbit/s. Lm channels are TCH channels with less transmission bandwidth than Bm channels and transport speech signals of half the bit rate (TCH/H) or bit streams for data services with 6 or 3.6 kbit/s.

1.2. Signaling channels

The control and management of a cellular network demands a very high signaling effort. Even when there is no active connection, signaling information (for example, location update information) is

permanently transmitted over the air interface. The GSM signaling channels offer a continuous, packet-oriented signaling service to MSs in order to enable them to send and receive messages at any time over the air interface to the BTS. Following ISDN terminology, the GSM signaling channels are also called Dm channels (mobile D channel). They are further divided into **Broadcast Channel (BCH)**, **Common Control Channel (CCCH)** and **Dedicated Control Channel (DCCH)** (see Table 4.1 previous page).

1. The unidirectional **BCHs** are used by the BSS to broadcast the same information to all MSs in a cell. The group of BCHs consists of three channels.

- Broadcast Control Channel (BCCH): On this channel, a series of information elements is broadcast to the MSs which characterize the organization of the radio network, such as radio channel configurations (of the currently used cell as well as of the neighboring cells), synchronization information (frequencies as well as frame numbering) and registration identifiers (LAI, CI, BSIC). In particular, this includes information about the structural organization (formats) of the CCCH of the local BTS. The BCCH is broadcast on the first frequency assigned to the cell (the so-called BCCH carrier).
- Frequency Correction Channel (FCCH): On the FCCH, information about correction of the transmission frequency is broadcast to the MSs; (frequency correction burst).
- Synchronization Channel (SCH): The SCH broadcasts information to identify a BTS, i.e. BSIC; see Chapter 3. The SCH also broadcasts data for the frame synchronization of a MS, i.e. Reduced Frame Number (RFN) of the TDMA frame;

FCCH and SCH are only visible within protocol Layer 1, since they are only needed for the operation of the radio subsystem. There is no access to them from Layer 2. In spite of this fact, the SCH messages contain data which are needed by Layer 3 for the administration of radio resources. These two channels are always broadcast together with the BCCH.

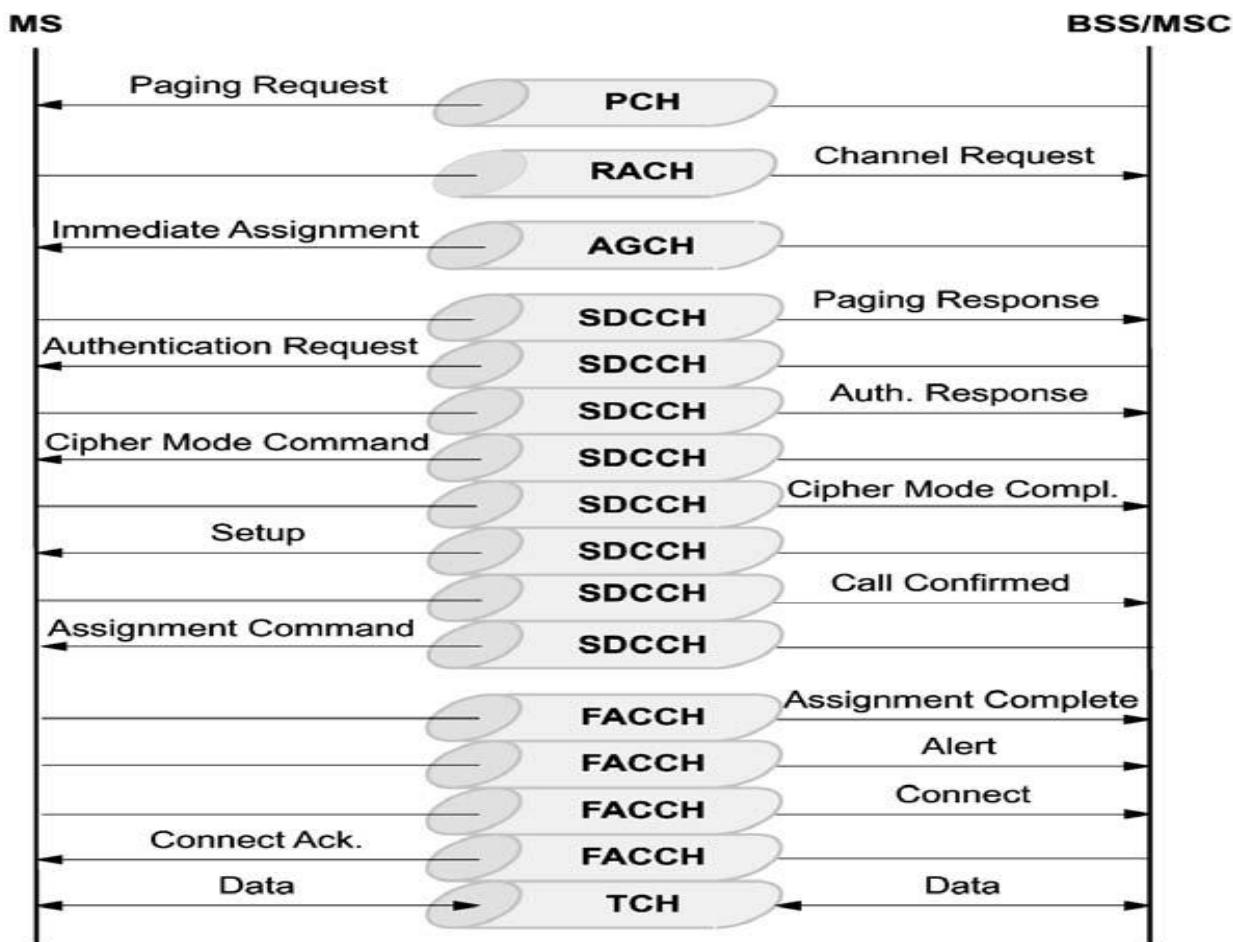
2. The **CCCH** is a point-to-multipoint signaling channel to deal with access management functions. This includes the assignment of dedicated channels and paging to localize a MS.

It comprises the following.

- Random Access Channel (RACH): The RACH is the uplink portion of the CCCH. It is accessed from the mobile stations in a cell without reservation in a competitive multiple-access mode using the principle of slotted Aloha (Bertsekas and Gallager, 1987), to ask for a dedicated signaling channel for exclusive use by one MS for one signaling transaction.
- Access Grant Channel (AGCH): The AGCH is the downlink part of the CCCH. It is used to assign an SDCCH or a TCH to a MS.
- Paging Channel (PCH): The PCH is also part of the downlink of the CCCH. It is used for paging to find specific MSs.
- Notification Channel (NCH): The NCH is used to inform MSs about incoming group and broadcast calls.

3. The last type of signaling channel, **the DCCH** is a bidirectional point-to-point signaling channel. An Associated Control Channel (ACCH) is also a dedicated control channel, but it is assigned only in connection with a TCH or an SDCCH. The group of Dedicated/Associated Control Channels (D/ACCH) comprises the following.

- Stand-alone Dedicated Control Channel (SDCCH): The SDCCH is a dedicated point-to-point signaling channel (DCCH) which is not tied to the existence of a TCH ('standalone'), i.e. it is used for signaling between a MS and the BSS when there is no active connection. The SDCCH is requested from the MS via the RACH and assigned via the AGCH. After the completion of the signaling transaction, the SDCCH is released and can be reassigned to another MS. Examples of signaling transactions which use an SDCCH are the updating of location information or parts of the connection setup until the connection is switched through (see Figure next page).



Logical channels and signaling (connection setup for an incoming call).

- Slow Associated Control Channel (SACCH): An SACCH is always assigned and used with a TCH or an SDCCH. The SACCH carries information for the optimal radio operation, e.g., commands for synchronization and transmitter power control and reports on channel measurements. Data must be transmitted continuously over the SACCH since the arrival of SACCH packets is taken as proof of the existence of the physical radio connection. When there is no signaling data to transmit, the MS sends a measurement report with the current results of the continuously conducted radio signal level measurements.
- Fast Associated Control Channel (FACCH): By using dynamic preemptive multiplexing on a TCH, additional bandwidth can be made available for signaling. The signaling channel created this way is called FACCH. It is only assigned in connection with a TCH, and its short-time usage goes at the expense of the user data transport.
- In addition to these channels, a Cell Broadcast Channel (CBCH) is defined, which is used to broadcast the messages of the Short Message Service Cell Broadcast (SMSCB). The CBCH shares a physical channel with the SDCCH.

1.3 Example: connection setup for incoming call

Figure 4.1 shows an example for an incoming call connection setup at the air interface. It is illustrated how the various logical channels are used in principle. The MS is called via the PCH and requests a signaling channel on the RACH. It obtains the SDCCH through an IMMEDIATE ASSIGNMENT message on the AGCH. Then follow authentication, start of ciphering and start of setup over the SDCCH. An ASSIGNMENT COMMAND message gives the traffic channel to the MS, which acknowledges its receipt on the FACCH of this traffic channel. The FACCH is also used to continue the connection setup.

1.4 Bit rates, block lengths and block distances

Below Table 4.2 gives an overview of the logical channels of Layer 1, the available bit rates, block lengths used and the intervals between transmissions of blocks. The 14.4 kbit/s data service has been standardized in further GSM standardization phases. Note that the logical channels can suffer from substantial transmission delays depending on the respective use of forward error correction (channel coding and interleaving).

Table 4.2 Logical channels of GSM protocol Layer 1.

Channel type	Net data throughput (kbit/s)	Block length (bits)	Block distance (ms)
TCH (full-rate speech)	13.0	182 + 78	20
TCH (half-rate speech)	5.6	95 + 17	20
TCH (data, 14.4 kbit/s)	14.5	290	20
TCH (data, 9.6 kbit/s)	12.0	60	5
TCH (data, 4.8 kbit/s)	6.0	60	10
TCH (data, up to 2.4 kbit/s)	3.6	72	10
FACCH full rate	9.2	184	20
FACCH half rate	4.6	184	40
SDCCH	598/765	184	3060/13
SACCH (with TCH)	115/300	168 + 16	480
SACCH (with SDCCH)	299/765	168 + 16	6120/13
BCCCH	598/765	184	3060/13
AGCH	$n \times 598/765$	184	3060/13
NCH	$m \times 598/765$	184	3060/13
PCH	$p \times 598/765$	184	3060/13
RACH	$r \times 27/765$	8	3060/13
CBCCH	598/765	184	3060/13

1.5 Combinations of logical channels

	B1	B2	B3	B4	B5	B6	B7	B8	B9
TCH/F									
TCH/H									
TCH/H									
BCCCH									
FCCH									
SCH									
CCCH									
SDCCH									
SACCH									
FACCH									

	M1	M2	M3	M4	M5	M6	M7	M8
TCH/F								$n+m$
TCH/H								
TCH/H								
BCCCH								
CCCH								
SDCCH								
SACCH								$n+m$
FACCH								

Channel combinations offered by the base station & Channel combinations used by the mobile station.

Not all logical channels can be used simultaneously at the radio interface. They can only be deployed in certain combinations and on certain physical channels. GSM has defined several channel configurations, which are realized and offered by the base stations (Table shown above left side). As already mentioned before, an SACCH is always allocated either with a TCH or with an SDCCH, which accounts for the attribute ‘associated’.

Depending on its current state, a MS can only use a subset of the logical channels offered by the base station. It uses the channels only in the combinations indicated in Table right side above. The combination M1 is used in the phase when no physical connection exists, i.e. immediately after the power-up of the MS or after a disruption due to unsatisfactory radio signal conditions. Channel combinations M2 and M3 are used by active MSs in standby mode. In phases requiring a dedicated signaling channel, a MS uses the combination M4, whereas M5 to M8 are used when there is a traffic channel up. M8 is a multislots combination (a MS transmits on several physical channels), where n denotes the number of bidirectional channels, and m denotes the number of unidirectional channels ($n = 1, \dots, 8$, $m = 0, \dots, 7$, $n + m = 1, \dots, 8$).

2. Physical channels

After discussing the logical channels and their tasks, we now deal with the physical channels, which transport the logical channels via the air interface. Here is to describe GSM modulation technique, multiplexing structure: GSM is a multicarrier TDMA system, i.e. it employs a combination of FDMA and TDMA for multiple access and also explanation of the radio bursts. Finally, describes the (optional) frequency hopping technique, which has been standardized to reduce interference.

2.1 GSM modulation technique



The modulation technique used on the radio channel is Gaussian Minimum Shift Keying (GMSK). GMSK belongs to a family of continuous-phase modulation procedures, which have the special advantages of a narrow transmitter power spectrum with low adjacent channel interference, on the one hand, and a constant amplitude envelope, on the other hand, which allows use of simple amplifiers in the transmitters without special linearity requirements (class C amplifiers). Such amplifiers are especially inexpensive to manufacture, have high degree of efficiency and therefore allow longer operation on a battery charge (David and Benkner, 1996; Watson, 1993).

The digital modulation procedure for the GSM air interface comprises several steps for the generation of a high-frequency signal from channel-coded and enciphered data blocks. The data d_i arrives at the modulator with a bit rate of $1625/6 = 270.83$ kbit/s (gross data rate) and are first differential-coded:

$$\hat{d}_i = (d_i + d_{i-1}) \bmod 2, \quad d_i \in (0; 1).$$

From this differential data, the modulation data are formed, which represents a sequence of Dirac pulses:

$$a_i = 1 - 2\hat{d}_i$$

This bipolar sequence of modulation data is fed into the transmitter filter – also called a frequency filter – to generate the phase $\varphi(t)$ of the modulation signal. The impulse response $g(t)$ of this linear filter is defined by the convolution of the impulse response $h(t)$ of a Gaussian low-pass with a rectangular step function:

$$g(t) = h(t) * \text{rect}(t/T),$$

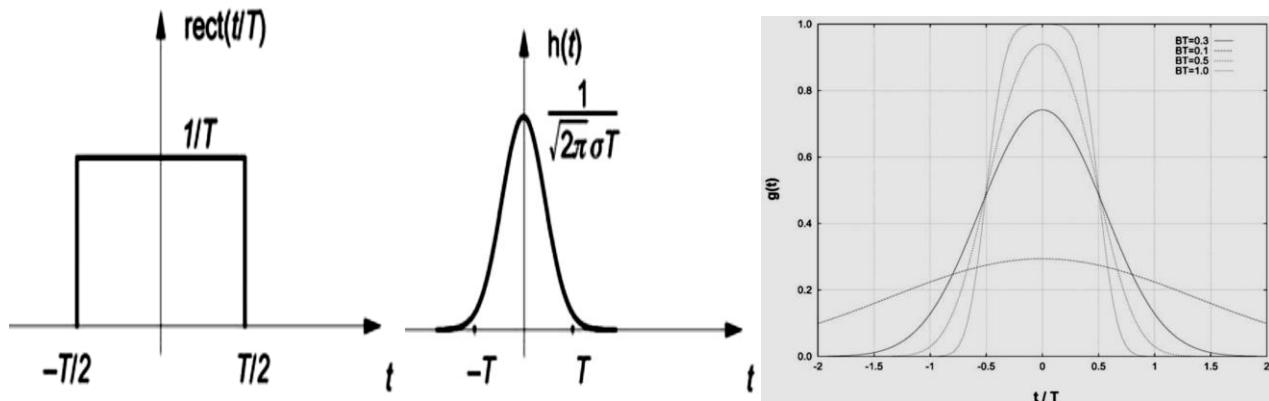
$$\text{rect}(t/T) = \begin{cases} 1/T & \text{for } |t| < T/2, \\ 0 & \text{for } |t| \geq T/2, \end{cases}$$

$$h(t) = \frac{1}{\sqrt{2\pi}\sigma T} \exp\left(\frac{-t^2}{2\sigma^2 T^2}\right), \quad \sigma = \frac{\sqrt{\ln 2}}{2\pi BT}, \quad BT = 0.3.$$

In the equations above, B is the 3 dB bandwidth of the filter $h(t)$ and T the bit duration of the incoming bit stream. The rectangular step function and the impulse response of the Gaussian lowpass are shown in Figure on next page, and the resulting impulse response $g(t)$ of the transmitter filter is given in Figure on next page for some values of BT . Note that with decreasing BT the impulse response becomes broader. For $BT \rightarrow \infty$ it converges to the $\text{rect}()$ function.

In essence, this modulation consists of a Minimum Shift Keying (MSK) procedure, where the data are filtered through an additional Gaussian lowpass before Continuous Phase Modulation (CPM) with the rectangular filter (David and Benkner, 1996). Accordingly it is called GMSK. The Gaussian lowpass filtering has the effect of additional smoothing, but also of broadening the impulse response $g(t)$. This means that, on the one hand, the power spectrum of the signal is made narrower, but, on the other hand, the individual

impulse responses are ‘smeared’ across several bit durations, which leads to increased intersymbol interference. This partial-response behavior has to be compensated for in the receiver by means of an equalizer (David and Benkner, 1996).



Impulse responses for the building blocks of the GMSK transmitter filter & Impulse response $g(t)$ of the frequency filter (transmitter filter).

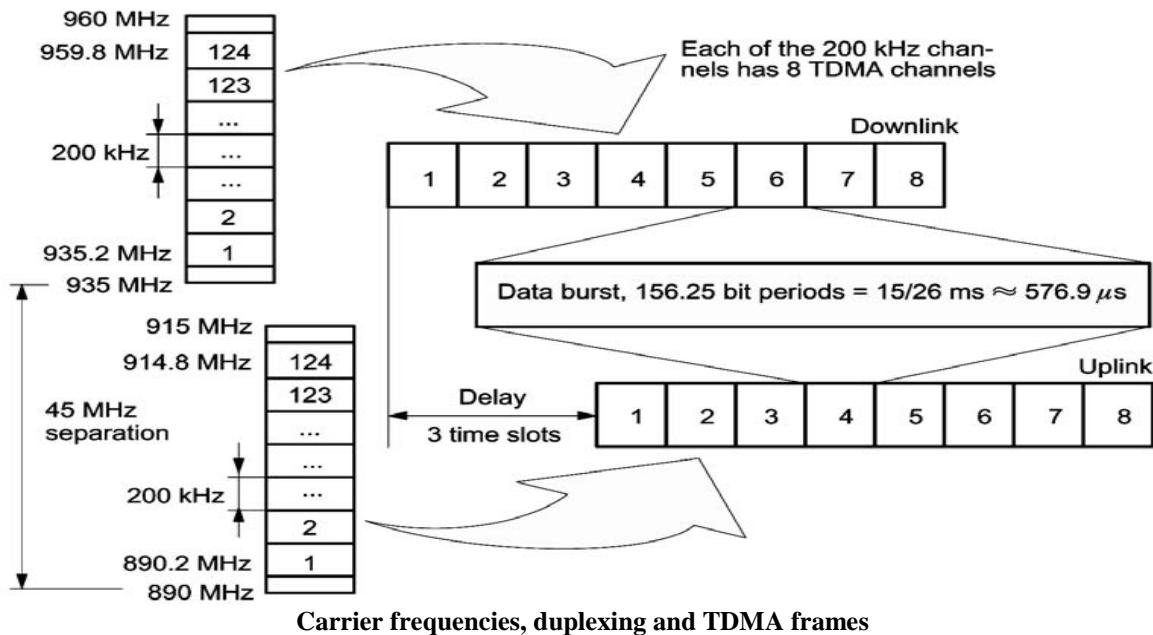
The phase of the modulation signal is the convolution of the impulse response $g(t)$ of the frequency filter with the Dirac impulse sequence a_i of the stream of modulation data: with the modulation index at $\eta = 1/2$, i.e. the maximal phase shift is $\pi/2$ per bit duration.

$$\varphi(t) = \sum_i a_i \pi \eta \int_{-\infty}^{t-iT} g(u) du$$

Accordingly, GSM modulation is designated as 0.3-GMSK with a $\pi/2$ phase shift. The phase $\varphi(t)$ is now fed to a phase modulator. The modulated high-frequency carrier signal can then be represented by the following expression, where E_c is the energy per bit of the modulated data rate, f_0 the carrier frequency and φ_0 is a random phase component staying constant during a burst:

$$x(t) = \sqrt{\frac{2E_c}{T}} \cos(2\pi f_0 t + \varphi(t) + \varphi_0)$$

2.2 Multiple access, duplexing and bursts

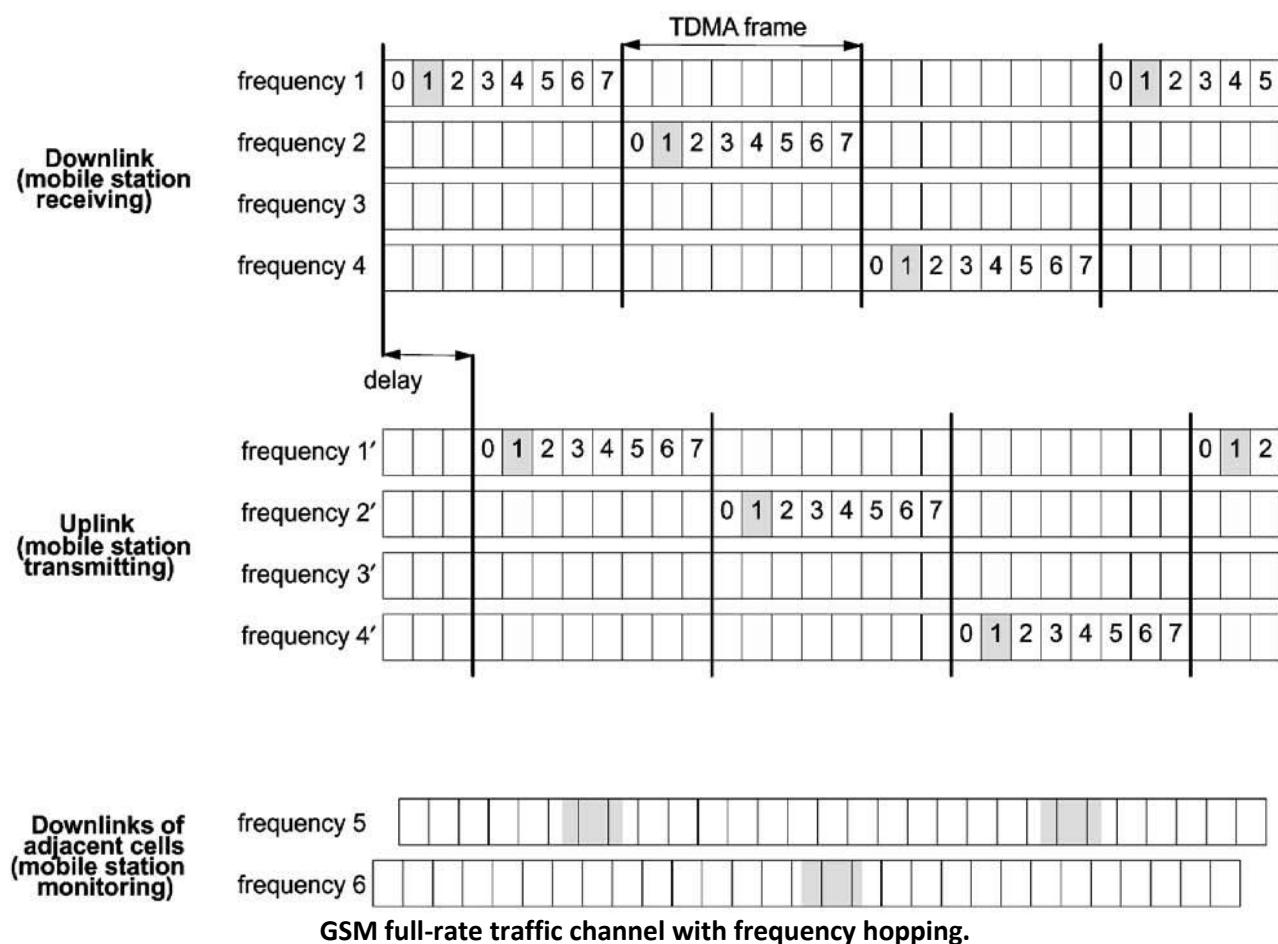


At the physical layer (OSI Layer 1), GSM uses a combination of FDMA and TDMA for multiple access. Two frequency bands 45 MHz apart have been reserved for GSM operation (Figure previous page): 890–915 MHz for transmission from the MS, i.e. uplink, and 935–960 MHz for transmission from the base station, i.e. downlink. Each of these bands of 25 MHz width is divided into 124 single carrier channels of 200 kHz width. This variant of FDMA is also called Multi-Carrier (MC). In each of the uplink/downlink bands there remains a guardband of 200 kHz. Each Radio Frequency Channel (RFCH) is uniquely numbered, and a pair of channels with the same number form a duplex channel with a duplex distance of 45 MHz

A subset of the frequency channels, the Cell Allocation (CA), is allocated to a base station, i.e. to a cell. One of the frequency channels of the CA is used for broadcasting the synchronization data (FCCH and SCH) and the BCCH. Therefore, this channel is also called the BCCH carrier. Another subset of the cell allocation is allocated to a MS, the Mobile Allocation (MA). The MA is used among others for the optional frequency hopping procedure.

Countries or areas which allow more than one mobile network to operate in the same area of the spectrum must have a licensing agency which distributes the available frequency number space (e.g. the Federal Communication Commission in the USA or the ‘Bundesnetzagentur’ in Germany), in order to avoid collisions and to allow the network operators to perform independent network planning. Here is an example for a possible division: Operator A uses RFCH 2–13, 52–81 and 106–120, whereas operator B receives RFCH 15–50 and 83–103, in which case RFCH 1, 14, 51, 82, 104, 105 and 121–124 are left unused as additional guard bands.

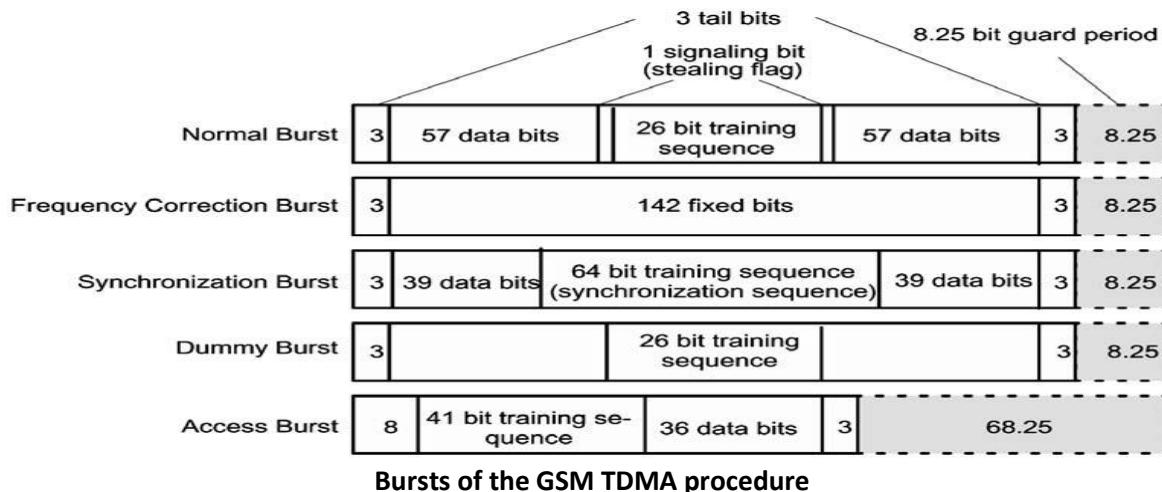
Each of the 200 kHz channels is divided into eight time slots and thus carries eight TDMA channels. The eight time slots together form a TDMA frame (Figure Previous). The TDMA frames of the uplink are transmitted with a delay of three time slots with regard to the downlink (see Figure below). A MS uses the same time slots in the uplink as in the downlink, i.e. the time slots with the same number (TN). Owing to the shift of three time slots, a MS does not have to send at the same time as it receives, and therefore does not need a duplex unit. This reduces the high-frequency requirements for the front end of the mobile and allows it to be manufactured as a less expensive and more compact unit.



So in addition to the separation into uplink and downlink bands – FDD with a distance of 45 MHz – the GSM access procedure contains a TDD component. Thus, the MS does not need its own high-frequency duplexing unit, which again reduces cost as well as energy consumption.

Each time slot of a TDMA frame lasts for a duration of 156.25 bit periods and, if used, contains a data burst. The time slot lasts $15/26 \text{ ms} = 576.9 \mu\text{s}$; so a frame takes 4.615 ms. The same result is also obtained from the GMSK procedure, which realizes a gross data transmission rate of 270.83 kbit/s per carrier frequency.

BURST: There are five kinds of burst



1. Normal Burst (NB):

The NB is used to transmit information on traffic and control (except RACH) channels. The individual bursts are separated from each other by guard periods during which no bits are transmitted. At the start and end of each burst are three tail bits which are always set to logical '0'. These bits fill a short time span during which transmitter power is ramped up or ramped down and during which no data transmission is possible.

Furthermore, the initial zero bits are also needed for the demodulation process. The Stealing Flags (SFs) are signaling bits which indicate whether the burst contains traffic data or signaling data. They are set to allow use of single time slots of the TCH in preemptive multiplexing mode, e.g. when, during a handover, fast transmission of signaling data on the FACCH is needed. This causes a loss of user data, i.e. these time slots are 'stolen' from the traffic channel, hence the name SF. In addition to the synchronization and signaling bits (Figure 4.7), the NB also contains two blocks of 57 bits each of error-protected and channel-coded user data separated by a 26-bit midamble. This midamble consists of predefined, known bit patterns, the training sequences, which are used for channel estimation to optimize reception with an equalizer and for synchronization. With the help of these training sequences, the equalizer eliminates or reduces the intersymbol interferences which are caused by propagation time differences of the multipath propagation.

Time differences of up to $16 \mu\text{s}$ can be compensated for. Eight different training sequences are defined for the NB which are designated by the Training Sequence Code (TSC). Initially, the TSC is obtained when the BCC is obtained, which is transmitted as part of the BSIC (see Chapter 3). Beyond that, training sequences can be individually assigned to mobile stations. In this case the TSC is contained in the Layer 3 message of the channel assignment (TCH or SDCCH). In this way the base station tells a MS which training sequence it should use with NBs of a specific traffic channel.

2. Frequency Correction Burst (FB):

This burst is used for the frequency synchronization of a MS. The repeated transmission of FBs is also called the FCCH. Tail bits as well as data bits are all set to 0 in the FB. Owing to the GSM modulation procedure (0.3- GMSK) this corresponds to broadcasting an unmodulated carrier with a frequency shift of 1625/24 kHz above the nominal carrier frequency. This signal is periodically transmitted by the base station on the BCCH carrier. It allows time synchronization with the TDMA frame of a MS as well as the exact tuning to the carrier frequency. Depending on the stability of its own reference clock, the mobile can periodically resynchronize with the base station using the FCCH.

3. Synchronization Burst (SB):

This burst is used to transmit information which allows the MS to synchronize time-wise with the BTS. In addition to a long midamble, this burst contains the running number of the TDMA frame, the RFN and the BSIC; the RFN is covered in section 4.3. Repeated broadcasting of SBs is considered as the SCH.

4. Dummy Burst (DB):

This burst is transmitted on one frequency of the CA, when no other bursts are to be transmitted. The frequency channel used is the same as that which carries the BCCH, i.e. it is the BCCH carrier. This ensures that the BCCH transmits a burst in each time slot which enables the MS to perform signal power measurements of the BCCH, a procedure also known as quality monitoring.

5. Access Burst (AB):

This burst is used for random access to the RACH without reservation. It has a guard period significantly longer than the other bursts. This reduces the probability of collisions, since the MSs competing for the RACH are not (yet) time-synchronized.

A single user gets one-eighth or 33.9 kbit/s of the gross data rate of 270.83 kbit/s. considering a normal burst, 9.2 kbit/s are used for signaling and synchronization, i.e. tail bits, SFs and training sequences, including guard periods. The remaining 24.7 kbit/s are available for the transmission of (raw) user or control data on the physical layer.

2.3 Optional frequency hopping

Mobile radio channels suffer from frequency-selective interferences, e.g. frequency-selective fading due to multipath propagation phenomena. This selective frequency interference can increase with the distance from the base station, especially at the cell boundaries and under unfavorable conditions. **Frequency hopping procedures change the transmission frequencies periodically and thus average the interference over the frequencies in one cell.**

This leads to a further improvement in the Signal-to-Noise Ratio (SNR) to a high enough level for good speech quality, so that conversations with acceptable quality can be conducted. GSM systems achieve a good speech quality with a SNR of about 11 dB. With frequency hopping a value of 9 dB is sufficient. GSM provides for an optional frequency hopping procedure which changes to a different frequency with each burst; this is known as slow frequency hopping. The resulting hopping rate is about 217 changes per second, corresponding to the TDMA frame duration.

The frequencies available for hopping, the hopping assignment, are taken from the CA. The principle is illustrated in Figure previous figure on page 16, showing the time slot allocations for a full-rate TCH. The exact synchronization is determined by several parameters: the MA, a Mobile Allocation Index Offset (MAIO), a Hopping Sequence Number (HSN) and the TDMA Frame Number (FN); see section 4.3. The use of frequency hopping is an option left to the network operator, which can be decided on an individual cell basis. Therefore, a MS must be able to switch to frequency hopping if a base station notices adverse conditions and decides to activate frequency hopping.

2.4 Summary

A physical GSM channel is defined by a sequence of frequencies and a sequence of TDMA frames. The RFCH sequence is defined by the frequency hopping parameters, and the temporal sequence of time slots of a physical channel is defined as a sequence of frame numbers and the time slot number within the frame. Frequencies for the uplink and downlink are always assigned as a pair of frequencies with a 45 MHz duplex

separation. As shown above, GSM uses a series of parameters to define a specific physical channel of a base station. Summarizing, these parameters are:

- MAIO;

- HSN;
- TSC;
- Time Slot Number (TN);
- MA, also known as RFCH Allocation;
- Type of logical channel carried on this physical channel;
- The number of the logical subchannel (if used) – the Subchannel Number (SCN).

Within a logical channel, there can be several subchannels (e.g. substrate multiplexing of the same channel type). The TDMA frame sequence can be derived from the type of the channel and the logical subchannel if present.

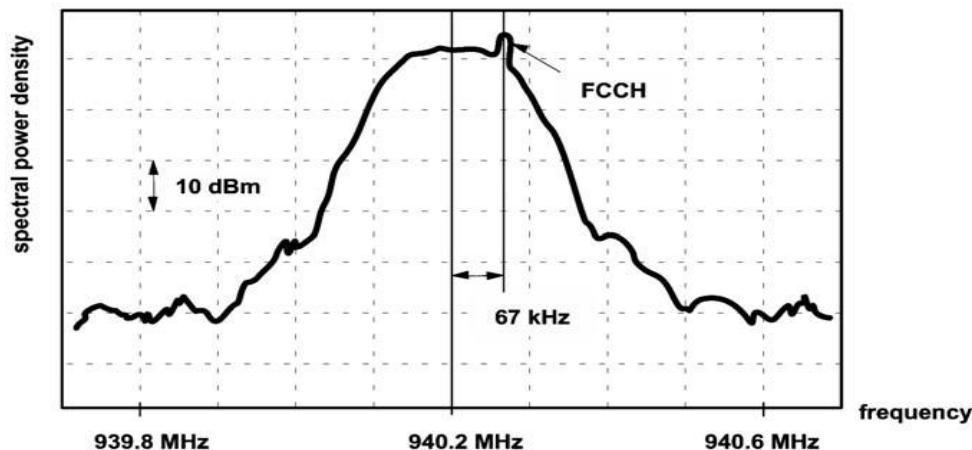
3. Synchronization

For the successful operation of a mobile radio system, synchronization between MSs and the base station is necessary. Two kinds of synchronization are distinguished: frequency synchrony and time synchrony of the bits and frames.

Frequency synchronization is necessary so that transmitter and receiver frequencies agree. The objective is to compensate for tolerances of the less-expensive and, therefore, less-stable oscillators in the mobile stations by obtaining an exact reference from the base station and to follow it.

Bit and frame synchrony are important in two regards for TDMA systems. First, the propagation time differences of signals from different MSs have to be adjusted, so that the transmitted bursts are received synchronously with the time slots of the base station and that bursts in adjacent time slots do not overlap and interfere with each other. Second, synchrony is needed for the frame structure since there is a higher-level frame structure superimposed on the TDMA frames for multiplexing logical signaling channels onto one physical channel. The synchronization procedures defined for GSM are explained in the following section.

3.1 Frequency and clock synchronization



Typical power spectrum of a BCCH carrier.

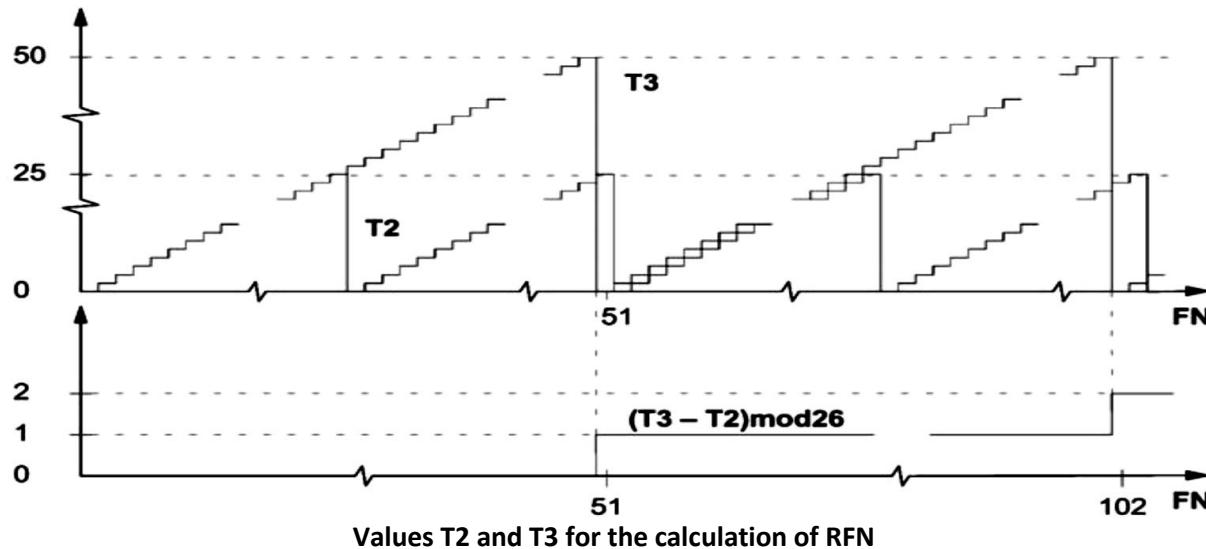
A GSM base station transmits signals on the frequency carrier of the BCCH which allow a MS to synchronize with the base station. Synchronization means on the one hand the time wise synchronization of the MS and base station with regards to bits and frames, and on the other hand tuning the MS to the correct transmitter and receiver frequencies.

For this purpose, the BTS provides the following signals (Figure Burst mode):

- SCH with SBs and extra-long training sequences, which facilitate synchronization;
- FCCH with FBs.

Owing to the 0.3-GMSK modulation procedure used in GSM, a data sequence of logical '0' generates a pure sine wave signal, i.e. broadcasting of the FB corresponds to an unmodulated carrier (frequency channel) with a frequency shift of 1625/24 kHz (\approx 67.7 kHz) above the nominal carrier frequency (Figure 4.8). In this way, the MS can keep exactly synchronized by periodically monitoring the FCCH. On the other hand, if the frequency of the BCCH is still unknown, the MS can search for the channel with the highest signal level. This channel is with all likelihood a BCCH channel, because DBs must be transmitted on all unused time slots in this channel, whereas not all time slots are always used on other carrier frequencies. Using the FCCH sine wave signal allows identification of a BCCH and synchronization of an MSs Oscillator.

For the time synchronization, TDMA frames in GSM are cyclically numbered modulo $2^{11} = 2048 (= 26 \times 51 \times 21)$ with the FN. One cycle generates the so-called hyper frame structure which comprises 2^{11} TDMA frames. This long numbering cycle of TDMA frames is used to synchronize the ciphering algorithm at the air interface. Each base station BTS periodically transmits the RFN on the SCH. With each SB the mobiles thus receive information about the number of the current TDMA frame. This enables each MS to be time-synchronized with the base station.



The RFN has a length of 19 bits. It consists of three fields: T1 (11 bits), T2 (5 bits) and T3' (3 bits). These three fields are defined by (with div designating integer division)

$$T1 = FN \text{ div } (26 \times 51) [0-2047],$$

$$T2 = FN \text{ mod } 26 [0-25],$$

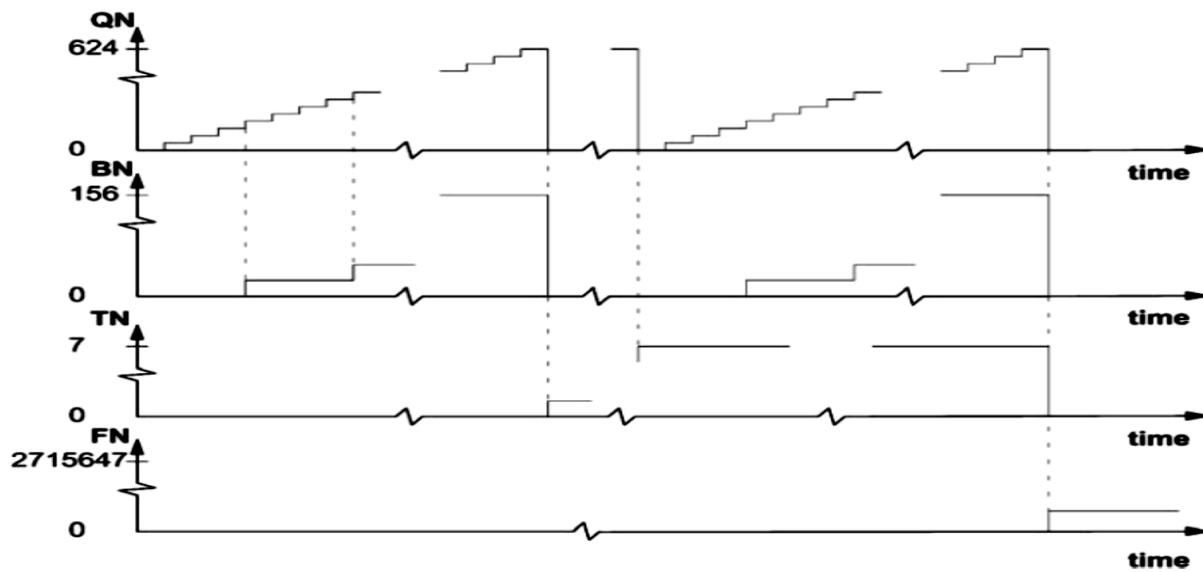
$$T3' = (T3 - 1) \text{ div } 10 [0-4], \quad \text{with } T3 = FN \text{ mod } 51 [0-50].$$

The sequences of running values of T2 and T3 are illustrated in Figure above. The value crucial for the reconstruction of the FN is the difference between the two fields. The time synchronization of a MS and its time slots, TDMA frames and control channels is based on a set of counters which run continuously, independent of MS or base station transmission. Once these counters have been started and initialized correctly, the MS is in a synchronized state with the base station. The following four counters are kept for this purpose:

- Quarter bit counter counting the Quarter Bit Number (QN);
- Bit counter counting the Bit Number (BN);
- Time slot counter counting the TN;
- frame counter counting the FN.

Owing to the bit and frame counting, these counters are of course interrelated, namely in such a way that the subsequent counter counts the overflows of the preceding counter. The following principle is used (Figure below): QN is incremented every 12 or 13 μ s; BN is obtained from this by integer division

(BN = QN div 4). With each transition from 624 to 0 TN is incremented, and each overflow of TN increments the frame counter FN by 1.

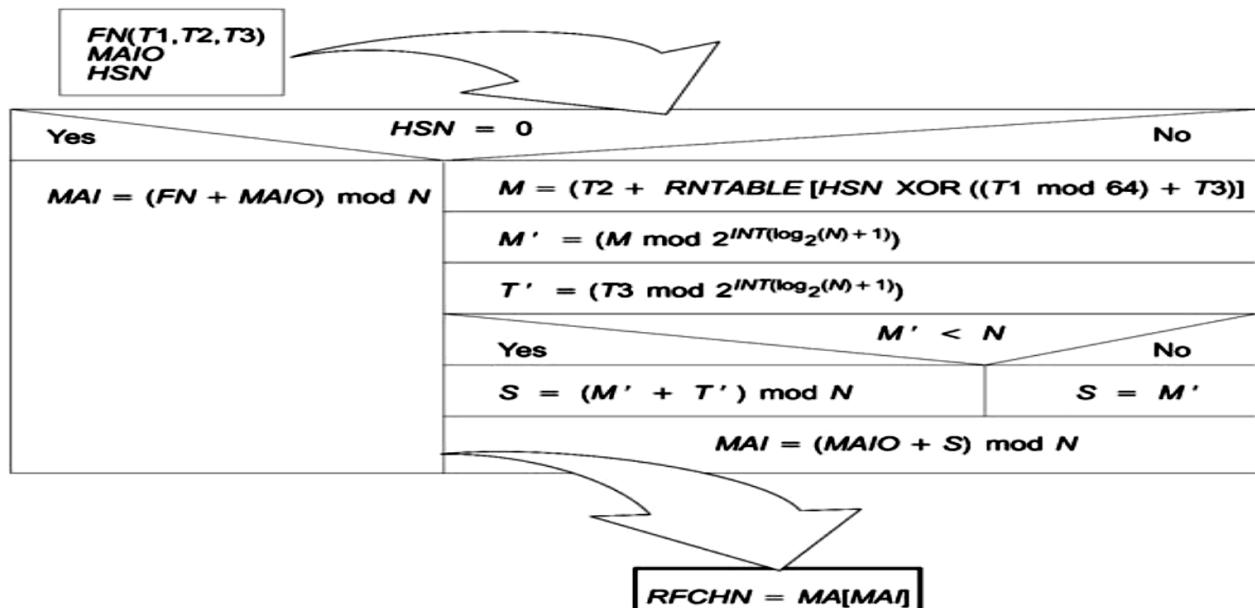


Synchronization timers, simplified: the TDMA frame duration is 156.25 bit times.

The timers can be reset and restarted when receiving an SB. The quarter bit counter is set by using the timing of the training sequence of the burst, whereas the TN is reset to 0 with the end of the burst. The FN can then be calculated from the RFN transmitted on the SCH:

$$FN = 51 \times ((T3 - T2) \bmod 26) + T3 + 51 \times 26 \times T1, \quad \text{with } T3 = 10 \times T3' + 1.$$

It is important to recalculate T3 from T3_ although, because of the binary representation, only the integer part of the division by 10 is taken into account. If the optional frequency hopping procedure is used, an additional mapping of the TDMA frame number onto the frequency to be used is required in addition to the evaluation of the synchronization signals from the FCCH and SCH. One has to obtain the index number of the frequency channel on which the current burst has to be transmitted from the MA table. This process uses a predefined RFNTABLE, the FN and a HSN (Figure on next page). The MA holds N frequencies, with a maximum value of 64 for N. With this procedure, every burst is sent on a different frequency in a cyclic way.



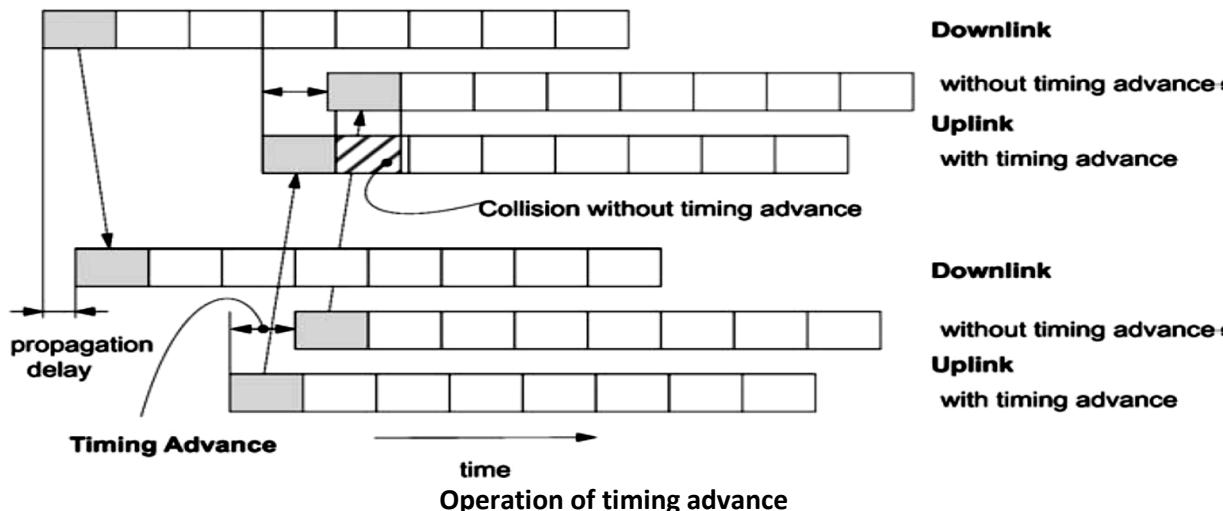
Generation of the GSM frequency hopping sequence

3.2 Adaptive frame synchronization

The MS can be anywhere within a cell, which means the distance between MS and base station may vary. Thus, the signal propagation times between MS and base station vary. Owing to the mobility of the subscribers, the bursts received at the base would be offset. The TDMA procedure cannot tolerate such time shifts, since it is based on the exact synchronization of transmitted and received data bursts. Bursts transmitted by different MSs in adjacent time slots must not overlap when received at the base station by more than the guard period, even if the propagation times within the cell are very different.

To avoid such collisions, the start of transmission time from the MS is advanced in proportion to the distance from the base station. The process of adapting the transmissions from the MSs to the TDMA frame is called adaptive frame alignment.

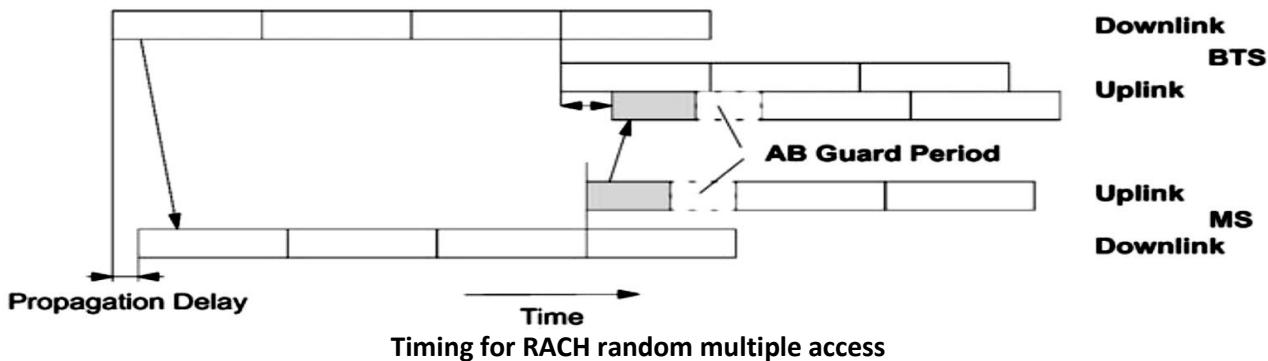
For this purpose, the parameter Timing Advance (TA) in each SACCH Layer 1 protocol block is used. The MS receives the TA value it must use from the base station on the SACCH downlink; it reports the actually used value on the SACCH uplink. There are 64 steps for the timing advance which are coded as 0 to 63. One step corresponds to one bit period. Step 0 means no timing advance, i.e. the frames are transmitted with a time shift of three slots or 468.75 bit durations with regards to the downlink. At step 63, the timing of the uplink is shifted by 63 bit durations, such that the TDMA frames are transmitted on the uplink only with a delay of 405.75 bit durations. So the required adjustment always corresponds to twice the propagation time or is equal to the round-trip delay (Figure below).



In this way, the available range of values allows a compensation over a maximum propagation time of 31.5 bit periods ($\approx 113.3 \mu s$). This corresponds to a maximum distance between mobile and base station of 35 km. A GSM cell may therefore have a maximum diameter of 70 km. The distance from the base station or the currently valid TA value for a MS is therefore an important handover criterion in GSM networks.

The adaptive frame alignment technique is based on continuous measurement of propagation delays by the base station and corresponding timing advance activity by the MS. In the case of an (unreserved) random access to the RACH, a channel must first be established. The base station has in this case not yet had the opportunity to measure the distance to the MS and to transmit a corresponding timing advance command. If a MS transmits an access burst in the current time slot, it uses a timing advance value of 0 or a default value.

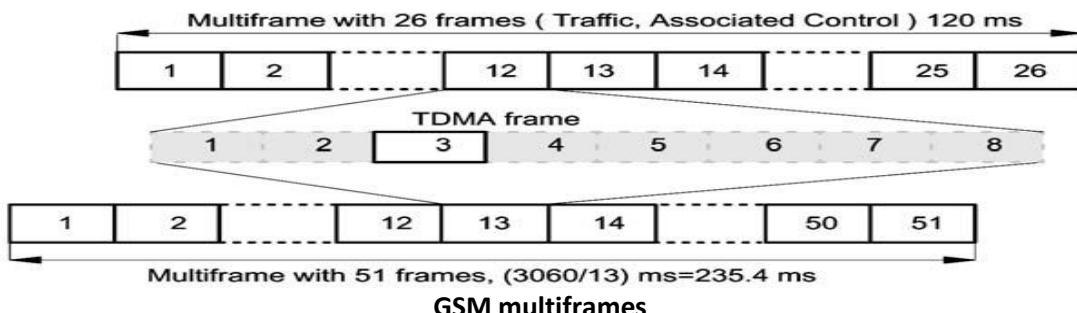
To minimize collisions with subsequent time slots at the base station, the access burst (AB) has to be correspondingly shorter than the time slot duration (Figure on next page). This explains the long guard period of the AB of 68.25 bit periods, which can compensate for the propagation delay if a MS sends an AB from the boundary of a cell of 70 km diameter.



4. Mapping of logical onto physical channels

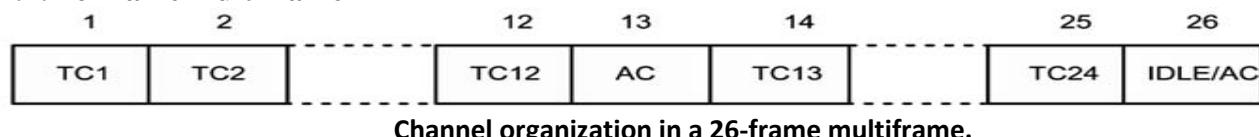
The mapping of logical channels onto physical channels has two components: mapping in frequency and mapping in time. The mapping of a logical channel onto a physical channel in the frequency domain is based on the FN, the frequencies allocated to base stations and MSs – CA and MA – and the rules for the optional frequency hopping. In the time domain, logical channels are transported in the corresponding time slots of the physical channel. They are mapped onto physical channels in certain time-multiplexed combinations, where they can occupy a complete physical channel or just part of a physical channel. Whereas user payload data is allocated a dedicated full-rate or half-rate channel, logical signaling (control) channels have to share a physical channel.

The logical channels are organized by the definition of complex superstructures on top of the TDMA frames, forming so-called multiframe, superframes and hyperframes (Figure 4.14). For the mapping of logical onto physical channels, we are interested in the multiframe domain. These multiframes allow us to map (logical) subchannels onto physical channels.



Two kinds of multiframe are defined (Figure below): a multiframe consisting of 26 TDMA frames (predominantly payload – speech and data – frames) and a multiframe of 51 TDMA frames (predominantly signaling frames). Each hyperframe is divided into 2048 superframes. With its long cycle period of 3 h 28 min 53.760 s, it is used for the synchronization of user data encryption. A superframe consists of 1326 consecutive TDMA frames which therefore lasts for 6.12 s, the same length as 51 multiframe of 26 TDMA frames or 26 multiframe of 51 TDMA frames. These multiframe are again used to multiplex the different logical channels onto a physical channel as shown below.

4.1: 26-frame multiframe



Each 26 subsequent TDMA frames form a multiframe which multiplexes two logical channels, a TCH and the SACCH, onto the physical channel (Figure 4.16). This process uses only one time slot per TDMA frame for the corresponding multiframe (e.g. time slot 3 in Figure GSM Multiframe), since a physical channel consists of just one time slot per TDMA frame.

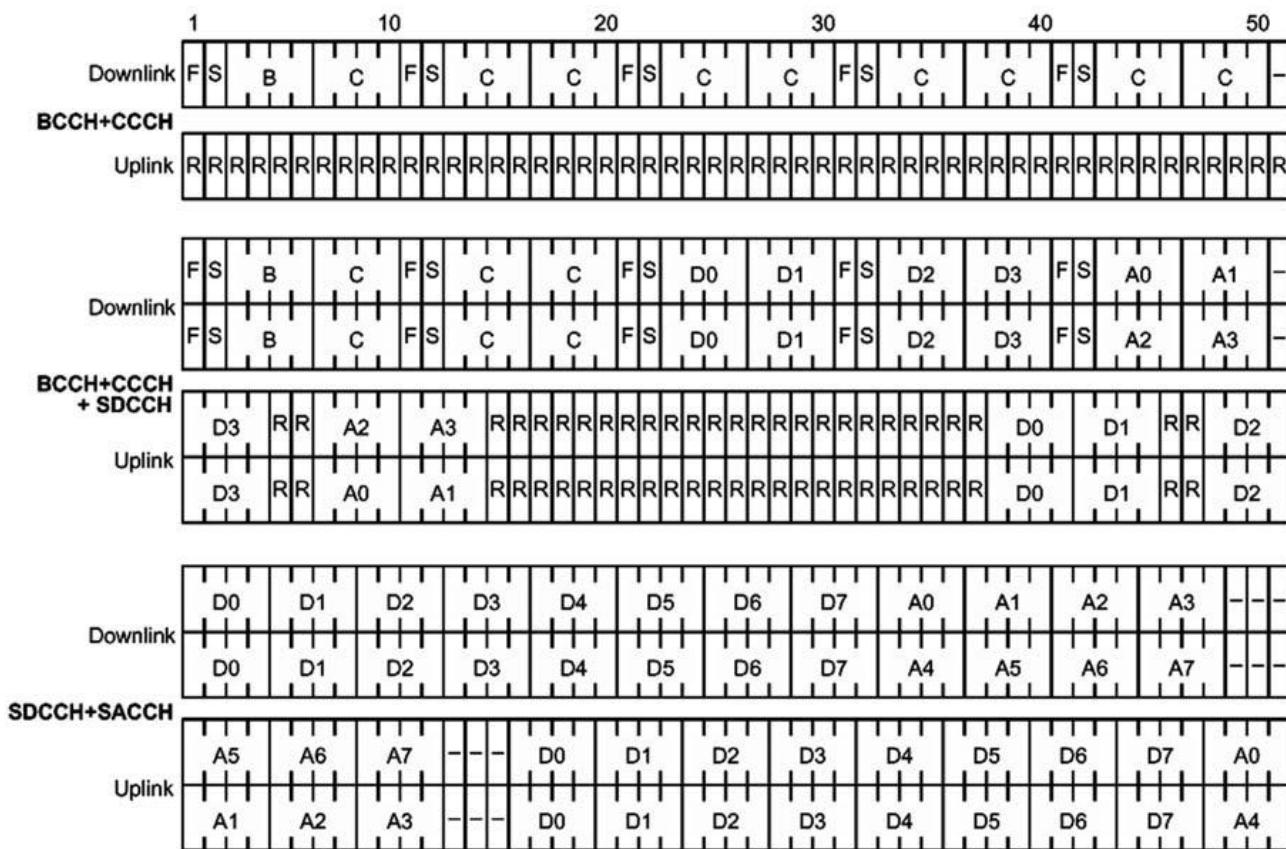
In addition to the 24 TCH frames for user data, this multiframe also contains an Associated Control (AC) frame for signaling data (SACCH data). One frame (the 26th) remains unused in the case of a full-rate TCH (IDLE/AC); it is reserved for the introduction of two half-rate TCHs; then the 26th frame will be used to carry the SACCH channels of the other half.

The data of the FACCH is transmitted by occupying one-half of the bits in eight consecutive bursts, by ‘stealing’ these bits from the TCH. For this purpose, the SFs of the normal bursts are set. A subscriber has available a gross data rate of $271/8 = 33.9$ kbit/s (section 4.2). Of this budget, 9.2 kbit/s are for signaling, synchronization and guard periods of the burst. Of the remaining 24.7 kbit/s, in the case of the 26-frame multiframe, 22.8 kbit/s are left for the coded and enciphered user data of a full-rate channel and 1.9 kbit/s remain for the SACCH and IDLE.

4.2: 51-frame multiframe

For the transmission of the control channels which are not associated with a TCH (all except FACCH and SACCH), a multiframe is formed consisting of 51 consecutive TDMA frames (Figure – Burst mode). According to the channel configuration, the multiframe is used differently. In each case, multiframe of 51 TDMA frames serve the purpose of mapping several logical channels onto a physical channel.

Furthermore, some of these control channels are unidirectional, which results in different structures for uplink and downlink. For some configurations, two adjacent multiframes are required to map all of the logical channels. Some examples are illustrated in Figure 4.17 below.



F: TDMA frame for frequency correction bursts

S: TDMA frame for synchronization bursts

B: TDMA frame for BCH

C: TDMA frame for CCCH

R: TDMA frame for RACH

D: TDMA frame for SDCCH

A: TDMA frame for SACCH

Channel organization in a 51-frame multiframe.

They correspond to the combinations B2, B3 and B4 in Table 4.3 whereas for channels SDCCH and SACCH some four or eight logical subchannels have been defined ($D_0, D_1, \dots, A_0, A_1, \dots$). One of the

frequency channels of the CA of a base station is used to broadcast synchronization data (FCCH and SCH) and the BCCH. Since the base station has to transmit in each time slot of the BCCH carrier to enable a continuous measurement of the BCCH carrier by the MS, a DB is transmitted in all time slots with no traffic.

On time slot 0 of the BCCH carrier, only two combinations of logical channels may be transmitted, the combinations B2 or B3 from Table 4.3: (BCCH + CCCH + FCCH + SCH + SDCCH + SACCH or BCCH + CCCH + FCCH + SCH). No other time slot of the CA must carry this combination of logical channels.

As one can see in Figure previous page, in the time slot 0 of the BCCH carrier of a base station (downlink) the frames 1, 11, 21, . . . are FCCH frames, and the subsequent frames 2, 12, 22, . . . form SCH frames. Frames 3, 4, 5 and 6 of the 51-frame BCCH multiframe transport the appropriate BCCH information, whereas the remaining frames may contain different combinations of logical channels. Once the MS has synchronized by using the information from FCCH and SCH, it can determine from the information in the FCCH and SCH how the remainder of the BCCH is constructed. For this purpose, the base station radio resource management periodically transmits a set of messages to all MSs in this cell.

These system information messages comprise six types, of which only types 1–4 are of interest here. Using FN, one can determine which type is to be sent in the current time slot by calculating a Type Code (TC):

$$\mathbf{TC = (FN \text{ div } 51) \bmod 8.}$$

Table 4.5 shows how the TC determines the type of the system information message to be sent within the current multiframe.

Table 4.5 Mapping of the frame number onto a BCCH message.

TC	System information message
0	Type 1
1	Type 2
2, 6	Type 3
3, 7	Type 4
4, 5	Any (optional)

Of the parameters contained in such a message, the following are of special interest. BS_CC_CHANS determines the number of physical channels which support a CCCH. The first CCCH is transmitted in time slot 0, the second in time slot 2, the third in time slot 4 and the fourth in time slot 6 of the BCCH carrier. Another parameter, BS_CCCH_SDCCH_COMB, determines whether the DCCHs SDCCH (0-3) and SACCH (0-3) are transmitted together with the CCCH on the same physical channel. In this case, each of these dedicated control channels consists of four subchannels.

Each of the CCCHs of a base station is assigned a group CCCH_GROUP of MSs. MSs are allowed random access (RACH) or receive paging information (PCH) only on the CCCH assigned to this group. Furthermore, a MS needs only to listen for paging information on every Nth block of the PCH. The number N is determined by multiplying the number of paging blocks per 51-frame multiframe of a CCCH with the parameter BS_PA_MFRMS designating the number of multiframes between paging frames of the same paging group (PAGING_GROUP).

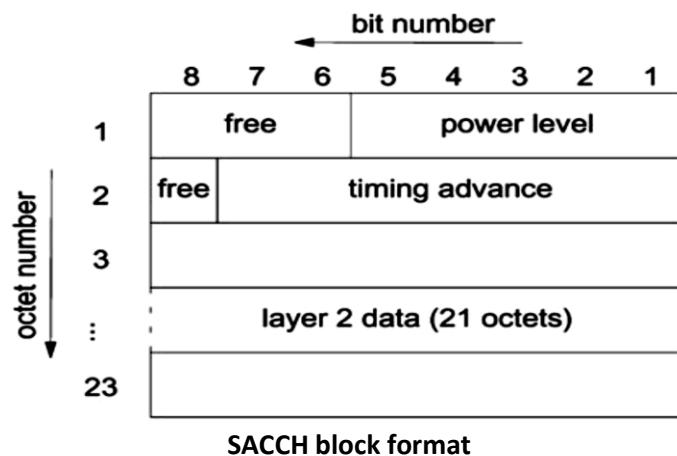
In cells with high traffic, in particular, the CCCH and paging groups serve to subdivide traffic and to reduce the load on the individual CCCHs. For this purpose, there is a simple algorithm which allows each MS to calculate its respective CCCH_GROUP and PAGING_GROUP from its IMSI and parameters BS_CC_CHANS, BS_PA_MFRMS and N.

5. Radio subsystem link control

The radio interface is characterized by another set of functions of which we discuss only the most important in the following. One of these functions is the control of the radio link: radio subsystem link control, with the main activities of received-signal quality measurement (quality monitoring) for cell selection and handover preparation, and of transmitter power control.

If there is no active connection, i.e. if the MS is at rest, the BSS has no tasks to perform. The MS, however, is still committed to continuously observing the BCCH carrier of the current and neighboring cells, so that it would be able to select the cell in which it can communicate with the highest probability. If a new cell needs to be selected, a location update may become necessary.

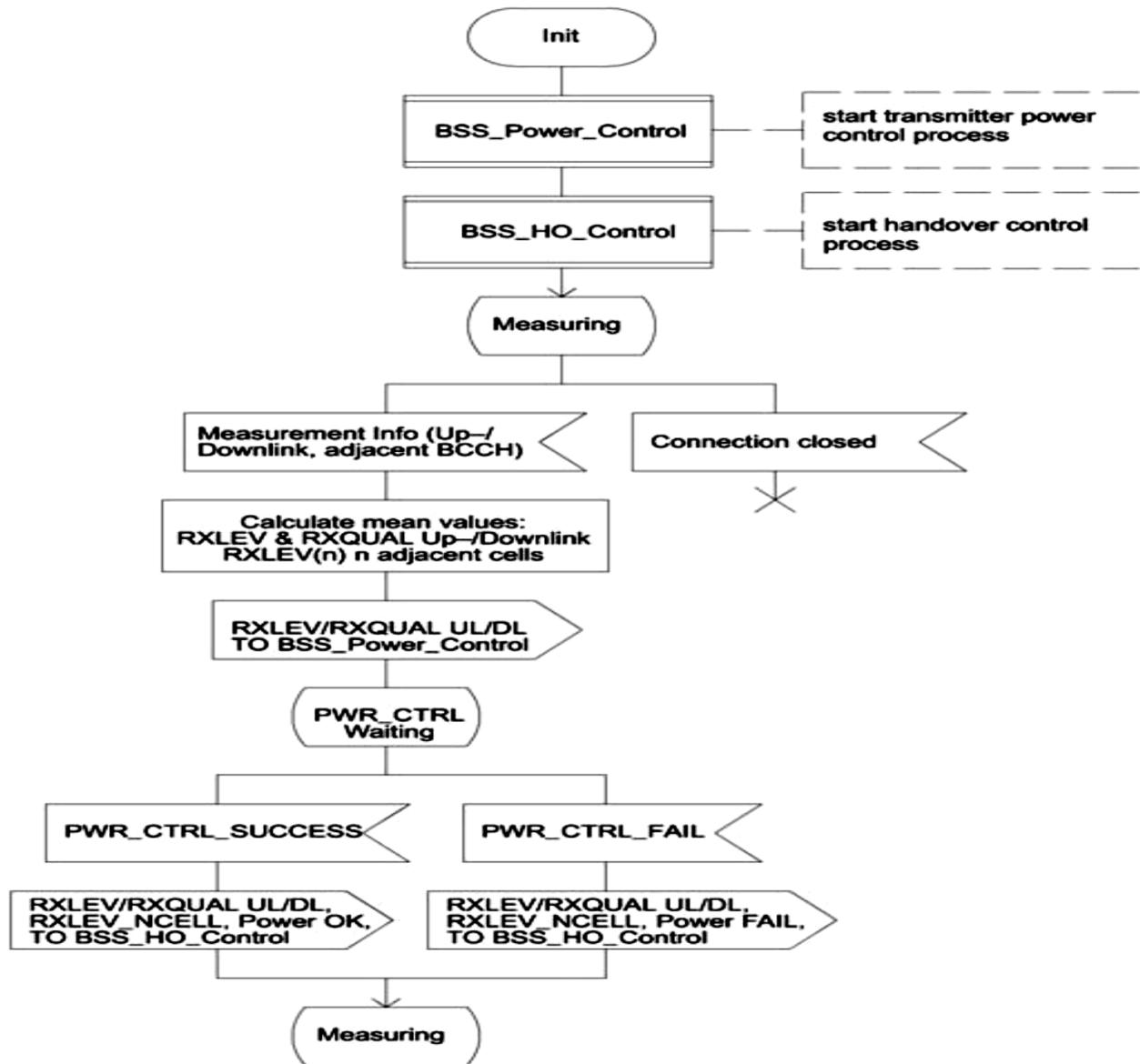
During a connection (TCH or SDCCH), the functions of channel measurement and power control serve to maintain and optimize the radio channel; this also includes adaptive frame alignment and frequency hopping. Both need to be done until the current base can hand over the current connection to the next base station.



These link control functions are performed over the SACCH channel. Two fields are defined in an SACCH block (Figure 4.18) for this purpose, the power level and the TA. On the downlink, these fields contain values as assigned by the BSS. On the uplink, the MS inserts its currently used values. The quality monitoring measurement values are transmitted in the data part of the SACCH block.

The following illustrates the basic operation of the radio subsystem link control at the BSS side for an existing connection; the detailed explanation of the respective functions is given later. In principle, the radio link control can be subdivided into three tasks: measurement collection and processing, transmitter power control, and handover control.

In the example of Figure on page 27 (next page), the process BSS_Link_Control starts at initialization the processes BSS_Power_Control and BSS_HO_Control and then enters a measurement loop, which is only left when the connection is terminated. In this loop, measurement data are periodically received (every 480 ms) and current mean values are calculated. At first, these measurement data are supplied to the transmitter power control to adapt the power of MS and BSS to a new situation if necessary. Thereafter, the measurement data and the result of the power control activity are supplied to the handover process, which can then decide whether a handover is necessary or not.



Principal operation of the radio subsystem link control.

5.1 Channel measurement

The task of radio subsystem link control in the MS includes identification of the reachable base stations and measurement of their respective received signal level and channel quality (quality monitoring task). In idle mode, these measurements serve to select the current base station, whose PCH is then periodically examined and on whose RACH desired connections can be requested.

During a connection, i.e. on a TCH or SDCCH with respective SACCH/FACCH, this measurement data are transmitted on the SACCH to the base station as a measurement report/measurement info. These reports serve as inputs for the handover and power control algorithms.

The measurement objects are, on the one hand, the uplink and downlink of the current channel (TCH or SDCCH) and, on the other hand, the BCCH carriers which are continuously broadcast with constant power by all BTSs in all time slots. It is especially important to keep the transmitter power of the BCCH carriers constant to allow comparisons between neighboring base stations. A list of neighboring base station's BCCH carrier frequencies, called the BCCH Allocation (BA) is supplied to each mobile by its current BTS, to enable measurement of all cells which are candidates for a handover. The cell identity is broadcast

as the BSIC on the BCCH. Furthermore, up to 36 BCCH carrier frequencies and their BSICs can be stored on the SIMcard. In principle, GSM uses two parameters to describe the quality of a channel: the Received Signal Level (RXLEV), measured in dBm, and the Received Signal Quality (RXQUAL), measured as bit error ratio as a percentage before error correction (Tables 4.6 and 4.7 below). The received signal power is measured continuously by MSs and base stations in each received burst within a range of -110 to -48 dBm. The respective RXLEV values are obtained by averaging.

Table 4.6 Measurement range of the received signal level. Table 4.7 Measurement range of the bit error ratio.

Received signal level (dBm)			Bit error ratio (%)		
Level	From	To	Level	From	To
RXLEV_0	-	-110	RXQUAL_0	-	0.2
RXLEV_1	-110	-109	RXQUAL_1	0.2	0.4
:	:	:	RXQUAL_2	0.4	0.8
RXLEV_62	-49	-48	RXQUAL_3	0.8	1.6
RXLEV_63	-48	-	RXQUAL_4	1.6	3.2
			RXQUAL_5	3.2	6.4
			RXQUAL_6	6.4	12.8
			RXQUAL_7	12.8	-

The bit error ratio before error correction can be determined in a variety of ways. For example, it can be estimated from information obtained from channel estimation for equalization from the training sequences, or the number of erroneous (corrected) bits can be determined through repeated coding of the decoded, error-corrected data blocks and comparison with the received data. Since the data before error correction is presented as blocks of 456 bits, the bit error ratio can only be given with a quantizing resolution of 2×10^{-3} . Again, the value of RXQUAL is determined from this information by averaging.

5.1.1. Channel measurement during idle mode

In idle mode the MS must always stay aware of its environment. The main purpose is to be able to assign a MS to a cell, whose BCCH carrier it can decode reliably. If this is the case, the MS is able to read system and paging information. If there is a desire to set up a connection, the MS can most likely communicate with the network. There are two possible starting situations:

- The MS has no a priori knowledge about the network at hand, especially which BCCH carrier frequencies are in use;
- The MS has a stored list of BCCH carriers.

In the first case, the more unfavorable of the two, the mobile has to search through all of the 124 GSM frequencies, measure their signal power level and calculate an average from at least five measurements. The measurements of the individual carriers should be evenly distributed over an interval of 3–5 s. After at most 5 s, a minimum of 629 measurement values are available that allow the 124 RXLEV values to be determined. The carriers with the highest RXLEV values are very likely BCCH carriers, since continuous transmission is required on them. Final identification occurs with the FB of the FCCH. Once the received BCCH carriers have been found, the MS starts to synchronize with each of them and reads the system information, beginning with the BCCH with the highest RXLEV value.

This orientation concerning the current location can be accelerated considerably, if a list of BCCH carriers has been stored on the SIM card. Then the MS tries first to synchronize with some known carrier. Only if it cannot find any of the stored BCCH carrier frequencies, it does start with the normal BCCH search. A MS can store several lists for the recently visited networks.

5.1. 2. Channel measurement during a connection

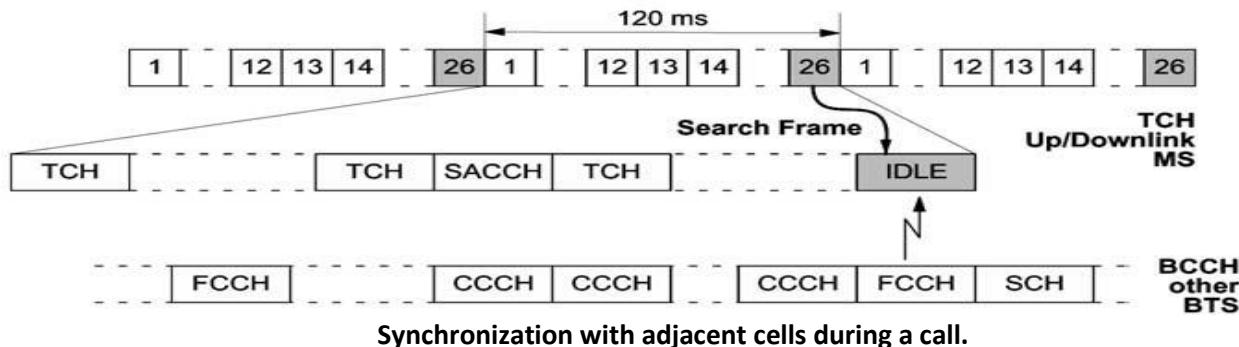
During a traffic (TCH) or signaling (SDCCH) connection, the channel measurement of the MS occurs over an SACCH interval, which comprises 104 TDMA frames in the case of a TCH channel (480 ms) or 102 TDMA frames (470.8 ms) in the case of an SDCCH channel.

For the channel at hand, two parameters are determined: the received signal level RXLEV and the signal quality RXQUAL. These two values are averaged over an SACCH interval (480 or 470.8 ms) and transmitted to the base station on the SACCH as a measurement report/measurement info. In this way the downlink quality of the channel assigned to the MS can be judged. In addition to these measurements of the downlink by the MS, the base station also measures the RXLEV and RXQUAL values of the respective uplink.

In order to make a handover decision, information about possible handover targets must be available. For this purpose, the MS has to observe continuously the BCCH carriers of up to six neighboring base stations. The RXLEV measurements of the neighboring BCCH carriers are performed during the MSs unused time slots (see Figure – Frequency hopping). The BCCH measurement results of the six strongest signals are included in the measurement report transmitted to the BSS.

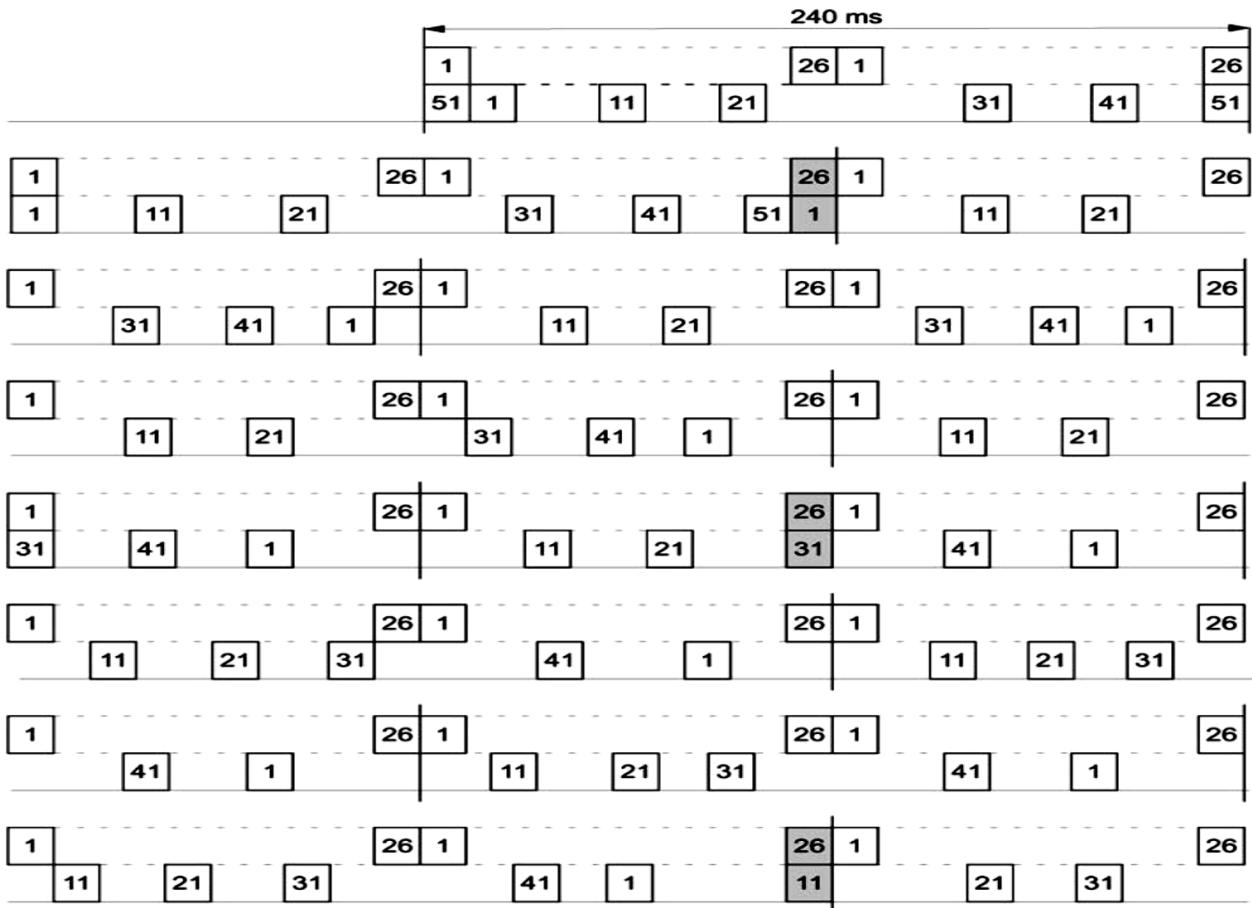
However, the received signal power level and the frequency of a BCCH carrier alone are not a sufficient criterion for a successful handover. Owing to the frequency reuse in cellular networks, and especially in the case of small clusters, it is possible that a cell can receive the same BCCH carrier from more than one neighboring cell, i.e. there exist several neighboring cells which use the same BCCH carrier. It is therefore necessary, to also know the identity (BSIC) of each neighboring cell. Simultaneously with the signal level measurement, the MS has to synchronize with each of the six neighboring BCCHs and read at least the SCH information.

For this purpose, one must first search for the FCCH burst of the BCCH carrier; then the SCH can be found in the next TDMA frame. Since the FCCH/SCH/BCCH is always transmitted in time slot 0 of the BCCH carrier, the search during a conversation for FCCHs can only be conducted in unused frames, i.e. in the case of a full-rate TCH in the IDLE frame of the multiframe (frame number 26 in Figure on page 23 and Figure below).



These free frames are therefore also known as search frames. Therefore, there are exactly four search frames within an SACCH block of 480 ms (four 26-frame multiframe of 120 ms). The MS has to examine the surrounding BCCH carriers for FCCH bursts, in order to synchronize with them and to decode the SCH. However, how can one search for synchronization points exactly within these frames during synchronized operation?

This is possible because the actual traffic channel and the respective BCCH carriers use different multiframe formats. Whereas the traffic channel uses the 26-frame multiframe format, time slot 0 of the BCCH carrier with the FCCH/SCH/BCCH is carried on a 51-frame multiframe format. This ratio of the different multiframe formats has the effect that the relative position of the search frames (frame 26 in a TCH multiframe) is shifting with regard to the BCCH multiframe by exactly one frame each 240 ms (Figure on next page).



Principle of FCCH search during the search frame.

Figuratively speaking, the search frame is travelling along the BCCH multiframe in such a way that at most after 11 TCH multiframe (= 1320 ms) a frequency correction burst of a neighboring cell becomes visible in a search frame. In this way, the MS is able to determine the BSIC for the respective RXLEV measurement value. Only BCCH carrier measurements whose identity can be established without doubt are included in the measurement report to the base station.

The base station can now make a handover decision based on these values, on the distance of the MS, and on the momentary interference of unused time slots. The algorithm for handover decisions has not been included in the GSM standard. The network operators may use algorithms which are optimized for their network or the local situation. GSM only gives a basic proposal which satisfies the minimum requirements for a handover decision algorithm. This algorithm defines threshold values, which must be violated in one or the other direction to arrive at a safe handover decision and to avoid so-called Ping-Pong handovers, which oscillate between two cells. Although the decision algorithm is part of radio subsystem link control, its discussion is postponed and it is treated together with handover signaling.

5.2 Transmission power control

Power classes (Table 4.8 on next page) are used for classification of base stations and MSs. The transmission power can also be controlled adaptively. As part of the radio subsystem link control, the MS's transmitter power is controlled in steps of 2 dBm.

The GSM transmitter power control has the purpose of limiting the MS's transmitter power to the minimum necessary level, in such a way that the base station receives signals from different MSs at approximately the same power level. Sixteen power control steps are defined for this purpose: step 0 (43 dBm = 20 W) to step 15 (13 dBm). Starting with the lowest, step 15, and the base station can increment the transmitter power of the MS in steps of 2 dBm up to the maximum power level of the respective power class of the MS. Similarly, the transmitter power of the base station can be controlled in steps of 2 dBm,

with the exception of the BCCH carrier of the base station, which must remain constant to allow comparative measurements of neighboring BCCH carriers by the MSs.

Transmission power control is based on the measurement values RXLEV and RXQUAL, for which one has defined upper and lower thresholds for uplink and downlink (Table 4.9 below). Network management defines the adjustable parameters P and N. If the values of P for the last N calculated mean values of the respective criterion (RXLEV or RXQUAL) are above or below the respective threshold value, the BSS can adjust the transmitter power (Figure below).

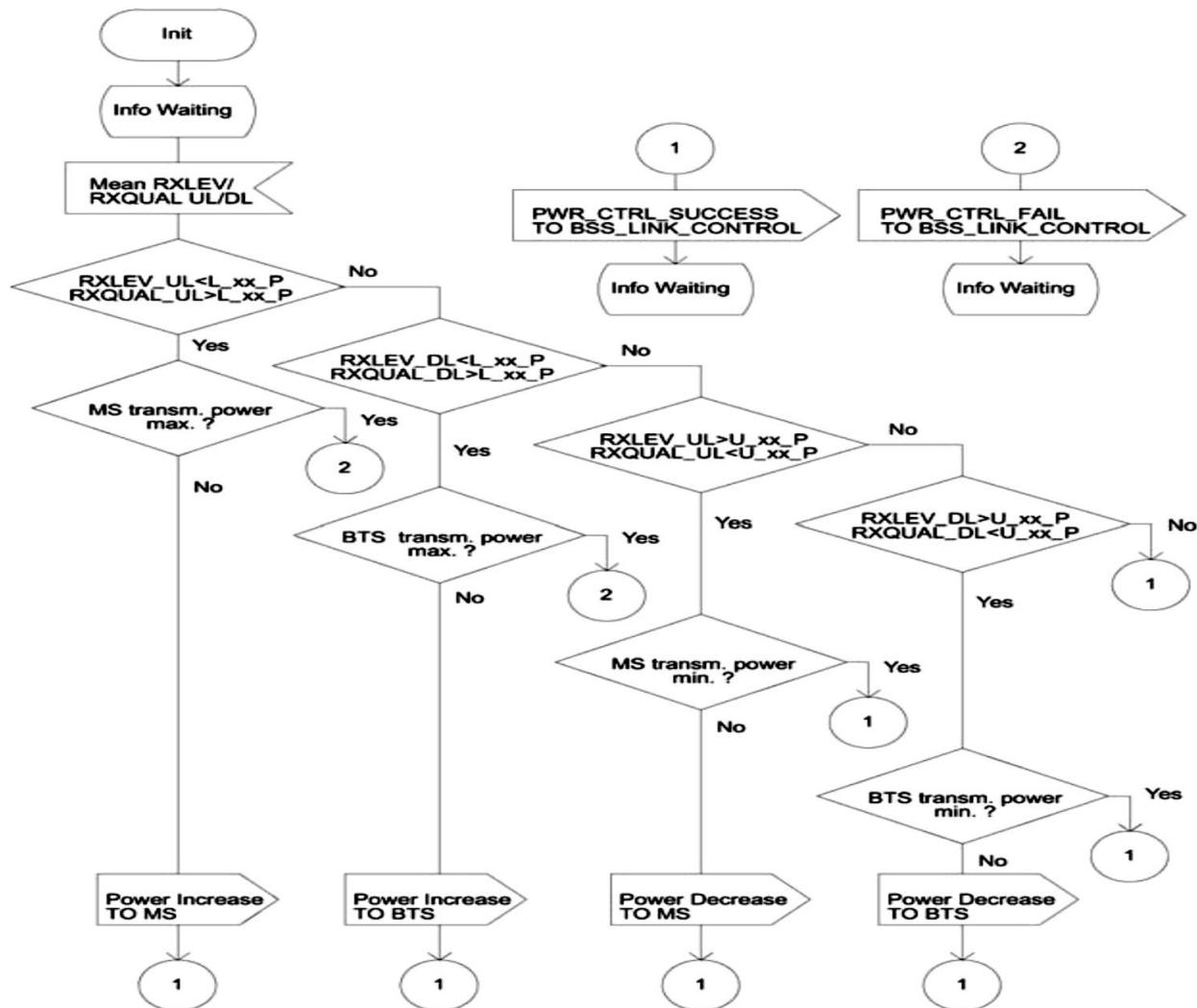
If the thresholds $U_{xx_UL_P}$ of the uplink are exceeded, the transmission power of the MS is reduced; in the other case, if the signal level is below the threshold $L_{xx_UL_P}$, the MS is ordered to increase its transmitter power. In an analogous way, the transmitter power of the base station can be adjusted, when the criteria for the downlink are exceeded in either direction.

Table 4.8 GSM power classes.

Power class	Maximum peak transmission power (W)	
	Mobile station (dBm)	Base station
1	20 (43)	320
2	<8 (39)	160
3	<5 (37)	80
4	<2 (33)	40
5	<0.8 (29)	20
6	-	10
7	-	5
8	-	2.5

Table 4.9 Thresholds for transmitter power control.

Threshold parameter	Typical value (dBm)	Meaning
$L_{RXLEV_UL_P}$	-103 to -73	Threshold for raising of transmission power in uplink or downlink
$L_{RXLEV_DL_P}$	-103 to -73	
$L_{RXQUAL_UL_P}$	-	
$L_{RXQUAL_DL_P}$	-	
$U_{RXLEV_UL_P}$	-	Threshold for reducing of transmission power in uplink or downlink
$U_{RXLEV_DL_P}$	-	
$U_{RXQUAL_UL_P}$	-	
$U_{RXQUAL_DL_P}$	-	

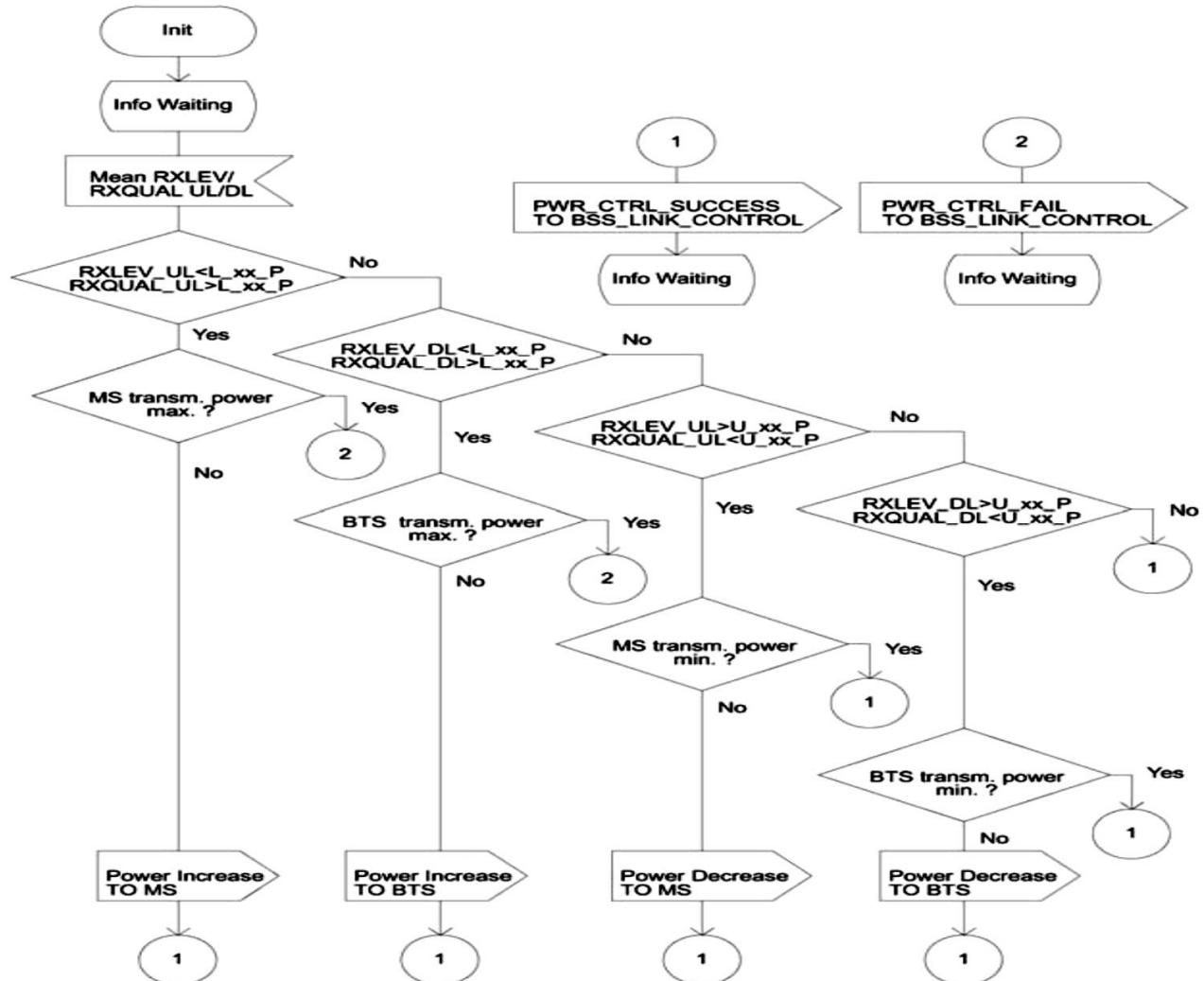


Schematic operation of transmitter power control

Even if the MS or base station signal levels stay within the thresholds, the current RXLEV/RXQUAL values can cause a change to another channel of the same or another cell based on the handover thresholds (next chapter). For this reason, checking for transmitter thresholds is immediately followed by a check of the handover thresholds as the second part of the radio subsystem link control (Figures 4.19 and 6.17 next chapter). If one of the threshold values is exceeded in either direction and the transmitter power cannot be adjusted accordingly, i.e. the respective transmitter power has reached its maximum or minimum value, this is an overriding cause for handover (PWR_CTRL_FAIL; Table) which the BSS must communicate immediately to the MSC.

5.3 Disconnection due to radio channel failure

The quality of a radio channel can vary considerably during an existing connection, or it can even fail in the case of shadowing. This should not lead to immediate disconnection, since such failures are often of short duration. For this reason GSM has a special algorithm within the Radio Subsystem Link Control which continuously checks for connectivity. It consists of recognizing a radio link failure by the inability to decode signaling information on the SACCH. This connectivity check is done both in the MS as well as in the base station. The connection is not immediately terminated, but is delayed so that only repeated consecutive failures (erroneous messages) represent a valid disconnect criterion. On the downlink, the MS must check the frequency of erroneous, nondecodable messages on the SACCH. The error protection on the SACCH has very powerful error correction capabilities and thus guarantees a very low probability of 10^{-10} for nonrecognized, wrongly corrected bits in SACCH messages.



Schematic operation of transmitter power control.

In this way, erroneous SACCH messages supply a measure for the quality of the downlink, which is already quite low when errors on the SACCH cannot be corrected any more. If a consecutive number of SACCH messages is erroneous, the link is considered bad and the connection is terminated. For this purpose, a counter S has been defined which is incremented by 2 with each arrival of an error-free message, and decremented by 1 for each erroneous SACCH message (Figure on previous page). When the counter reaches the value S = 0, the downlink is considered as failing, and the connection is terminated. This failure is signaled to the upper layers, Mobility Management (MM), which can start a call reestablishment procedure.

The maximum value RADIO_LINK_TIMEOUT for the counter S therefore determines the interval length during which a channel has to fail before a connection is terminated. After assignment of a dedicated channel (TCH or SDCCH), the MS starts the checking process by initializing the counter S with this value (Figure 4.23), which can be set individually per cell and is broadcast on the BCCH.

The corresponding checks are also conducted on the uplink. In both cases, however, this requires continuous transmission of data on the SACCH, i.e. when no signaling data have to be sent, filling data are transmitted. On the uplink, current measurement reports are transmitted, whereas the downlink carries system information of Type 5 and Type 6

5.4 Cell selection and operation in power conservation mode

5.4.1 Cell selection and cell reselection

A MS in idle mode must periodically measure the receivable BCCH carriers of the base stations in the area and calculate mean values RXLEV(n) from this data (section 4.5.1). Based on these measurements, the MS selects a cell, namely that with the best reception, i.e. the MS is committed to this cell. This is called ‘camping’ on this cell. In this state, accessing a service becomes possible, and the MS listens periodically to the PCH. Two criteria are defined for the automatic selection of cells: the path loss criterion C1 and the reselection criterion C2. The path loss criterion serves to identify cell candidates for camping. For such cells, C1 has to be greater than zero. At least every 5 s, a MS has to recalculate C1 and C2 for the current and neighboring cells. If the path loss criterion of the current cell falls below zero, the path loss to the current base station has become too large. A new cell has to be selected, which requires use of the criterion C2. If one of the neighboring cells has a value of C2 greater than zero, it becomes the new current cell. The cell selection algorithm uses two further threshold values, which are broadcast on the BCCH:

- The minimum received power level RXLEV_ACCESS_MIN (typically –98 to –106 dBm) required for registration into the network of the current cell;
- The maximum allowed transmitter power MS_TXPWR_CCH (typically 31–39 dBm) allowed for transmission on a control channel (RACH) before having received the first power control command.

In consideration of the maximal transmitter power P of a mobile station, the path loss criterion C1 is now defined using the minimal threshold RXLEV_ACCESS_MIN for network access and the maximal allowed transmitter power MS_TXPWR_MAX_CCH:

$$C1(n) = (RXLEV(n) - RXLEV_ACCESS_MIN$$

$$- \max(0, (MS_TXPWR_MAX_CCH - P)))..$$

The values of the path loss criterion C1 are determined for each cell for which a value RXLEV (n) of a BCCH carrier can be obtained. The cell with the lowest path loss can thus be determined using this criterion. It is the cell for which C1 > 0 has the largest value. During cell selection, the MS is not allowed to enter power conservation mode (Discontinuous Transmission (DTX), section 4.5.4).

A prerequisite for cell selection is that the cell considered belongs to the home Public Land Mobile Network (PLMN) of the MS or that access to the PLMN of this cell is allowed. Beyond that, a limited service mode has been defined with restricted service access, which still allows emergency calls if nothing else. In limited service mode, a MS can be camping on any cell but can only make emergency calls. Limited service

mode exists when there is no SIM card in the MS, when the IMSI is unknown in the network or the IMEI is barred from service, but also if the cell with the best value of C1 does not belong to an allowed PLMN.

Once a MS is camping on a cell and is in idle mode, it should keep observing all of the BCCH carriers whose frequencies, the BA, are broadcast on the current BCCH. Having left idle mode, e.g. if a TCH has been assigned, the MS monitors only the six strongest neighboring BCCH carriers. A list of these six strongest neighboring BCCH carriers has already been prepared and kept up to date in idle mode. The BCCH of the camped-on cell must be decoded at least every 30 s. At least once every 5 min, the complete set of data from the six strongest neighboring BCCH carriers has to be decoded, and the BSIC of each of these carriers has to be checked every 30 s. This allows the MS to stay aware of changes in its environment and to react appropriately. In the worst case, conditions have changed so much that a new cell to camp on needs to be selected (cell reselection). For this cell reselection, a further criterion C2, the reselection criterion, has been defined:

$$C2(n) = C1(n) + \text{CELL_RESELECT_OFFSET} - (\text{TEMPORARY_OFFSET} \times H(\text{PENALTY_TIME} - T)),$$

with $H(x) = \begin{cases} 0 & \text{for } x < 0, \\ 1 & \text{for } x \geq 0. \end{cases}$

The interval T in this criterion is the time passed since the MS observed the cell n for the first time with a value of $C1 > 0$. It is set back to 0 when the path loss criterion C1 falls to $C1 < 0$. The parameters CELL_RESELECT_OFFSET, TEMPORARY_OFFSET, and PENALTY_TIME are announced on the BCCH. However, as a default, they are set to 0.

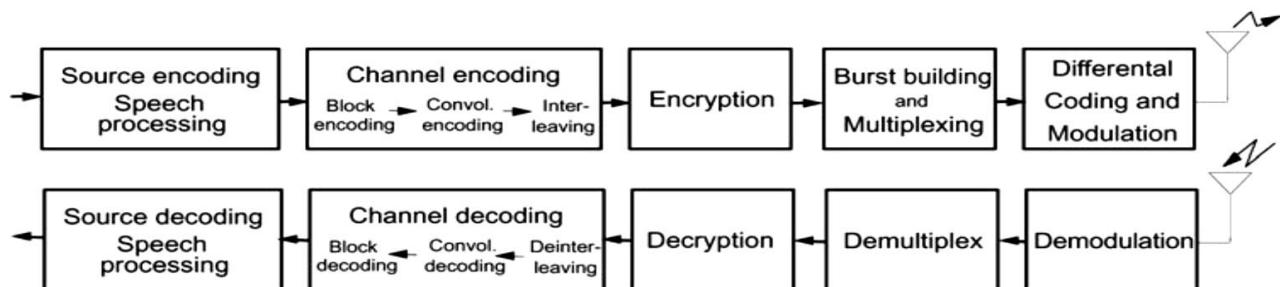
Otherwise, the criterion C2 introduces a time hysteresis for cell reselection. It tries to ensure that the MS is camping on the cell with the highest probability of successful communication. One exception for cell reselection is the case when a new cell belongs to another location area. In this case C2 must not only be larger than zero, but $C2 > \text{CELL_RESELECT_HYSTERESIS}$ to avoid too frequent location updates.

Discontinuous reception

To limit power consumption in idle mode and thus increase battery life in standby mode, the MS can activate the Discontinuous Reception (DRX) mode. In this mode, the receiver is turned on only for the phases of receiving paging messages and is otherwise in the power conservation mode which still maintains synchronization with BCCH signals through internal timers. In this DRX mode, measurement of BCCH carriers is performed only during unused time slots of the paging blocks.

6 Channel coding, source coding and speech processing

The previous sections explained the basic functions of the physical layer at the air interface, e.g. the definition of logical and physical channels, modulation, multiple access techniques, duplexing, and the definition of bursts. The following sections discuss several additional functions that are performed to transmit the data in an efficient, reliable way over the radio channel: source coding and speech processing, channel coding and burst mapping. Security related functions, such as encryption and authentication are discussed next chapter.



Basic elements of GSM transmission chain on the physical layer at the air interface.

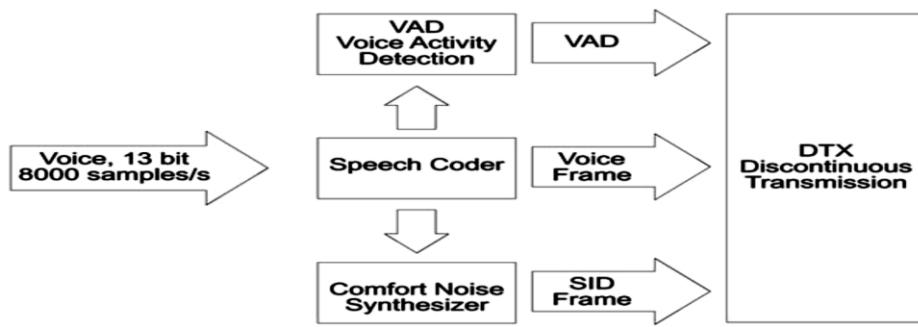
Figure above gives a schematic overview of the basic elements of the GSM transmission chain. The stream of sampled speech data is fed into a source encoder, which compresses the data by removing

unnecessary redundancy. The resulting information bit sequence is passed to the channel encoder. Its purpose is to add, in a controlled manner, some redundancy to the information sequence. This redundancy serves to protect the data against the negative effects of noise and interference encountered in the transmission through the radio channel. On the receiver side, the introduced redundancy allows the channel decoder to detect and correct transmission errors. GSM uses a combination of block and convolutional coding.

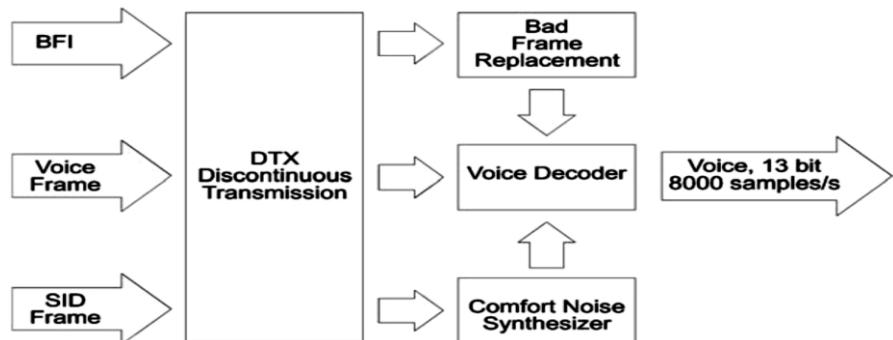
Moreover, an interleaving scheme is used to deal with burst errors that occur over multipath and fading channels. Next, the encoded and interleaved data are encrypted to guarantee secure and confident data transmission. The encryption technique as well as the methods for subscriber authentication and secrecy of the subscriber identity. The encrypted data are subsequently mapped to bursts, which are then multiplexed as explained in previous sections. Finally the stream of bits is differential coded and modulated. After transmission, the demodulator processes the signal, which was corrupted by the noisy channel. It attempts to recover the actual signal from the received signal. The next steps are demultiplexing and decryption. The channel decoder attempts to reconstruct the original information sequence and, as a final step, the source decoder tries to reconstruct the original source signal.

7. Source coding and speech processing

Source coding reduces redundancy in the speech signal and thus results in signal compression, which means that a significantly lower bit rate is achieved than needed by the original speech signal. The speech coder/decoder is the central part of the GSM speech processing function, both at the transmitter (Figure below shown) as well as at the receiver (Figure below shown).



Schematic representation of speech functions at the transmitter



Schematic representation of speech functions at the receiver.

The functions of the GSM speech coder and decoder are usually combined in one building block called the codec (COder/DECoder). The analog speech signal at the transmitter is sampled at a rate of 8000 samples per second and the samples are quantized with a resolution of 13 bits. This corresponds to a bit rate of 104 kbit/s for the speech signal. At the input to the speech codec, a speech frame containing 160

samples of 13 bits arrives every 20 ms. The speech codec compresses this speech signal into a source-coded speech signal of 260-bit blocks at a bit rate of 13 kbit/s. Thus, the GSM speech coder achieves a compression ratio of 1 to 8. The source coding procedure is briefly explained in the following;

A further ingredient of speech processing at the transmitter is the recognition of speech pauses, called Voice Activity Detection (VAD). The voice activity detector decides, based on a set of parameters delivered by the speech coder, whether the current speech frame (20 ms) contains speech or a speech pause. This decision is used to turn off the transmitter amplifier during speech pauses, under control of the Discontinuous Transmission (DTX) block.

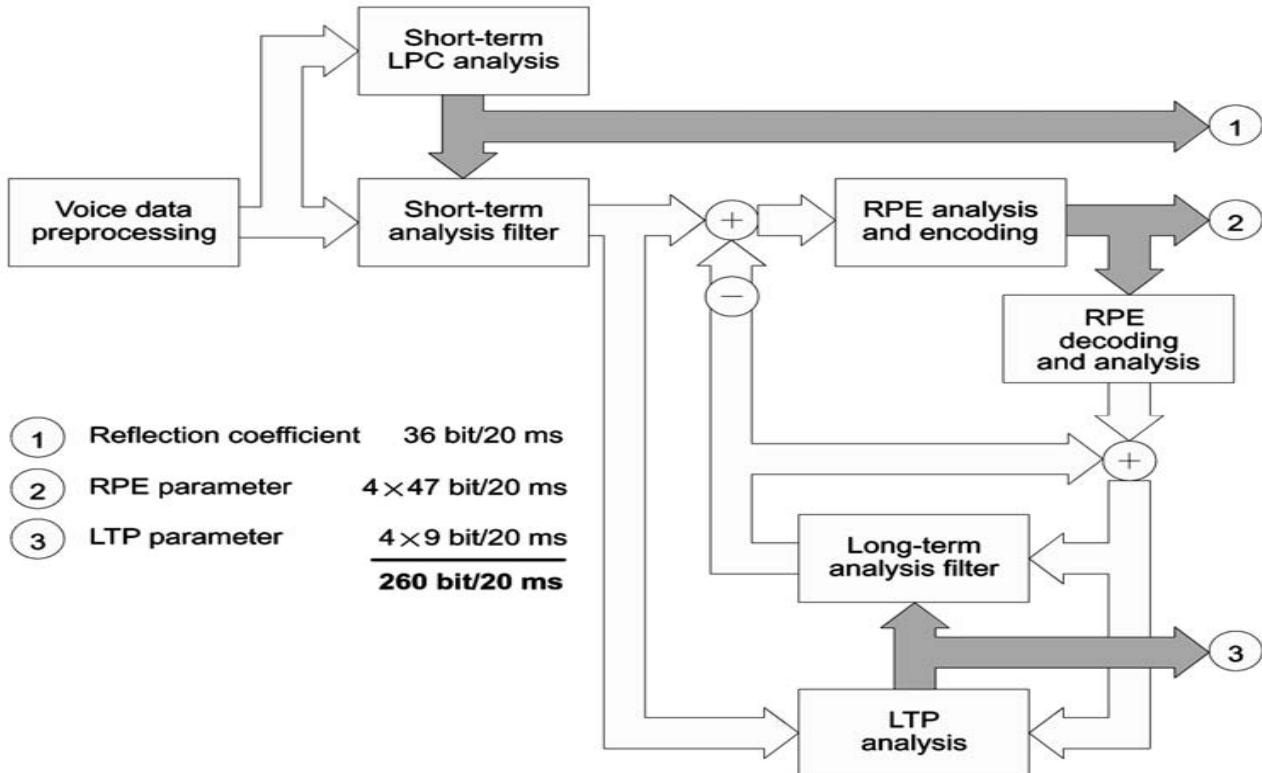
The discontinuous transmission mode takes advantage of the fact, that during a normal telephone conversation, both parties rarely speak at the same time, and thus each directional transmission path has to transport speech data only half of the time. In DTX mode, the transmitter is only activated when the current frame indeed carries speech information. This decision is based on the VAD signal of speech pause recognition. The DTX mode can reduce the power consumption and, hence, prolong the battery life. In addition, the reduction of transmitted energy also reduces the level of interference and thus improves the spectral efficiency of the GSM system. The missing speech frames are replaced at the receiver by a synthetic background noise signal called comfort noise (Figure previous page).

The parameters for the comfort noise synthesizer are transmitted in a special silence descriptor (SID) frame. This silence descriptor is generated at the transmitter from continuous measurements of the (acoustic) background noise level. It represents a speech frame which is transmitted at the end of a speech burst, i.e. at the beginning of a speech pause. In this way, the receiver recognizes the end of a speech burst and can activate the comfort noise synthesizer with the parameters received in the SID frame.

The generation of this artificial background noise in DTX mode prevents the audible background noise transmitted with normal speech bursts from suddenly dropping to a minimal level at a speech pause. This modulation of the background noise would have a very disturbing effect on the human listener and would significantly deteriorate the subjective speech quality. Insertion of comfort noise is a very effective countermeasure to compensate for this so-called noise-contrast effect.

Another loss of speech frames can occur, when bit errors caused by a noisy transmission channel cannot be corrected by the channel coding protection mechanism, and the block is received at the codec as a speech frame in error, which must be discarded. Such bad speech frames are flagged by the channel decoder with the Bad Frame Indication (BFI). In this case, the respective speech frame is discarded and the lost frame is replaced by a speech frame which is predictively calculated from the preceding frame. This technique is called error concealment. Simple insertion of comfort noise is not allowed. If 16 consecutive speech frames are lost, the receiver is muted to acoustically signal the temporary failure of the channel.

The speech compression takes place in the speech coder. The GSM speech coder uses a procedure known as Regular Pulse Excitation Long-Term Prediction (RPE-LTP). This procedure belongs to the family of hybrid speech coders. This hybrid procedure transmits part of the speech signal as the amplitude of a signal envelope, a pure wave form encoding, whereas the remaining part is encoded into a set of parameters. The receiver reconstructs these signal parts through speech synthesis (vocoder technique). Examples of envelope encoding are Pulse Code Modulation (PCM) or Adaptive Delta Pulse Code Modulation (ADPCM). A pure vocoder procedure is Linear Predictive Coding (LPC). The GSM procedure RPE-LTP as well as Code Excited Linear Predictive Coding (CELP) represent mixed (hybrid) approaches (David and Benkner, 1996; Natwig, 1998; Steele, 1992).



Simplified block diagram of the GSM speech coder.

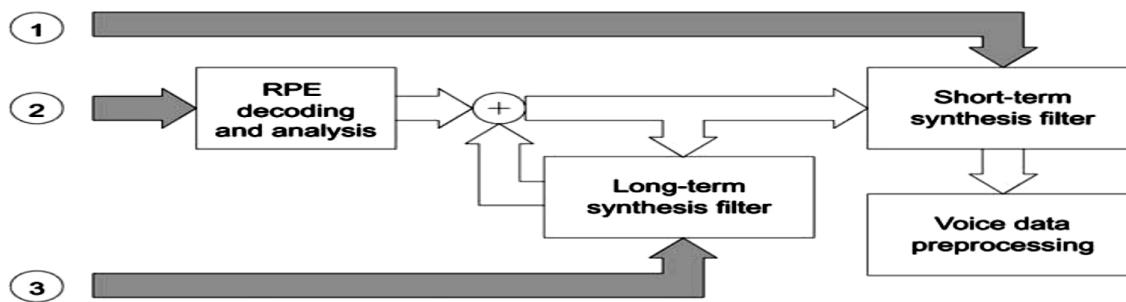
A simplified block diagram of the RPE-LTP coder is shown in Figure above. Speech data generated with a sampling rate of 8000 samples per second and 13-bit resolution arrive in blocks of 160 samples at the input of the coder. The speech signal is then decomposed into three components: a set of parameters for the adjustment of the short-term analysis filter (LPC) also called reflection coefficients; an excitation signal for the RPE part with irrelevant portions removed and highly compressed; and finally a set of parameters for the control of the LTP long-term analysis filter. The LPC and LTP analyses supply 36 filter parameters for each sample block, and the RPE coding compresses the sample block to 188 bits of RPE parameters. This results in the generation of a frame of 260 bits every 20 ms, equivalent to a 13 kbit/s GSM speech signal rate.

The speech data preprocessing of the coder (Figure above) removes the DC portion of the signal if present and uses a preemphasis filter to emphasize the higher frequencies of the speech spectrum. The preprocessed speech data are run through a nonrecursive lattice filter (LPC filter; Figure above) to reduce the dynamic range of the signal. Since this filter has a ‘memory’ of about 1 ms, it is also called short-term prediction filter. The coefficients of this filter, called reflection coefficients, are calculated during LPC analysis and transmitted in a logarithmic representation as part of the speech frame, Log Area Ratios (LARs).

Further processing of the speech data is preceded by a recalculation of the coefficients of the long-term prediction filter (LTP analysis in Figure above). The new prediction is based on the previous and current blocks of speech data. The resulting estimated block is finally subtracted from the block to be processed, and the resulting difference signal is passed on to the RPE coder.

After LPC and LTP filtering, the speech signal has been redundancy reduced, i.e. it already needs a lower bit rate than the sampled signal; however, the original signal can still be reconstructed from the calculated parameters. The irrelevance contained in the speech signal is reduced by the RPE coder. This irrelevance represents speech information that is not needed for the understandability of the speech signal,

since it is hardly noticeable to human hearing and thus can be removed without loss of quality. On the one hand, this results in a significant compression (factor $160 \times 13/188 \approx 11$); on the other hand, it has the effect that the original signal cannot be reconstructed uniquely.



Simplified block diagram of the GSM speech decoder.

Figure above summarizes the reconstruction of the speech signal from RPE data, as well as the long-term and short-term synthesis from LTP and LPC filter parameters. In principle, at the receiver site, the functions performed are the inverse of the functions of the encoding process. The irrelevance reduction only minimally affects the subjectively perceived speech quality, since the main objective of the GSM codec is not just the highest possible compression but also good subjective speech quality. To measure the speech quality in an objective manner, a series of tests were performed on a large number of candidate systems and competing codecs.

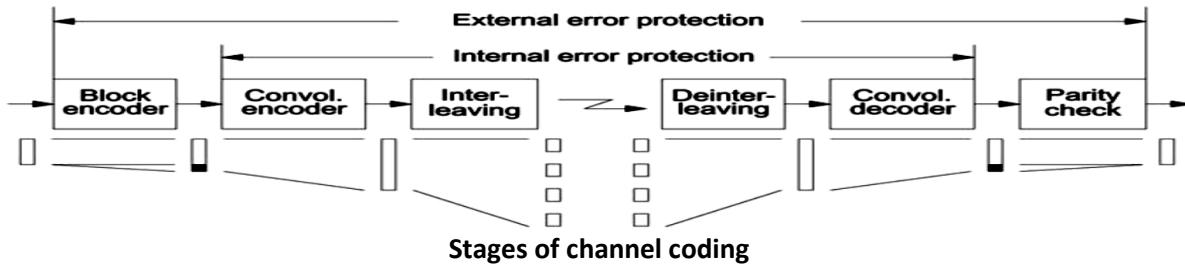
The base for comparison used is the Mean Opinion Score (MOS), ranging from MOS = 1, meaning quality is very bad or unacceptable, to MOS = 5, meaning quality is very good or fully acceptable. A series of coding procedures were discussed for the GSM system; they were examined in extensive hearing tests for their respective subjective speech quality (Natwig, 1998). Table 4.10 below gives an overview of these test results; it includes as reference also ADPCM and frequency-modulated analog transmission. The GSM codec with the RPELTP procedure generates a speech quality with an MOS value of about 4 for a wide range of different inputs.

Table 4.10 MOS results of codec hearing tests (Natwig, 1998).

CODEC	Process	Bit rate (in kbit/s)	MOS
FM	Frequency Modulation	—	1.95
SBC-ADPCM	Subband-CODEC – Adaptive Delta-PCM	15	2.92
SBC-APCM	Subband-CODEC – Adaptive PCM	16	3.14
MPE-LTP	Multi-Pulse Excited LPC-CODEC – Long-Term Prediction	16	3.27
RPE-LPC	Regular-Pulse Excited LPC-CODEC	13	3.54
RPE-LTP	Regular Pulse Excited LPC-CODEC – Long-Term Prediction	13	≈4
ADPCM	Adaptive Delta Modulation	32	≥4

8. Channel coding

The heavily varying properties of the mobile radio channel (see Chapter 2) often result in a very high bit error ratio, of the order of 10^{-3} to 10^{-1} . The highly compressed, redundancy-reduced source coding makes speech communication with acceptable quality almost impossible; moreover, it makes reasonable data communication impossible. Suitable error correction procedures are therefore necessary to reduce the bit error probability into an acceptable range of about 10^{-5} to 10^{-6} . Channel coding, in contrast to source coding, adds redundancy to the data stream to enable detection and correction of transmission errors. It is the modern high-performance coding and error correction techniques which essentially enable the implementation of a digital mobile communication system.



The GSM system uses a combination of several procedures: in addition to a block code, which generates parity bits for error detection, a convolutional code generates the redundancy needed for error correction. Furthermore, sophisticated interleaving of data over several blocks reduces the damage done by burst errors. The individual steps of channel coding are shown in Figure above.

- Calculation of parity bits (block code) and addition of fill bits;
- Error protection coding through convolutional coding;
- Interleaving.

Finally, the coded and interleaved blocks are enciphered, distributed across bursts, modulated and transmitted on the respective carrier frequencies. The sequence of data blocks that arrives at the input of the channel encoder is combined into blocks, partially supplemented by parity bits (depending on the logical channel) and then complemented to a block size suitable for the convolutional encoder. This involves appending zero bits at the end of each data block, which allow a defined resetting procedure of the convolutional encoder (zero termination) and thus a correct decoding decision. Finally, these blocks are run through the convolutional encoder. The ratio of uncoded to coded block length is called the rate of the convolutional code. Some of the redundancy bits generated by the convolutional encoder are deleted again for some of the logical channels.

This procedure is known as puncturing, and the resulting code is a punctured convolutional code (Begin and Haccoun, 1994; Hagenauer et al., 1990; Kallel, 1995). Puncturing increases the rate of the convolutional code, so it reduces the redundancy per block to be transmitted, and lowers the bandwidth requirements, such that the convolution-encoded signal fits into the available channel bit rate. The convolution-encoded bits are passed to the interleaver which shuffles various bit streams. At the receiving site, the respective inverse functions are performed: deinterleaving, convolutional decoding and parity checking. Depending on the position within the transmission chain (Figure above shown), one distinguishes between external error protection (block code) and internal protection (convolutional code).

The basic unit for all coding procedures is the data block. For example, the speech coder delivers a sequence of data blocks to the channel encoder. Depending on the logical channel, the length of the data block is different; after convolutional coding at the latest, data from all channels are transformed into units of 456 bits. Such a block of 456 bits transports a complete speech frame or a protocol message in most of the signaling channels, except for the RACH and SCH channels. The starting points are the blocks delivered to the input of the channel encoder from the protocol processing in higher layers (Figure shown in next page).

Speech traffic channels

One block of the full-rate speech codec consists of 260 bits of speech data, i.e. each block contains 260 information bits, which must be encoded. They are graded into two classes (class I, 182 bits; class II, 78 bits) which have different sensitivity against bit errors. Class I includes speech bits that have a greater impact on speech quality and hence must be better protected. Speech bits of class II, however, are less important. They are therefore transmitted without convolutional coding, but are included in the interleaving process. The individual sections of a speech frame are therefore protected to differing degrees against transmission errors (Unequal Error Protection (UEP)). In the case of a half-rate speech codec, data blocks of 112 information bits are input into the channel encoder. Of these, 95 bits belong to class I and 17 bits belong to class II. Again, one data block corresponds to one speech frame.

Data traffic channels

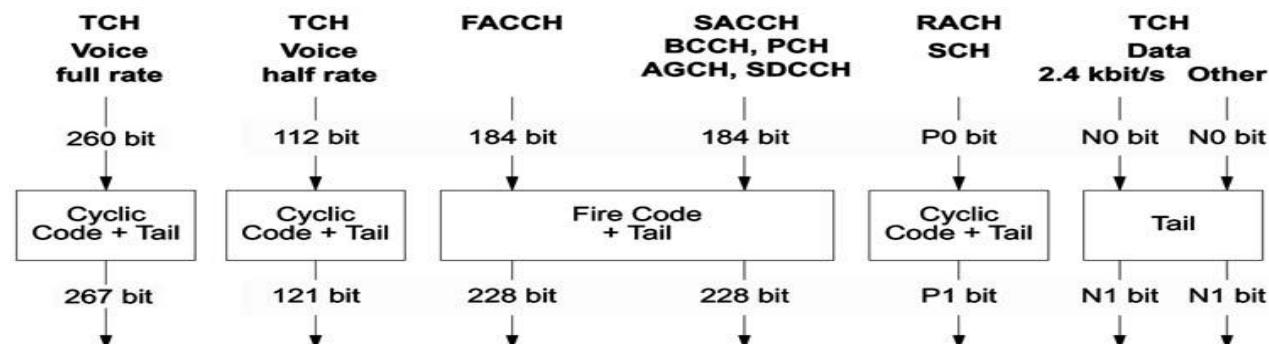
Blocks of traffic channels for data services have a length of N_0 bits, the value of N_0 being a function of the data service bit rate. We take for example the 9.6 kbit/s data service on a full rate traffic channel (TCH/F9.6). Here, a bit stream organized into blocks of 60 information bits arrives every 5 ms at the input of the encoder. Four subsequent blocks are combined for the encoding process.

Signaling channels

The data streams of most of the signaling channels are constructed of blocks of 184 bits each; with the exception of the RACH and SCH which supply blocks of length P_0 to the channel coder. The block length of 184 bits results from the fixed length of the protocol message frames of 23 octets on the signaling channels. The channel coding process maps pairs of subblocks of 57 bits onto the bursts such that it can fill a normal data burst NB (Figure burst mode).

8.1 External error protection: block coding

The block coding stage in GSM has the purpose of generating parity bits for a block of data, which allow the detection of errors in this block. In addition, these blocks are supplemented by fill bits (tail bits) to a block length suitable for further processing. Since block coding is the first or external stage of channel coding, the block code is also known as external protection. Figure below gives a brief overview showing which codes are used for which channels. In principle, only two kinds of codes are used: a Cyclic Redundancy Check (CRC) and a fire code.



Overview of block coding for logical channels (also see Table 4.11).

Table 4.11 Error protection coding and interleaving of logical channels.

Channel type	Abbreviation	Block distance (ms)	Bits per block			Convolution code rate	Encoded bits per block	Interleaver depth
			Data	Parity	Tail			
TCH, full rate, speech	TCH/FS	20	260				456	8
Class I			182	3	4	1/2	378	
Class II			78	0	0	-	78	
TCH, half rate, speech	TCH/H/S	20	112				228	4
Class I			95	3	6	104/211	211	
Class II			17	0	0	-	17	
TCH, full rate, 14.4 kbit/s	TCH/F14.4	20	290	0	4	294/456	456	19
TCH, full rate, 9.6 kbit/s	TCH/F9.6	5	4 × 60	0	4	244/456	456	19
TCH, full rate, 4.8 kbit/s	TCH/F4.8	10	60	0	16	1/3	228	19
TCH, half rate, 4.8 kbit/s	TCH/H4.8	10	4 × 60	0	4	244/456	456	19
TCH, full rate, 2.4 kbit/s	TCH/F2.4	10	2 × 36	0	4	1/6	456	8
TCH, half rate, 2.4 kbit/s	TCH/H2.4	10	2 × 36	0	4	1/3	228	19
FACCH, full rate	FACCH/F	20	184	40	4	1/2	456	8
FACCH, half rate	FACCH/H	40	184	40	4	1/2	456	6
SDCCH, SACCH			184	40	4	1/2	456	4
BCCH, NCH, AGCH, PCH		235	184	40	4	1/2	456	4
RACH		235	8	6	4	1/2	36	1
SCH			25	10	4	1/2	78	1
CBCCH		235	184	40	4	1/2	456	4

8.1.1 Block coding for speech traffic channel

As mentioned above, speech data occurs on the TCH in speech frames (blocks) of 260 bits for TCH/F and 112 bits for TCH/H, respectively. The bits belonging to class I are error-protected, whereas the bits of class II and are not protected. A 3-bit CRC code is calculated for the first 50 bits of class I (in the case of TCH/F). The generator polynomial for this CRC is

$$G_{\text{CRC}}(x) = x^3 + x + 1.$$

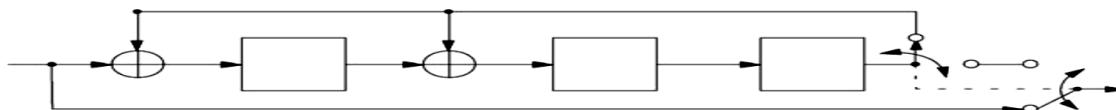
In the case of a TCH/H speech channel, the most significant 22 bits of Class I are protected by 3 parity bits, using the same generator polynomial.

We now explain the block coding process in more detail with focus on the TCH/F speech codec. Since cyclic codes are easily generated with a feedback shift register, they are often defined directly with this register representation. Figure 4.31 shows such a shift register with storage locations (delay elements) and modulo-2 adders. For initialization, the register is primed with the first three bits of the data block. The other data are shifted bitwise into the feedback shift register; after the last data bit has been shifted out of the register, the register contains the check sum bits, which are then appended to the block. The operation of this shift register can be easily explained if the bit sequences are also represented as polynomials as with the generating function. The first 50 bits of a speech frame D0, D1, ..., D49 are denoted as

$$D(x) = D_{49}x^{49} + D_{48}x^{48} + \dots + D_1x + D_0.$$

If this data sequence is shifted through the register of Figure shown below, after the register was primed with D47, D48, D49 followed by 50 shift operations, then the check sum bits R(x) correspond to the remainder, which is left by dividing the data sequence x3D(x) (supplemented by three zero bits) by the generator polynomial:

$$R(x) = \text{Remainder} \left[\frac{x^3 D(x)}{G_{\text{CRC}}(x)} \right]$$



Feedback shift register for CRC.

In the case of error-free transmission, the codeword C '(x) = x³D(x) + R(x) is therefore divisible by GCRCC(x) without remainder. However, since the check sum bits R(x) are transmitted in inverted form, the division yields a remainder:

$$S(x) = \text{Remainder} \left[\frac{C(x)}{G_{\text{CRC}}(x)} \right] = \text{Remainder} \left[\frac{x^3 D(x) + \bar{R}(x)}{G_{\text{CRC}}(x)} \right] = x^2 + x + 1.$$

This is equivalent to shifting the whole codeword C(x) through an identical shift register on the decoder side, after priming it with C50, C51, and C52. After shifting in the last check sum bit (50 shift operations), this register should contain a 1. If this is not the case, the block contains erroneous bits. Inversion of the parity bits avoids the generation of null code-words, i.e. bursts which contain only zeros cannot occur on the traffic channel. The speech data d(k) (k = 1, ..., 182) of class I of a block are combined with the parity bits p(k) (k = 1, 2, 3) and fill bits to form a new block u(k) (k = 1, ..., 189)

$$u(k) = \begin{cases} d(2k) & k = 1, \dots, 90, \\ d(2 \times (184 - k) + 1) & k = 94, \dots, 184, \\ p((k - 91) + 1) & k = 91, 92, 93, \\ 0 & k = 185, \dots, 189. \end{cases}$$

The bits in even or odd positions are shifted to the upper or lower half of the block, respectively, and separated by the three check sum bits; in addition, the order of the odd bits is reversed. Finally, the block is filled to 189 bits. Combination with the speech bits of class II yields a block of 267 bits, which serves as input to the convolutional coder. This enormous effort is taken because of the high compression rate and sensitivity against bit errors of the speech data. A speech frame in which the bits of class I have been

recognized as erroneous can therefore be reported as erroneous to the speech codec using the BFI;. In order to maintain a constantly good speech quality, speech frames recognized as faulty are discarded, and the last correctly received frame is repeated, or an extrapolation of received speech data is performed.

8.1.2 Block coding for data traffic channels

Block coding of traffic channels is somewhat simpler for data services. In this case, no parity bits are determined. Blocks of length N0 arriving at the input of the encoder are supplemented by fill bits to a size of N1 suitable for further coding. Table 4.12 below gives an overview of the different block lengths, which depend on the data rate and channel type, i.e. whether the channel is a full-rate (TCH/Fxx) or half-rate (TCH/Hxx) channel.

Table 4.12 Block formation for data traffic channels.

Data channel	N0	Tail bits	=	N1
TCH/F14.4	290	+	4	294
TCH/F9.6	4×60	+	4	244
TCH/F4.8	$(2 \times) 60$	+	$(2 \times) 16$	$(2 \times) 76$
TCH/H4.8	4×60	+	4	244
TCH/F2.4	2×36	+	4	76
TCH/H2.4	$(2 \times) 2 \times 36$	+	$(2 \times) 4$	$(2 \times) 76$

The 9.6 kbit/s data service is only offered on a full-rate traffic channel. The data comes in blocks of 60 bits to the channel encoder (every 5 ms). Four blocks each are combined and supplemented by four appended tail bits (zero bits). In the case of a nontransparent data service, these four blocks make up exactly one protocol frame of the Radio Link Protocol (RLP; 240 bits).

The procedures for other data services are similar. As shown in Table 4.11 previous table, for the 4.8 and 2.4 kbit/s services, blocks of 60 or 36 bit length arrive every 10 ms. Subsequent blocks are combined and are then supplemented with tail bits (zero bits) to form blocks of 76 or 244 bits, respectively. The bit stream for the 14.4 kbit/s data service (TCH/F14.4) is offered to the encoder in blocks of 290 information bits every 20 ms. here, four tail bits are added, resulting in 294 bits.

8.1.3 Block coding for signaling channels

The majority of the signaling channels (SACCH, FACCH, SDCCH, BCCH, PCH and AGCH) use an extremely powerful block code for error detection. This is a so-called fire code, i.e. a shortened binary cyclic code which appends 40 redundancy bits to the 184-bit data block. Its pure error-detection capability is sufficient to let undetected errors go through only with a probability of 2^{-40} . (A fire code can also be used for error correction, but here it is used only for error detection.) Error detection with the fire code in the SACCH channel is used to verify connectivity (Figure 4.23), and is used, if indicated, for decisions regarding breaking a connection. The fire code can be defined in the same way as the CRC by way of a generator polynomial:

$$TG_F(x) = (x^{23} + 1)(x^{17} + x^3 + 1).$$

The check sum bits RF(s) of this code are calculated in such a way that a 40-bit remainder SF(x) is left after dividing the codeword CF(x) by the generator polynomial GF(x). In the case of no errors, the remainder contains only '1' bits:

$$\begin{aligned} SF(x) &= \text{Remainder} \left[\frac{CF(x)}{GF(x)} \right] = \text{Remainder} \left[\frac{x^{40} D_F(x) + R_F(x)}{GF(x)} \right] \\ &= x^{39} + x^{38} + \dots + x^2 + x + 1. \end{aligned}$$

The codeword generated with the redundancy bits of the fire code is supplemented with '0' bits to a total length of 228 bits, which are then delivered to the convolutional coder. Another approach has been used for error detection in the RACH channel. The very short random access burst in the RACH allows only a data block length of P0 = 8 bits, which is supplemented in a cyclic code by six redundancy bits. The corresponding generator polynomial is

$$G_{RACH}(x) = x^6 + x^5 + x^3 + x^2 + x + 1.$$

In the AB, the MS also has to indicate a target base station. The BSIC of the respective base station is used for this purpose. The six bits of the BSIC are added to the six redundancy bits modulo-2, and the

resulting sequence is inserted as the redundancy of the data block. The total codeword to be convolution-coded for the RACH thus has a length of 18 bits; i.e. four fill bits ('0') are also added in the RACH to this block. In exactly the same way, block coding is performed for the handover access burst, which is in principle also a random access burst.

The SCH channel, as an important synchronization channel, uses somewhat more elaborate error protection than the RACH channel. The SCH data blocks have a length of 25 bits and receive, in addition to the fill bits, another 10 bits of redundancy for error detection through a cyclic code with somewhat better error detection capability than on the RACH:

$$G_{SCH}(x) = x^{10} + x^8 + x^6 + x^5 + x^4 + x^2 + 1$$

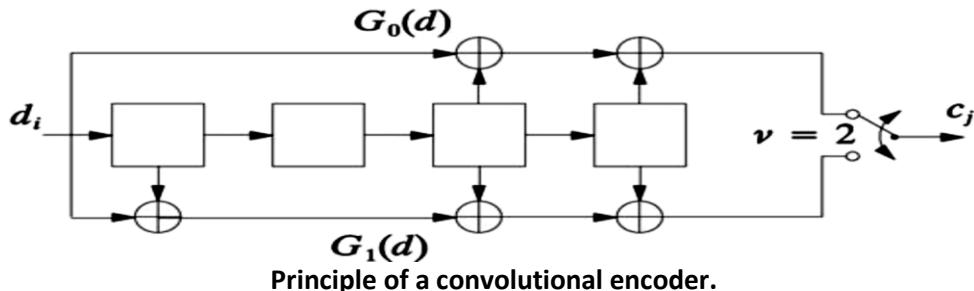
Thus the length of the codeword delivered to the channel coder in the SCH channel is 39 bits. Table 4.13 below summarizes the block parameters of the RACH and SCH channels. Table 4.14 presents an overview of the cyclic codes used in GSM.

Table 4.13 Block lengths for the RACH and SCH channels.

Data channel	P0	Parity bits	Tail bits	P1
RACH	8	+6	+4	= 18
SCH	25	+10	+4	= 39

8.2 Internal error protection: convolutional coding

After block coding has supplemented the data with redundancy bits for error detection (parity bits), added fill bits and thus generated sorted blocks, the next stage is the calculation of additional redundancy for error correction to correct the transmission errors caused by the radio channel. The internal error correction of GSM is based exclusively on convolutional codes.



Convolutional codes (Johannesson and Zigangirov, 1999) can also be defined using shift registers and generator polynomials. Figure above illustrates a possible convolutional encoder realization. It basically consists of a shift register with modulo-2 adders and K storage locations (here K = 4). One data/information symbol d_i is read into the shift register per tact interval. A symbol consists of k (here k = 1) data/information bits, each of which is moved into the shift register. A data symbol could also consist of more than one bit ($k > 1$), but this is not implemented in GSM.

The symbol read is combined with up to K of its predecessor symbols d_{i-1}, \dots, d_{i-K} in several modulo-2 additions. The results of these operations are given to the interleaver as coded user payload symbols c_j . The value K determines the number of predecessor symbols to be combined with a data symbol and is therefore also called the memory of the convolutional encoder. The number v of combinatorial rules (here $v = 2$) determines the number of coded bits in a code symbol C_j generated for each input symbol d_i . In Figure above shown, the combinatorial results are scanned from top to bottom to generate the code symbol C_j . The combinatorial rules are defined by the generator polynomial $G_i(d)$. It is important to note that a specific convolutional code can be generated by various encoders. Thus, it must be carefully distinguished between code properties and encoder properties.

As mentioned in section block coding appends at least four zero bits to each block. These bits not only serve as fill bits at the end of a block, but they are also important for the channel coding procedure. Shifted at the end of each block into the encoder, these bits serve to reset the encoder into the defined starting position (zero termination of the encoder), such that in principle adjacent data blocks can be coded independently of each other. The rate r of a convolutional code indicates how many data (information) bits are processed for each coded bit. Consequently, $1/r$ is the number of coded bits per information bit. This rate is the essential measure of the redundancy produced by the code and, hence, its error correction capability:

$$r = \frac{k}{v}, \quad \text{here: } r = \frac{1}{v} = \frac{1}{2}.$$

The code rate is therefore determined by the number of bits k per input data symbol and the number of combinatorial rules v which are used for the calculation of a code symbol. In combination with the memory K , the code rate r determines the error-correction capability of the code. In a simplified way: with decreasing r and increasing K , the number of corrigible errors per codeword increase, and, thus, the error-correction capabilities of the code are improved. The encoding procedure is expressed in the combinatorial operations (modulo-2 additions). These coding rules can be described with polynomials. In the case of the convolutional encoder of Figure previous, the two generator polynomials are

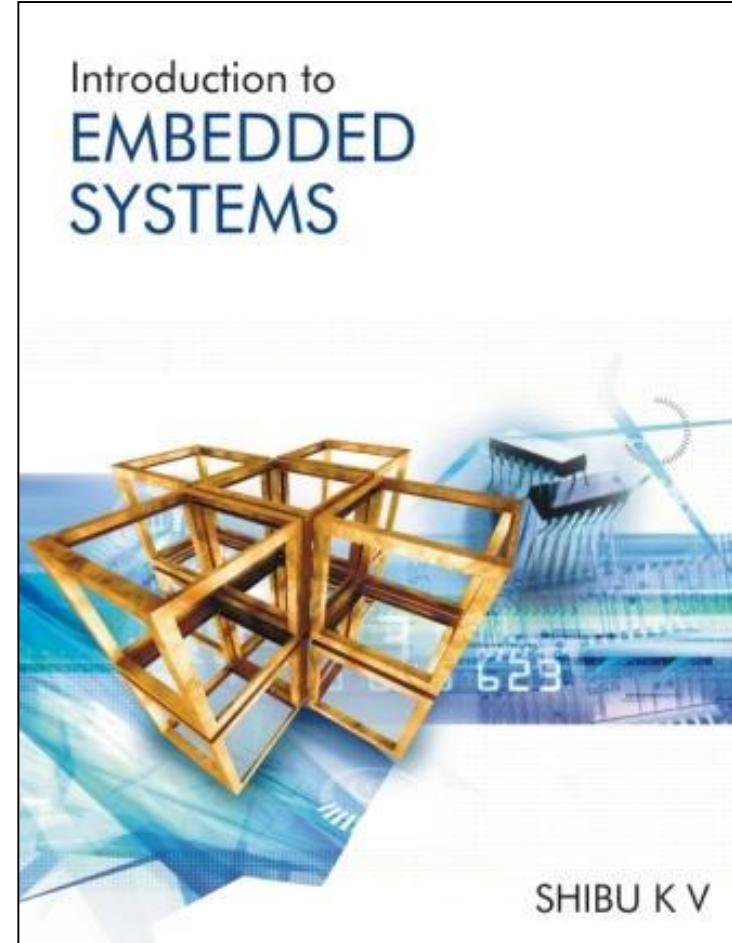
$$G_0(d) = d^4 + d^3 + 1,$$

$$G_1(d) = d^4 + d^3 + d + 1.$$

Embedded System Design

VI Semester

**Mr. Nagesh H B and
Harish L
ACSCE**



Introduction to Embedded System

What is Embedded System?

An Electronic/Electro mechanical system which is designed to perform a specific function and is a combination of both hardware and firmware (Software)

E.g. Electronic Toys, Mobile Handsets, Washing Machines, Air Conditioners, Automotive Control Units, Set Top Box, DVD Player etc...

Embedded Systems are:

- Unique in character and behavior
- With specialized hardware and software

Embedded Systems Vs General Computing Systems

General Purpose System	Embedded System
A system which is a combination of generic hardware and General Purpose Operating System for executing a variety of applications	A system which is a combination of special purpose hardware and embedded OS for executing a specific set of applications
Contain a General Purpose Operating System (GPOS)	May or may not contain an operating system for functioning
Applications are alterable (programmable) by user (It is possible for the end user to re-install the Operating System, and add or remove user applications)	The firmware of the embedded system is pre-programmed and it is non-alterable by end-user (There may be exceptions for systems supporting OS kernel image flashing through special hardware settings)
Performance is the key deciding factor on the selection of the system. Always 'Faster is Better'	Application specific requirements (like performance, power requirements, memory usage etc) are the key deciding factors
Less/not at all tailored towards reduced operating power requirements, options for different levels of power management.	Highly tailored to take advantage of the power saving modes supported by hardware and Operating System
Response requirements are not time critical	For certain category of embedded systems like mission critical systems, the response time requirement is highly critical
Need not be deterministic in execution behavior	Execution behavior is deterministic for certain type of embedded systems like 'Hard Real Time' systems

Introduction to Embedded System

Classification of Embedded Systems:

- Based on Generation
- Based on Complexity & Performance Requirements
- Based on deterministic behavior
- Based on Triggering

Introduction to Embedded System

Embedded Systems - Classification based on Generation

First Generation: The early embedded systems built around 8bit microprocessors like 8085 and Z80 and 4bit microcontrollers

Second Generation: Embedded Systems built around 16bit microprocessors and 8 or 16bit microcontrollers, following the first generation embedded systems

Third Generation: Embedded Systems built around high performance 16/32 bit Microprocessors/controllers, Application Specific Instruction set processors like Digital Signal Processors (DSPs), and Application Specific Integrated Circuits (ASICs)

Fourth Generation: Embedded Systems built around System on Chips (SoCs), Re-configurable processors and multicore processors

Introduction to Embedded System

Embedded Systems - Classification based on Complexity & Performance

Small Scale: The early embedded systems built around 8bit microprocessors like 8085 and Z80 and 4bit microcontrollers

Medium Scale: Embedded Systems built around 16bit microprocessors and 8 or 16bit microcontrollers, following the first generation embedded systems

Large Scale/Complex: Embedded Systems built around high performance 16/32 bit Microprocessors/controllers, Application Specific Instruction set processors like Digital Signal Processors (DSPs), and Application Specific Integrated Circuits (ASICs)

Introduction to Embedded System

Embedded Systems - Classification based on Deterministic Behaviour.

- The classification based on these are applicable for “**Real Time**” systems.
- The application/task execution behavior for an embedded systems can be either deterministic or non-deterministic.
- Based on execution behavior , real time embedded systems are classified into Hard and Soft.

Introduction to Embedded System

Embedded Systems - Classification based on trigger.

- Embedded systems are Reactive in nature{like process control system in industrial control applications) can be classified based on the trigger.
- Reactive systems can be either event triggered or time triggered.

Introduction to Embedded System

Major Application Areas of Embedded Systems

- Consumer Electronics: Camcorders, Cameras etc.
- Household Appliances: Television, DVD players, Washing machine, Fridge, Microwave Oven etc.
- Home Automation and Security Systems: Air conditioners, sprinklers, Intruder detection alarms, Closed Circuit Television Cameras, Fire alarms etc.
- Automotive Industry: Anti-lock breaking systems (ABS), Engine Control, Ignition Systems, Automatic Navigation Systems etc.
- Telecom: Cellular Telephones, Telephone switches, Handset Multimedia Applications etc.
- Computer Peripherals: Printers, Scanners, Fax machines etc.
- Computer Networking Systems: Network Routers, Switches, Hubs, Firewalls etc.
- Health Care: Different Kinds of Scanners, EEG, ECG Machines etc.
- Measurement & Instrumentation: Digital multi meters, Digital CROs, Logic Analyzers PLC systems etc.
- Banking & Retail: Automatic Teller Machines (ATM) and Currency counters, Point of Sales (POS)
- Card Readers: Barcode, Smart Card Readers, Hand held Devices etc.

Introduction to Embedded System

Purpose of Embedded Systems

Each Embedded Systems is designed to serve the purpose of any one or a combination of the following tasks.

- Data Collection/Storage/Representation
- Data Communication
- Data (Signal) Processing
- Monitoring
- Control
- Application Specific User Interface

Introduction to Embedded System

Purpose of Embedded Systems – Data Collection/Storage/Representation

- ✓ Performs acquisition of data from the external world.
- ✓ The collected data can be either analog or digital
- ✓ Data collection is usually done for storage, analysis, manipulation and transmission
- ✓ The collected data may be stored directly in the system or may be transmitted to some other systems or it may be processed by the system or it may be deleted instantly after giving a meaningful representation



Digital Camera for Image capturing/storage/display

Photo Courtesy of Casio -Model EXILIM ex-Z850

(www.casio.com)

Introduction to Embedded System

Purpose of Embedded Systems – Data Communication

- ✓ Embedded Data communication systems are deployed in applications ranging from complex satellite communication systems to simple home networking systems
- ✓ Embedded Data communication systems are dedicated for data communication
- ✓ The data communication can happen through a wired interface (like Ethernet, RS-232C/USB/IEEE1394 etc) or wireless interface (like Wi-Fi, GSM,/GPRS, Bluetooth, ZigBee etc)
- ✓ Network hubs, Routers, switches, Modems etc are typical examples for dedicated data transmission embedded systems



Wireless Network Router for Data Communication

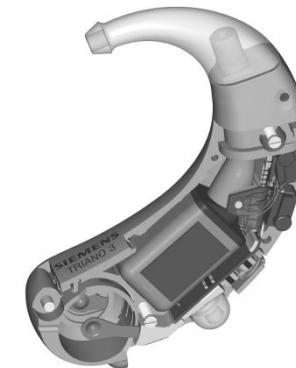
Photo Courtesy of Linksys (www.linksys.com).

A division of CISCO system

Introduction to Embedded System

Purpose of Embedded Systems – Data (Signal) Processing

- ✓ Embedded systems with Signal processing functionalities are employed in applications demanding signal processing like Speech coding, synthesis, audio video codec, transmission applications etc
- ✓ Computational intensive systems
- ✓ Employs Digital Signal Processors (DSPs)



Digital hearing Aid employing Signal Processing Technique
Siemens TRIANO 3 Digital hearing aid;
Siemens Audiology Copyright © 2005

Introduction to Embedded System

Purpose of Embedded Systems – Monitoring

- ✓ Embedded systems coming under this category are specifically designed for monitoring purpose
- ✓ They are used for determining the state of some variables using input sensors
- ✓ They cannot impose control over variables.
- ✓ Electro Cardiogram (ECG) machine for monitoring the heart beat of a patient is a typical example for this
- ✓ The sensors used in ECG are the different Electrodes connected to the patient's body
- ✓ Measuring instruments like Digital CRO, Digital Multi meter, Logic Analyzer etc used in Control & Instrumentation applications are also examples of embedded systems for monitoring purpose



Patient Monitoring system

Photo courtesy of Philips Medical Systems
(www.medical.philips.com/)

Introduction to Embedded System

Purpose of Embedded Systems – Control

- ✓ Embedded systems with control functionalities are used for imposing control over some variables according to the changes in input variables
- ✓ Embedded system with control functionality contains both sensors and actuators
- ✓ Sensors are connected to the input port for capturing the changes in environmental variable or measuring variable
- ✓ The actuators connected to the output port are controlled according to the changes in input variable to put an impact on the controlling variable to bring the controlled variable to the specified range
- ✓ Air conditioner for controlling room temperature is a typical example for embedded system with ‘Control’ functionality
- ✓ Air conditioner contains a room temperature sensing element (sensor) which may be a thermistor and a handheld unit for setting up (feeding) the desired temperature
- ✓ The air compressor unit acts as the actuator. The compressor is controlled according to the current room temperature and the desired temperature set by the end user.



Air Conditioner for controlling room temperature

Photo Courtesy of Electrolux Corporation

(www.electrolux.com.au)

Introduction to Embedded System

Purpose of Embedded Systems – Application Specific User Interface

- ✓ Embedded systems which are designed for a specific application
- ✓ Contains Application Specific User interface (rather than general standard UI) like key board, Display units etc
- ✓ Aimed at a specific target group of users
- ✓ Mobile handsets, Control units in industrial applications etc are examples for this



Patient Monitoring system

Photo courtesy of Philips Medical Systems
(www.medical.philips.com/)

Introduction to Embedded System

‘Smart’ running shoes from Adidas – The Innovative bonding of Life Style with Embedded Technology

- ✓ Shoe developed by Adidas, which constantly adapts its shock-absorbing characteristics to customize its value to the individual runner, depending on running style, pace, body weight, and running surface
- ✓ It contains sensors, actuators and a microprocessor unit which runs the algorithm for adapting the shock-absorbing characteristics of the shoe
- ✓ A ‘Hall effect sensor’ placed at the top of the “cushioning element” senses the compression and passes it to the Microprocessor
- ✓ A micro motor actuator controls the cushioning as per the commands from the MPU, based on the compression sensed by the ‘Hall effect sensor’

What an innovative bonding of Embedded Technology with Real life needs !!!😊



Electronics-enabled “Smart” running shoes from Adidas

Photo Courtesy of Adidas – Salomon AG
www.adidas.com

Module 4 - Roaming and handover

Mobile application part interfaces

The main benefit for the mobile subscribers that the international standardization of GSM has brought is that they can move freely not only within their home networks but also in international GSM networks and that at the same time they can even gain access to the special services they subscribed to at home provided that there are agreements between the operators. The functions needed for this free roaming are called roaming or mobility functions. They rely mostly on the GSM-specific extension of the SS#7. The MAP procedures relevant for roaming are first the location registration/update, IMSI attach/detach, requesting subscriber data for call setup and paging. In addition, the MAP contains functions and procedures for the control of SS and handover, for subscriber management, for IMEI management, for authentication and identification management, as well as for the user data transport of the SMS. MAP entities for roaming services reside in the MSC, HLR and VLR. The corresponding MAP interfaces are defined as B (MSC-VLR), C (MSC-HLR), D (HLRVLR), E (MSC-MSC) and G (VLR-VLR).

At the subscriber interface, the MAP functions correspond to the functions of MM, i.e. the MM messages and procedures of the Um interface are translated into the MAP protocols in the MSC. The most important functions of GSM MM are location registration with the PLMN and location updating to report the current location of a MS, as well as the identification and authentication of subscribers. These actions are closely interrelated. During registration into a GSM network, during the location updating procedure, and also during the setup of a connection, the identity of a mobile subscriber must be determined and verified (authentication).

The MM data are the foundation for creating the functions needed for routing and switching of user connections and for the associated services. For example, they are requested for routing an incoming call to the current MSC or for localizing a MS before paging is started. In addition to mobility data management, information about the configuration of SS is requested or changed, e.g. the currently valid target number for unconditional call forwarding in the HLR or VLR registers.

1. Location registration and location update

The location registration and location update are shown in Fig 4.1, 4, 2, 4, 3. Before a MS can be called or gain access to services, the subscriber has to register with the mobile network (PLMN). This is usually the home network where the subscriber has a service contract. However, the subscriber can equally register with a foreign network provider in whose service area they are currently visiting, provided that there is a roaming agreement between the two network operators. Registration is only required if there is a change of networks, and therefore a VLR of the current network has not yet issued a TMSI to the subscriber. This means the subscriber has to report to the current network with their IMSI and receives a new TMSI by executing a location registration procedure. This TMSI is stored by the MS in its nonvolatile SIM storage, such that even after a power down and subsequent power-up only a normal location updating procedure is required.

The sequence of operations for registration is presented schematically in Figure 4.1. After a subscriber has requested registration at their current location by sending a LOCATION UPDATE REQUEST with their IMSI and the current location area (LAI), first the MSC instructs the VLR with a MAP message UPDATE LOCATION AREA to register the MS with its current LAI. In order for this registration to be valid, the identity of the subscriber has to be checked first, i.e. the authentication procedure is executed. For this purpose, the authentication parameters have to be requested from the AUC through the HLR. The precalculated sets of security parameters (K_c , RAND, SRES) are not usually transmitted individually to the respective VLR. In most cases, several complete sets are kept at hand for several authentications. Each set of parameters, however, can only be used once, i.e. the VLR must continually update its supply of security parameters (AUTHENTICATION PARAMETER REQUEST).

After successful authentication, the subscriber is assigned a new MSRN, which is stored with the LAI in the HLR, and a new TMSI is also reserved for this subscriber; this is TMSI reallocation. To encrypt the user data, the base station needs the ciphering key K_c , which it receives from the VLR by way of the MSC with the command START CIPHERING. After ciphering of the user data has begun, the TMSI is sent in encrypted form to the MS. Simultaneously with the TMSI assignment, the correct and successful registration into the PLMN is acknowledged (LOCAPDATE ACCEPT). Finally, the MS acknowledges the correct reception of the TMSI (TMSI REALLOCATION COMPLETE). While the location information is being updated, the VLR is obtaining additional information about the subscriber, e.g. the MS category or configuration parameters for SS. For this purpose, the insert subscriber data procedure is defined (INSERT SUBSCRIBER DATA message in Figure 4.1). It is used for registration or location updating in the HLR to transmit the current data of the subscriber profile to the VLR. In general, this MAP procedure can always be used when the profile parameters are changed, e.g. if the subscriber reconfigures a SS such as unconditional forwarding. The changes are communicated immediately to the VLR with the insert subscriber data procedure.

The location update procedure is executed if the MS recognizes (by reading the LAI broadcast on the BCCH) that it is in a new location area, which leads to updating the location information in the HLR record. Alternatively, the location update can also occur periodically, independent of the current location. For this purpose, a time interval value is broadcast on the BCCH, which prescribes the time between location updates. The main objective of this location update is to know the current location for incoming calls or short messages, so that the call or message can be directed to the current location of the MS. The difference between the location update procedure and the location registration procedure is that in the first case the MS has already been assigned a TMSI. The TMSI is unique only in connection with an LAI, and both are kept together in the nonvolatile storage of the SIM card. With a valid TMSI, the MS also keeps a current ciphering key K_c for encryption of user data (Figure 6.2), although this key is renewed during the location update procedure. This key is recalculated by the MS based on the random number RAND used for authentication, whereas on the network side it is calculated in the AUC and made available in the VLR. Corresponding to the location update procedure, there is a MM procedure at the air interface of the MM category ‘specific’.

In addition to the location updating proper, it contains three blocks which are realized at the air interface by three procedures of the category ‘common’: the identification of the subscriber, the authentication and the start of ciphering on the radio channel. In the course of

location updating, the MS also receives a new TMSI, and the current location is updated in the HLR. Figure 4.2 illustrates the standard case of a location update. The MS has entered a new LA, or the timer for periodic location updating has expired, and the MS requests to update its location information. It is assumed that the new LA still belongs to the same VLR as the previous LA, so only a new TMSI needs to be assigned. This is the most frequent case. However, if it is not quite so crucial to keep the subscriber identity confidential, it is possible to avoid assigning a new TMSI. In this case, only the location information is updated in the HLR/VLR. The new TMSI is transmitted to the MS in enciphered form together with the acknowledgement of the successful location update. The location update is complete after acknowledgement by the MS. After execution of the authentication, the VLR can complete its database and replace the ‘consumed’ 3-tuple (RAND, SRES, Kc) by another one requested from the HLR/AUC.

If location change involves both LA and VLR, the location update procedure is somewhat more complicated (Figure 4.3). In this case, the new VLR has to request the identification and security data for the MS from the old VLR and store them locally. Only in emergency cases, if the old VLR cannot be determined from the old LAI or if the old TMSI is not known in the VLR, the new VLR may request the IMSI directly from the MS (identification procedure). Only after a MS has been identified through the IMSI from the old VLR and after the security parameters are available in the new VLR, is it possible for the MS to be authenticated and registered in the new VLR, for a new TMSI to be assigned, and for the location information in the HLR to be actualized. After successful registration in the new VLR (LOCATION UPDATE ACCEPT) the HLR instructs the old VLR to cancel the invalid location data in the old VLR (CANCEL LOCATION).

In the examples shown (Figures 4.1–4.3), the location information is stored as a MSRN in the HLR. The MSRN contains the routing information for incoming calls and this information is used to route incoming calls to the current MSC. In this case, the HLR receives the routing information already at the time of the location update. Alternatively, at location update time, the HLR may just store the current MSC and/or VLR number in connection with an LMSI, such that routing information is only determined at the time of an incoming call.

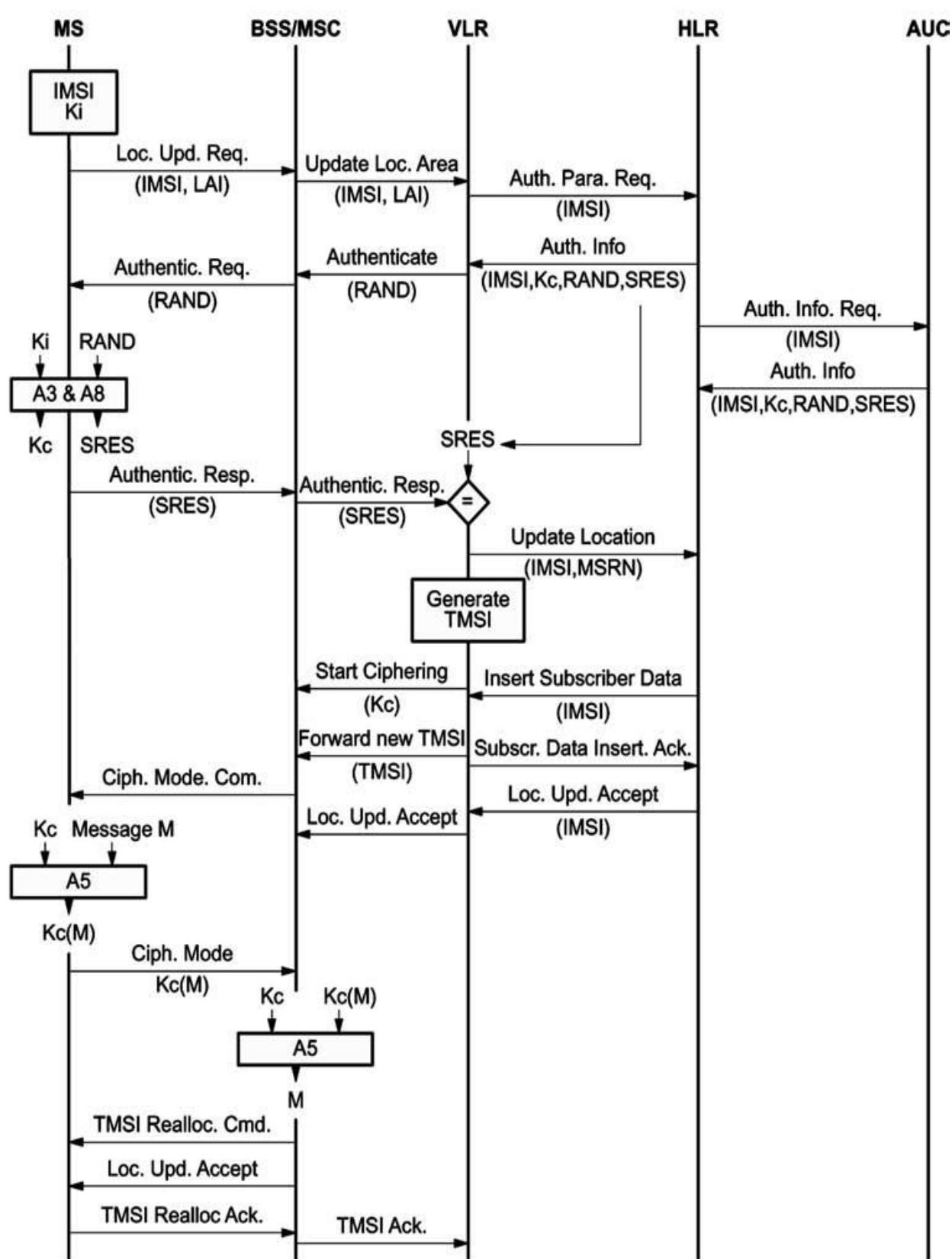


Figure 4.1 Overview of the location registration procedure.

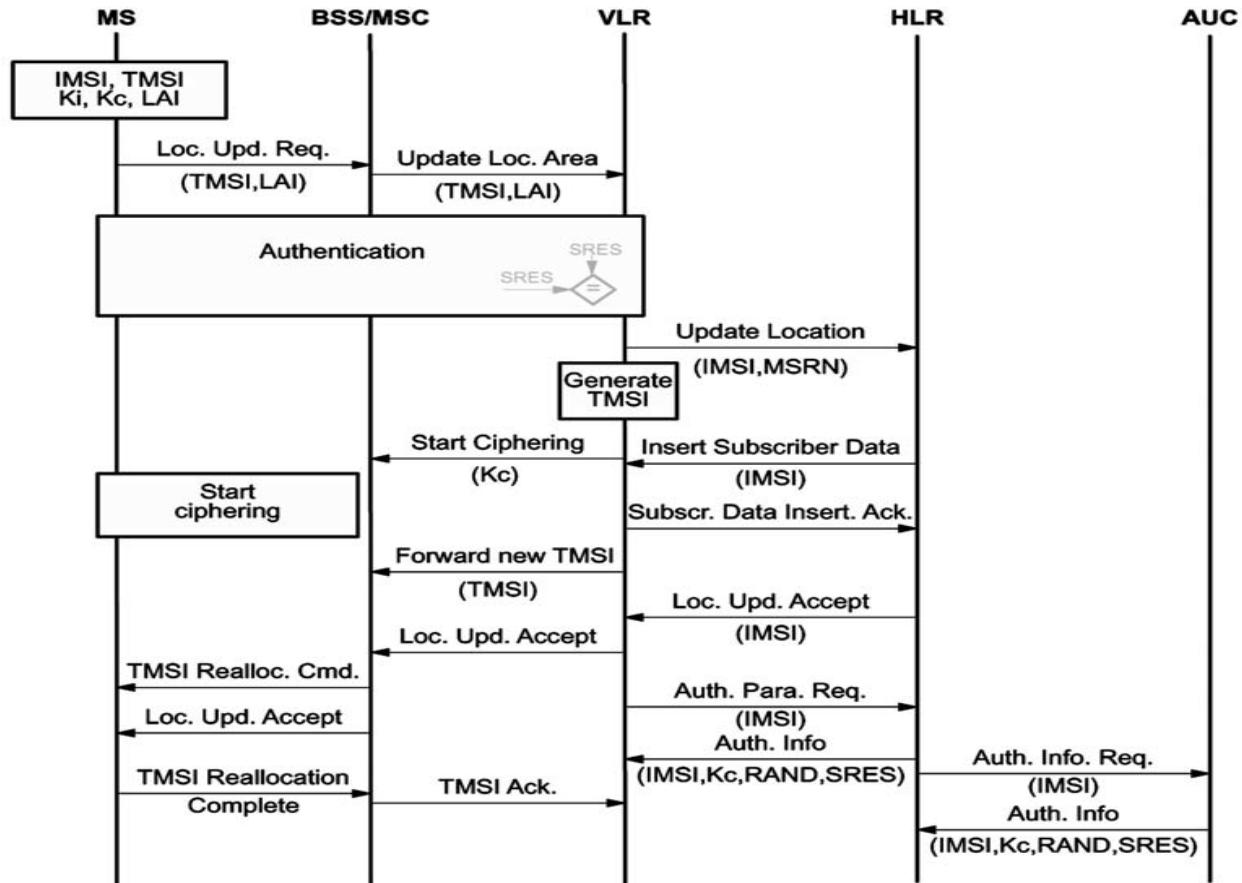


Figure 4.2 Overview of the location updating procedure.

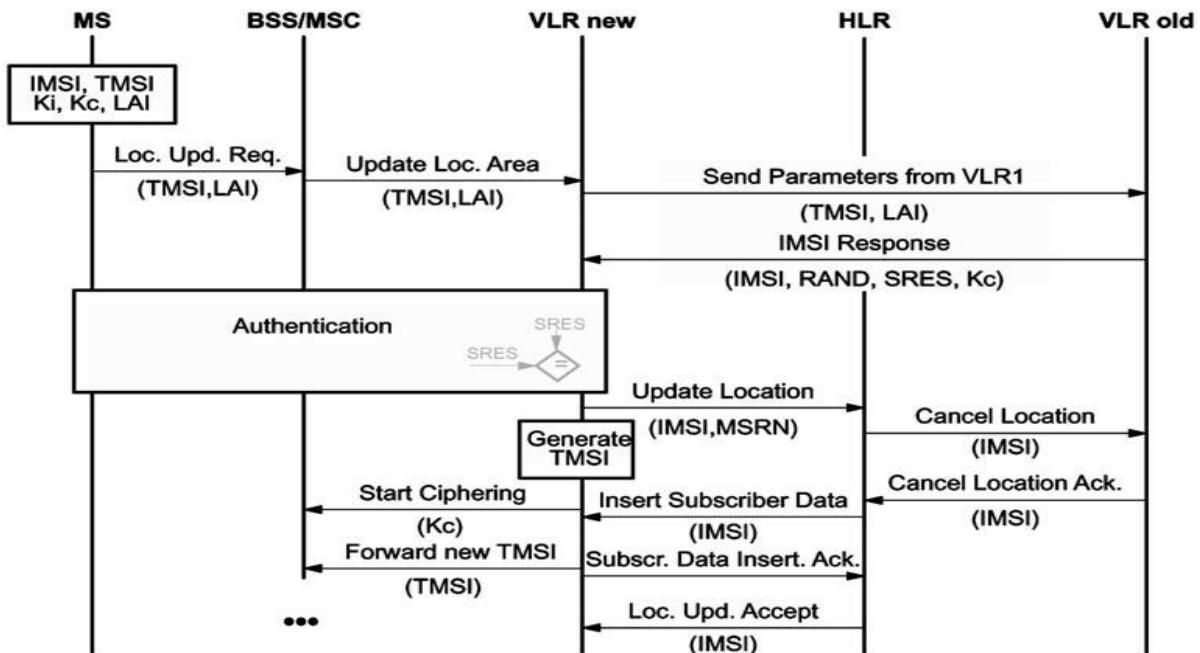


Figure 4.3 Location update after changing the VLR area.

2. Connection establishment and termination

a. Routing calls to MSs

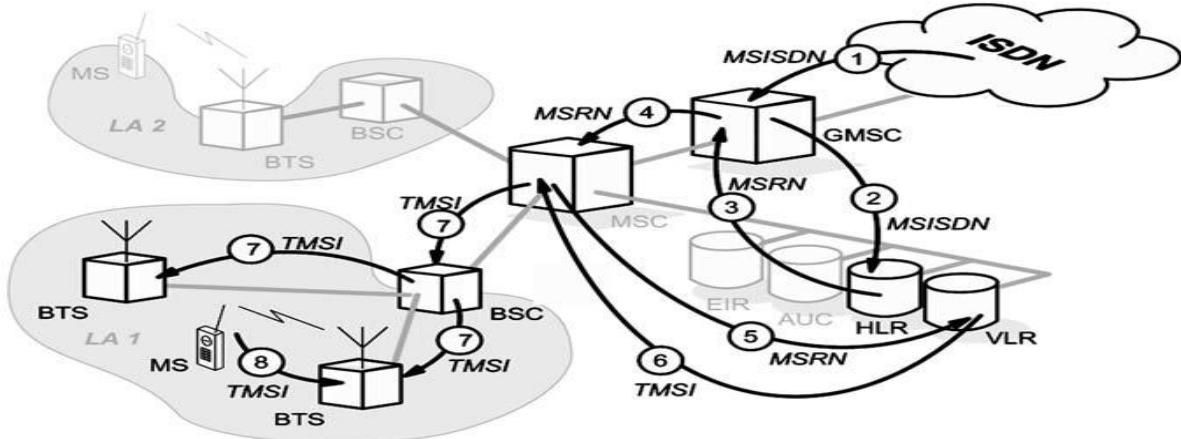


Fig 4.4 Principle of routing calls to mobile subscribers

The number dialed to reach a mobile subscriber (MSISDN) contains no information at all about the current location of the subscriber. In order to establish a complete connection to a mobile subscriber, however, one must determine the current location and the locally responsible switch (MSC). In order to be able to route the call to this switch, the routing address to this subscriber (MSRN) has to be obtained. This routing address is assigned temporarily to a subscriber by its currently associated VLR. At the arrival of a call at the GMSC, the HLR is the only entity in the GSM network which can supply this information, and therefore it must be interrogated for each connection setup to a mobile subscriber. The principal sequence of operations for routing to a mobile subscriber is shown in Figure 6.4. An ISDN switch recognizes from the MSISDN that the called subscriber is a mobile subscriber, and therefore can forward the call to the GMSC of the subscriber's home PLMN based on the CC and NDC in the MSISDN (1). This GMSC can now request the current routing address (MSRN) for the mobile subscriber from the HLR using the MAP (2,3). By way of the MSRN the call is forwarded to the local MSC (4), which determines the TMSI of the subscriber (5,6) and initiates the paging procedure in the relevant location area (7). After the MS has responded to the paging call (8), the connection can be switched through. Several variants for determining the route and interrogating the HLR exist, depending on how the MSRN was assigned and stored, whether the call is national or international and depending on the capabilities of the associated switching centers.

Effect of the MSRN assignment on routing

There are two ways to obtain the MSRN:

- obtaining the MSRN at location update;
- obtaining the MSRN on a per call basis.

For the first variant, an MSRN for the MS is assigned at the time of each location update which is stored in the HLR. In this way the HLR is in a position to immediately supply the routing information needed to switch a call through to the local MSC. The second variant requires that the HLR has at least identification for the currently responsible VLR. In this case, when routing information is requested from the HLR, the HLR first has to obtain the MSRN from the VLR. This MSRN is assigned on a per call basis, i.e. each call involves a new MSRN assignment.

Placement of the protocol entities for HLR interrogation

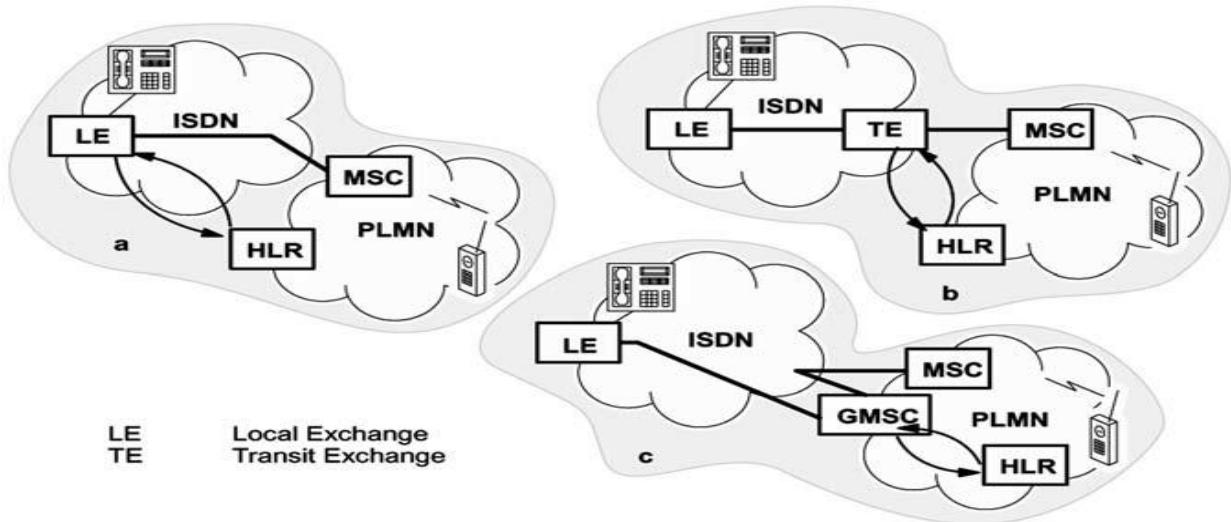


Fig: 4.5 Routing variants for national MSISDN.

Depending on the capabilities of the associated switches and the called target (national or International MSISDN), there are different routing procedures. In general, the local switching center analyzes the MSISDN. Owing to the NDC, this analysis of the MSISDN allows the separation of the mobile traffic from other traffic. The case that mobile call numbers are integrated into the numbering plan of the fixed network is currently not provided.

In the case of a national number, the local exchange recognizes from the NDC that the number is a mobile ISDN number. The fixed network and home PLMN of the called subscriber reside in the same country. In the ideal case, the local switch can interrogate the HLR responsible for this MSISDN (HLR in the home PLMN of the subscriber) and obtain the routing information (Figure 4.5(a)). The connection can then be switched through via fixed connections of the ISDN directly to the MSC.

If the local exchange does not have the required protocol intelligence for the interrogation of the HLR, the connection can be passed on preliminarily to a transit exchange, which then assumes the HLR interrogation and routing determination to the current MSC (Figure 4.5(b)). If the fixed network is not at all capable of performing a HLR interrogation, the connection has to be directed through a GMSC. This GMSC connects through to the current MSC (Figure 4.5(c)). For all three cases, the MS could also reside in a foreign PLMN (roaming); the connection is then made through international lines to the current MSC after interrogating the HLR of the home PLMN.

In the case of an international call number, the local exchange recognizes only the international CC and directs the call to an ISC. Then the ISC can recognize the NDC of the mobile network and process the call accordingly. Figures 4.6 and 4.7 show examples for the processing of routing information. An international call to a mobile subscriber involves at least three networks: the country from which the call originates; the country with the home PLMN of the subscriber, Home PLMN (H-PLMN); and the country in which the mobile subscriber is currently roaming, Visited PLMN (V-PLMN). The traffic between countries is routed through ISCs. Depending on the capabilities of the ISC, there are several routing variants for

international calls to mobile subscribers. The difference is determined by the entity that performs the HLR interrogation, resulting in differently occupied line capacities. If the ISC performs the HLR interrogation, the routing to the current MSC is performed either by the ISC of the originating call or by the ISC of the mobile subscriber's H-PLMN (Figure 4.6). If no ISC can process the routing, again a GMSC has to become involved, either a GMSC in the country where the call originates or the GMSC of the H-PLMN (Figure 4.7).

For the routing procedures explained here, it does not matter which kind of subscriber is calling, i.e. the subscriber may be in the fixed network or in the mobile network. However, for calls from mobile subscribers, the HLR interrogation is usually performed at the local exchange (MSC).

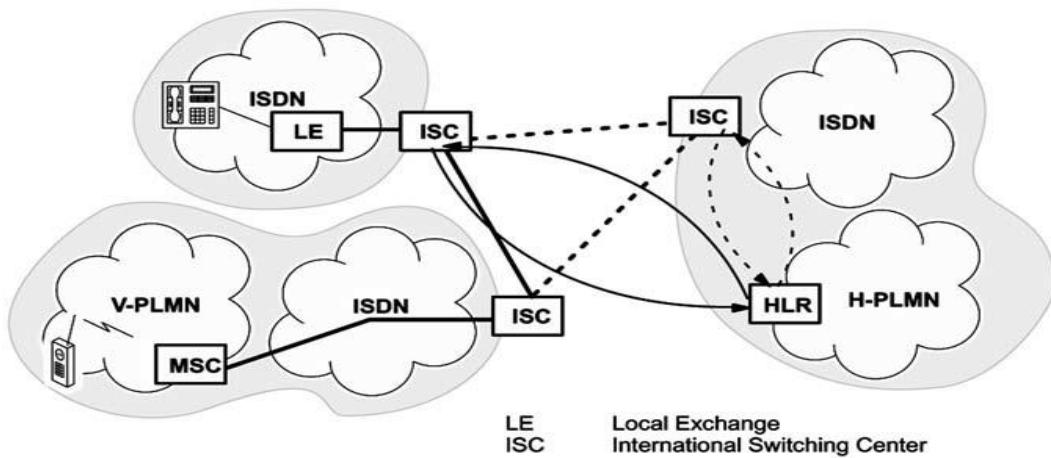


Fig 4.6 Routing for international MSISDN (HLR interrogation from ISC).

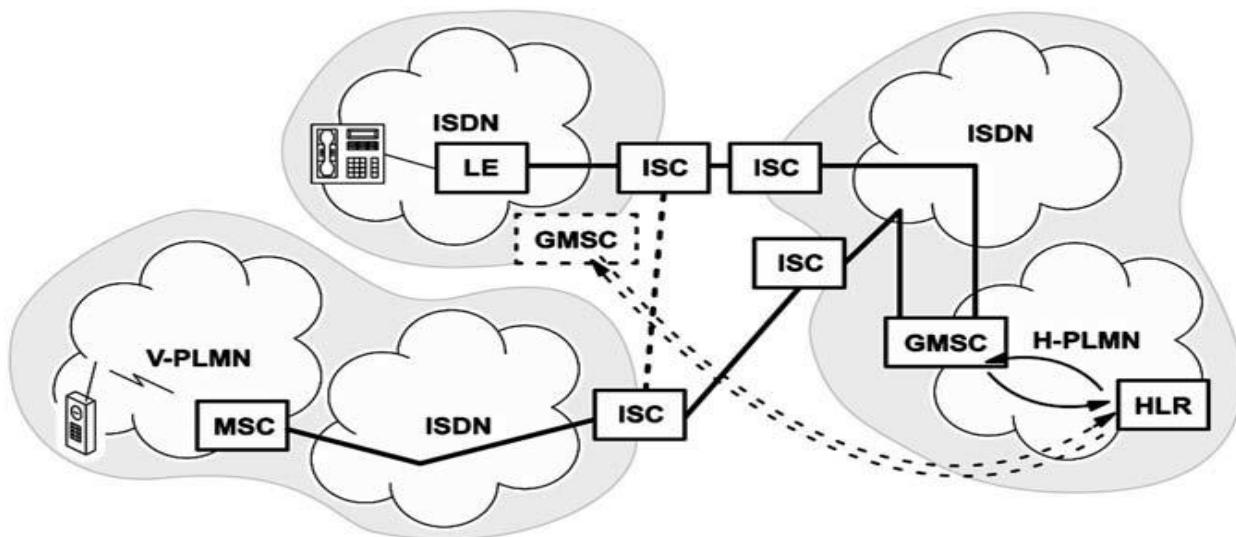


Fig 4.7 Routing through GMSC for international MSISDN

b. Call establishment and corresponding MAP procedures

1. Outgoing connection setup

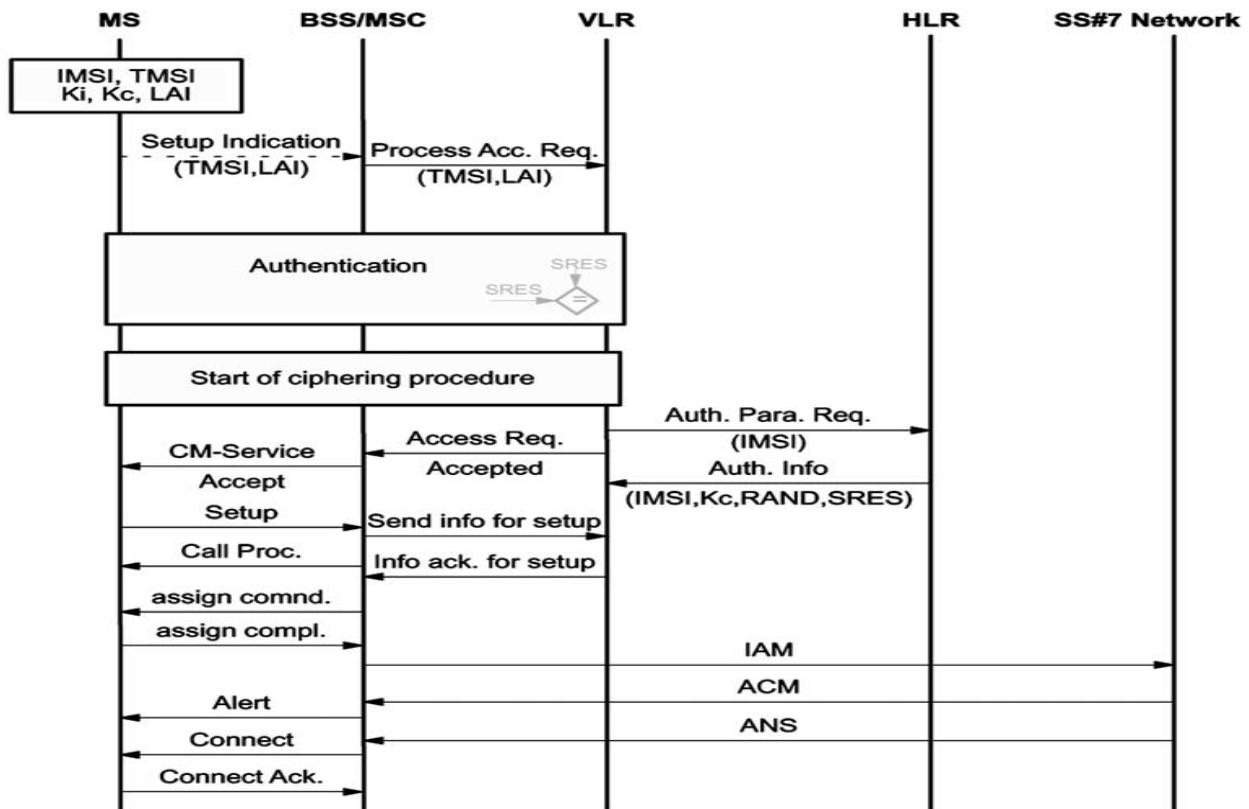


Fig 4.8 Overview of an outgoing call setup

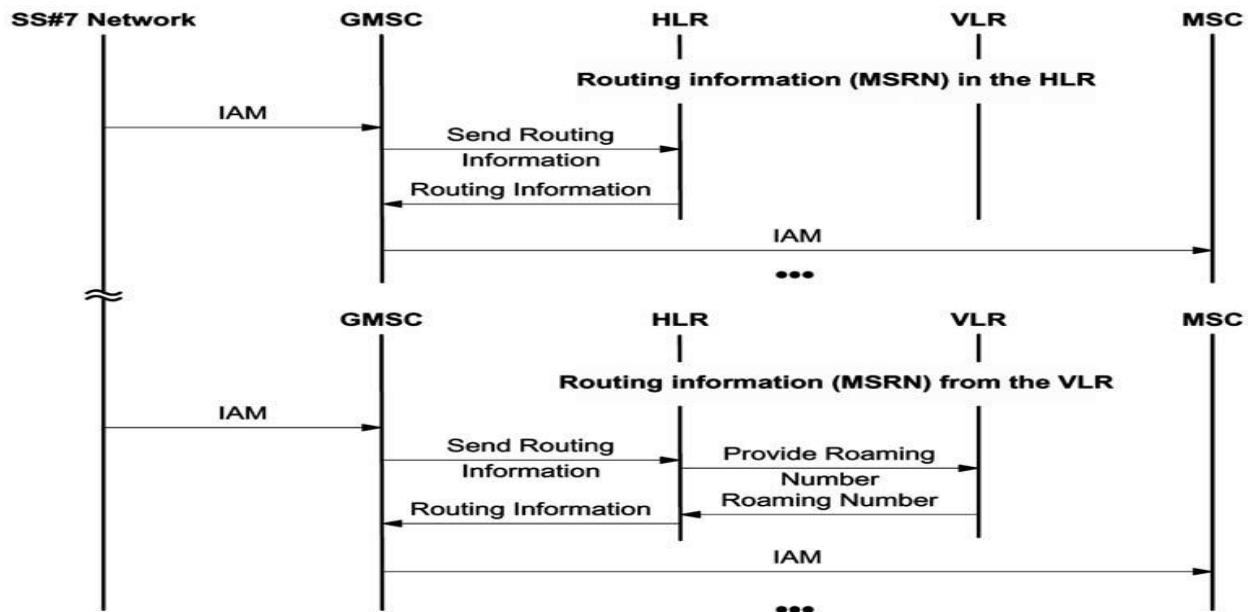


Fig 4.9 Interrogation of routing information for an incoming call

For outgoing connection setup (Figure 4.8), first the MS announces its connection request to The MSC with a SETUP INDICATION message, which is a pseudo-message. It is generated between the MMentity of the MSC and the MAP entity, when the MSC receives the message CM-SERVICE REQUEST from the MS, which indicates in this way the request for a MM connection (Figure 5.27). Then the MSC signals to the VLR that the MS identified by the temporary TMSI in the location area LAI has requested service access (PROCESS ACCESS REQUEST) which is an implicit request for a random number RAND from the VLR, to be able to start the authentication of the MS. This random number is transmitted to the MS, its response with authentication result SRES is returned to the VLR, which now examines the authenticity of the MS's identity (compare authentication at registration; Figure 4.1).

After successful authentication, the ciphering process is started on the air interface, and this way the MM connection between MS and MSC has been completely established (CMSERVICE ACCEPT). Subsequently, all signaling messages can be sent in encrypted form. Only now the MS reports the desired calling target. While the MS is informed with a CALL PROCEEDING message that processing of its connection request has been started, the MSC reserves a channel for the conversation and assigns it to the MS (ASSIGN). The connection request is signaled to the remote network exchange through the signaling system SS#7 with the ISUP message IAM (Bocker, 1990). When the remote exchange answers (ACM), the delivery of the call can be indicated to the MS (ALERT). Finally, when the called partner goes off-hook, the connection can be switched through (CONNECT, ANS, CONNECT ACKNOWLEDGE).

2. Incoming connection setup

For incoming connection setup, it is necessary to determine the exact location of a MS in order to route the call to the currently responsible MSC. A call to a MS is therefore always routed to an entity which is able to interrogate the HLR for temporary routing information and to use it to forward the call. Usually, this entity is a GMSC of the home PLMN of the MS. Through this HLR interrogation, the GMSC obtains the current MSRN of the MS and forwards it to the current MSC (Figure 4.9). Depending on whether the MSRN is stored in the HLR or first has to be obtained from the VLR, two variants of the HLR interrogation exist. In the first case, the interrogated HLR can supply the MSRN immediately (ROUTING INFORMATION). In the second case, the HLR has only received and stored the current VLR address during location update. Therefore, the HLR first has to request the current routing information from the VLR before the call can be switched through to the local MSC.

Call processing is interrupted again in the local MSC in order to determine the exact location of the MS within the MSC area (SEND INFO FOR SETUP, Figure 4.10). The current LAI is stored in the location registers, but an LA can comprise several cells. Therefore, a broadcast (paging call) in all cells of the LA is used to determine the exact location, i.e. cell, of the MS. Paging is initiated from the VLR using the MAP (PAGE MS) and transformed by the MSC into the paging procedure at the air interface. When a MS receives a paging call, it responds directly and thus allows the current cell to be determined. Thereafter, the VLR instructs the MSC to authenticate the MS and to start ciphering on the signaling channel. Optionally, the VLR can execute a reallocation of the TMSI (TMSI reallocation procedure) during call setup. Only at this point, after the network internal connection has been established (section 5.4.4), can the connection setup proper be processed (command COMPLETE CALL

from VLR to MSC). The MS is told about the connection request with a SETUP message, and after answering CALL COMPLETE it receives a channel. After ringing (ALERT) and going off-hook, the connection is switched through (CONNECT, CONNECT, ACKNOWLEDGE), and this fact is also signaled to the remote exchange (ACM, ANS).

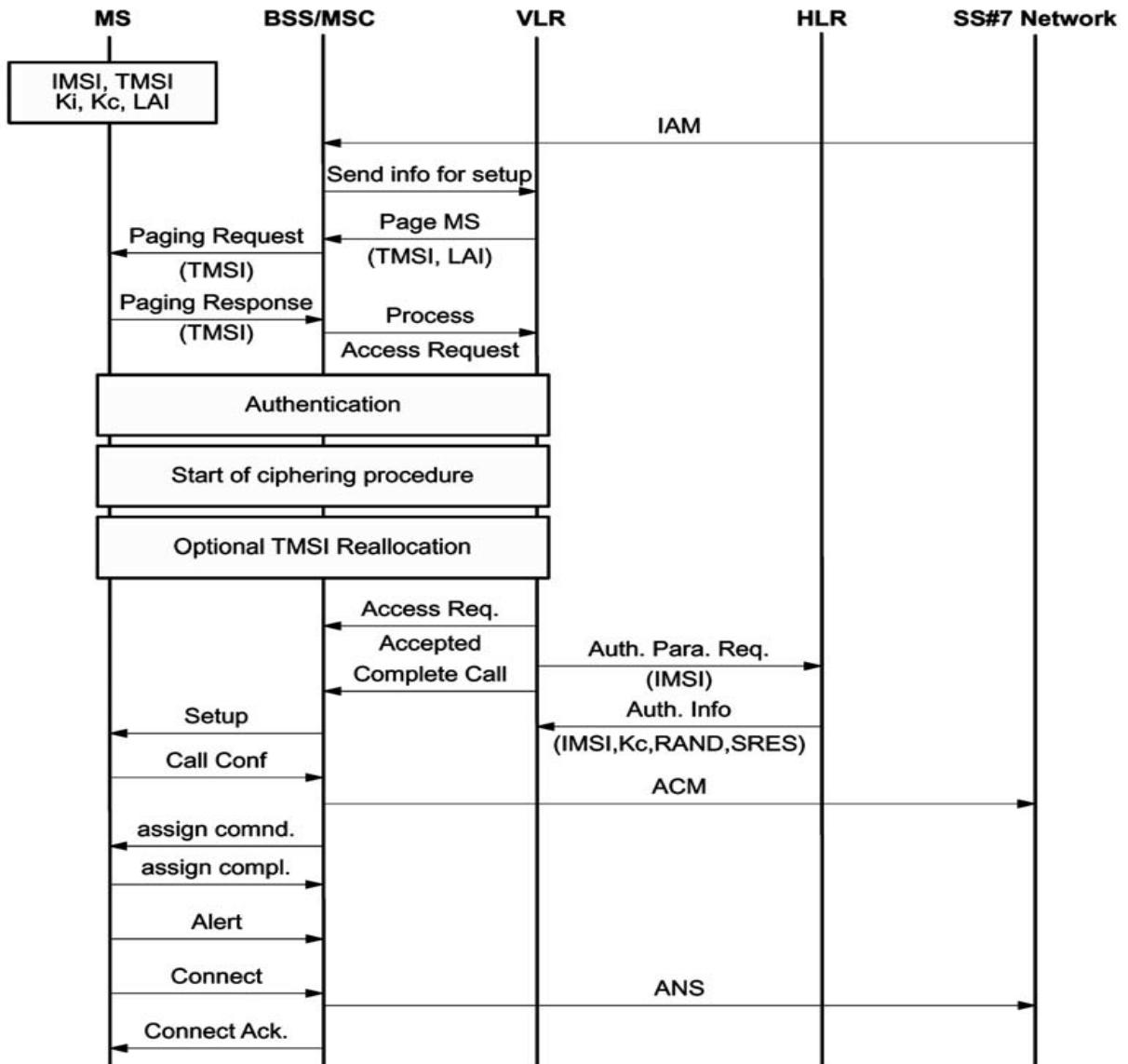


Fig 4.10 Overview of an incoming call setup

3. Call termination

At the air interface, a given call can be terminated either by the mobile equipment or by the network. The taking down of the connection is initiated at the Um interface by means of the CC messages DISCONNECT, RELEASE and RELEASE COMPLETE. This is followed by an explicit release of occupied radio resources (CHANNEL RELEASE). On the network side, the connection between the involved switching centers (MSC, etc.) is terminated using the ISUP messages REL and RLC in the SS#7 network (Figure 4.11). After the connection has been taken down, information about charges (CHARGING INFORMATION) is stored in the VLR or HLR

using the MAP. This charging data can also be required for an incoming call, e.g. if roaming charges are due because the called subscriber is not in their home PLMN.

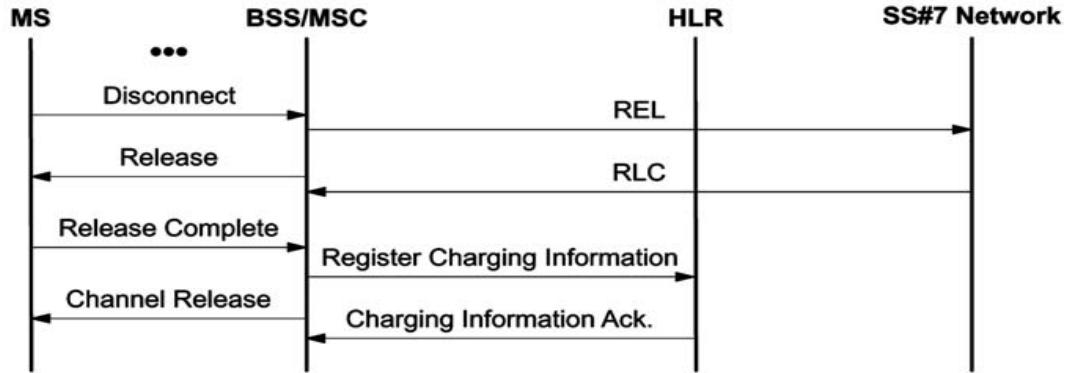


Fig 4.11 Mobile-initiated call termination and storing of charging information.

c. MAP procedures and routing for short messages

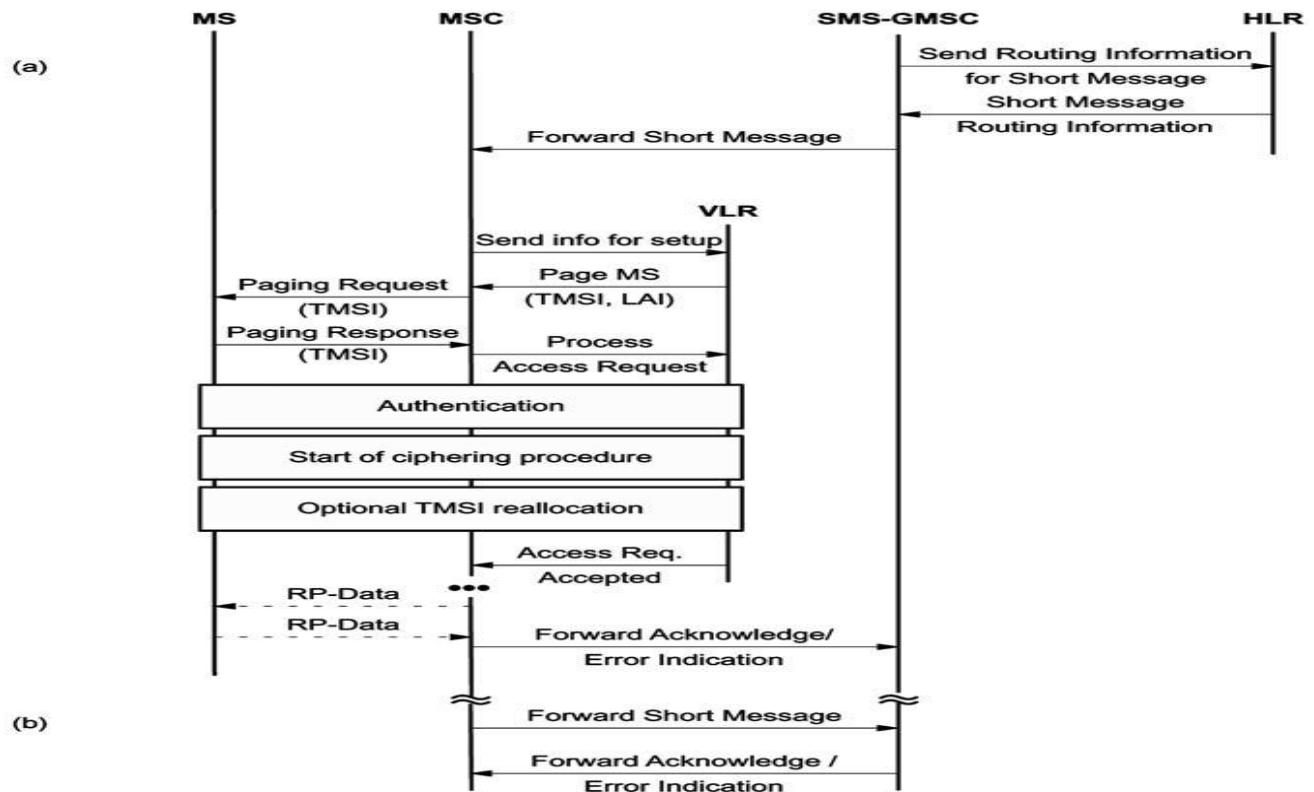


Fig 4.12 Forwarding short messages in a PLMN.

A connectionless relay protocol has been defined for the transport of short messages at the air interface, which has a counterpart in the network in a store-and-forward operation for short messages. This forwarding of transport PDUs of the SMS uses MAP procedures. For an incoming short message which arrives from the SMS-SC at a SMS-GMSC, the exact location of the MS is the first item that needs to be determined just as for an incoming call. The current

MSC of the MS is first obtained with a HLR interrogation (SHORT MESSAGE ROUTING INFORMATION; Figure 4.12(a)). The short message is then passed to this MSC (FORWARD SHORT MESSAGE) and is locally delivered after paging and SMS connection setup. Success or failure are reported to the SMS-GMSC in another MAP message (FORWARD ACKNOWLEDGEMENT/ERROR INDICATION) which\ then informs the service center. In the reverse case, for an outgoing short message, no routing interrogation is needed, since the SMS-GMSC is known to all MSCs, so the message can be passed immediately to the SMS-GMSC (Figure 4.12(b)).

3. Handover

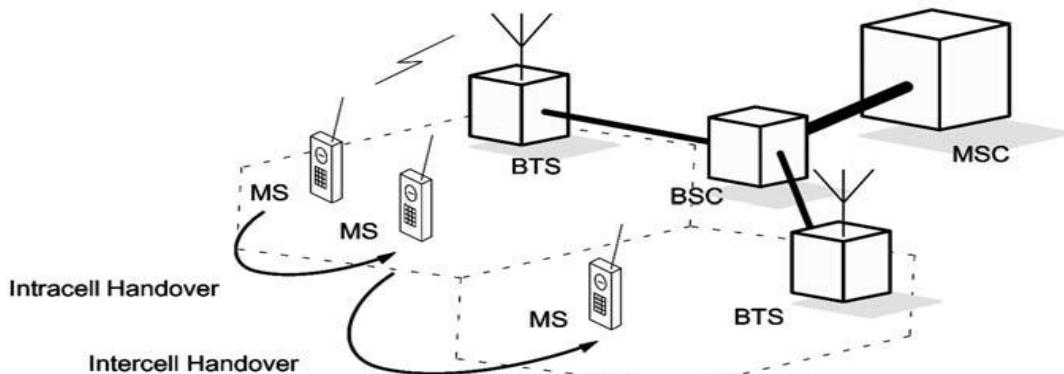


Fig 4.13 Intracell and intercell handover

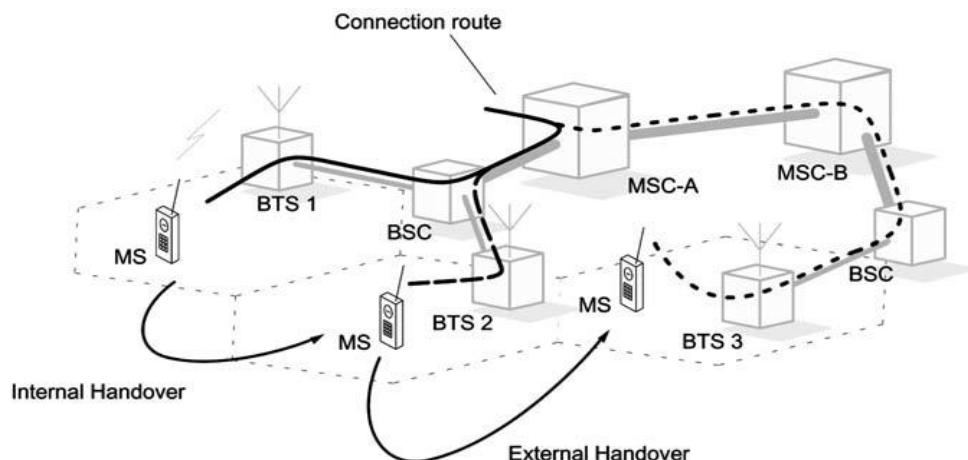


Fig 4.14 Internal and external handover

Handover is the transfer of an existing voice connection to a new base station. There are different reasons for the handover to become necessary. In GSM, a handover decision is made by the network, not the MS, and it is based on BSS criteria (received signal level, channel quality, distance between MS and BTS) and on network operation criteria (e.g. current traffic load of the cell and ongoing maintenance work).

The functions for preparation of handover are part of the radio subsystem link control.\ Above all; this includes the measurement of the channel. Periodically, a MS checks the signal

field strength of its current downlinks as well as those of the neighboring base stations, including their BSICs. The MS sends measurement reports to its current base station (quality monitoring); section 4.5.1. On the network side, the signal quality of the uplink is monitored, the measurement reports are evaluated, and handover decisions are made. As a matter of principle, handovers are only performed between base stations of the same PLMN. Handovers between BSSs in different networks are not allowed. Two kinds of handover are distinguished (Figure 4.13).

- **Intracell handover:** for administrative reasons or because of channel quality (channel selective interferences), the MS is assigned a new channel within the same cell. This decision is made locally by the RR of the BSS and is also executed within the BSS.
- **Intercell handover:** the connection to a MS is transferred over the cell boundary to a new BTS. The decision about the time of handover is made by the RR protocol module of the network based on measurement data from MSs and BSSs. The MSC, however, can participate in the selection of the new cell or BTS. The intercell handover occurs most often when it is recognized from weak signal field strength and bad channel quality (high bit error ratio) that a MS is moving near the cell boundary. However, an intercell handover can also occur due to administrative reasons, say for traffic load balancing. The decision about such a network-directed handover is made by the MSC, which instructs the BSS to select candidates for such a handover.

Two cases need to be distinguished between with regard to participation of network\ components in the handover, depending on whether the signaling sequences of a handover execution also involve a MSC. Since the RR module of the network resides in the BSC , the BSS can perform the handover without participation of the MSC. Such handovers occur between cells which are controlled by the same BSC and are called internal handover. They can be performed independently by the BSS; the MSC is only informed about the successful execution of internal handovers. All other handovers require participation of at least one MSC, or their BSSMAP and MAP parts, respectively. These handovers are known as external handovers.

Participating MSCs can act in the role of MSC-A or MSC-B. MSC-A is the MSC which performed the initial connection setup, and it keeps the MSC-A role and complete control (anchor MSC) for the entire life of the connection. A handover is therefore, in general, the extension of the connection from the anchor MSC-A to another MSC (MSC-B). In this case, the mobile connection is passed from MSC-A to MSC-B with MSC-A keeping the ultimate control over the connection. An example is presented in Figure 4.14.

A MS occupies an active connection via BTS1 and moves into the next cell. This cell of BTS2 is controlled by the same BSC so that an internal handover is indicated. The connection is now carried from MSC-A over the BSC and the BTS2 to the MS; the connections of BTS1 (radio channel and ISDN channel between BTS and BSC) were taken down. As the MS moves on to the cell handled by BTS3, it enters a new BSS which requires an external handover. In addition, this BSS belongs to another MSC, which now has to play the role of MSC-B. Logically, the connection is extended from MSC-A to MSC-B and carried over the BSS to the MS. At the next change of the MSC, the connection element between MSC-A and MSC-B is taken down, and a connection to the new MSC from MSC-A is set up. Then the new MSC takes over the role of MSC-B.

4. Services

Classical GSM services

1 Teleservices

Voice

Voice services had to be implemented by each operator in the start-up phase (E1) by 1991. In this category, two teleservices were distinguished: regular telephone service (TS11) and emergency service (TS12). For the transmission of the digitally coded speech signals, both services use a bidirectional, symmetric, full-duplex, point-to-point connection, which is set up on user demand. The sole difference between TS11 and TS12 teleservices is that regular service requires an international IWF, whereas the emergency service stays within the boundaries of a national network.

Fax transmission

As a teleservice for the second implementation phase (E2), the implementation of a transparent fax service (TS61) for Group 3 fax was planned. The fax service is called transparent because it uses a transparent bearer service for the transmission of fax data. The coding and transmission of the facsimile data uses the fax protocol according to the ITU-T recommendation T30. The network operator also has the option to implement TS61 on a nontransparent bearer service in order to improve the transmission quality. TS61 is transmitted over a traffic channel that is alternately used for voice or fax. Another optional alternative is designated as a fax transfer with automatic call acceptance (TS61). This service can be offered by a network operator when multinumbering is used as the interworking solution. In the case of multinumbering, a subscriber is assigned several MSISDN numbers, and a separate interworking profile is stored for each of them. In this way a specific teleservice can be associated with each MSISDN, the fax service being one of them. If a mobile subscriber is called on their GSM-fax number, the required resources in the IWF of the MSC as well as in the MS can be activated; in the case of TS61, fax calls arrive with the same number as voice calls (no multinumbering) and have to be switched over to fax reception manually.

2 Popular GSM services: SMS and MMS

Next to voice services, which have been the main focus of GSM since the start, SMS has proven to be extremely popular and successful in GSM. To follow this success, similar but enhanced services have been standardized, such as the Enhanced Messaging Service (EMS) and the Multimedia Messaging Service (MMS). A brief description of these services follows in the next sections.

a. SMS

One of the most important services in GSM systems today, in terms of popularity at the user side as well as in terms of revenue generation at the provider side, is the capability to receive or send short messages at the MS: SMS, TS21 and TS22. This service was supposed to be offered in the third phase (E3) at the latest from 1996 on all GSM networks. TS21 is the

point-to-point version of the SMS, which allows a single station to be sent a message of up to 160 characters. Conversely, TS22 has been defined as an optional implementation of the capability to send short messages from a MS. The combinations of SMS with other added value services, e.g. mailbox systems with automatic notification of newly arrived messages or the transmission by short message of incurred charges clearly show how the services offered by GSM networks go significantly beyond the services offered in fixed networks.

For SMS, the network operator has to establish a service center which accepts short messages from the fixed network and processes them in a store-and-forward mode. The interface has not been specified and can be by DTMF signaling, special order, email, fax, etc. The delivery can be time-shifted and is of course independent of the current location of the MS. Conversely, a service center can accept short messages from MSs which can also be forwarded to subscribers in the fixed network, for example by fax or email. The transmission of short messages uses a connectionless, protected, packet-switching protocol. The reception of a message must be acknowledged by the MS or the service center; in the case of failure, retransmission occurs. TS21 and TS22 are the only teleservices which can be used simultaneously with other services, i.e. short messages can also be received or transmitted during an ongoing call.

A further variation of the SMS is the cell broadcast service TS23, SMSCB. SMSCB messages are broadcast only in a limited region of the network. They can only be received by MSs in idle mode, and reception is not acknowledged. A MS itself cannot send SMSCB messages. With this service, messages contain a category designation, so that MSs can select categories of interest which they want to receive and store. The maximum length of SMSCB messages is 93 characters, but by using a special reassembly mechanism, the network can transmit longer messages of up to 15 subsequent SMSCB messages.

Table 7.1 SMS, EMS and MMS.

Service	Introduced	Payload size	Content
SMS	1995	160 byte	Text
EMS	2000	160 byte	Text, pictures, animations
MMS	2001	100 kbyte	Text, voice, pictures, photos, video

b. EMS

EMS is an extension of SMS. SMS was limited to text messages only. However, as ring tones and pictures gained a lot of popularity, EMS was introduced. EMS was developed by major GSM manufacturers as an open 3GPP standard. It allows unicolor pictures with 16×16 or 32×32 pixels to be sent and the pictures to be modified in the handset. Picture sequences can comprise six pictures. Fonts can be formatted in EMS and tones of three octaves can be included, from the pitch of C to the pitch of B++. The duration of tones can be 150, 225, 300 or 450 ms and up to 80 notes can be included in one EMS. However, before EMS had a chance to become popular, the MMS standard, which is much richer, took over.

c. MMS

MMS is similar to SMS or EMS, however has much higher capabilities in terms of size and flexibility. The MMS standard was developed by a consortium of industry partners and has

become a 3GPP standard. In addition to pure text, MMS is capable of transmitting pictures, melodies and multimedia sequences of different kinds. MMS can transmit up to 100 kbyte of data and can handle AMR-coded speech, pictures (e.g. JPEG or GIF), music and even video. From the network point of view, a MMS-Center, called MMS-C, is required, which is responsible for storing, converting and forwarding MMS data. The MMS-C also stores information about the preferences of users as well as their terminal capabilities. Therefore, it is possible to avoid the transmission of MMS messages to a terminal which cannot deal with the specific format. Instead, the MMS message can possibly be transformed at the MMS-C into a format which the receiving terminal can handle. A comparison between SMS, EMS and MMS is provided in Table 7.1.

MMS architecture

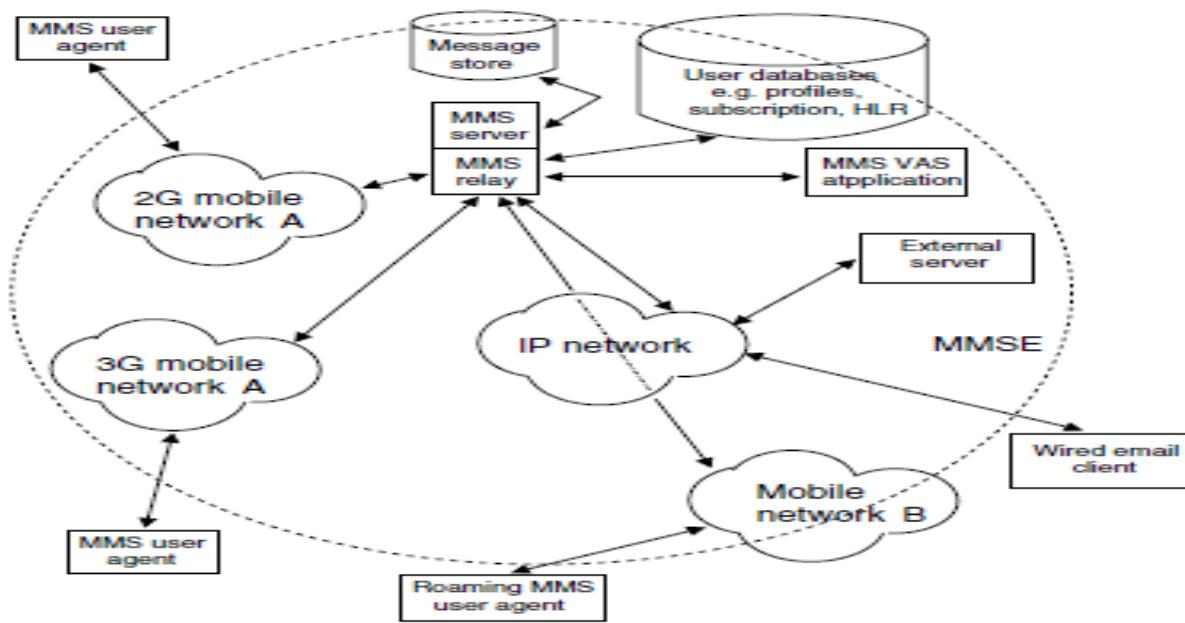


Figure 7.1 Multimedia Messaging Service Network Architecture (MMSNA).

The Multimedia Messaging Service Network Architecture (MMSNA) (Figure 7.1) comprises all required elements to provide a complete MMS to a user. This includes interworking between service providers. The Multimedia Messaging Service Environment (MMSE) is a collection of MMS-specific network elements which is controlled by a single administration. In the case of roaming, the visited network is part of the MMSE of the user in question, while subscribers of a different service provider are part of a different MMSE. An important role is taken on by the MMS relay/server, which is responsible for storing and handling both incoming and outgoing messages. It is also in charge of the transfer of messages between different messaging systems.

The MMS relay/server can either be a single logical element or it may be separated into two single elements, a MMS relay and a MMS server. The MMS relay/server must also be able to generate charging data, when receiving Multimedia Messages (MMs) from or delivering MMs to other elements of the architecture. The MMS user database contains user-related information,

such as subscription, configuration and capability data. The MMS user database can be either a single entity or it can be distributed. The MMS user agent is an application layer function, providing the users with the ability to view, compose and handle MMs. The MMS user agent resides on the mobile device. Finally, the MMS Value Added Services (VAS) applications offer value-added services to MMS users. Several MMS VAS applications may be connected to a MMSE.

5. Improved data services in GSM: GPRS, HSCSD and EDGE

The main data service used in modern GSM systems is the GPRS. Therefore, the major part of this chapter is devoted to GPRS. However, HSCSD is also of importance. HSCSD was available earlier than GPRS and was therefore chosen by some network providers in order to provide higher data rate services in GSM as fast as possible and it is still used in many networks. A third important aspect of providing improved data rates in GSM systems is EDGE, which helps to increase the data rate of both GPRS and HSCSD by allowing for higher-order modulation schemes when the signal strength is sufficiently high.

1. GPRS

Packet data transmission has already been standardized in GSM phase 2, offering access to the Packet Switched Public Data Network (PSPDN); see Appendix A. However, on the air interface such access occupies a complete circuit switched traffic channel for the entire call period. In the case of bursty traffic (e.g. Internet traffic), such access leads to a highly inefficient resource utilization. It is obvious that in this case, packet switched bearer services result in a much better utilization of the traffic channels. This is because a packet channel will only be allocated when needed and will be released after the transmission of the packets. With this principle, multiple users can share one physical channel (statistical multiplexing).

In order to address these inefficiencies, GPRS has been developed in GSM phase 2+. It offers a genuine packet-switched bearer service for GSM also at the air interface. GPRS is thus a huge improvement and simplification of the wireless access to packet data networks. Networks based on IP (e.g. the global Internet or private/corporate intranets) and X.25 networks are supported. In order to introduce GPRS to existing GSM networks, several modifications and enhancements must be made in the network infrastructure as well as in the MS's.

Users of GPRS benefit from higher data rates and shorter access times. In conventional GSM, the connection setup takes several seconds and rates for data transmission are restricted to 9.6 kbit/s. GPRS, in practice, offers almost ISDN-like data rates up to approximately 40– 50 kbit/s and session establishment times below one second. Furthermore, GPRS supports a more user-friendly billing than that offered by circuit-switched data services. In circuit switched services, billing is based on the duration of the connection. This is unsuitable for applications with bursty traffic, since the user must pay for the entire airtime even for idle periods when no packets are sent (e.g. when the user reads a Web page). In contrast to this, with packet-switched services, billing can be based on the amount of transmitted data (e.g. Mbyte) and the Quality of Service (QoS). The advantage for the user is that they can be ‘online’ over a long period of time

but will be billed mainly based on the transmitted data volume. The network operators can utilize their radio resources in a more efficient way and simplify the access to external data networks.

a. System architecture of GPRS

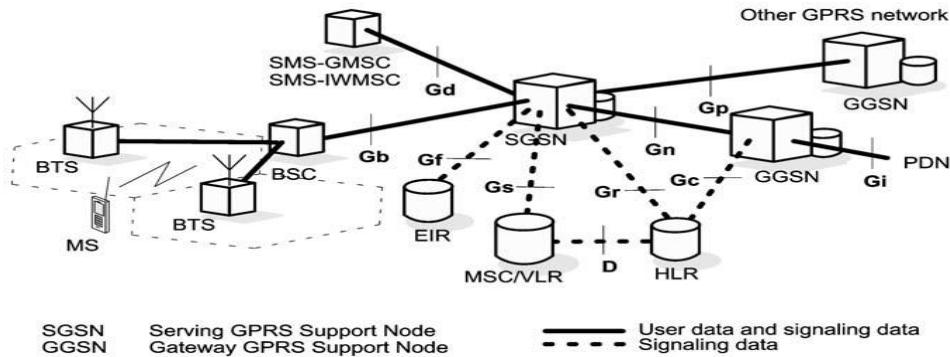


Fig 4.15 GPRS system architecture and interfaces

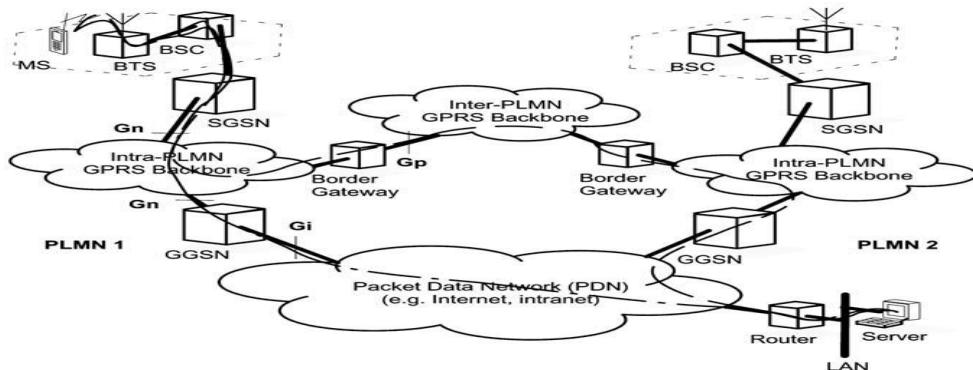


Fig 4.16 GPRS system architecture, interfaces and routing example

In order to integrate GPRS into the existing GSM architecture (Chapter 3), a new class of network nodes, called GPRS Support Nodes (GSNs), has been introduced. GSNs are responsible for the delivery and routing of data packets between the MSs and external Packet Data Networks (PDNs). Figure 8.1 illustrates the resulting system architecture.

A Serving GPRS Support Node (SGSN) delivers data packets from and to the MSs within its service area. Its tasks include packet routing and transfer, functions for attach/detach of MSs and their authentication, and logical link management. The location register of the SGSN stores location information (e.g. current cell, current VLR) and user profiles (e.g. IMSI, address used in the packet data network) of all GPRS users registered with this SGSN.

A Gateway GPRS Support Node (GGSN) acts as an interface to external packet data networks (e.g. to the Internet). It converts GPRS packets coming from the SGSN into the appropriate Packet Data Protocol (PDP) format (i.e. IP or X.25) and sends them out on the corresponding external network. In the other direction, the PDP address of incoming data packets (e.g. the IP destination address) is converted into the GSM address of the destination user. The readdressed

packets are sent to the responsible SGSN. For this purpose, the GGSN stores the current SGSN addresses and profiles of registered users in its location register. In general, there is a many-to-many relationship between the SGSNs and the GGSNs: a GGSN is the interface to an external network for several SGSNs; an SGSN may route its packets to different GGSNs. Figure 4.14 also shows the interfaces between the GPRS support nodes and the GSM network.

The Gb interface connects the BSC with the SGSN. Via the Gn and the Gp interfaces, user and signaling data are transmitted between the GSNs. The Gn interface is used if SGSN and GGSN are located in the same PLMN, whereas the Gp interface is used if they are in different PLMNs. All GSNs are connected via an IP-based GPRS backbone network. Within this backbone, the GSNs encapsulate the PDN packets and transmit (tunnel) them using the so-called GPRS Tunneling Protocol (GTP). In principle, we can distinguish between two kinds of GPRS backbones.

- Intra-PLMN backbones are IP-based networks owned by the GPRS network provider connecting the GSNs of the GPRS network.
- Inter-PLMN backbone networks connect GSNs of different GPRS networks. They are installed if there is a roaming agreement between two GPRS network providers. Figure 4.15 shows, how two intra-PLMN backbone networks of different PLMNs are connected with an inter-PLMN backbone. The gateways between the PLMNs and the external inter-PLMN backbone are called Border Gateways (BGs). Their main task is to perform security functions in order to protect the private intra-PLMN backbones against unauthorized users and attacks. The illustrated routing example is explained later.

The Gn and Gp interfaces are also defined between two SGSNs. This allows the SGSNs to exchange user profiles when a MS moves from one SGSN area to another. Across the Gf interface, the SGSN may query and check the IMEI of a MS trying to register with the network. The Gi interface connects the PLMN with external PDNs. In the GPRS standard, interfaces to IP (IPv4 and IPv6) and X.25 networks are supported. GPRS also adds some more entries to the GSM registers. For MM, the user's entry in the HLR is extended with a link to its current SGSN. Moreover, their GPRS-specific profile and current PDP address(es) are stored. The Gr interface is used to exchange this information between HLR and SGSN. For example, the SGSN informs the HLR about the current location of the MS. When a MS registers with a new SGSN, the HLR will send the user profile to the new SGSN. In a similar manner, the signaling path between GGSN and HLR (Gc interface) may be used by the GGSN to query the location and profile of a user who is unknown to the GGSN.

In addition, the MSC/VLR may be extended with functions and register entries which allow efficient coordination between packet-switched (GPRS) and conventional circuit switched GSM services. Examples for this are combined GPRS and GSM location updates and combined attachment procedures. Moreover, paging requests of circuit-switched GSM calls can be performed via the SGSN. For this purpose, the Gs interface connects the registers of SGSN and MSC/VLR. Finally, it is worth mentioning that it is possible to exchange messages of the SMS via GPRS. The Gd interface interconnects SMS-GMSC with the SGSN.

b. Services

Bearer services and supplementary services

The bearer services of GPRS offer end-to-end packet switched data transfer to mobile subscribers. Currently, a Point-to-Point (PTP) service is specified, which comes in two variants: a connectionless mode (PTP Connectionless Network Service (PTP-CLNS), e.g. for IP) and a connection-oriented mode (PTP Connection Oriented Network Service (PTPCONS), e.g. for X.25).

It is possible to use IP multicast routing protocols (see, e.g., Sahasra buddhe and Mukherjee (2000)) over GPRS. Packets addressed to an IP multicast group will then be routed to all group members. Furthermore, SMS messages can be sent and received over GPRS. Based on these standardized services, GPRS providers may offer additional nonstandardized services. Examples are access to information databases, messaging services (via store-and-forward mailboxes) and transaction services (e.g. credit card validations and electronic monitoring/surveillance systems). The most important application scenario, however, is the wireless access to the World Wide Web and to corporate intranets as well as e-mail communication.

Quality of service

The QoS requirements for the variety of mobile data applications, in which GPRS is used as transmission technology, are very diverse (for example, compare the requirements of real-time video conferencing with those of e-mail transfer with respect to packet delay and error-free transmission). Support of different QoS classes is therefore an important feature to support a broad variety of applications but still preserve radio and network resources in an efficient way. Moreover, QoS classes enable providers to offer different billing options.

The billing can be based on the amount of transmitted data, the service type itself and the QoS profile. At the moment, four QoS parameters are defined in GPRS: service precedence, reliability, delay and throughput. Using these parameters, QoS profiles can be negotiated between the mobile user and the network for each session, depending on the QoS demand and the currently available resources.

The service precedence is the priority of a service (in relation to other services). There exist three levels of priority: high, normal and low. In the case of a heavy traffic load, for example, packets of low priority will be discarded first. The reliability indicates the transmission characteristics required by an application. Three reliability classes are defined (see Table 8.1), which guarantee certain maximum values for the probability of packet loss, packet duplication, mis-sequencing and packet corruption (i.e. undetected error in a packet).

The delay parameters define maximum values for the mean delay and the 95th percentile delay (see Table 8.2). The latter is the maximum delay guaranteed in 95% of all transfers. Here, ‘delay’ is defined as the end-to-end transfer time between two communicating MSs or between a MS and the Gi interface to an external network, respectively. This includes all delays within the GPRS network, e.g., the delay for request and assignment of radio resources, transmission over

the air interface and the transit delay in the GPRS backbone network. Delays outside the GPRS network, e.g. in external transit networks, are not taken

Table 8.1 Probability of various outcomes with the three reliability classes.

Class	Lost packet	Duplicated packet	Out of sequence packet	Corrupted packet
1	10^{-9}	10^{-9}	10^{-9}	10^{-9}
2	10^{-4}	10^{-5}	10^{-5}	10^{-6}
3	10^{-2}	10^{-5}	10^{-5}	10^{-2}

Table 8.2 Delay classes.

Class	128 byte packet		1024 byte packet	
	Mean delay (s)	95% delay (s)	Mean delay (s)	95% delay (s)
1	<0.5	<1.5	<2	<7
2	<5	<25	<15	<75
3	<50	<250	<75	<375
4	Best effort	Best effort	Best effort	Best effort

into account. Table 8.2 lists the four defined delay classes and their parameters for a 128 byte and 1024 byte packet, respectively.

Finally, the throughput parameter specifies the maximum/peak bit rate and the mean bit rate.

2. Session management, mobility management and routing

Attachment and detachment procedure

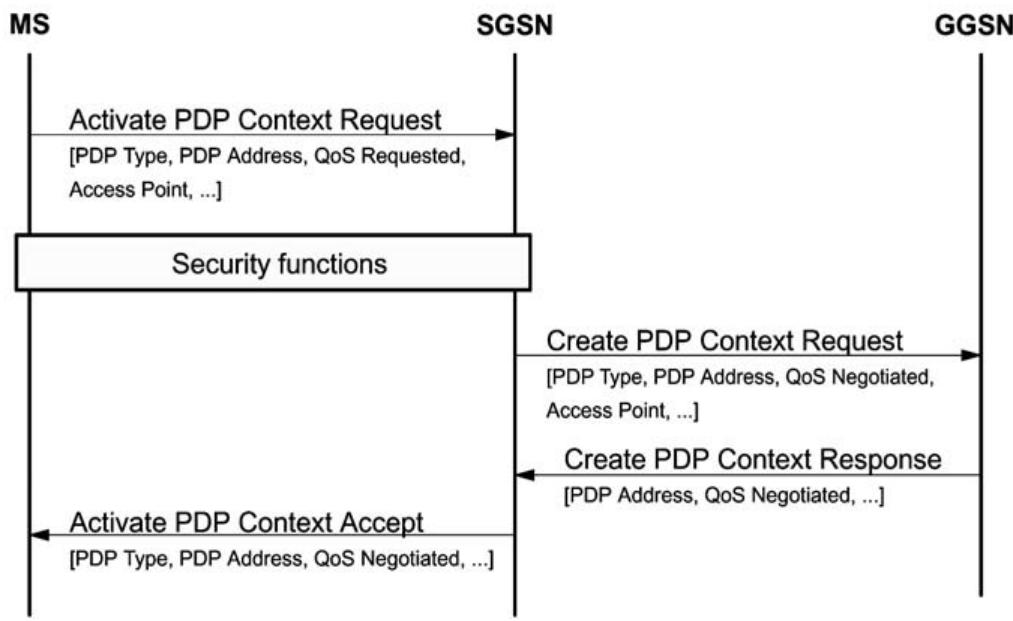


Fig 4.16 PDP context activation

Before a MS can use GPRS services, it must attach to the network (similar to the IMSI attach used for circuit-switched GSM services). The MS's ATTACH REQUEST message is sent

to the SGSN. The network then checks whether the user is authorized, copies the user profile from the HLR to the SGSN, and assigns a Packet Temporary Mobile Subscriber Identity (PTMSI) to the user. This procedure is called GPRS attach. For MSs using both circuit switched and packet-switched services, it is possible to perform combined GPRS/IMSI attach procedures. The disconnection from the GPRS network is called GPRS detach. It can be initiated by the MS or by the network.

Session management and PDP context

To exchange data packets with external PDNs after a successful GPRS attach, a MS must apply for an address used in the PDN. In general, this address is called a PDP address. In case the PDN is an IP network, this will be an IP address. For each session, a so-called PDP context is created, which describes the characteristics of the session. It contains the PDP type (e.g. IPv4), the PDP address assigned to the MS (e.g. an IP address), the requested QoS class and the address of a GGSN that serves as the access point to the external network. This context is stored in the MS, the SGSN and the GGSN. Once a MS has an active PDP context, it is ‘visible’ to the external network and can send and receive data packets. The mapping between the two addresses (PDP \leftrightarrow GSM address) makes the transfer of data packets between MS and GGSN possible.

The allocation of a PDP address can be static or dynamic. In the first case, the MS permanently owns a PDP address, which has been assigned by the network operator of the user’s home PLMN. Using a dynamic addressing concept, a PDP address is assigned upon\ activation of a PDP context, i.e. each time a MS attaches to the network it will in general get a new PDP address, and after its GPRS detach this PDP address will again be available to other MSs. The PDP address can be assigned by the user’s home-PLMN operator (dynamic PDP address). The GGSN is responsible for the allocation and deactivation of the addresses. Figure 4.16 shows the PDP context activation procedure initialized by the MS. Using the message ACTIVATE PDP CONTEXT REQUEST, the MS informs the SGSN about the requested PDP context. If a dynamic address is requested, the parameter PDP ADDRESS will be left empty. Afterwards, the usual GSM security functions (e.g. authentication of the user) are performed. If access is granted, the SGSN will send a CREATE PDP CONTEXT REQUEST to the affected GGSN. The GGSN creates a new entry in its PDP context table, which enables the GGSN to route data packets between the SGSN and the external PDN. It confirms this to the SGSN with a message CREATE PDP CONTEXT RESPONSE, which also contains the dynamic PDP address (if needed). Finally, the SGSN updates its PDP context table and confirms the activation of the new PDP context to the MS (ACTIVATE PDP CONTEXT ACCEPT). It is also worth mentioning that the GPRS standard supports anonymous PDP context activation, which is useful for special applications such as prepaid services. In such a session, the user (i.e. the IMSI) using the PDP context remains unknown to the network. Security functions as shown in Figure 8.3 are skipped. Only dynamic address allocation is possible in this case.

Routing

In Figure 4.14 we give an example of how packets can be routed in GPRS. We assume that the packet data network is an IP network. A GPRS MS located in PLMN1 sends IP packets to a Web server connected to the Internet. The SGSN which the MS is registered with encapsulates the IP packets coming from the MS, examines the PDP context, and routes them through the GPRS backbone to the appropriate GGSN. The GGSN decapsulates the IP packets

and sends them out on the IP network, where IP routing mechanisms transfer the packets to the access router of the destination network.

The latter delivers the IP packets to the host. Let us assume that the MS's home-PLMN is PLMN2 and that its IP address has been assigned from the PLMN2 address space – either in a dynamic or static way. When the Web server now addresses IP packets to the MS, they are routed to the GGSN of PLMN2 (the home-GGSN of the MS). This is because the MS's IP address has the same network prefix as the IP address of its home-GGSN. The GGSN queries the HLR and obtains the information that the MS is currently located in PLMN1. In the following, it encapsulates the incoming IP\ packets and tunnels them through the inter-PLMN GPRS backbone to the appropriate SGSN in PLMN1. The SGSN decapsulates the packets and delivers them to the MS.

Location management

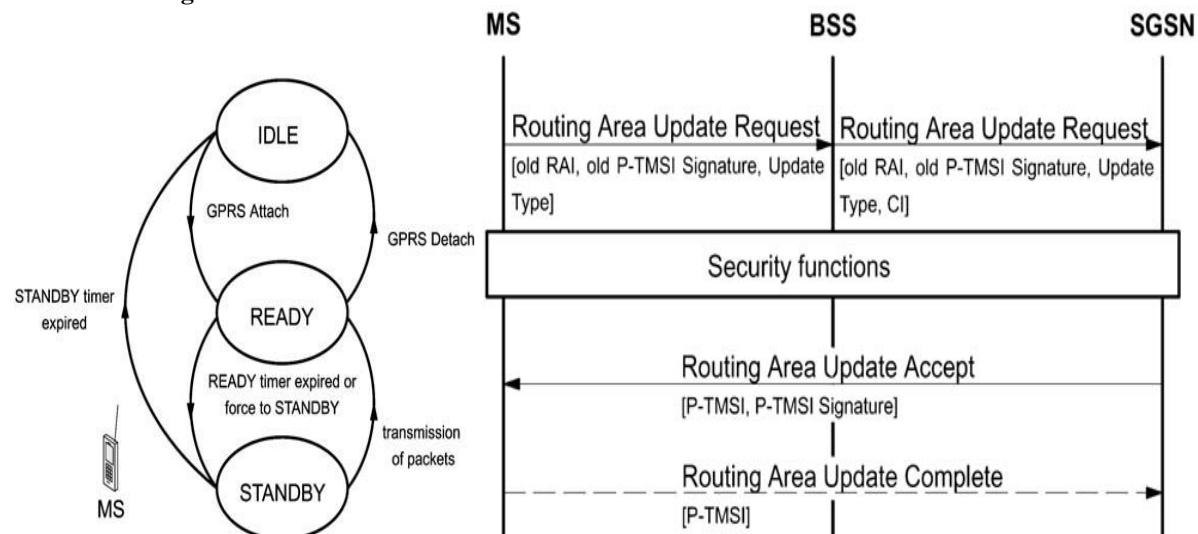


Fig 4.17 State model of a GPRS MS and Intra-SGSN routing area update

As in circuit-switched GSM, the main task of location management is to keep track of the user's current location, so that incoming packets can be routed to their MS. For this purpose, the MS frequently sends location update messages to its SGSN. How often should a MS send such a message? If it updates its current location (e.g. its cell) rather seldom, the network must perform a paging process in order to search the MS when packets are coming in. This will result in a significant delivery delay. On the other hand, if location updates happen very often, the MS's location is well known to the network (and thus the packets can be delivered without any additional paging delay), but quite a lot of uplink radio bandwidth and battery power is used for MM in this case. Thus, a good location management strategy must be a compromise between these two extreme methods.

For this reason, a state model for GPRS MSs has been defined (shown in Figure 4.17). In IDLE state the MS is not reachable. Performing a GPRS attach, it turns into READY state. With a GPRS detach it may deregister from the network and fall back to IDLE state, and all PDP contexts will be deleted. The STANDBY state will be reached when a MS does not send any packets for a long period of time, and therefore the READY timer (which was started at GPRS

attach and is reset for each incoming and outgoing transmission) expires. The location update frequency depends on the state in which the MS currently is. In IDLE state, no location updating is performed, i.e. the current location of the MS is unknown. If a MS is in READY state, it will inform its SGSN of every movement to a new cell. For the location management of a MS in STANDBY state, a GSM LA is divided into so-called Routing Areas (RAs). In general, an RA consists of several cells. The SGSN will only be informed, when a MS moves to a new RA; cell changes will not be indicated. To find out the current cell of a MS that is in STANDBY state, paging of the MS within a certain RA must be performed (section 8.1.7). For MSs in READY state, no paging is necessary. Whenever a MS moves to a new RA, it sends a ROUTING AREA UPDATE REQUEST to its assigned SGSN (Figure 8.5). The message contains the Routing Area Identity (RAI) of its old RA. The BSS adds the CI of the new cell to the request, from which the SGSN can derive the new RAI.

Two different scenarios are possible:

- Intra-SGSN routing area updates (Figure 4.17);
- Inter-SGSN routing area updates (Figure 4.18).

In the Intra-SGSN case, the MS has moved to an RA which is assigned to the same SGSN as the old RA. In this case, the SGSN has already stored the necessary user profile and can immediately assign a new P-TMSI (ROUTING AREA UPDATE ACCEPT). Since the routing context does not change, there is no need to inform other network elements, such as GGSN or HLR.

In the inter-SGSN case, the new RA is administered by a different SGSN than the old RA. The new SGSN realizes that the MS has entered its area and requests the old SGSN to send the PDP contexts of the user (SGSN CONTEXT REQUEST, SGSN CONTEXT RESPONSE, SGSN CONTEXT ACKNOWLEDGE). Afterward, the new SGSN informs the involved GGSNs about the user's new routing context (UPDATE PDP CONTEXT REQUEST, UPDATE PDP CONTEXT RESPONSE). In addition, the HLR and (if needed) the MSC/VLR are informed about the user's new SGSN number (UPDATE LOCATION, . . . , UPDATE LOCATION ACKNOWLEDGE; LOCATION UPDATE REQUEST, LOCATION UPDATE ACCEPT).

In addition to pure RA updates, there also exist combined RA/LA updates. They are performed whenever a MS using GPRS as well as conventional GSM services moves to a new LA. The MS sends a ROUTING AREA UPDATE REQUEST to the SGSN and uses a parameter update type to indicate that an LA update is needed. The message is then forwarded from the SGSN to the VLR.

To sum up, we can say that GPRS mobility management consists – as with GSM mobility management – of two levels: micro mobility management tracks the current RA or cell of the user; macro mobility management keeps track of the user's current SGSN and stores it in the HLR, VLR and GGSN.

3. Protocol architecture

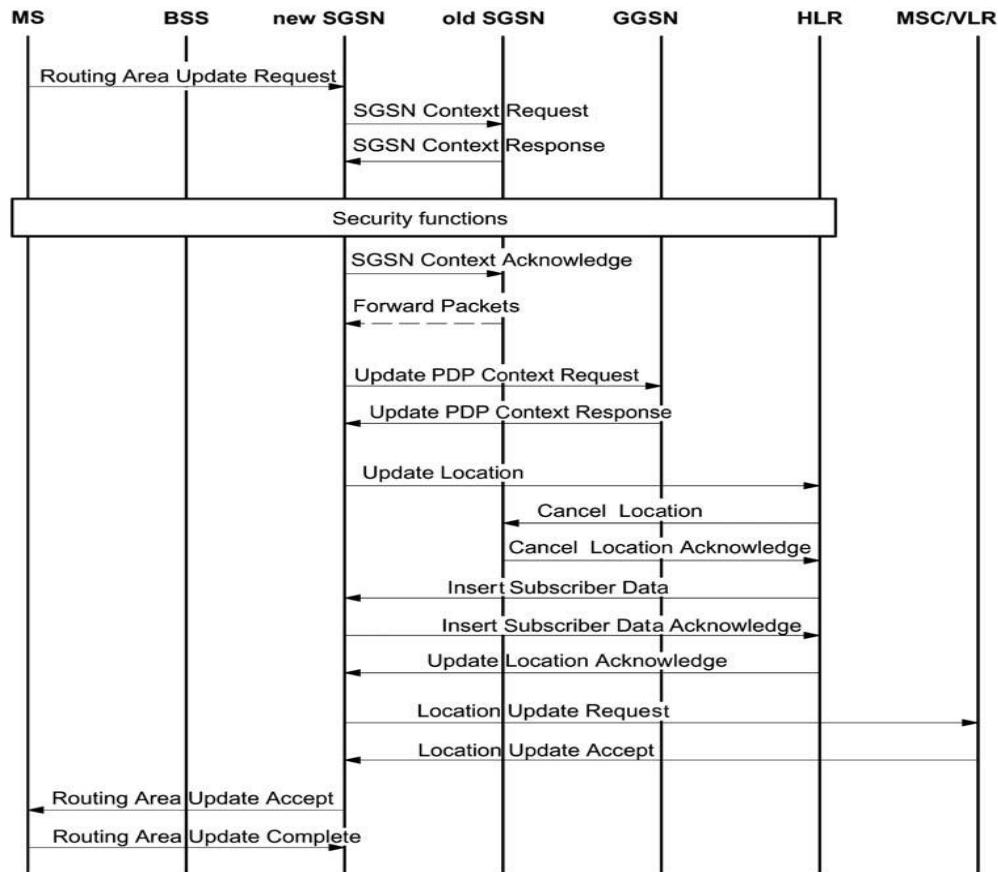


Fig 4.17 Inter-SGSN routing area update

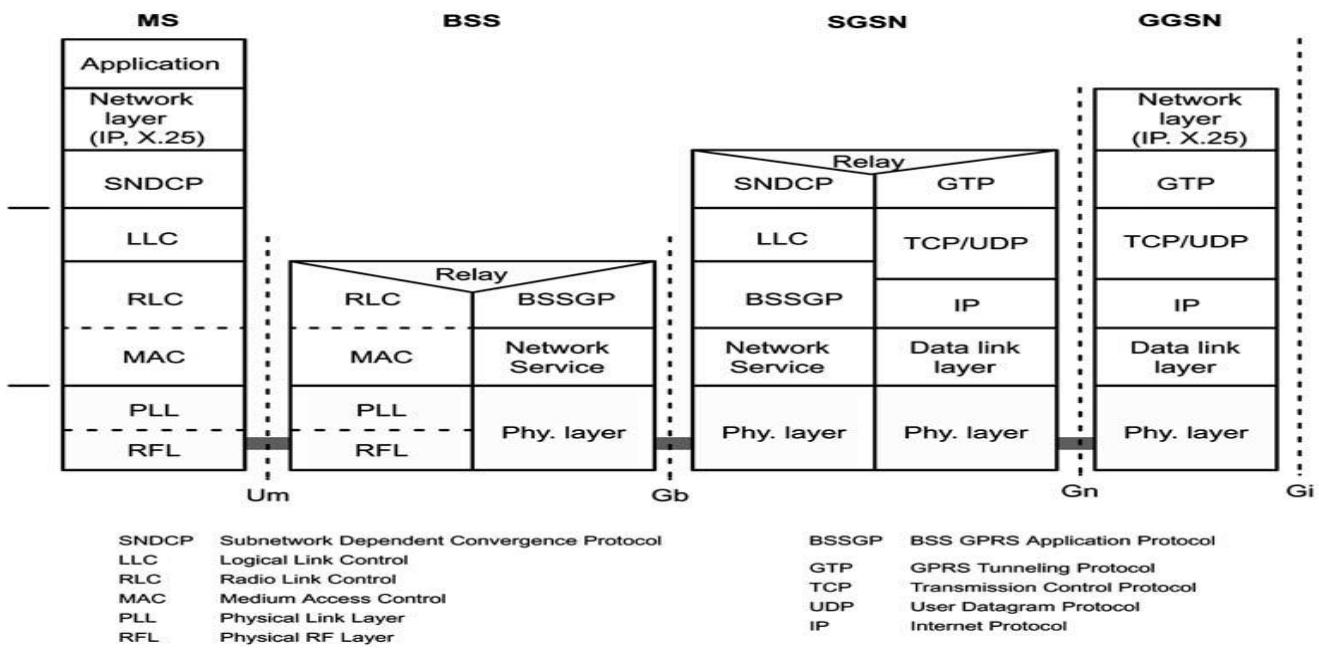


Fig 4.18 Protocol architecture: transmission plane.

Transmission plane

Figure 4.17 illustrates the protocol architecture of the GPRS transmission plane. The protocols offer transmission of user data and its associated signaling (e.g. for flow control, error detection and error correction). An application running in the GPRS-MS (e.g. a browser) uses IP or X.25 in the network layer. GPRS backbone: SGSN–GGSN. As mentioned earlier in this chapter, IP and X.25 packets are transmitted encapsulated within the GPRS backbone network. This is done using the GTP, i.e. GTP packets carry the user's IP or X.25 packets. GTP is defined both between GSNs within the same PLMN (Gn interface) and between GSNs of different PLMNs (Gp interface).

GTP contains procedures in the transmission plane as well as in the signaling plane. In the transmission plane, GTP employs a tunnel mechanism to transfer user data packets. In the signaling plane, GTP specifies a tunnel control and management protocol. The signaling is used to create, modify and delete tunnels. A Tunnel Identifier (TID), which is composed of the IMSI of the user and a Network Layer Service Access Point Identifier (NSAPI), uniquely indicates a PDP context. Below GTP, the standard protocols Transmission Control Protocol (TCP) or User Datagram Protocol (UDP) are employed to transport the GTP packets within the backbone network. TCP is used for X.25 (since X.25 expects a reliable end-to end connection) and UDP is used for access to IP-based networks (which do not expect reliability in the network layer or below). In the network layer, IP is employed to route the packets through the backbone. Ethernet, ISDN or Asynchronous Transfer Mode (ATM)-based protocols may be used below IP. To summarize, in the GPRS backbone we have an IP/X.25- over-GTP-over-UDP/TCP-over-IP protocol architecture.

Air interface. In the following we consider the air interface (Um) between MS and BSS or SGSN, respectively.

Subnetwork dependent convergence protocol. The Subnetwork Dependent Convergence Protocol (SNDCP) is used to transfer packets of the network layer (IP and X.25 packets) between the MSs and their SGSN. Its functionality includes:

- multiplexing of several PDP contexts of the network layer onto one virtual logical connection of the underlying Logical Link Control (LLC) layer; and
- Segmentation of network layer packets onto one frame of the underlying LLC layer and reassembly on the receiver side.

Moreover, SNDCP offers compression and decompression of user data and redundant header information (e.g. TCP/IP header compression).

Data link layer. The data link layer is divided into two sublayers:

- LLC layer (between MS and SGSN); and
- RLC/Medium Access Control (MAC) layer (between MS and BSS).

The LLC layer provides a reliable logical link between a MS and its assigned SGSN. Its functionality is based on the LAPDm protocol (which is a protocol similar to HDLC and has

been explained in section 5.3.1). LLC includes in-order delivery, flow control, error detection, retransmission of packets (ARQ) and ciphering functions. It supports variable frame lengths and different QoS classes, and besides point-to-point also point-to-multipoint transfer is possible. A logical link is uniquely addressed with a Temporary Logical Link Identifier (TLLI). Within one RA the mapping between TLLI and IMSI is unique. However, the user's identity remains confidential, since the TLLI is derived from the P-TMSI of the user.

The RLC/MAC layer has two functions. The purpose of the RLC layer is to establish a reliable link between the MS and the BSS. This includes the segmentation and reassembly of LLC frames into RLC data blocks and ARQ of uncorrectable blocks. The MAC layer controls the access attempts of MSs on the radio channel. It is based on a slotted-aloha principle (section 4.1). The MAC layer employs algorithms for contention resolution of access attempts, statistical multiplexing of channels and a scheduling and prioritizing scheme, which takes into account the negotiated QoS. On the one hand, the MAC protocol allows that a single MS simultaneously uses several physical channels (several time slots of the same TDMA frame). On the other hand, it also controls the statistical multiplexing, i.e. it controls how several MSs can access the same physical channel (the same time slot of successive TDMA frames).

Physical layer: The physical layer between MS and BSS can be divided into the two sublayers: Physical Link Layer (PLL) and Physical RF Layer (RFL). The PLL provides a physical channel between the MS and the BSS. Its tasks include channel coding (i.e. detection of transmission errors, forward error correction and indication of uncorrectable code words), interleaving and detection of physical link congestion. The RFL, which operates below the PLL, includes modulation and demodulation.

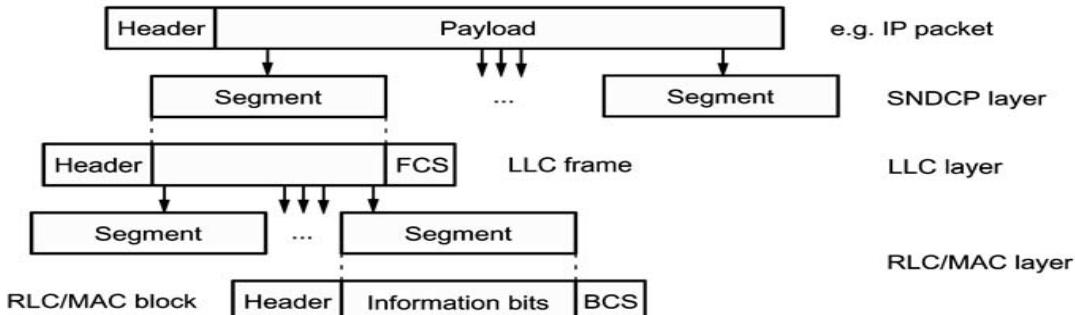


Fig 4.19 Data flow and segmentation between the protocol layers in the MS

To summarize this section, Figure 4.19 illustrates the data flow between the protocol layers in the MS. Packets of the network layer (e.g. IP packets) are passed down to the SNDCP layer, where they are segmented to LLC frames. After adding header information and a Frame Check Sequence (FCS) for error protection, these frames are segmented into one or several RLC data blocks. Those are then passed down to the MAC layer. One RLC/MAC block contains a MAC and RLC header, the RLC payload ('information bits') and a Block Check Sequence (BCS) at the end.

BSS–SGSN interface. At the Gb interface, the BSS GPRS Application Protocol (BSSGP) is defined on Layer 3. It is derived from the BSSMAP, which has been explained in section 5.3.1. The

BSSGP delivers routing and QoS-related information between BSS and SGSN. The underlying Network Service (NS) protocol is based on the frame relay protocol.

Routing and conversion of addresses

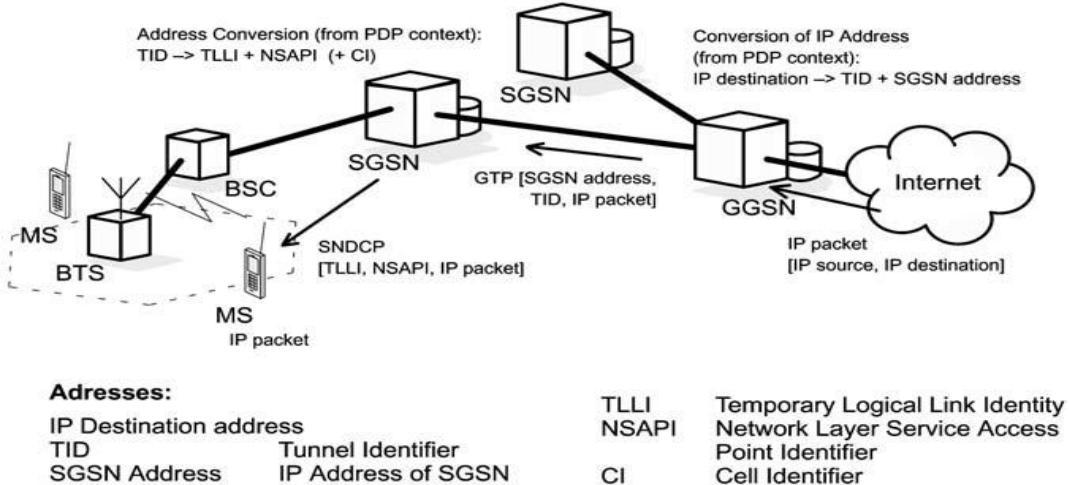


Fig 4.20 Routing and address conversion: incoming IP packet (mobile-terminated data transfer)

Figure 4.20 roughly illustrates the transfer of an incoming IP packet. It arrives at the GGSN, is then routed through the GPRS backbone to the responsible SGSN and finally to the MS. Using the PDP context, the GGSN determines from the IP destination address a TID and the IP address of the relevant SGSN. Between GGSN and the SGSN, the GTP is employed. The SGSN derives the TLLI from the TID and finally transfers the IP packet to the MS. The so-called NSAPI is part of the TID. It maps a given IP address to the corresponding PDP context. An NSAPI/TLLI pair is unique within one RA. Figure 4.21 gives a similar example with an outgoing (mobile originated) IP packet.

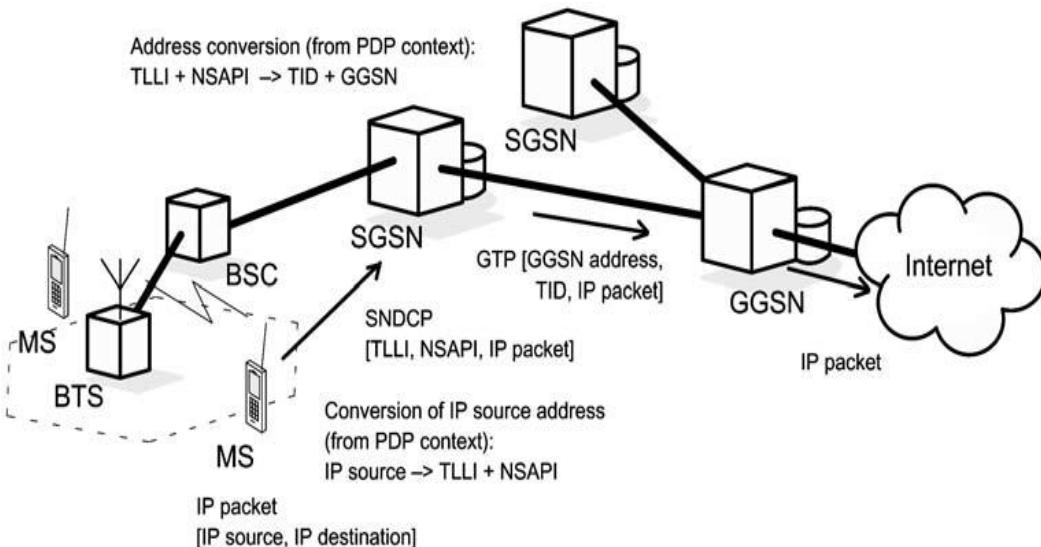


Fig 4.21 Routing and address conversion: outgoing IP packet (mobile-originated data transfer).

4. Signaling plane

The protocol architecture of the signaling plane comprises protocols for control and support of the functions of the transmission plane, e.g., for the execution of GPRS attach and detach, PDP context activation, the control of routing paths and the allocation of network resources. Between MS and SGSN (Figure 4.22), the GPRS Mobility Management and Session Management (GMM/SM) protocol is responsible for mobility and session management. It includes functions for GPRS attach/detach PDP context activation, routing area updates and security procedures. The signaling architecture between SGSN and the registers HLR, VLR and EIR (Figure 4.23) uses protocols known from conventional GSM and partly extends them with GPRS-specific functionality. Between SGSN and HLR as well as between SGSN and EIR, an enhanced MAP is employed. The exchange of MAP messages is accomplished over the TCAP, the SCCP and the MTP.

The BSS Application Part (BSSAP+) includes functions of GSM's BSSAP. It is applied to transfer signaling information between the SGSN and the VLR (Gs interface). This includes, in particular, signaling of the mobility management when coordination of GPRS and conventional GSM functions is necessary (e.g. for combined GPRS and nonGPRS location update, combined GPRS/IMSI attach or paging of a MS via GPRS for an incoming GSM call).

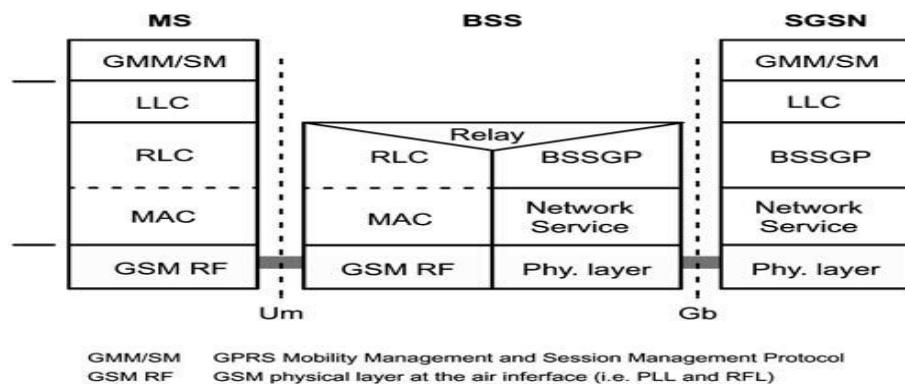


Fig4.22 signaling plane: MS-SGSN

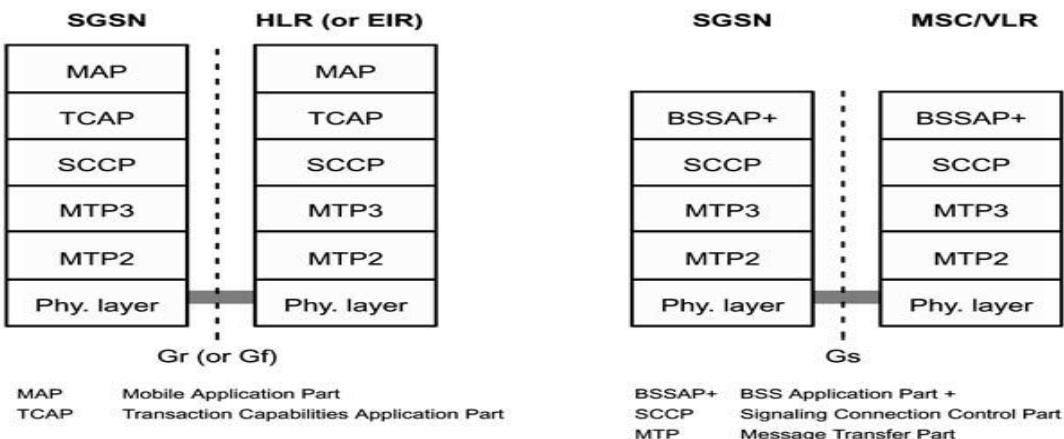


Fig 4.23 Signaling plane: SGSN-HLR, SGSN-EIR and SGSN-MSC/VLR.

5. Interworking with IP networks

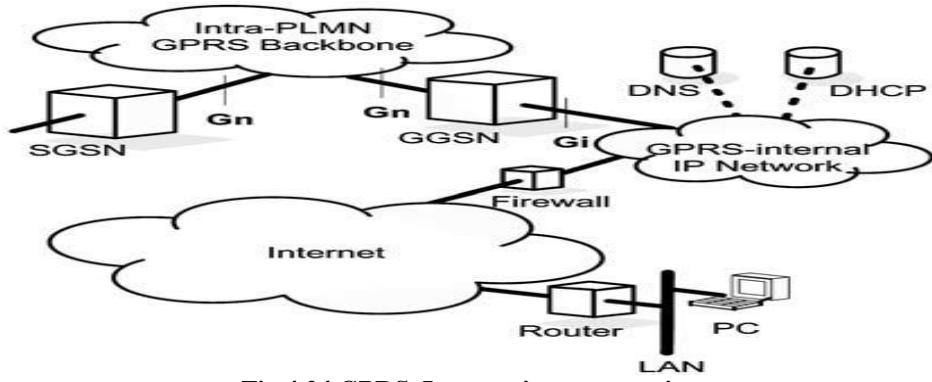


Fig 4.24 GPRS–Internet interconnection

Figure 4.24 gives an example of how a GPRS network is interconnected with the Internet. From outside, i.e. from an external IP network's point of view, the GPRS network looks like any other IP Subnetwork and the GGSN looks like a usual IP router. As explained in section 8.1.3, each MS obtains an IP address after its GPRS attach, which is valid for the duration of the session. The network provider has reserved a certain number of IP addresses, and can dynamically assign these addresses to active MSs. To do so, the network provider may install a Dynamic Host Configuration Protocol (DHCP) server in its network. This server automatically manages the available address space. The address resolution between IP address and GSM address is performed by the GGSN, using the appropriate PDP context.

6. Air interface

The enhanced air interface of GPRS offers higher data rates and a packet-oriented transmission. It is therefore considered one of the key aspects in GPRS. In this section, we explain how several MSs can share one physical channel (multiple access) and how the assignment of radio resources between circuit-switched GSM services and GPRS services is controlled. Afterwards, the logical channels and their mapping onto physical channels (using multiframe) is presented. Finally, GPRS channel coding concludes this chapter.

Multiple access and radio resource management

On the physical layer, GPRS uses the GSM combination of FDMA and TDMA with eight time slots per TDMA frame). However, several new methods are used for channel allocation and multiple access. They have a significant impact on the performance of GPRS. In circuit-switched GSM, a physical channel (i.e. one time slot of successive TDMA frames) is permanently allocated for a particular MS during the entire call period (no matter whether data are transmitted or not). Moreover, it is assigned in the uplink as well as in the downlink.

GPRS enables a far more flexible resource allocation scheme for packet transmission. A GPRS MS can transmit on several of the eight time slots within the same TDMA frame (multislot operation). The number of time slots which a MS is able to use is called a multislot class. In addition, uplink and downlink are allocated separately, which saves radio resources, especially for asymmetric traffic (e.g. Web browsing).

A cell supporting GPRS must allocate physical channels for GPRS traffic. In other words, the radio resources of a cell are shared by all MSs (GSM and GPRS) located in this cell. The mapping of physical channels to either GPRS or circuit-switched GSM services can be performed in a dynamic way. A physical channel which has been allocated for GPRS transmission is denoted as a Packet Data Channel (PDCH). The number of PDCHs can be adjusted according to the current traffic demand (capacity on demand principle). For example, physical channels not currently in use by GSM calls can be allocated as PDCHs for GPRS to increase the QoS for GPRS. When there is a resource demand for GSM calls, PDCHs may be deallocated.

As already mentioned, physical channels for packet-switched transmission (PDCHs) are only allocated for a particular MS when this MS sends or receives data packets, and they are released after the transmission. With this dynamic channel allocation principle, multiple MSs can share one physical channel. For bursty traffic this results in a much more efficient usage of the radio resources. The channel allocation is controlled by the BSC. To prevent collisions, the network indicates which channels are currently available in the downlink. An Uplink State Flag (USF) in the header of downlink packets shows which MS is allowed to use this channel in the uplink. The allocation of PDCHs to a MS also depends on its multislot class and the QoS of the session.

Logical channels

Table 8.3 lists the packet data logical channels defined in GPRS. As with logical channels in conventional GSM, they can be divided into two categories: traffic channels and signaling (control) channels. The signaling channels can further be divided into packet broadcast control, packet common control, and packet dedicated control channels.

The Packet Data Traffic Channel (PDTCH) is employed for the transfer of user data. It is assigned to one MS (or, in the case of PTM, to multiple MSs). One MS can use several PDTCHs simultaneously. The Packet Broadcast Control Channel (PBCCH) is a unidirectional point-to-multipoint signaling channel from the BSS to the MSs. It is used by the BSS to broadcast information about the organization of the GPRS radio network to all GPRS MSs of a cell. In addition to system information about GPRS, the PBCCH should also broadcast important system information about circuit-switched services, so that a GSM/GPRS MS does not need to listen to the BCCH.

The Packet Common Control Channel (PCCCH) transports signaling information for functions of the network access management, i.e. for allocation of radio channels, medium access control and paging. Four sub-channels are defined:

- The Packet Random Access Channel (PRACH) is used by the MSs to request one or more PDTCH;
- The Packet Access Grant Channel (PAGCH) is used to allocate one or more PDTCH to a MS;
- The Packet Paging Channel (PPCH) is used by the BSS to find the location of a MS (paging) prior to downlink packet transmission;
- The Packet Notification Channel (PNCH) is used to inform MSs of incoming PTM messages.

Figure 4.25 shows the principle of the uplink channel allocation (mobile-originated packet transfer). A MS requests a channel by sending a PACKET CHANNEL REQUEST on the PRACH or RACH. The BSS answers on the PAGCH or AGCH, respectively. Once the PACKET CHANNEL REQUEST is successful; a so-called Temporary Block Flow (TBF) is established. With that, resources (e.g. PDTCH and buffers) are allocated for the MS, and data transmission can start. During transfer, the USF in the header of downlink blocks indicates to other MSs that this uplink PDTCH is already in use. On the receiver side, a Temporary Flow Identifier (TFI) helps to reassemble the packet. Once all data has been transmitted, the TBF and the resources are released again. Figure 4.25 illustrates the paging procedure of a mobile station (mobile-terminated packet transfer). The packet dedicated control channel is a bidirectional point-to-point signaling channel. It contains the following channels.

- The Packet Associated Control Channel (PACCH) is always allocated in combination with one or more PDTCH. It transports signaling information related to one specific MS (e.g. power control information).
- The Packet Timing Advance Control Channel (PTCCH) is used for adaptive frame synchronization. The MS sends over the uplink part of the PTCCH, the PTCCH/U, ABs to the BTS. From the delay of these bursts, the correct value for the TA can be derived; This value is then transmitted in the downlink part, the PTCCH/D, to inform the MS.

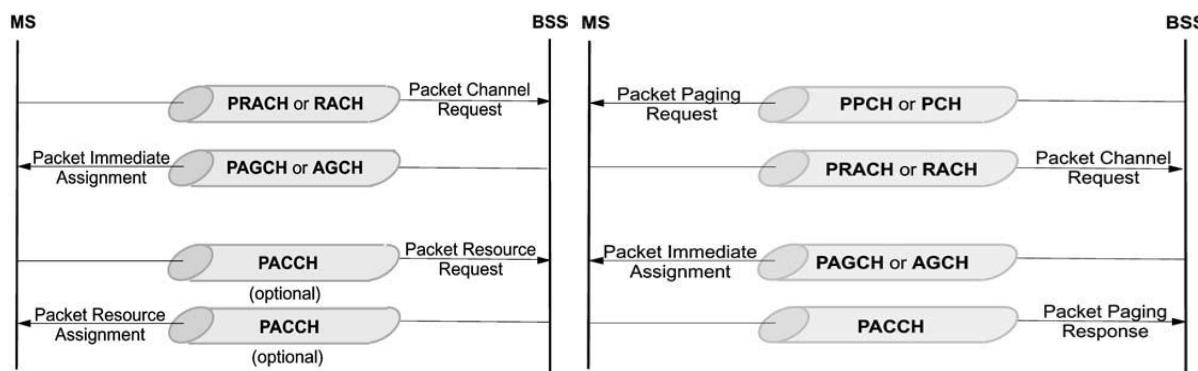


Fig 4.25 Uplink channel allocation (mobile-originated packet transfer) and Paging (mobile-terminated packet transfer)

Table 8.3 Logical channels in GPRS.

Group		Channel	Function	Direction
Traffic channels	Packet data traffic channel	PDTCH	Packet data traffic	MS ↔ BSS
Signaling channels	Packet broadcast control channel	PBCCH	Packet broadcast control	MS ← BSS
	Packet common control channel (PCCCH)	PRACH	Packet random access	MS → BSS
		PAGCH	Packet access grant	MS ← BSS
		PPCH	Packet paging	MS ← BSS
		PNCH	Packet notification	MS ← BSS
Packet dedicated control channels		PACCH	Packet associated control	MS ↔ BSS
		PTCCH	Packet timing advance control	MS ↔ BSS

Mapping of packet data logical channels onto physical channels

Table 8.5 Combinations of logical GPRS channels.

	B10	B11	B12	B13
PDTCH				
PBCCH				
PCCCH				
PACCH				
PTCCH				

Table 8.6 Channel combinations used by the MS.

	M19	M10
PDTCH		✓+✓
PBCCH		
PCCCH		
PACCH		
PTCCH		

We know that the mapping of logical GSM channels onto physical channels has two components: mapping in frequency and mapping in time. The mapping in frequency is based on the TDMA frame number and the frequencies allocated to the BTS and the MS. The mapping in time is based on the definition of complex multiframe structures on top of the TDMA frames.

A multiframe structure for PDCHs consisting of 52 TDMA frames (each with eight time slots). The corresponding time slots of a PDCH of four consecutive TDMA frames form one radio block (blocks B0–B11). Two TDMA frames are reserved for\ transmission of the PTCCH, and the remaining two frames are IDLE frames. A multiframe has thus duration of approximately 240 ms (52×4.615 ms).

A radio block consists of 456 bits. The mapping of the logical channels onto the blocks B0–B11 of the multiframe can vary from block to block and is controlled by parameters which are broadcast on the PBCCH. The GPRS recommendations define which time slots may be used by a logical channel. In addition to the 52-multiframe structure, which can be used by all logical GPRS channels, a 51-multiframe structure is also defined. It is used for PDCHs carrying only the logical channels PCCCH and PBCCH (channel combination B13 in Table 8.5). In the downlink, it consists of 10 blocks each of 4 frames (B0-B9) and 10 IDLE frames. In the uplink, it has 51 random access frames. Its duration is 235.4 ms.

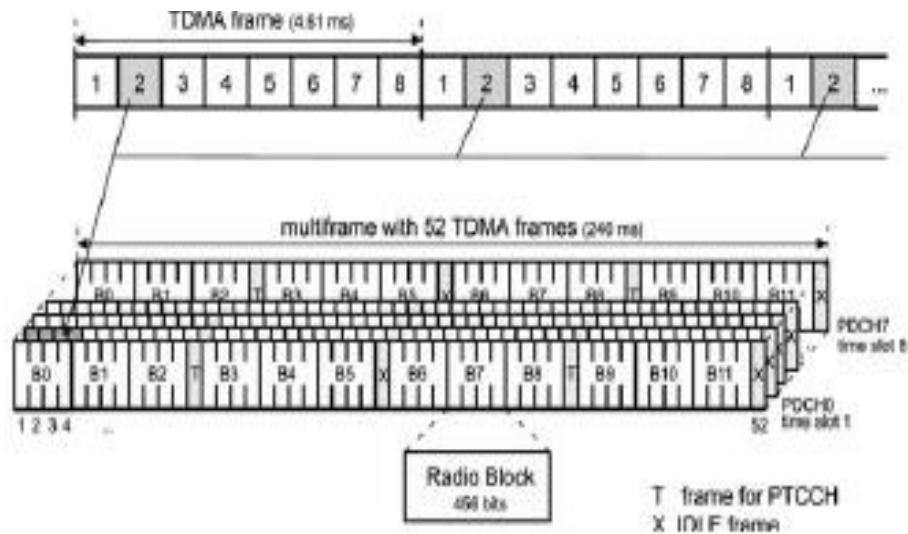


Figure 8.16 Multiframe structure with 52 TDMA frames.

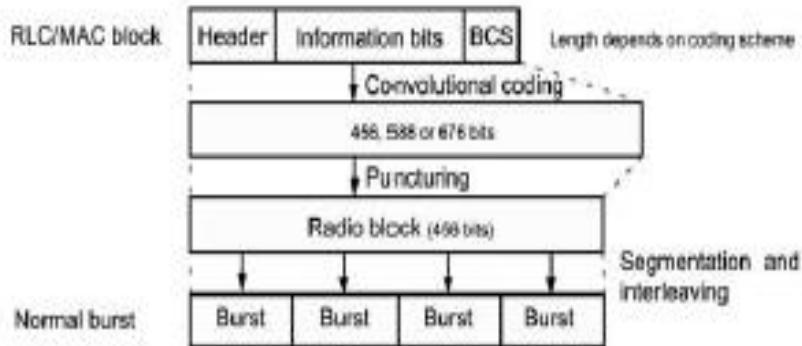


Figure 8.17 Physical layer at the air interface: channel coding, interleaving and formation of bursts (continued from Figure 8.8).

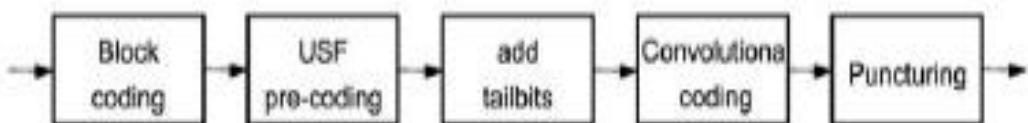


Figure 8.18 Encoding of GPRS data blocks.

Channel coding

Figure 8.17 shows how a block of the RLC/MAC layer (compare with Figure 8.8) is encoded and mapped onto four bursts. Channel coding is used to protect the transmitted data packets against errors and perform forward error correction. The channel coding technique in

GPRS is quite similar to that employed in conventional GSM. An outer block coding, an inner convolutional coding and an interleaving scheme are used.

Let us employ coding scheme CS-2. First of all, the 271 information bits of an RLC/MAC block (268 bits plus 3 bits USF; Table 8.4) are mapped to 287 bits using a systematic block encoder, i.e. 16 parity bits are added. These parity bits are denoted as a BCS. The USF pre-encoding maps the first 3 bits of the block (i.e. the USF) to 6 bits in a systematic way. Afterwards, 4 zero bits (tail bits) are added at the end of the entire block. The tail bits are needed for the termination of the subsequent convolutional coding.

Let us employ coding scheme CS-2. First of all, the 271 information bits of an RLC/MAC block (268 bits plus 3 bits USF; Table 8.4) are mapped to 287 bits using a systematic block encoder, i.e. 16 parity bits are added. These parity bits are denoted as a BCS. The USF pre-encoding maps the first 3 bits of the block (i.e. the USF) to 6 bits in a systematic way. Afterwards, 4 zero bits (tail bits) are added at the end of the entire block. The tail bits are needed for the termination of the subsequent convolutional coding.

For the convolutional coding, a nonsystematic rate-1/2 encoder with memory 4 is used, which is defined by the generator polynomials

$$G_0(d) = 1 + d^3 + d^4,$$

$$G_1(d) = 1 + d + d^3 + d^4.$$

This is the same encoder as used in conventional GSM. A possible encoder realization is shown in Figure 4.32. At the output of the convolutional encoder a codeword of length 588 bits results. Following this, 132 bits are punctured, resulting in a radio block of length 456 bits. Thus, we obtain a code rate of the convolutional encoder (including the puncturing) of

$$r = \frac{6 + 268 + 16 + 4}{456} \approx \frac{2}{3}.$$

Coding scheme CS-1 is equivalent to the coding of the SACCH. A systematic fire code is used for block coding (see section 4.8.1, first paragraph). There is no pre-coding of the USF bits. The convolutional coding is performed with the known rate-1/2 encoder, however, this time the output sequence is not punctured. Using CS-4, the 3 USF bits are mapped to 12 bits, and no convolutional coding is applied.

For the coding of the traffic channels (PDTCH), one of the four coding schemes is chosen, depending on the quality of the signal. The two SFs in a NB (Figure 4.7) are used to indicate which coding scheme is used. Under very bad channel conditions, CS-1 yields a data rate of only 9.05 kbit/s per time slot, but a very reliable coding. Under good channel conditions, convolutional coding is skipped (CS-4), and we achieve a data rate of 21.4 kbit/s per time slot. Thus, we obtain a theoretical maximum data rate of 171.2 kbit/s per TDMA frame. In practice, multiple users share the time slots and, thus, a much lower bit rate is available to the individual user. Moreover, the quality of the radio channel will not always allow us to use CS-4 (or CS-4 is not supported by the mobile terminal or by the network operator). The data rate available to the user depends (among other things) on the current total traffic load in the cell (i.e. the number of users and their traffic characteristics), the used coding scheme,

7. Authentication and ciphering

The security principles inside the GPRS network are almost equivalent to those used in conventional GSM (section 5.6). Security functions in the GPRS network:

- protect against unauthorized use of services (by authentication and service request validation);
- provide data confidentiality (using ciphering); and
- provide confidentiality of the subscriber identity.

As in GSM, two keys are used: the subscriber authentication key K_i and the cipher key K_c . The main difference is that the SGSN, not the MSC, which handles authentication. Moreover, a special GPRS ciphering algorithm (A5) has been defined, which is optimized for encryption of packet data.

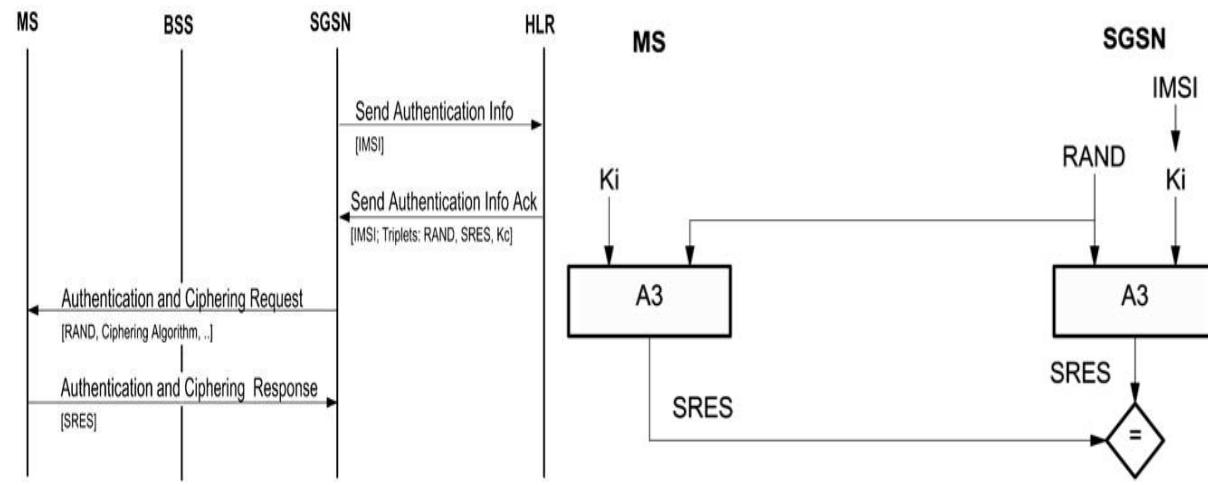


Fig 4.27 Subscriber authentication in GPRS and Principle of subscriber authentication in GPRS

User authentication

Figures 4.27 illustrate the GPRS authentication process. The standard GSM algorithms are used to generate security data. The algorithm A3 calculates the signature response (SRES) from the subscriber authentication key (K_i) and a random number (RAND). If the SGSN does not have authentication sets for a user (K_c , RAND, SRES), it requests them from the HLR by sending a message SEND AUTHENTICATION INFO. The HLR responds with a SEND AUTHENTICATION INFO ACK which includes the security data. Now, the SGSN offers a random number RAND to the MS (AUTHENTICATION AND CIPHERING REQUEST). The MS calculates SRES and transmits it back to the SGSN (AUTHENTICATION AND CIPHERING RESPONSE). If the SRES of the MS is equal to the SRES calculated (or maintained) by the SGSN, the user is authenticated and is allowed to use the network.

Ciphering

The ciphering functionality is performed in the LLC layer between MS and SGSN. Thus, the ciphering scope reaches from the MS all the way to the SGSN (and vice versa), whereas in conventional GSM the scope is only between MS and BTS/BSC.

As in GSM ciphering, the algorithm A8 generates the cipher key K_c from the key K_i and a random number RAND (see Figure 5.53). K_c is then used by the GPRS Encryption Algorithm (GEA) for data encryption (algorithm A5). Note that the key K_c which is handled by the SGSN is independent of the key K_c handled by the MSC for conventional GSM services. A MS may

thus have more than one Kc key. The MS and the SGSN start ciphering after the message AUTHENTICATION AND CIPHERING RESPONSE is sent or received, respectively. Afterwards, GPRS user data and signaling during data transfer are transmitted in an encrypted manner.

Subscriber identity confidentiality

As in GSM, the identity of the subscriber is confidential. This is done by using temporary identities on the radio channel. In particular, the user's IMSI is not transmitted unencrypted; instead a Packet Temporary Mobile Subscriber Identity (P-TMSI) is assigned to each user by the SGSN. This address is temporary and is only valid and unique in the service area of this SGSN. From the P-TMSI, a TLLI can be derived. The mapping between these temporary identities and the IMSI is stored only in the MS and in the SGSN.

8. HSCSD

As the name implies, the High Speed Circuit Switched Data Service is, in contrast to GPRS, circuit-switched. That is, the user has a fixed data rate bearer available for the duration of the data connection. This is independent of the amount of data actually transmitted, such that the connection has to be paid for, even during periods, in which no data are transmitted, according to the respective higher layer services used. This means that HSCSD is specifically useful for applications which demand for a fixed data rate.

The advantage of this, however, is that the data rate is guaranteed during the connection time, in case a transparent bearer service is applied. On the other hand, the QoS is secured, if a nontransparent bearer service is applied. HSCSD supports both options. Just as in GPRS, also HSCSD allows for the parallel use of several, say n , traffic channels to provide higher data rates. Figure 8.21 depicts an example with $n = 2$, in which timeslots TS1 and TS2 are used for one HSCSD connection, both in uplink and downlink.

The principal restriction for the number of timeslots n is given by the requirement that they all have to reside on the same frequency channel. Therefore, the standard allows for up to $n = 8$ timeslots or channels to be assigned to one user. This gives us the maximal data rate achievable: using eight channels at once, each carrying a full-rate traffic channel TCH/F9.6, a sum data rate of 76.8 kbit/s could be achieved. However, it should be recalled that, even though GSM applies FDD, uplink and downlink timeslots of the same TCH have an offset, such that the mobile terminal can perform transmission and reception subsequently. This because terminals with the ability to perform transmit and receive operations in parallel would imply much higher complexity at the terminal and thus would make terminals much more expensive. Therefore, earlier versions of the HSCSD standard only allowed for up to $n = 4$ timeslots to be used at once. This would present us with a data rate of 38.4 kbit/s, when four TCH/F9.6 channels are used.

In fact, by applying a different coding scheme, a maximum data rate of 57.6 kbit/s is achieved with $n = 4$. Figure 8.21 shows, for the example of $n = 2$, how the required operations in the mobile terminal can still be performed sequentially, as was originally intended in classical GSM: the mobile is receiving in timeslots TS1 and TS2 of the downlink frame. Owing to the

time offset of three timeslots between the uplink and the downlink frames, the terminal has already completed reception, when it has to start transmitting on the uplink, also using TS1 and TS2, but now on the uplink frames. After completion of transmission, and before having to receive data again in the following TS1 and TS2 on the downlink, there is still enough time to monitor BCCH carriers of neighboring cells, which is important, e.g., for handover issues.

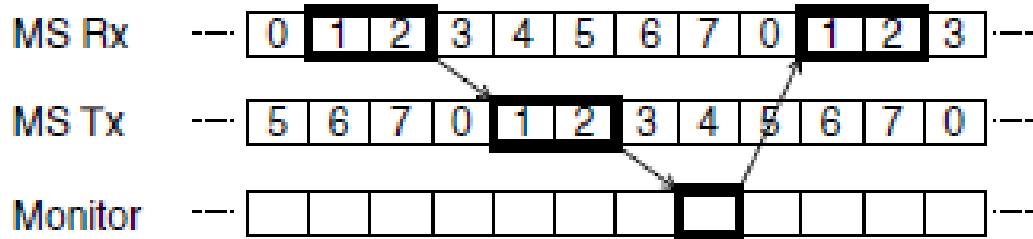


Figure 8.21 Example HSCSD channel occupation with $n = 2$.

When assigning timeslots, the HSCSD service takes the so-called multislot class of the MS into consideration. A list of the multislot classes considered for HSCSD is shown in Table 8.8. It lists the class name, the maximum allowed number of downlink timeslots (Rx), uplink timeslots (Tx), and sum of downlink and uplink timeslots. For instance, multislot class 3 terminals might use up to two timeslots for reception (downlink) and up to two timeslots for transmission (uplink). However, the sum of uplink and downlink timeslots must never be larger than three. When a mobile of a certain multislot class initiates a HSCSD connection, the RRM must consider the restrictions of the respective multislot class, when assigning timeslots. The last column in Table 8.8 refers to the type of the MS: type 1 terminals do not have the ability to transmit and receive at the same time. Therefore, the RRM must ensure the appropriate selection of timeslots. In contrast, type 2 mobiles do have the ability to transmit and receive in parallel, which gives the RRM more options to allocate timeslots for HSCSD.

Architecture

From an architectural point of view, HSCSD does not demand many changes as compared with standard GSM data services (see Appendix A). The GSM architecture for HSCSD support is depicted in Figure 8.22. At the Um interface between MS and BTS up to $n = 8$ TCH/F channels are used, which are forwarded transparently via the Abis interface from the BTS to the BSC. Here the data from all parallel channels are multiplexed on a single 64 kbit/s connection and transmitted over the A interface to the MSC. The n full rate channels are considered independent of each other and are treated individually for the purpose of, e.g., air interface error control. However, logically they belong to the same HSCSD configuration and are controlled as one radio link by the network for the purpose of cellular operations, such as handover. This requires new BSS functionality. The main difference to the standard GSM data services is in a combining and splitting functionality that is demanded at the MS and the MSC, combining and splitting, respectively, the multiple data streams that are transmitted between both entities. This functionality is provided in the Terminal Adoption Function (TAF) at the MS and in the IWF at the MSC, respectively.

Table 8.8 HSCSD MS multislot classes.

Multislot class	Maximum number of slots				Type
	Rx	Tx	Sum		
1	1	1	2		1
2	2	1	3		1
3	2	2	3		1
4	3	1	4		1
5	2	2	4		1
6	3	2	4		1
7	3	3	4		1
8	4	1	5		1
9	3	2	5		1
10	4	2	5		1
11	4	3	5		1
12	4	4	5		1
13	3	3	NA		2
14	4	4	NA		2
15	5	5	NA		2
16	6	6	NA		2
17	7	7	NA		2
18	8	8	NA		2

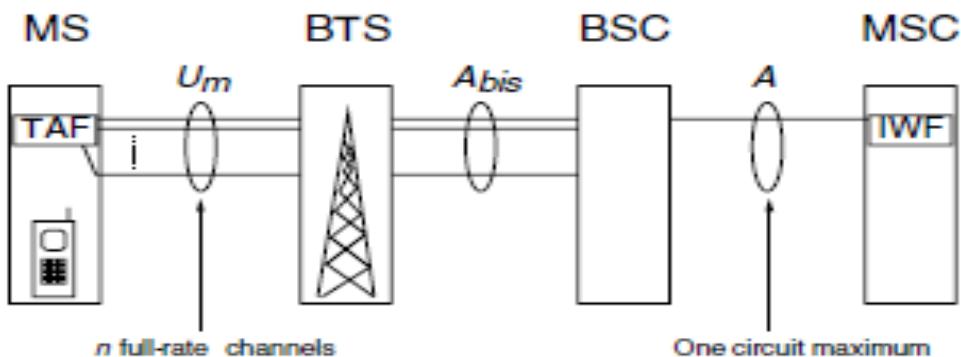


Figure 8.22 GSM architecture for HSCSD support.

Air interface

As described above at the air interface, a HSCSD connection comprises n traffic channels (TCH). All n channels use the same hopping sequences, if frequency hopping is applied. Also the training sequence assigned, is the same for all n traffic channels. However, each traffic channel is assigned an independent SACCH signaling channel, allowing for independent signal level and quality reporting on each timeslot, in turn, allowing for individual power control on each timeslot. This is done, since the interference level can be different on each timeslot such that different power levels might be required in order to secure the required signal quality on every single timeslot. The HSCSD connection has just one FACCH control channel assigned for management purposes. For all HSCSD channels, the same channel coding is applied, however, for nontransparent services, different channel codings can be used for uplink and downlink, respectively. For each HSCSD channel, different ciphering keys are used, derived from K_c . The channel assignment configuration can be either symmetric or asymmetric and channels can be

allocated either on consecutive or on nonconsecutive timeslots of the same carrier, as long as the restrictions of the given multislots class of the MS are considered.

Symmetric connections comprise a bidirectional FACCH as well as co-allocated bidirectional TCH/F and SACCH channels. In contrast, asymmetric connections can have unidirectional and/or bidirectional TCH/f and SACCH channels in addition to the bidirectional FACCH. Unidirectional channels in HSCSD are always downlink channels only. Both symmetric and asymmetric HSCSD channels have a bidirectional main channel, which carries the FACCH.

HSCSD resource allocation and capacity issues

When initiating a HSCSD connection, the user indicates during setup the maximum number of traffic channels, the multislots class of the terminal, acceptable channel codings, and a wanted fixed network user rate. In the case of a nontransparent connection, the wanted air interface user rate is also indicated. The connection requirements can be symmetric or asymmetric, considering uplink and downlink. These parameters are then used by the network to allocate appropriate resources and to setup the demanded HSCSD connection. The minimum channel requirement is always one TCH/F. This means that transparent and nontransparent connections can be established with any number of TCH/F from one up to the specified maximum number. When the user rate requirements cannot be met, the network will give priority to fulfill the air interface user requirement in downlink direction, possibly using an asymmetric configuration. The network can use dynamic resource allocation for nontransparent HSCSD connections, as long as the allocated channel configurations are always in line with the limiting values defined by the MS and with the multislots class of the terminal. In the case of transparent HSCSD connections, dynamic resource allocation is only allowed if the air interface user data rate is kept constant. The network performs the change of channel allocation configurations during the HSCSD connection, by means of resource upgrading and resource downgrading procedures. If indicated by the MS during call setup, the MS may issue a request an upgrade or downgrade of service level anytime during the HSCSD connection.

Obviously, HSCSD can increase the data rate of a single user. This, however, is achieved at the cost of assigning more frequency resources to a single user. This implies that from the point of view of a network provider, resources are more frequently occupied. A single HSCSD user using $n = 4$ TCHs will occupy the resources of four potential voice connections. Therefore, when many users utilize HSCSD at once in a single cell, the blocking performance in that cell will deteriorate. This will influence the radio resource management strategy that a network provider will use: if possible, the provider might try to assign more frequency carriers to cells in areas where the frequent use of HSCSD seems likely, for instance at airports or in business areas of a city. Also, the provider has the option of restricting the number of HSCSD connections in a cell in favor of voice connections.

Another interesting aspect of the HSCSD radio resource management is the handover. When a handover of the HSCSD connection to a neighboring cell becomes necessary, it might not be possible to find n free TCHs on a single frequency channel in the target cell. In this case, two options are possible: either the HSCSD connection is simply dropped or the connection is resumed with a lower number of TCHs. For this purpose, the resource downgrading procedure is

applied, as described above for the case of dynamic resource allocation. With this concept, the probability of a handover being blocked and thus a HSCSD call being dropped can be reduced. Later, a resource upgrading procedure can be applied if an appropriate resource becomes available, and the original channel allocation configuration can be recreated.

9. EDGE

As discussed above in this chapter, HSCSD and GPRS achieve higher data rates because a MS can use several time slots of the same TDMA frame and partly because new coding schemes are employed. The EDGE2 system goes one step further, by improving the spectral efficiency on the physical layer on a single timeslot (Furuskär et al., 1999). Technically speaking, EDGE can be considered mainly as an air interface improvement. However, in effect it is a system concept that is used in order to introduce new bearer services into GSM systems. In this context it is interesting to note that both GPRS and EDGE have also been standardized for the North American cellular network TDMA-136 (GPRS-136 and GPRS- 136HS EDGE). Within the GSM context, EDGE is used to improve the existing data services with a focus on GPRS and HSCSD, which become Enhanced GPRS (EGPRS) and Enhanced Circuit Switched Data (ECSD), respectively, when combined with EDGE technology.

The EDGE concept

A classical GSM system is designed and planned in a worst-case fashion: the radio network planning is carried out such that there is a high probability that all users in the network will experience a minimum signal quality that is sufficient for low error probability with a fixed modulation and error coding scheme. In fact, the main restrictions for the radio network planning come from those users that are located at the cell edges, far away from base stations. So, it can be said that the GSM network is designed for the cell edge users. However, users located closer to the base stations are likely to experience signal quality levels that are much better than required for the standard GSM modulation and coding schemes. To this end, EDGE introduces several additional combinations of modulation and coding schemes, which allow terminals to adapt their data rates to their individual signal quality levels. For this purpose a link adaptation technique is introduced with EDGE, which dynamically chooses a modulation and coding scheme according to the current radio channel conditions.

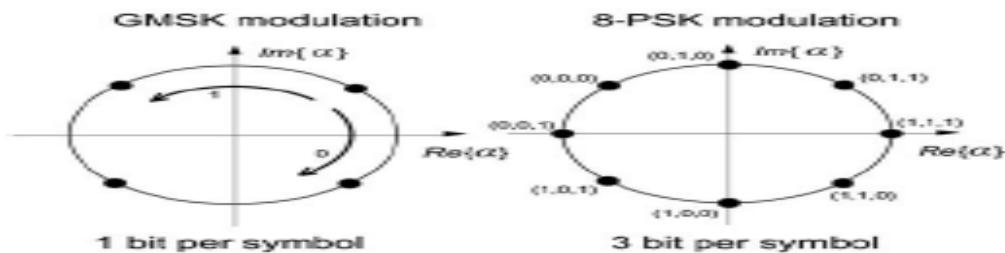
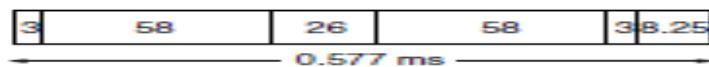


Figure 8.23 Symbol space constellations for GMSK and 8-PSK.



8.24 The EDGE burst has a training sequence of 26 symbols in the middle, 3 symbols at either end, and 8.25 guard symbols at the end. The burst carries 2 > s.

EDGE physical layer, modulation and coding

For EDGE, in addition to the GMSK modulation scheme used in GSM, an 8-Phase Shift Keying (8-PSK) scheme is available, which achieves an approximately three times higher data rate per time slot and hence a higher spectral efficiency. Using GMSK, one data bit d_i is mapped to one symbol a_i (see section 4.2.1); with 8-PSK, three data bits d_i are combined into one symbol a_i and transmitted together. Figure 8.23 shows the symbol constellations in the complex plane and the associated bit sequences. As opposed to GMSK, 8-PSK does not have a constant envelope and therefore puts higher requirements on new transceivers in BTSs and MSs.

In order to achieve compatibility with the GSM system, most EDGE physical layer parameters are the same as in GSM: the carrier spacing is 200 kHz and the TDMA frame structure remains unchanged. The burst format for the 8-PSK modulated transmission is also similar to the standard GSM frames (Figure 8.24): A burst comprises a 26-symbol training sequence as midamble, three tail symbols at the beginning and end of the frame, and 8.25 guard symbols at the end. Before and after the midamble, the frame carries 58 data symbols, each symbol representing three bits according to the 8-PSK modulation applied. Several different coding schemes with different code rates can be combined with the two different modulation schemes available (Tables 8.9 and 8.10) (Molkdar et al., 2002). In order to select the optimal scheme, EDGE applies a link adaptation based on the current channel quality.

In addition to the additional modulation and coding schemes, EDGE also introduces a code combining technique called incremental redundancy, also known as Hybrid Automatic Repeat Request (HARQ). In the incremental redundancy mode, the first RLC data block can be transmitted with some or even no redundancy added. If the data block cannot be decoded correctly, in the next retransmission, more redundancy is sent, applying a different puncturing scheme on the same RLC block. The erroneous blocks are stored, such that they can be combined with each retransmission required, until the RLC can be decoded correctly. Also, the RLC/MAC layer has been enhanced with appropriate RLC/burst mapping functionality. The improved RLC/MAC layer allows for resegmentation to enable for retransmission with different coding schemes, independent coding of RLC/MAC headers and an increased ACK/NACK window size.

As for the segmentation, each LLC PDU is broken into 20 ms RLC data blocks, to match the TDMA burst structure of GSM. Depending on the selected coding scheme by the link adaptation, the number of bits that fit into a RLC data block varies. The required RLC/MAC headers are appended to the user data before transmission. The content of the RLC/MAC headers includes a Block Sequence Number (BSN), and a Coding and Puncturing Scheme (CPS) indicator, that is required for the code combining process in the HARQ procedure. Then, check sequences for the user data (BCS) and for the header (HCS) are added to form a RLC radio block. The RLC radio block is then passed to the physical layer. There the user data and the header are coded separately and finally mapped to two or four TDMA bursts, depending on the coding scheme. The modulation and coding scheme can be changed for each RLC block, i.e. typically every four TDMA bursts. However, the modifications will be based on channel quality monitoring and changes will, in fact, be more seldom, depending on the channel measurement and reporting frequency.

EDGE: effects on the GSM system architecture

It is mainly the improved data rate of EDGE that imposes new requirements also for the GSM/GPRS network architecture: the main bottleneck for EDGE in GSM is the Abis interface between BTS and BSC. In standard GSM, this interface supports only 16 kbit/s per traffic channel. However, EDGE can support close to 64 kbit/s for one traffic channel. Therefore, EDGE requires the allocation of multiple Abis slots to one traffic channel. In future 3G network architectures, this requirement might also be fulfilled by applying ATM or IP-based solutions.

The A interface between BSC and MSC can handle 64 kbit/s already in standard GSM and therefore does not present a problem. Other than that, the GSM/GPRS network architecture is not affected by the introduction of EDGE. This is due to the fact that EDGE is foremost just an enhanced air interface technology. Two different types of EDGE mobile terminals are considered in the standard: terminals that are capable of 8-PSK modulation on the downlink only, and terminals which provide 8-PSK capability on both the uplink and downlink. The first class of terminals can benefit from higher data rates on the downlink, which is still considered to be the link of higher importance, due to the popularity of services which are heavy on the downlink data amount, such as browsing and file downloads.

On the other hand, the second class of terminals can support higher bandwidth on both links and of course makes a more efficient use of the scarce frequency resource, even when nonsymmetric services are considered. The necessity to have this capability information in the network has some minor implications on control plane layers: mobility management modifications are related to the introduction of EGPRS capability information of the respective terminal. These include the multislot class as well as the EDGE modulation capabilities (downlink or uplink and downlink) and in addition an 8-PSK power class. Some modifications are also required on the RRM layer for supporting, setting up, and maintaining EGPRS temporary block flows. In addition, signaling to support the radio link control, the link quality control, and measurement procedures are introduced. There is, however, no impact on session management.

10 ECSD and EGPRS

EDGE can be used to improve both GPRS and HSCSD data services in GSM systems. In combination with EDGE, GPRS becomes EGPRS. Likewise, HSCSD becomes ECSD when enhanced by EDGE. The different achievable data rates per timeslot for the different combinations of modulation and coding schemes are summarized in Tables 8.9 and 8.10, respectively. The table shows which combinations of modulation scheme and code rate can be applied in EGPRS and ECSD, respectively. The highest data rate is achieved in EGPRS when the 8-PSK modulation scheme is combined with a code rate of one. This, however, demands extremely good channel conditions, since the code rate of one implies that there is in fact no error protection. So it will, in fact, rarely be possible to apply this mode, unless the respective application can tolerate packet losses to a good extent. The data rates that can be achieved with EGPRS and ECSD will be obviously multiples of the values in the tables when several timeslots are combined.

The link adaptation for both EGPRS and ECSD requires appropriate signaling. For this purpose, existing signaling mechanisms are applied, specifically the RR channel mode modify procedure, the assignment procedure, and the intra-cell handover procedure.

Table 8.9 EGPRS transmission modes.

Channel name	Code rate	Modulation	Bitrate per timeslot (kbit/s)
MCS-1	0.53	GMSK	8.8
MCS-2	0.66	GMSK	11.2
MCS-3	0.85	GMSK	14.8
MCS-4	1	GMSK	17.6
MCS-5	0.37	8-PSK	22.4
MCS-6	0.49	8-PSK	29.6
MCS-7	0.76	8-PSK	44.8
MCS-8	0.92	8-PSK	54.4
MCS-9	1	8-PSK	59.2

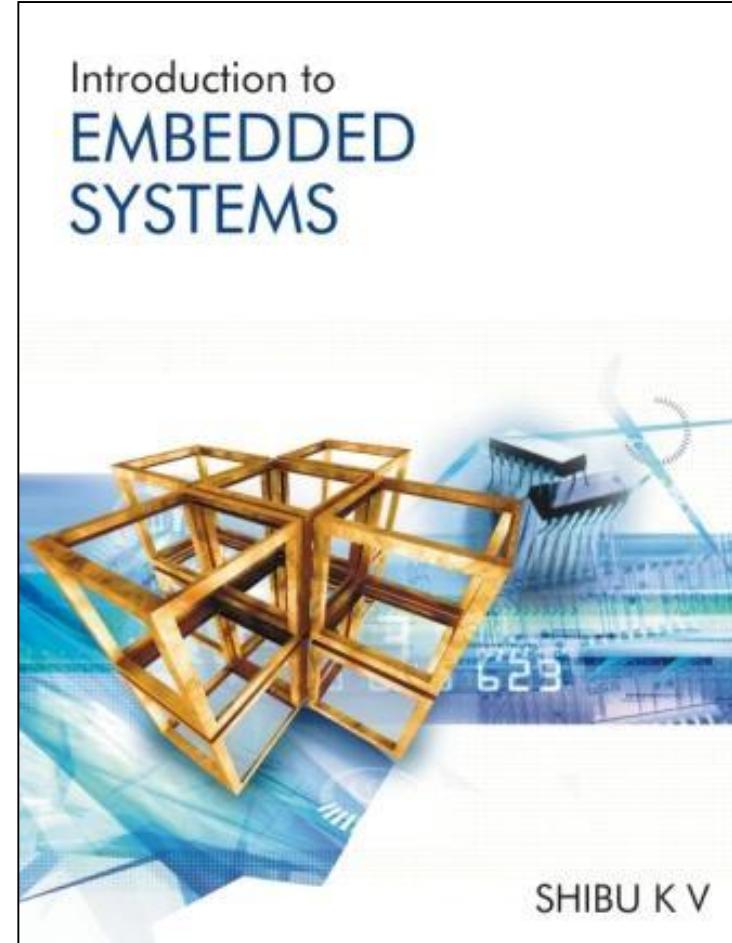
Table 8.10 ECSD transmission modes.

Channel name	Code rate	Modulation	Bitrate per timeslot (kbit/s)
TCH/F2.4	0.16	GMSK	3.6
TCH/F4.8	0.26	GMSK	6
TCH/F9.6	0.53	GMSK	12
TCH/F14.4	0.64	GMSK	14.5
ECSD TCS-1	0.42	8-PSK	29
ECSD TCS-2	0.46	8-PSK	32
ECSD TCS-3	0.56	8-PSK	38.8

Embedded System Design

VI Semester

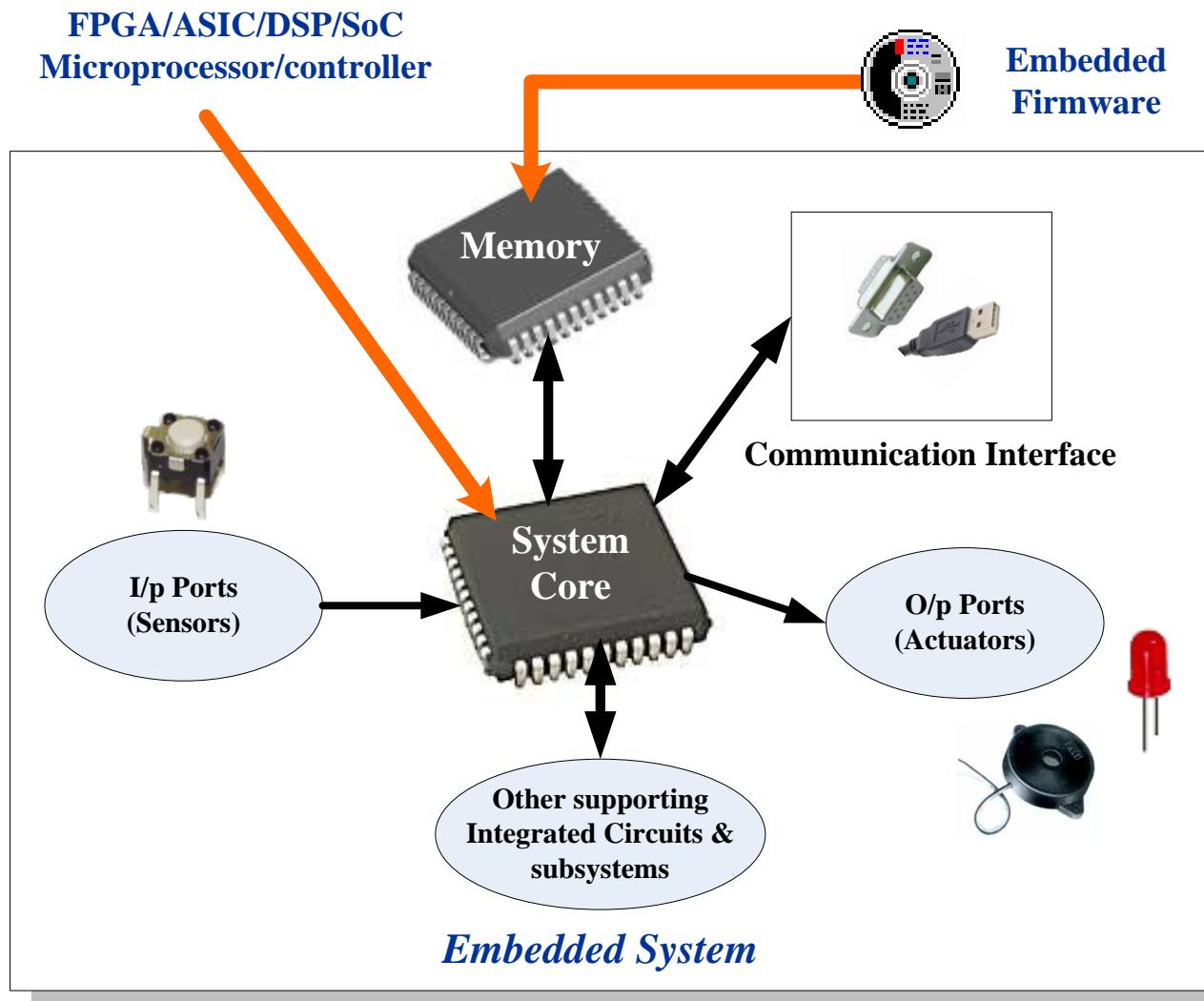
**Mr. Nagesh H B
Harish L
ACSCE**



Module 4

The Typical Embedded System

The Typical Embedded System



The Typical Embedded System

The Core of the Embedded Systems

The core of the embedded system falls into any one of the following categories.

- General Purpose and Domain Specific Processors**
 - Microprocessors
 - Microcontrollers
 - Digital Signal Processors
- Programmable Logic Devices (PLDs)**
- Application Specific Integrated Circuits (ASICs)**
- Commercial off the shelf Components (COTS)**

The Typical Embedded System

Microprocessor

- ✓ A silicon chip representing a Central Processing Unit (CPU), which is capable of performing arithmetic as well as logical operations according to a pre-defined set of Instructions, which is specific to the manufacturer
- ✓ In general the CPU contains the Arithmetic and Logic Unit (ALU), Control Unit and Working registers
- ✓ Microprocessor is a dependant unit and it requires the combination of other hardware like Memory, Timer Unit, and Interrupt Controller etc for proper functioning.
- ✓ Intel claims the credit for developing the first Microprocessor unit Intel 4004, a 4 bit processor which was released in Nov 1971

The Typical Embedded System

General Purpose Processor (GPP) Vs Application Specific Instruction Set Processor (ASIP)

- ✓ General Purpose Processor or GPP is a processor designed for general computational tasks
- ✓ GPPs are produced in large volumes and targeting the general market. Due to the high volume production, the per unit cost for a chip is low compared to ASIC or other specific ICs
- ✓ A typical general purpose processor contains an Arithmetic and Logic Unit (ALU) and Control Unit (CU)
- ✓ Application Specific Instruction Set processors (ASIPs) are processors with architecture and instruction set optimized to specific domain/application requirements like Network processing, Automotive, Telecom, media applications, digital signal processing, control applications etc.
- ✓ ASIPs fill the architectural spectrum between General Purpose Processors and Application Specific Integrated Circuits (ASICs)
- ✓ The need for an ASIP arises when the traditional general purpose processor are unable to meet the increasing application needs
- ✓ Some Microcontrollers (like Automotive AVR, USB AVR from Atmel), System on Chips, Digital Signal Processors etc are examples of Application Specific Instruction Set Processors (ASIPs)
- ✓ ASIPs incorporate a processor and on-chip peripherals, demanded by the application requirement, program and data memory

The Typical Embedded System

Microcontroller

- ✓ A highly integrated silicon chip containing a CPU, scratch pad RAM, Special and General purpose Register Arrays, On Chip ROM/FLASH memory for program storage, Timer and Interrupt control units and dedicated I/O ports
- ✓ Microcontrollers can be considered as a super set of Microprocessors
- ✓ Microcontroller can be general purpose (like Intel 8051, designed for generic applications and domains) or application specific (Like Automotive AVR from Atmel Corporation. Designed specifically for automotive applications)
- ✓ Since a microcontroller contains all the necessary functional blocks for independent working, they found greater place in the embedded domain in place of microprocessors
- ✓ Microcontrollers are cheap, cost effective and are readily available in the market
- ✓ Texas Instruments TMS 1000 is considered as the world's first microcontroller

The Typical Embedded System

Microprocessor Vs Microcontroller

Micropocessor	Microcontroller
A silicon chip representing a Central Processing Unit (CPU), which is capable of performing arithmetic as well as logical operations according to a pre-defined set of Instructions	A microcontroller is a highly integrated chip that contains a CPU, scratch pad RAM, Special and General purpose Register Arrays, On Chip ROM/FLASH memory for program storage, Timer and Interrupt control units and dedicated I/O ports
It is a dependent unit. It requires the combination of other chips like Timers, Program and data memory chips, Interrupt controllers etc for functioning	It is a self contained unit and it doesn't require external Interrupt Controller, Timer, UART etc for its functioning
Most of the time general purpose in design and operation	Mostly application oriented or domain specific
Doesn't contain a built in I/O port. The I/O Port functionality needs to be implemented with the help of external Programmable Peripheral Interface Chips like 8255	Most of the processors contain multiple built-in I/O ports which can be operated as a single 8 or 16 or 32 bit Port or as individual port pins
Targeted for high end market where performance is important	Targeted for embedded market where performance is not so critical (At present this demarcation is invalid)
Limited power saving options compared to microcontrollers	Includes lot of power saving features

The Typical Embedded System

Digital Signal Processors (DSPs)

- ✓ Powerful special purpose 8/16/32 bit microprocessors designed specifically to meet the computational demands and power constraints of today's embedded audio, video, and communications applications
- ✓ Digital Signal Processors are 2 to 3 times faster than the general purpose microprocessors in signal processing applications
- ✓ DSPs implement algorithms in hardware which speeds up the execution whereas general purpose processors implement the algorithm in firmware and the speed of execution depends primarily on the clock for the processors
- ✓ DSP can be viewed as a microchip designed for performing high speed computational operations for ‘addition’, ‘subtraction’, ‘multiplication’ and ‘division’
- ✓ A typical Digital Signal Processor incorporates the following key units
 - ✓ Program Memory
 - ✓ Data Memory
 - ✓ Computational Engine
 - ✓ I/O Unit
- ✓ Audio video signal processing, telecommunication and multimedia applications are typical examples where DSP is employed

The Typical Embedded System

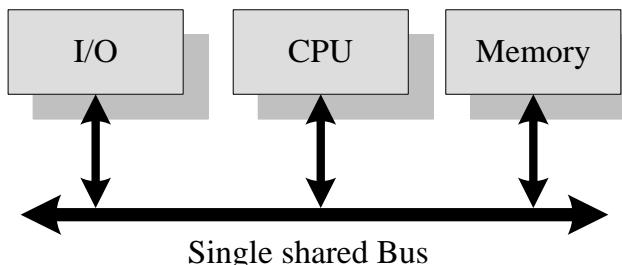
RISC V/s CISC Processors/Controllers

RISC	CISC
Lesser no. of instructions	Greater no. of Instructions
Instruction Pipelining and increased execution speed	Generally no instruction pipelining feature
Orthogonal Instruction Set (Allows each instruction to operate on any register and use any addressing mode)	Non Orthogonal Instruction Set (All instructions are not allowed to operate on any register and use any addressing mode. It is instruction specific)
Operations are performed on registers only, the only memory operations are load and store	Operations are performed on registers or memory depending on the instruction
Large number of registers are available	Limited no. of general purpose registers
Programmer needs to write more code to execute a task since the instructions are simpler ones	Instructions are like macros in C language. A programmer can achieve the desired functionality with a single instruction which in turn provides the effect of using more simpler single instructions in RISC
Single, Fixed length Instructions	Variable length Instructions
Less Silicon usage and pin count	More silicon usage since more additional decoder logic is required to implement the complex instruction decoding.
With Harvard Architecture	Can be Harvard or Von-Neumann Architecture

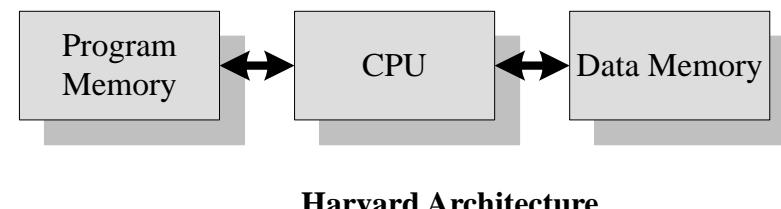
The Typical Embedded System

Harvard V/s Von-Neumann Processor/Controller Architecture

- ✓ The terms Harvard and Von-Neumann refers to the processor architecture design.
- ✓ Microprocessors/controllers based on the **Von-Neumann** architecture shares a single common bus for fetching both instructions and data. Program instructions and data are stored in a common main memory
- ✓ Microprocessors/controllers based on the **Harvard** architecture will have separate data bus and instruction bus. This allows the data transfer and program fetching to occur simultaneously on both buses
- ✓ With Harvard architecture, the data memory can be read and written while the program memory is being accessed. These separated data memory and code memory buses allow one instruction to execute while the next instruction is fetched (“Pre-fetching”)



Von-Neumann Architecture



Harvard Architecture

The Typical Embedded System

Harvard V/s Von-Neumann Processor/Controller Architecture

Harvard Architecture	Von-Neumann Architecture
Separate buses for Instruction and Data fetching	Single shared bus for Instruction and Data fetching
Easier to Pipeline, so high performance can be achieved	Low performance Compared to Harvard Architecture
Comparatively high cost	Cheaper
No memory alignment problems	Allows self modifying codes [†]
Since data memory and program memory are stored physically in different locations, no chances for accidental corruption of program memory	Since data memory and program memory are stored physically in same chip, chances for accidental corruption of program memory

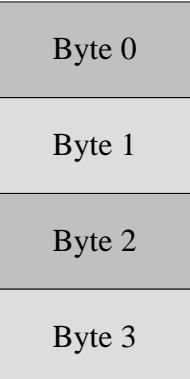
The Typical Embedded System

Big-endian V/s Little-endian processors

- ✓ Endianness specifies the order in which the data is stored in the memory by processor operations in a multi byte system (Processors whose word size is greater than one byte). Suppose the word length is two byte then data can be stored in memory in two different ways
 - ✓ Higher order of data byte at the higher memory and lower order of data byte at location just below the higher memory
 - ✓ Lower order of data byte at the higher memory and higher order of data byte at location just below the higher memory
- ✓ **Little-endian** means the lower-order byte of the data is stored in memory at the lowest address, and the higher-order byte at the highest address. (The little end comes first)
- ✓ **Big-endian** means the higher-order byte of the data is stored in memory at the lowest address, and the lower-order byte at the highest address. (The big end comes first.)

The Typical Embedded System

Big-endian V/s Little-endian processors

Base Address + 0	Byte 0		0x20000 (Base Address)
Base Address + 1	Byte 1		0x20001 (Base Address + 1)
Base Address + 2	Byte 2		0x20002 (Base Address + 2)
Base Address + 3	Byte 3		0x20003 (Base Address + 3)

Little-endian Operation

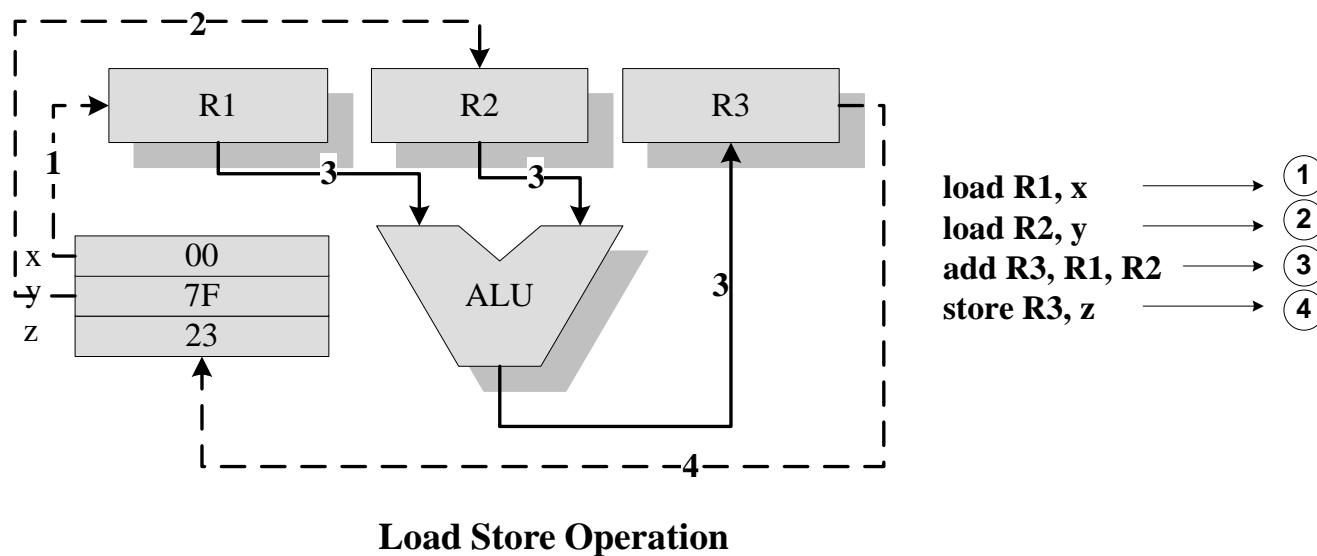
Base Address + 0	Byte 3		0x20000 (Base Address)
Base Address + 1	Byte 2		0x20001 (Base Address + 1)
Base Address + 2	Byte 1		0x20002 (Base Address + 2)
Base Address + 3	Byte 0		0x20003 (Base Address + 3)

Big-endian Operation

The Typical Embedded System

Load Store Operation & Instruction Pipelining

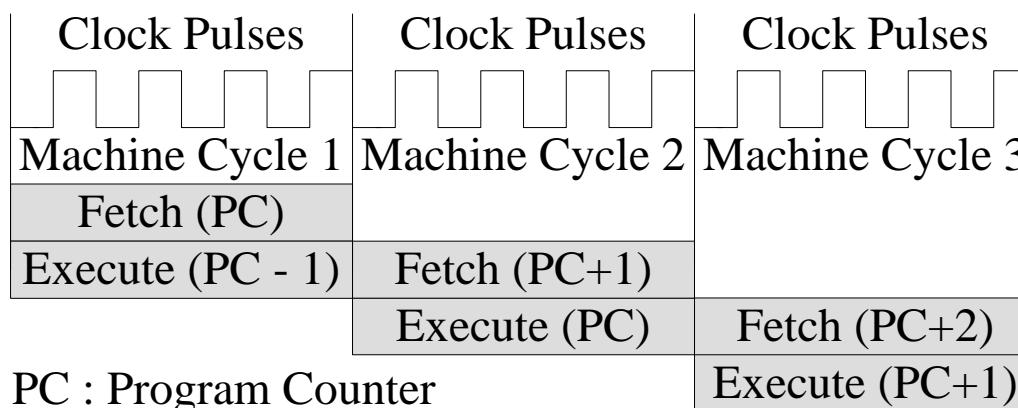
The RISC processor instruction set is orthogonal and it operates on registers. The memory access related operations are performed by the special instructions *load* and *store*. If the operand is specified as memory location, the content of it is loaded to a register using the *load* instruction. The instruction *store* stores data from a specified register to a specified memory location



The Typical Embedded System

Instruction Pipelining

- ✓ The conventional instruction execution by the processor follows the fetch-decode-execute sequence
 - ✓ The ‘fetch’ part fetches the instruction from program memory or code memory and the decode part decodes the instruction to generate the necessary control signals
 - ✓ The execute stage reads the operands, perform ALU operations and stores the result. In conventional program execution, the fetch and decode operations are performed in sequence
 - ✓ During the decode operation the memory address bus is available and if it possible to effectively utilize it for an instruction fetch, the processing speed can be increased
 - ✓ In its simplest form instruction pipelining refers to the overlapped execution of instructions



The Single stage pipelining concept

The Typical Embedded System

Application Specific Integrated Circuit (ASIC)

- ✓ A microchip designed to perform a specific or unique application. It is used as replacement to conventional general purpose logic chips.
- ✓ ASIC integrates several functions into a single chip and thereby reduces the system development cost
- ✓ Most of the ASICs are proprietary products. As a single chip, ASIC consumes very small area in the total system and thereby helps in the design of smaller systems with high capabilities/functionalities.
- ✓ ASICs can be pre-fabricated for a special application or it can be custom fabricated by using the components from a re-usable '*building block*' library of components for a particular customer application
- ✓ Fabrication of ASICs requires a non-refundable initial investment (Non Recurring Engineering (NRE) charges) for the process technology and configuration expenses
- ✓ If the Non-Recurring Engineering Charges (NRE) is born by a third party and the Application Specific Integrated Circuit (ASIC) is made openly available in the market, the ASIC is referred as Application Specific Standard Product (ASSP)

The Typical Embedded System

Programmable Logic Devices (PLDs)

- ✓ Logic devices provide specific functions, including device-to-device interfacing, data communication, signal processing, data display, timing and control operations, and almost every other function a system must perform.
- ✓ Logic devices can be classified into two broad categories - Fixed and Programmable. The circuits in a fixed logic device are permanent, they perform one function or set of functions - once manufactured, they cannot be changed
- ✓ Programmable logic devices (PLDs) offer customers a wide range of logic capacity, features, speed, and voltage characteristics - and these devices can be re-configured to perform any number of functions at any time
- ✓ Designers can use inexpensive software tools to quickly develop, simulate, and test their logic designs in PLD based design. The design can be quickly programmed into a device, and immediately tested in a live circuit
- ✓ PLDs are based on re-writable memory technology and the device is reprogrammed to change the design

The Typical Embedded System

Programmable Logic Devices (PLDs) – CPLDs and FPGA

- ✓ Field Programmable Gate Arrays (FPGAs) and Complex Programmable Logic Devices (CPLDs) are the two major types of programmable logic devices
- ✓ FPGAs offer the highest amount of logic density, the most features, and the highest performance.
- ✓ These advanced FPGA devices also offer features such as built-in hardwired processors (such as the IBM Power PC), substantial amounts of memory, clock management systems, and support for many of the latest, very fast device-to-device signaling technologies
- ✓ FPGAs are used in a wide variety of applications ranging from data processing and storage, to instrumentation, telecommunications, and digital signal processing
- ✓ CPLDs, by contrast, offer much smaller amounts of logic - up to about 10,000 gates
- ✓ CPLDs offer very predictable timing characteristics and are therefore ideal for critical control applications
- ✓ CPLDs such as the Xilinx **CoolRunner** series also require extremely low amounts of power and are very inexpensive, making them ideal for cost-sensitive, battery-operated, portable applications such as mobile phones and digital handheld assistants

The Typical Embedded System

Commercial off the Shelf Component (COTS)

- ✓ A Commercial off-the-shelf (COTS) product is one which is used ‘as-is’
- ✓ COTS products are designed in such a way to provide easy integration and interoperability with existing system components
- ✓ Typical examples for the COTS hardware unit are Remote Controlled Toy Car control unit including the RF Circuitry part, High performance, high frequency microwave electronics (2 to 200 GHz), High bandwidth analog-to-digital converters, Devices and components for operation at very high temperatures, Electro-optic IR imaging arrays, UV/IR Detectors etc
- ✓ A COTS component in turn contains a General Purpose Processor (GPP) or Application Specific Instruction Set Processor (ASIP) or Application Specific Integrated Chip (ASIC)/Application Specific Standard Product (ASSP) or Programmable Logic Device (PLD)
- ✓ The major advantage of using COTS is that they are readily available in the market, cheap and a developer can cut down his/her development time to a great extend

The Typical Embedded System

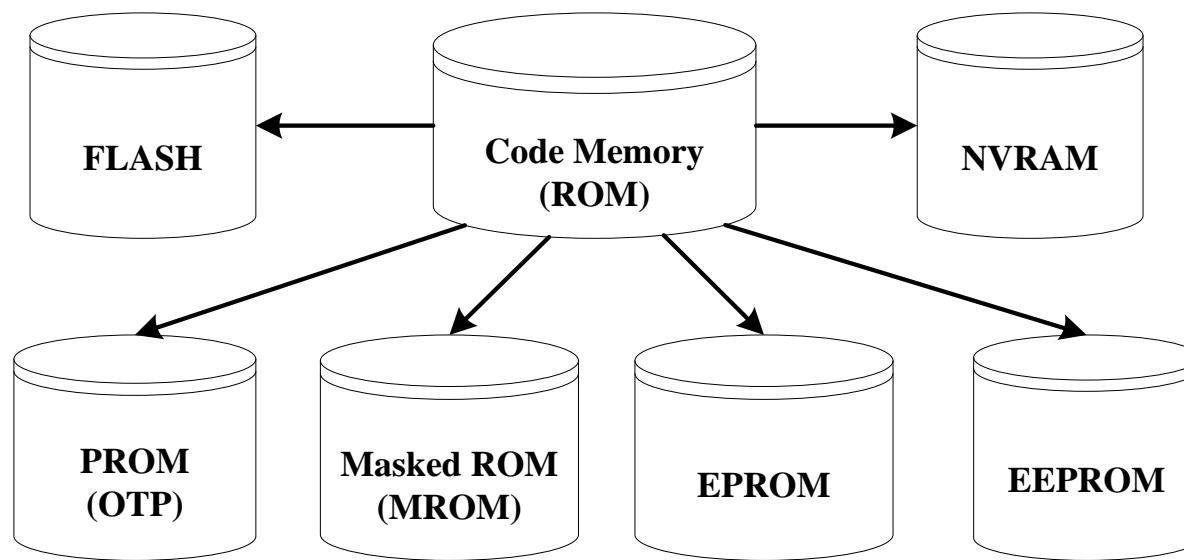
Memory

- ✓ Memory is an important part of an embedded system. The memory used in embedded system can be either Program Storage Memory (ROM) or Data memory (RAM)
- ✓ Certain Embedded processors/controllers contain built in program memory and data memory and this memory is known as on-chip memory

The Typical Embedded System

Memory – Program Storage Memory

- ✓ Stores the program instructions
- ✓ Retains its contents even after the power to it is turned off. It is generally known as Non volatile storage memory
- ✓ Depending on the fabrication, erasing and programming techniques they are classified into



The Typical Embedded System

Memory – Program Storage Memory – Masked ROM (MROM)

- ✓ One-time programmable memory. Uses hardwired technology for storing data. The device is factory programmed by masking and metallization process according to the data provided by the end user
- ✓ The primary advantage of MROM is low cost for high volume production. They are the least expensive type of solid state memory
- ✓ Different mechanisms are used for the masking process of the ROM, like
 - ✓ Creation of an enhancement or depletion mode transistor through channel implant
 - ✓ By creating the memory cell either using a standard transistor or a high threshold transistor. In the high threshold mode, the supply voltage required to turn ON the transistor is above the normal ROM IC operating voltage. This ensures that the transistor is always off and the memory cell stores always logic 0.
- ✓ The limitation with MROM based firmware storage is the inability to modify the device firmware against firmware upgrades. Since the MROM is permanent in bit storage, it is not possible to alter the bit information

The Typical Embedded System

Memory – Program Storage Memory – Programmable Read Only Memory (PROM) / (OTP)

- ✓ Unlike MROM it is not pre-programmed by the manufacturer
- ✓ PROM/OTP has *nichrome* or *polysilicon* wires arranged in a matrix, these wires can be functionally viewed as fuses
- ✓ It is programmed by a PROM programmer which selectively burns the fuses according to the bit pattern to be stored
- ✓ Fuses which are not blown/burned represents a logic “1” whereas fuses which are blown/burned represents a logic “0”. The default state is logic “1”
- ✓ OTP is widely used for commercial production of embedded systems whose proto-typed versions are proven and the code is finalized
- ✓ It is a low cost solution for commercial production. OTPs cannot be reprogrammed

The Typical Embedded System

Memory – Program Storage Memory – Erasable Programmable Read Only Memory (EPROM)

- ✓ Erasable Programmable Read Only (EPROM) memory gives the flexibility to re-program the same chip
- ✓ EPROM stores the bit information by charging the floating gate of an FET
- ✓ Bit information is stored by using an EPROM Programmer, which applies high voltage to charge the floating gate
- ✓ EPROM contains a quartz crystal window for erasing the stored information. If the window is exposed to Ultra violet rays for a fixed duration, the entire memory will be erased
- ✓ Even though the EPROM chip is flexible in terms of re-programmability, it needs to be taken out of the circuit board and needs to be put in a UV eraser device for 20 to 30 minutes

The Typical Embedded System

Memory – Program Storage Memory – Electrically Erasable Programmable Read Only Memory (EEPROM)

- ✓ Erasable Programmable Read Only (EPROM) memory gives the flexibility to re-program the same chip using electrical signals
- ✓ The information contained in the EEPROM memory can be altered by using electrical signals at the register/Byte level
- ✓ They can be erased and reprogrammed within the circuit
- ✓ These chips include a chip erase mode and in this mode they can be erased in a few milliseconds
- ✓ It provides greater flexibility for system design
- ✓ The only limitation is their capacity is limited when compared with the standard ROM (A few kilobytes).

The Typical Embedded System

Memory – Program Storage Memory – FLASH

- ✓ FLASH memory is a variation of EEPROM technology
- ✓ It combines the re-programmability of EEPROM and the high capacity of standard ROMs
- ✓ FLASH memory is organized as sectors (blocks) or pages
- ✓ FLASH memory stores information in an array of floating gate MOSFET transistors
- ✓ The erasing of memory can be done at sector level or page level without affecting the other sectors or pages
- ✓ Each sector/page should be erased before re-programming

The Typical Embedded System

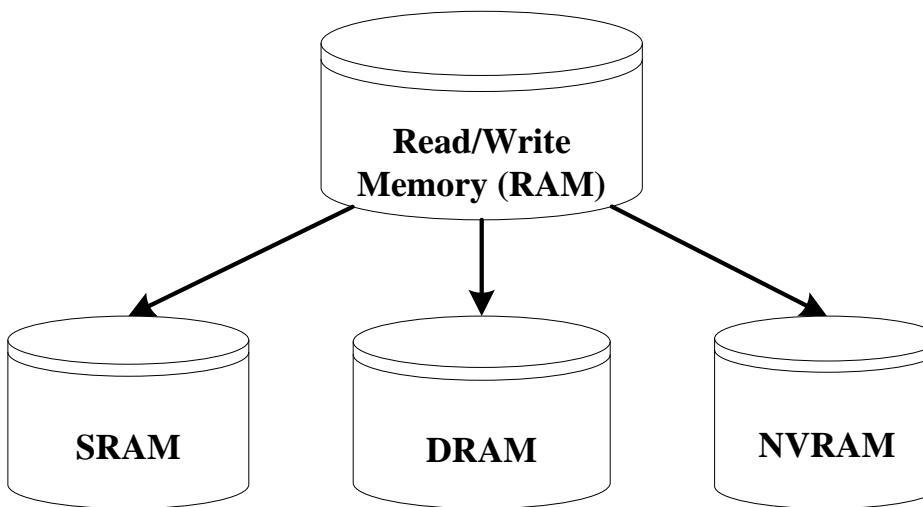
Memory – RAM – Non Volatile RAM (NVRAM)

- ✓ Random access memory with battery backup
- ✓ It contains Static RAM based memory and a minute battery for providing supply to the memory in the absence of external power supply
- ✓ The memory and battery are packed together in a single package
- ✓ NVRAM is used for the non volatile storage of results of operations or for setting up of flags etc
- ✓ The life span of NVRAM is expected to be around 10 years
- ✓ DS1744 from Maxim/Dallas is an example for 32KB NVRAM

The Typical Embedded System

Memory – Read-Write Memory/Random Access Memory (RAM)

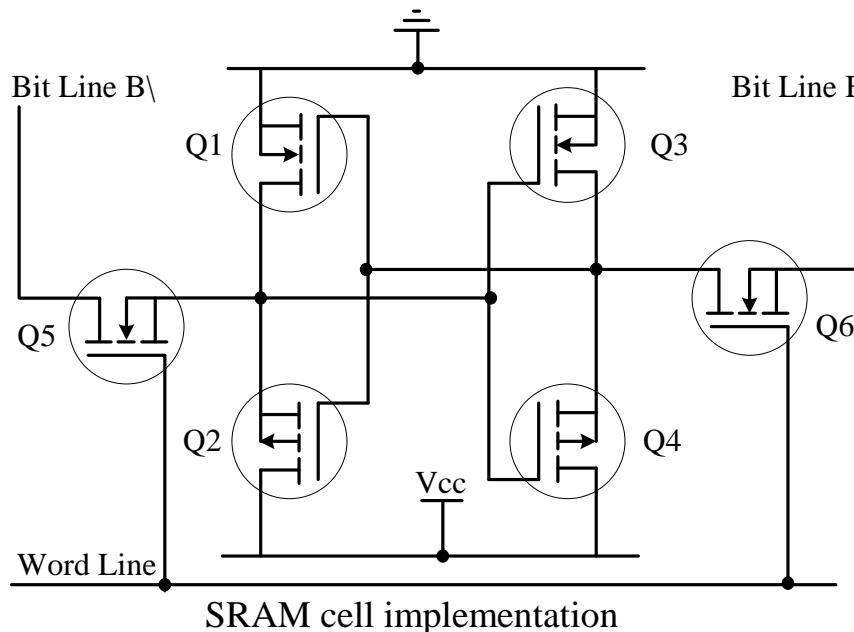
- ✓ RAM is the data memory or working memory of the controller/processor
- ✓ RAM is volatile, meaning when the power is turned off, all the contents are destroyed
- ✓ RAM is a direct access memory, meaning we can access the desired memory location directly without the need for traversing through the entire memory locations to reach the desired memory position (i.e. Random Access of memory location)



The Typical Embedded System

Memory – RAM – Static RAM (SRAM)

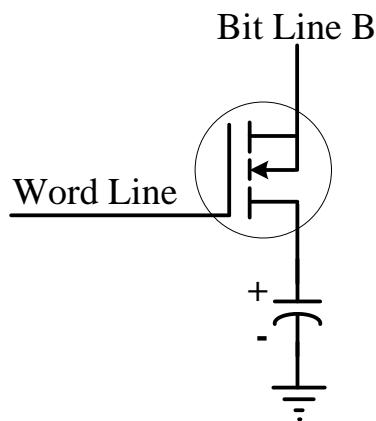
- ✓ Static RAM stores data in the form of Voltage. They are made up of flip-flops
- ✓ In typical implementation, an SRAM cell (bit) is realized using 6 transistors (or 6 MOSFETs). Four of the transistors are used for building the latch (flip-flop) part of the memory cell and 2 for controlling the access.
- ✓ Static RAM is the fastest form of RAM available. SRAM is fast in operation due to its resistive networking and switching capabilities



The Typical Embedded System

Memory – RAM – Dynamic RAM (DRAM)

- ✓ Dynamic RAM stores data in the form of charge. They are made up of MOS transistor gates
- ✓ The advantages of DRAM are its high density and low cost compared to SRAM
- ✓ The disadvantage is that since the information is stored as charge it gets leaked off with time and to prevent this they need to be refreshed periodically
- ✓ Special circuits called DRAM controllers are used for the refreshing operation. The refresh operation is done periodically in milliseconds interval



DRAM cell implementation

The Typical Embedded System

Memory – RAM – SRAM Vs DRAM)

SRAM Cell	DRAM Cell
Made up of 6 CMOS transistors (MOSFET)	Made up of a MOSFET and a capacitor
Doesn't Require refreshing	Requires refreshing
Low capacity (Less dense)	High Capacity (Highly dense)
More expensive	Less Expensive
Fast in operation. Typical access time is 10ns	Slow in operation due to refresh requirements. Typical access time is 60ns. Write operation is faster than read operation.

The Typical Embedded System

Sensors & Actuators

Sensor:

A transducer device which converts energy from one form to another for any measurement or control purpose. Sensors acts as input device

Eg. Hall Effect Sensor which measures the distance between the cushion and magnet in the Smart Running shoes from adidas

Actuator:

A form of transducer device (mechanical or electrical) which converts signals to corresponding physical action (motion). Actuator acts as an output device

Eg. Micro motor actuator which adjusts the position of the cushioning element in the Smart Running shoes from adidas

Introduction to Embedded System

‘Smart’ running shoes from Adidas – The Innovative bonding of Life Style with Embedded Technology

- ✓ Shoe developed by Adidas, which constantly adapts its shock-absorbing characteristics to customize its value to the individual runner, depending on running style, pace, body weight, and running surface
- ✓ It contains sensors, actuators and a microprocessor unit which runs the algorithm for adapting the shock-absorbing characteristics of the shoe
- ✓ A ‘Hall effect sensor’ placed at the top of the “cushioning element” senses the compression and passes it to the Microprocessor
- ✓ A micro motor actuator controls the cushioning as per the commands from the MPU, based on the compression sensed by the ‘Hall effect sensor’

What an innovative bonding of Embedded Technology with Real life needs !!!😊



Electronics-enabled “Smart” running shoes from Adidas

Photo Courtesy of Adidas – Salomon AG
www.adidas.com

The Typical Embedded System

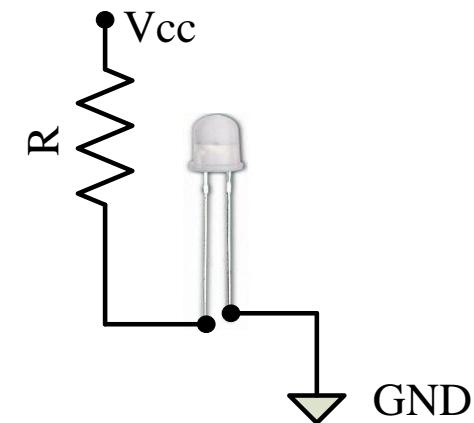
The I/O Subsystem

- ✓ The I/O subsystem of the embedded system facilitates the interaction of the embedded system with external world
- ✓ The interaction happens through the sensors and actuators connected to the Input and output ports respectively of the embedded system
- ✓ The sensors may not be directly interfaced to the Input ports, instead they may be interfaced through signal conditioning and translating systems like ADC, Optocouplers etc

The Typical Embedded System

The I/O Subsystem – I/O Devices - Light Emitting Diode (LED)

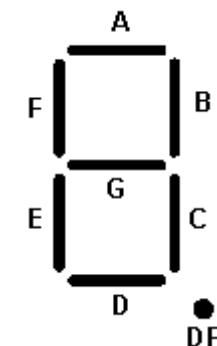
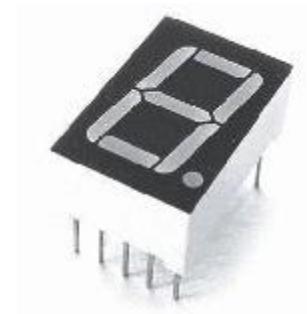
- ✓ Light Emitting Diode (LED) is an output device for visual indication in any embedded system
- ✓ LED can be used as an indicator for the status of various signals or situations. Typical examples are indicating the presence of power conditions like ‘Device ON’, ‘Battery low’ or ‘Charging of battery’ for a battery operated handheld embedded devices
- ✓ LED is a p-n junction diode and it contains an anode and a cathode. For proper functioning of the LED, the anode of it should be connected to +ve terminal of the supply voltage and cathode to the –ve terminal of supply voltage
- ✓ The current flowing through the LED must limited to a value below the maximum current that it can conduct. A resistor is used in series between the power supply and the resistor to limit the current through the LED



The Typical Embedded System

The I/O Subsystem – I/O Devices – 7-Segment LED Display

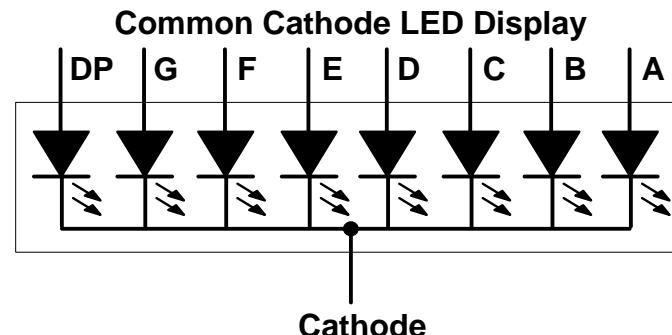
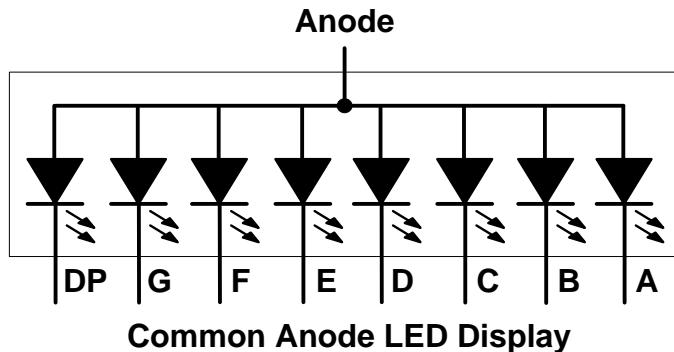
- ✓ The 7 – segment LED display is an output device for displaying alpha numeric characters
- ✓ It contains 8 light-emitting diode (LED) segments arranged in a special form. Out of the 8 LED segments, 7 are used for displaying alpha numeric characters
- ✓ The LED segments are named A to G and the decimal point LED segment is named as DP
- ✓ The LED Segments A to G and DP should be lit accordingly to display numbers and characters
- ✓ The 7 – segment LED displays are available in two different configurations, namely; Common anode and Common cathode
- ✓ In the Common anode configuration, the anodes of the 8 segments are connected commonly whereas in the Common cathode configuration, the 8 LED segments share a common cathode line



The Typical Embedded System

The I/O Subsystem – I/O Devices – 7-Segment LED Display

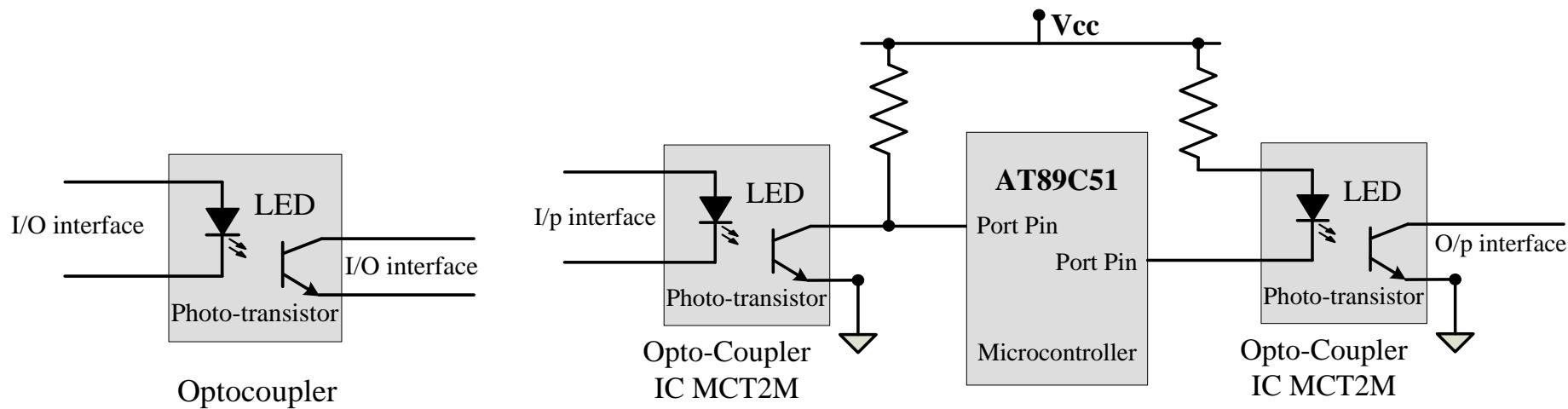
- ✓ Based on the configuration of the 7 – segment LED unit, the LED segment anode or cathode is connected to the Port of the processor/controller in the order ‘A’ segment to the Least significant port Pin and DP segment to the most significant Port Pin.
- ✓ The current flow through each of the LED segments should be limited to the maximum value supported by the LED display unit
- ✓ The typical value for the current falls within the range of 20mA
- ✓ The current through each segment can be limited by connecting a current limiting resistor to the anode or cathode of each segment



The Typical Embedded System

The I/O Subsystem – I/O Devices – Optocoupler

- ✓ Optocoupler is a solid state device to isolate two parts of a circuit. Optocoupler combines an LED and a photo-transistor in a single housing (package)
- ✓ In electronic circuits, optocoupler is used for suppressing interference in data communication, circuit isolation, High voltage separation, simultaneous separation and intensification signal etc
- ✓ Optocouplers can be used in either input circuits or in output circuits

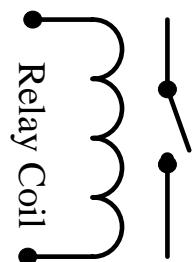


Optocoupler in input and output circuit

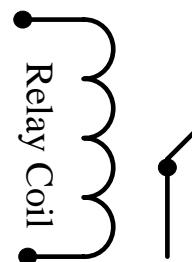
The Typical Embedded System

The I/O Subsystem – I/O Devices – Relay

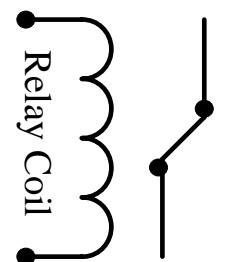
- ✓ An electro mechanical device which acts as dynamic path selectors for signals and power
- ✓ The ‘Relay’ unit contains a relay coil made up of insulated wire on a metal core and a metal armature with one or more contacts.
- ✓ ‘Relay’ works on electromagnetic principle. When a voltage is applied to the relay coil, current flows through the coil, which in turn generates a magnetic field. The magnetic field attracts the armature core and moves the contact point. The movement of the contact point changes the power/signal flow path



Single Pole Single Throw Normally Open



Single Pole Single Throw Normally Closed

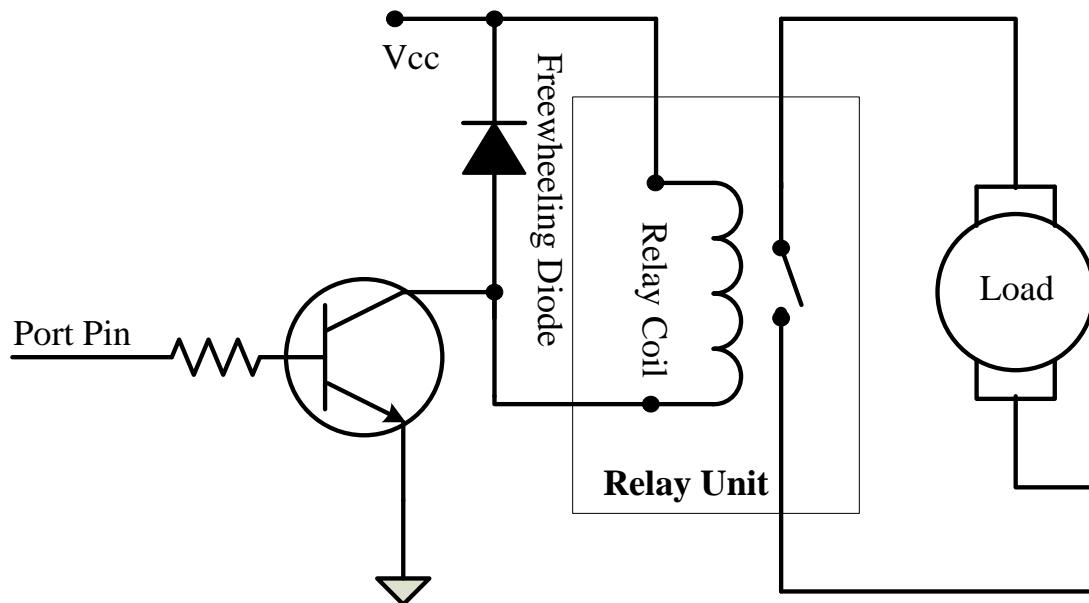


Single Pole Double Throw

The Typical Embedded System

The I/O Subsystem – I/O Devices – Relay Driver Circuit

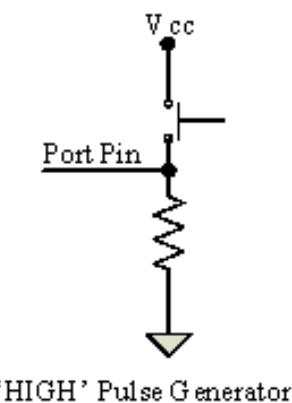
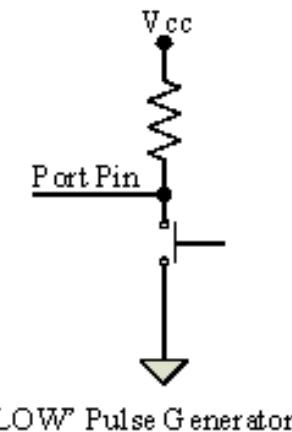
- ✓ The Relay is normally controlled using a relay driver circuit connected to the port pin of the processor/controller
- ✓ A transistor can be used as the relay driver. The transistor can be selected depending on the relay driving current requirements



The Typical Embedded System

The I/O Subsystem – I/O Devices – Push button switch

- ✓ Push Button switch is an input device
- ✓ Push button switch comes in two configurations, namely ‘Push to Make’ and ‘Push to Break’
- ✓ The switch is normally in the open state and it makes a circuit contact when it is pushed or pressed in the ‘Push to Make’ configuration
- ✓ In the ‘Push to Break’ configuration, the switch is normally in the closed state and it breaks the circuit contact when it is pushed or pressed
- ✓ The push button stays in the ‘closed’ (For Push to Make type) or ‘open’ (For Push to Break type) state as long as it is kept in the pushed state and it breaks/makes the circuit connection when it is released
- ✓ Push button is used for generating a momentary pulse



The Typical Embedded System

Communication Interface

- ✓ Communication interface is essential for communicating with various subsystems of the embedded system and with the external world
- ✓ For an embedded product, the communication interface can be viewed in two different perspectives; namely; Device/board level communication interface (Onboard Communication Interface) and Product level communication interface (External Communication Interface)
- ✓ Embedded product is a combination of different types of components (chips/devices) arranged on a Printed Circuit Board (PCB). The communication channel which interconnects the various components within an embedded product is referred as Device/board level communication interface (Onboard Communication Interface)
- ✓ Serial interfaces like I2C, SPI, UART, 1-Wire etc and Parallel bus interface are examples of ‘Onboard Communication Interface’
- ✓ The ‘Product level communication interface’ (External Communication Interface) is responsible for data transfer between the embedded system and other devices or modules
- ✓ The external communication interface can be either wired media or wireless media and it can be a serial or parallel interface. Infrared (IR), Bluetooth (BT), Wireless LAN (Wi-Fi), Radio Frequency waves (RF), GPRS etc are examples for wireless communication interface
- ✓ RS-232C/RS-422/RS 485, USB, Ethernet (TCP-IP), IEEE 1394 port, Parallel port, CF-II Slot, SDIO, PCMCIA etc are examples for wired interfaces

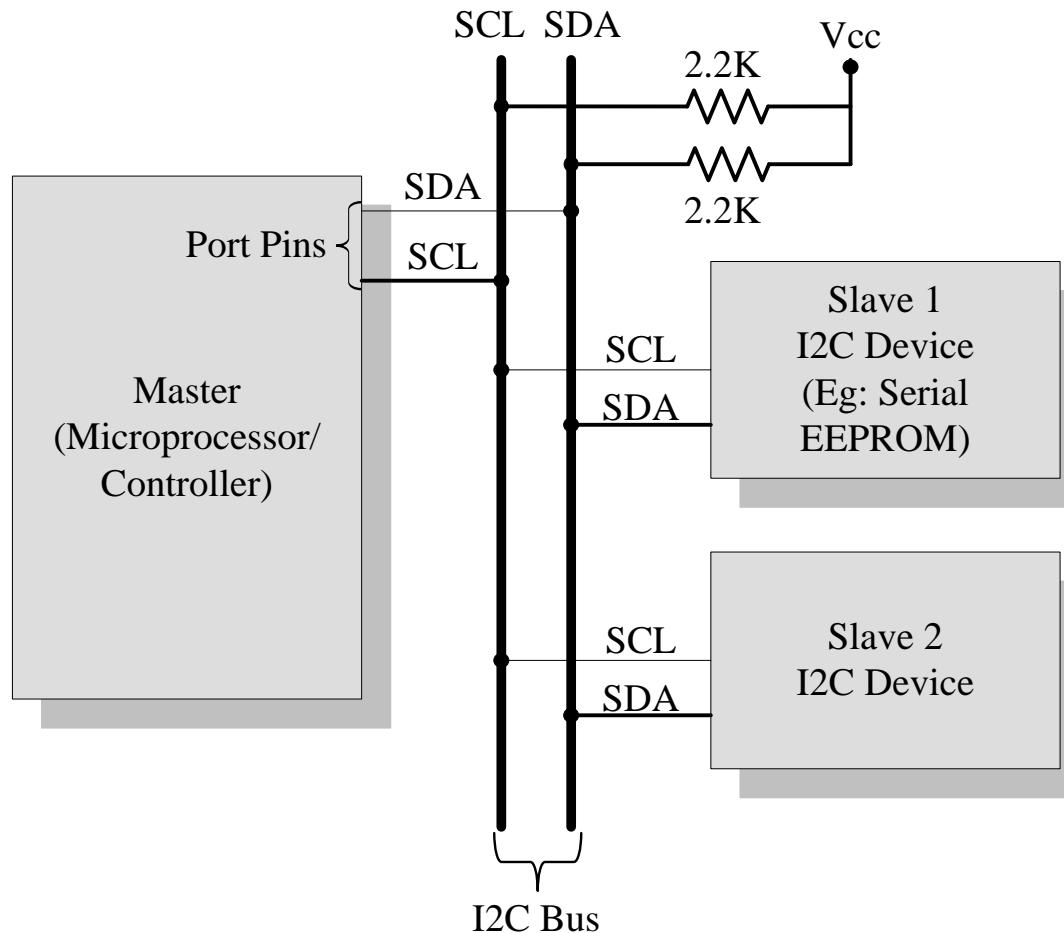
The Typical Embedded System

On-board Communication Interface - I2C

- ✓ Inter Integrated Circuit Bus (I2C - Pronounced 'I square C') is a synchronous bi-directional half duplex (one-directional communication at a given point of time) two wire serial interface bus
- ✓ The concept of I2C bus was developed by 'Philips Semiconductors' in the early 1980's. The original intention of I2C was to provide an easy way of connection between a microprocessor/microcontroller system and the peripheral chips in Television sets
- ✓ The I2C bus is comprised of two bus lines, namely; Serial Clock – SCL and Serial Data – SDA. SCL line is responsible for generating synchronization clock pulses and SDA is responsible for transmitting the serial data across devices.
- ✓ I2C bus is a shared bus system to which many number of I2C devices can be connected. Devices connected to the I2C bus can act as either 'Master' device or 'Slave' device
- ✓ The 'Master' device is responsible for controlling the communication by initiating/terminating data transfer, sending data and generating necessary synchronization clock pulses
- ✓ 'Slave' devices wait for the commands from the master and respond upon receiving the commands
- ✓ 'Master' and 'Slave' devices can act as either transmitter or receiver
- ✓ Regardless whether a master is acting as transmitter or receiver, the synchronization clock signal is generated by the 'Master' device only
- ✓ I2C supports multi masters on the same bus

The Typical Embedded System

On-board Communication Interface - I2C



The Typical Embedded System

On-board Communication Interface - I2C

The sequence of operation for communicating with an I2C slave device is:

1. Master device pulls the clock line (SCL) of the bus to ‘HIGH’
2. Master device pulls the data line (SDA) ‘LOW’, when the SCL line is at logic ‘HIGH’ (This is the ‘Start’ condition for data transfer)
3. Master sends the address (7 bit or 10 bit wide) of the ‘Slave’ device to which it wants to communicate, over the SDA line. Clock pulses are generated at the SCL line for synchronizing the bit reception by the slave device. The MSB of the data is always transmitted first. The data in the bus is valid during the ‘HIGH’ period of the clock signal
4. Master sends the Read or Write bit (Bit value = 1 Read Operation; Bit value = 0 Write Operation) according to the requirement
5. Master waits for the acknowledgement bit from the slave device whose address is sent on the bus along with the Read/Write operation command. Slave devices connected to the bus compares the address received with the address assigned to them
6. The Slave device with the address requested by the master device responds by sending an acknowledge bit (Bit value =1) over the SDA line
7. Upon receiving the acknowledge bit, master sends the 8bit data to the slave device over SDA line, if the requested operation is ‘Write to device’. If the requested operation is ‘Read from device’, the slave device sends data to the master over the SDA line
8. Master waits for the acknowledgement bit from the device upon byte transfer complete for a write operation and sends an acknowledge bit to the slave device for a read operation
9. Master terminates the transfer by pulling the SDA line ‘HIGH’ when the clock line SCL is at logic ‘HIGH’ (Indicating the ‘STOP’ condition)

The Typical Embedded System

On-board Communication Interface – Serial Peripheral Interface (SPI) Bus

The Serial Peripheral Interface Bus (SPI) is a synchronous bi-directional full duplex four wire serial interface bus. The concept of SPI is introduced by Motorola. SPI is a single master multi-slave system. It is possible to have a system where more than one SPI device can be master, provided the condition only one master device is active at any given point of time, is satisfied. SPI requires four signal lines for communication. They are:

Master Out Slave In (MOSI): Signal line carrying the data from master to slave device.
It is also known as Slave Input/Slave Data In (SI/SDI)

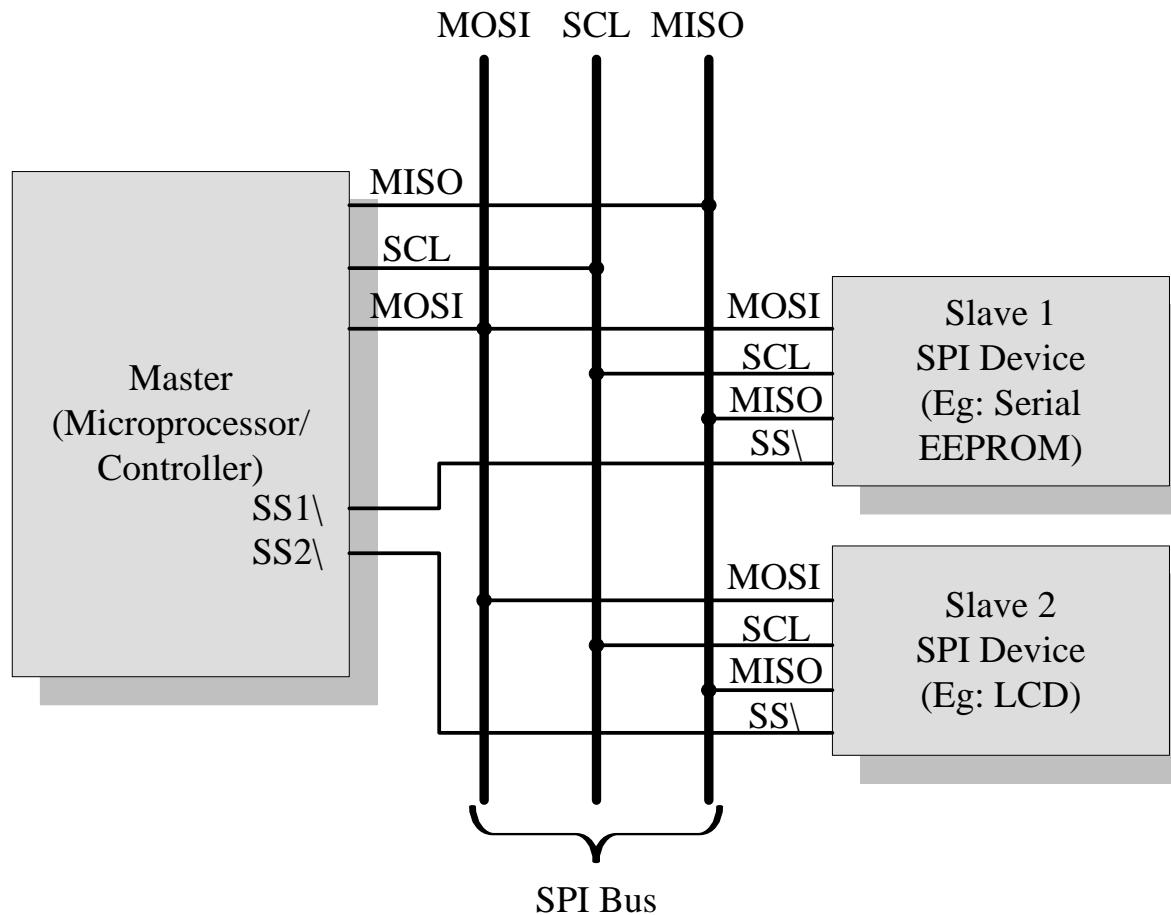
Master In Slave Out (MISO): Signal line carrying the data from slave to master device.
It is also known as Slave Output (SO/SDO)

Serial Clock (SCLK): Signal line carrying the clock signals

Slave Select (SS): Signal line for slave device select. It is an active low signal

The Typical Embedded System

On-board Communication Interface – Serial Peripheral Interface (SPI) Bus



The Typical Embedded System

On-board Communication Interface – Serial Peripheral Interface (SPI) Bus

- ✓ The master device is responsible for generating the clock signal. Master device selects the required slave device by asserting the corresponding slave device's slave select signal 'LOW'. The data out line (MISO) of all the slave devices when not selected floats at high impedance state
- ✓ The serial data transmission through SPI Bus is fully configurable. SPI devices contain certain set of registers for holding these configurations. The Serial Peripheral Control Register holds the various configuration parameters like master/slave selection for the device, baudrate selection for communication, clock signal control etc. The status register holds the status of various conditions for transmission and reception.
- ✓ SPI works on the principle of 'Shift Register'. The master and slave devices contain a special shift register for the data to transmit or receive. The size of the shift register is device dependent. Normally it is a multiple of 8. During transmission from the master to slave, the data in the master's shift register is shifted out to the MOSI pin and it enters the shift register of the slave device through the MOSI pin of the slave device. At the same time the shifted out data bit from the slave device's shift register enters the shift register of the master device through MISO pin

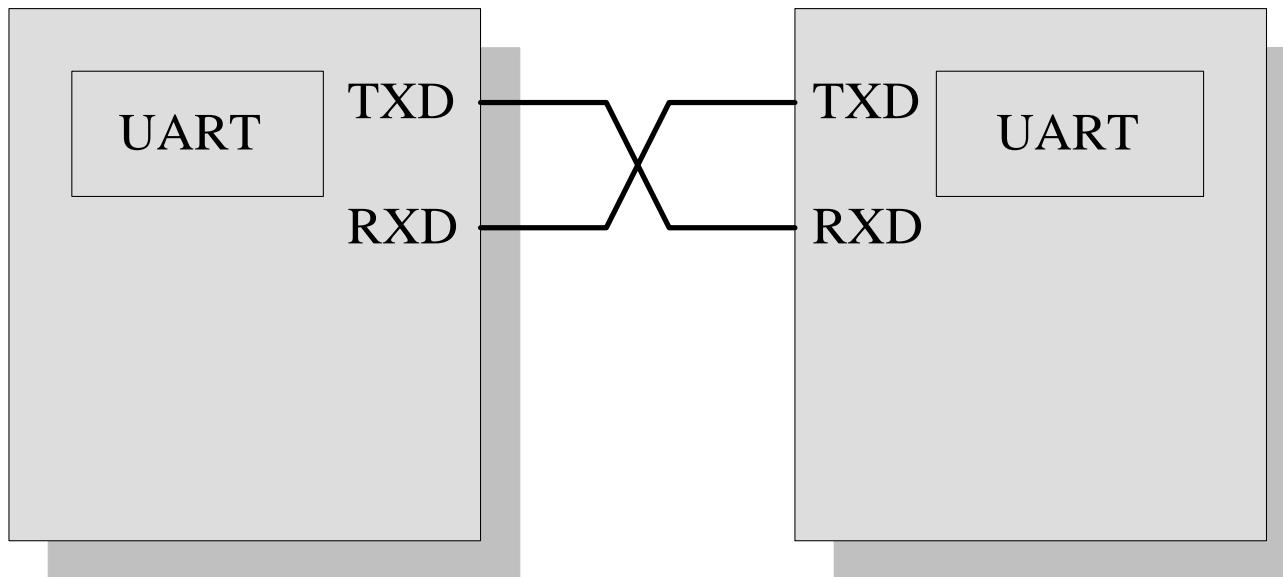
The Typical Embedded System

On-board Communication Interface – Universal Asynchronous Receiver Transmitter (UART)

- ✓ Universal Asynchronous Receiver Transmitter (UART) based data transmission is an asynchronous form of serial data transmission
- ✓ The serial communication settings (Baudrate, No. of bits per byte, parity, No. of start bits and stop bit and flow control) for both transmitter and receiver should be set as identical
- ✓ The start and stop of communication is indicated through inserting special bits in the data stream
- ✓ While sending a byte of data, a start bit is added first and a stop bit is added at the end of the bit stream. The least significant bit of the data byte follows the start bit.
- ✓ The ‘Start’ bit informs the receiver that a data byte is about to arrive. The receiver device starts polling its ‘receive line’ as per the baudrate settings
- ✓ If parity is enabled for communication, the UART of the transmitting device adds a parity bit
- ✓ The UART of the receiving device calculates the parity of the bits received and compares it with the received parity bit for error checking
- ✓ The UART of the receiving device discards the ‘Start’, ‘Stop’ and ‘Parity’ bit from the received bit stream and converts the received serial bit data to a word

The Typical Embedded System

On-board Communication Interface – Universal Asynchronous Receiver Transmitter (UART)



TXD: Transmitter Line
RXD: Receiver Line

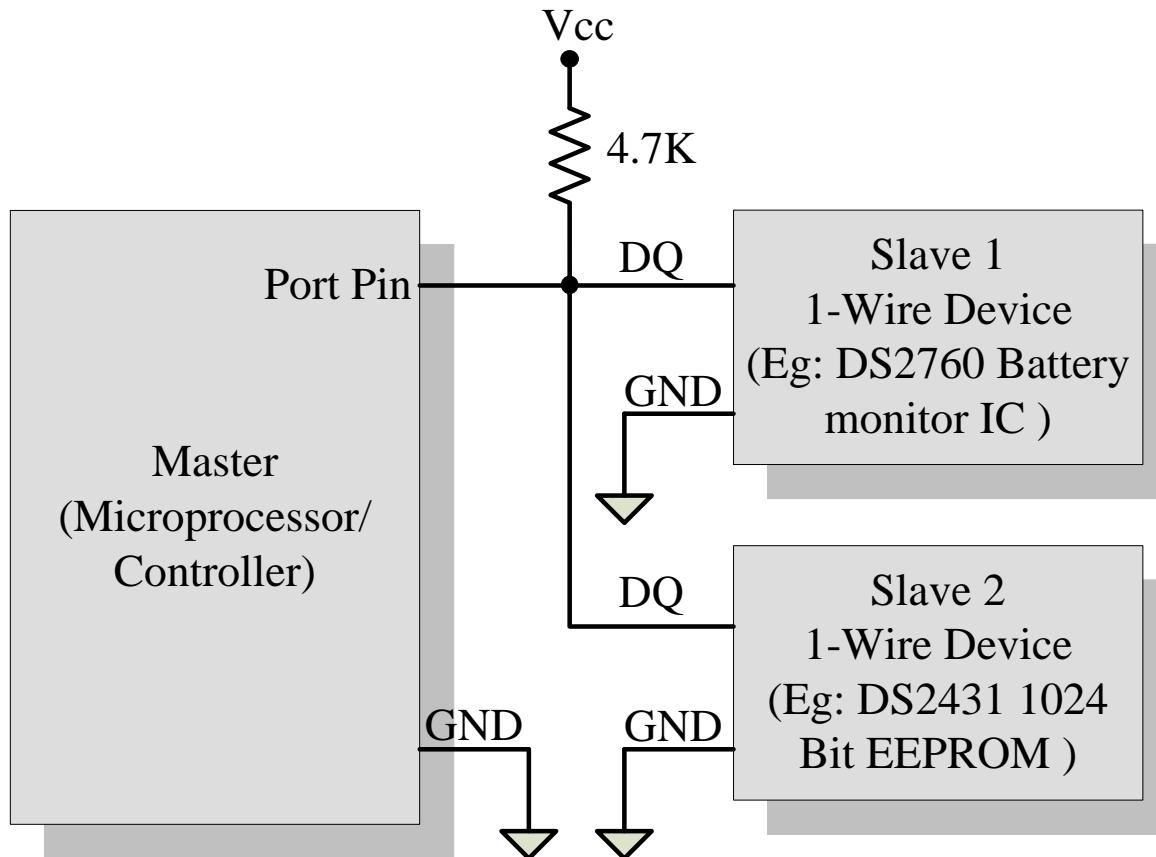
The Typical Embedded System

On-board Communication Interface – 1-Wire Interface

- ✓ An asynchronous half-duplex communication protocol developed by Maxim Dallas Semiconductor (<http://www.maxim-ic.com>)
- ✓ It is also known as Dallas 1-Wire® protocol. It makes use of only a single signal line (wire) called DQ for communication and follows the master-slave communication model
- ✓ The 1-Wire interface supports a single master and one or more slave devices on the bus
- ✓ The 1-Wire is capable of carrying power to the slave device apart from carrying the signals. Slave devices incorporate internal capacitor to generate power to operate the device from the 1-Wire
- ✓ Every 1-Wire device contains a globally unique 64 bit identification number stored within it. This unique identification number can be used for addressing individual devices present in the bus in case there are multiple slave devices connected to the 1-Wire bus
- ✓ The identifier has three parts: an 8 bit family code, a 48 bit serial number and an 8 bit CRC computed from the first 56 bits

The Typical Embedded System

On-board Communication Interface – 1-Wire Interface



The Typical Embedded System

On-board Communication Interface – 1-Wire Interface

The sequence of operation for communicating with a 1-Wire slave device is:

1. Master device sends a ‘Reset’ pulse on the 1-Wire bus.
2. Slave device(s) present on the bus respond with a ‘Presence’ pulse.
3. Master device sends a ROM Command (Net Address Command followed by the 64 bit address of the device). This addresses the slave device(s) to which it wants to initiate a communication
4. Master device sends a read/write function command to read/write the internal memory or register of the slave device.
5. Master initiates a Read data /Write data from the device or to the device

The Typical Embedded System

On-board Communication Interface – 1-Wire Interface

- ✓ All communication over the 1-Wire bus is master initiated
- ✓ The communication over the 1-Wire bus is divided into timeslots of 60 microseconds
- ✓ The ‘Reset’ pulse occupies 8 time slots. For starting a communication, the master asserts the reset pulse by pulling the 1-Wire bus ‘LOW’ for at least 8 time slots (480 μ s)
- ✓ If a ‘Slave’ device is present on the bus and is ready for communication it should respond to the master with a ‘Presence’ pulse, within 60 μ s of the release of the ‘Reset’ pulse by the master
- ✓ The slave device(s) responds with a ‘Presence’ pulse by pulling the 1-Wire bus ‘LOW’ for a minimum of 1 time slot (60 μ s)
- ✓ For writing a bit value of 1 on the 1-Wire bus, the bus master pulls the bus for 1 to 15 μ s and then releases the bus for the rest of the time slot
- ✓ A bit value of ‘0’ is written on the bus by master pulling the bus for a minimum of 1 time slot (60 μ s) and a maximum of 2 time slots (120 μ s)
- ✓ To Read a bit from the slave device, the master pulls the bus ‘LOW’ for 1 to 15 μ s
- ✓ If the slave wants to send a bit value ‘1’ in response to the read request from the slave, it simply releases the bus for the rest of the time slot
- ✓ If the slave wants to send a bit value ‘0’, it pulls the bus ‘LOW’ for the rest of the time slot.

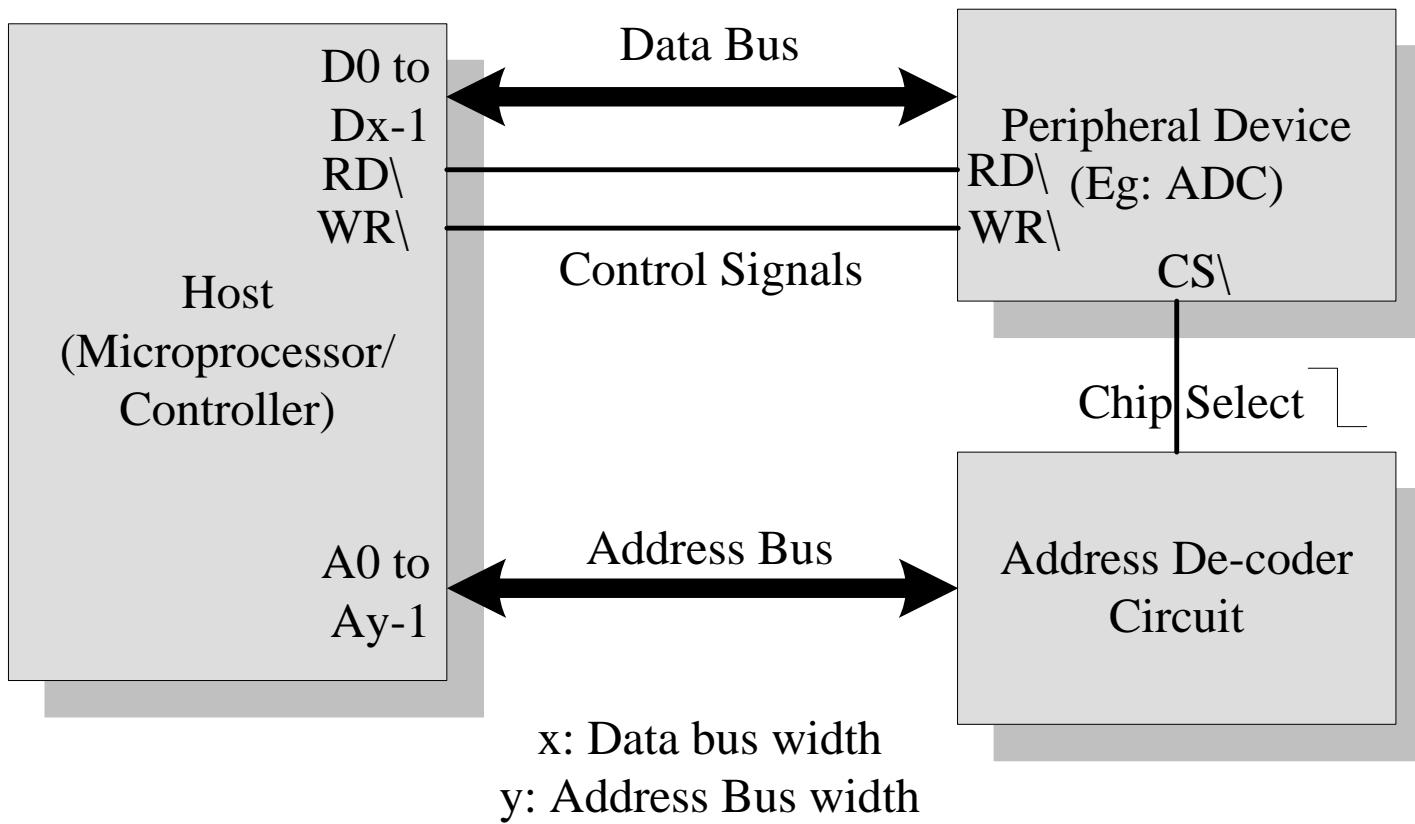
The Typical Embedded System

On-board Communication Interface – Parallel Interface

- ✓ Parallel interface is normally used for communicating with peripheral devices which are memory mapped to the host of the system
- ✓ The host processor/controller of the embedded system contains a parallel bus and the device which supports parallel bus can directly connect to this bus system
- ✓ The communication through the parallel bus is controlled by the control signal interface between the device and the host
- ✓ The ‘Control Signals’ for communication includes ‘Read/Write’ signal and device select signal
- ✓ The device normally contains a device select line and the device becomes active only when this line is asserted by the host processor
- ✓ The direction of data transfer (Host to Device or Device to Host) can be controlled through the control signal lines for ‘Read’ and ‘Write’
- ✓ Only the host processor has control over the ‘Read’ and ‘Write’ control signals

The Typical Embedded System

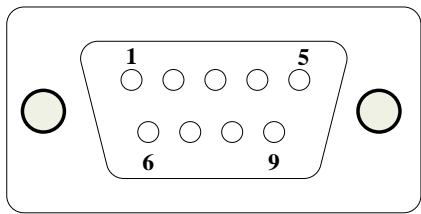
On-board Communication Interface – Parallel Interface



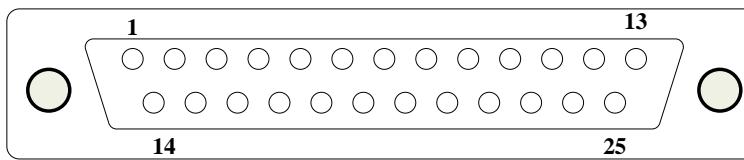
The Typical Embedded System

External Communication Interface – RS-232 C & RS-485

- ✓ RS-232 C (Recommended Standard number 232, revision C from the Electronic Industry Association) is a legacy, full duplex, wired, asynchronous serial communication interface.
- ✓ RS-232 extends the UART communication signals for external data communication.
- ✓ UART uses the standard TTL/CMOS logic (Logic ‘High’ corresponds to bit value 1 and Logic ‘LOW’ corresponds to bit value 0) for bit transmission whereas RS232 use the EIA standard for bit transmission. As per EIA standard, a logic ‘0’ is represented with voltage between +3 and +25V and a logic ‘1’ is represented with voltage between -3 and -25V. In EIA standard, logic ‘0’ is known as ‘Space’ and logic ‘1’ as ‘Mark’.
- ✓ The RS232 interface define various handshaking and control signals for communication apart from the ‘Transmit’ and ‘Receive’ signal lines for data communication. RS-232 supports two different types of connectors, namely; DB-9: 9-Pin connector and DB-25: 25-Pin connector.



DB-9



DB-25

The Typical Embedded System

External Communication Interface – RS-232 C & RS-485

Pin Name	Pin No: (For DB-9 Connector)	Description
TXD	3	Transmit Pin. Used for Transmitting Serial Data
RXD	2	Receive Pin. Used for Receiving serial Data
RTS	7	Request to send.
CTS	8	Clear To Send
DSR	6	Data Set ready
GND	5	Signal Ground
DCD	1	Data Carrier Detect
DTR	4	Data Terminal Ready
RI	9	Ring Indicator

The Typical Embedded System

External Communication Interface – RS-232 C & RS-485

- ✓ RS-232 is a point-to-point communication interface and the devices involved in RS-232 communication are called ‘Data Terminal Equipment (DTE)’ and ‘Data Communication Equipment (DCE)’
- ✓ If no data flow control is required, only TXD and RXD signal lines and ground line (GND) are required for data transmission and reception. The RXD pin of DCE should be connected to the TXD pin of DTE and vice versa for proper data transmission.
- ✓ If hardware data flow control is required for serial transmission, various control signal lines of the RS-232 connection are used appropriately. The control signals are implemented mainly for modem communication and some of them may be irrelevant for other type of devices
- ✓ The Request To Send (RTS) and Clear To Send (CTS) signals co-ordinate the communication between DTE and DCE. Whenever the DTE has a data to send, it activates the RTS line and if the DCE is ready to accept the data, it activates the CTS line
- ✓ The Data Terminal Ready (DTR) signal is activated by DTE when it is ready to accept data. The Data Set Ready (DSR) is activated by DCE when it is ready for establishing a communication link. DTR should be in the activated state before the activation of DSR
- ✓ The Data Carrier Detect (DCD) is used by the DCE to indicate the DTE that a good signal is being received
- ✓ Ring Indicator (RI) is a modem specific signal line for indicating an incoming call on the telephone line

The Typical Embedded System

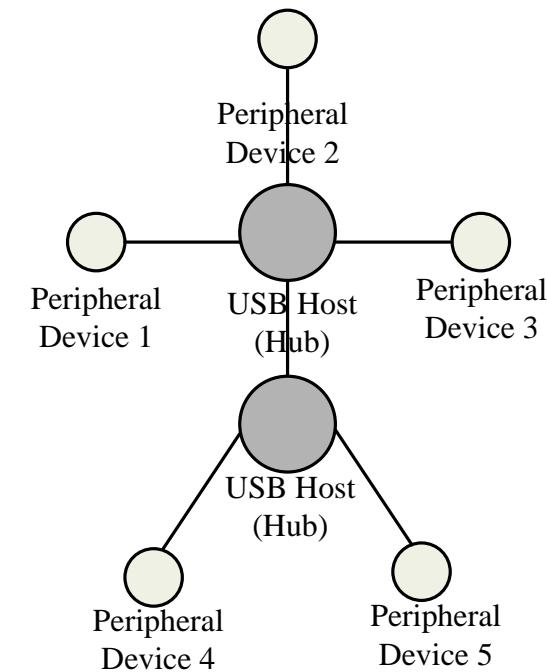
External Communication Interface – RS-232 C & RS-485

- ✓ As per the EIA standard RS-232 C supports baudrates up to 20Kbps (Upper limit 19.2Kbps) The commonly used baudrates by devices are 300bps, 1200bps, 2400bps, 9600bps, 11.52Kbps and 19.2Kbps
- ✓ The maximum operating distance supported in RS-232 communication is 50 feet at the highest supported baudrate.
- ✓ Embedded devices contain a UART for serial communication and they generate signal levels conforming to TTL/CMOS logic. A level translator IC like MAX 232 from Maxim Dallas semiconductor is used for converting the signal lines from the UART to RS-232 signal lines for communication. On the receiving side the received data is converted back to digital logic level by a converter IC. Converter chips contain converters for both transmitter and receiver
- ✓ RS-232 uses single ended data transfer and supports only point-to-point communication and not suitable for multi-drop communication
- ✓ RS-422 is another serial interface standard from EIA for differential data communication. It supports data rates up to 100Kbps and distance up to 400 ft
- ✓ RS-422 supports multi-drop communication with one transmitter device and receiver devices up to 10
- ✓ RS-485 is the enhanced version of RS-422 and it supports multi-drop communication with up to 32 transmitting devices (drivers) and 32 receiving devices on the bus. The communication between devices in the bus makes use of the ‘addressing’ mechanism to identify slave devices

The Typical Embedded System

External Communication Interface – Universal Serial Bus (USB)

- ✓ Universal Serial Bus (USB) is a wired high speed serial bus for data communication
- ✓ The USB communication system follows a star topology with a USB host at the center and one or more USB peripheral devices/USB hosts connected to it
- ✓ A USB host can support connections up to 127, including slave peripheral devices and other USB hosts
- ✓ USB transmits data in packet format. Each data packet has a standard format. The USB communication is a host initiated one
- ✓ The USB Host contains a host controller which is responsible for controlling the data communication, including establishing connectivity with USB slave devices, packetizing and formatting the data packet. There are different standards for implementing the USB Host Control interface; namely Open Host Control Interface (OHCI) and Universal Host Control Interface (UHCI)



The Typical Embedded System

External Communication Interface – Universal Serial Bus (USB)

- ✓ The Physical connection between a USB peripheral device and master device is established with a USB cable
- ✓ The USB cable supports communication distance of up to 5 meters
- ✓ The USB standard uses two different types of connectors namely ‘Type A’ and ‘Type B’ at the ends of the USB cable for connecting the USB peripheral device and host device
- ✓ ‘Type A’ connector is used for upstream connection (connection with host) and ‘Type B’ connector is used for downstream connection (connection with slave device)

Pin No:	Pin Name	Description
1	V _{BUS}	Carries power (5V)
2	D-	Differential data carrier line
3	D+	Differential data carrier line
4	GND	Ground signal line

The Typical Embedded System

External Communication Interface – Universal Serial Bus (USB)

- ✓ Each USB device contains a Product ID (PID) and a Vendor ID (VID)
- ✓ The PID and VID are embedded into the USB chip by the USB device manufacturer
- ✓ The VID for a device is supplied by the USB standards forum
- ✓ PID and VID are essential for loading the drivers corresponding to a USB device for communication
- ✓ USB supports four different types of data transfers, namely; Control, Bulk, Isochronous and Interrupt
- ✓ Control transfer is used by USB system software to query, configure and issue commands to the USB device
- ✓ Bulk transfer is used for sending a block of data to a device. Bulk transfer supports error checking and correction. Transferring data to a printer is an example for bulk transfer.
- ✓ Isochronous data transfer is used for real time data communication. In Isochronous transfer, data is transmitted as streams in real time. Isochronous transfer doesn't support error checking and re-transmission of data in case of any transmission loss
- ✓ Interrupt transfer is used for transferring small amount of data. Interrupt transfer mechanism makes use of polling technique to see whether the USB device has any data to send
- ✓ The frequency of polling is determined by the USB device and it varies from 1 to 255 milliseconds. Devices like Mouse and Keyboard, which transmits fewer amounts of data, uses Interrupt transfer.

The Typical Embedded System

External Communication Interface – IEEE 1394 (Firewire)

- ✓ A wired, isochronous high speed serial communication bus. It is also known as High Performance Serial Bus (HPSB)
- ✓ The research on 1394 was started by Apple Inc in 1985 and the standard for this was coined by IEEE.
- ✓ The Apple Inc's (www.apple.com) implementation of 1394 protocol is popularly known as ***Firewire***.
- ✓ ***i.LINK*** is the 1394 implementation from Sony Corporation (www.sony.net) and ***Lynx*** is the implementation from Texas Instruments (www.ti.com)
- ✓ 1394 supports peer-to-peer connection and point-to-multipoint communication allowing 63 devices to be connected on the bus in a tree topology
- ✓ The 1394 standard supports a data rate of 400 to 3200Mbits/Second
- ✓ IEEE 1394 uses differential data transfer and the interface cable supports 3 types of connectors, namely; 4-pin connector, 6-pin connector (alpha connector) and 9 pin connector (beta connector)
- ✓ The 6 and 9 pin connectors carry power also to support external devices. It can supply unregulated power in the range of 24 to 30V (The Apple implementation is for battery operated devices and it can supply a voltage in the range 9 to 12V)

The Typical Embedded System

External Communication Interface – IEEE 1394 (Firewire)

Pin Name	Pin No: (4 Pin Connector)	Pin No: (6 Pin Connector)	Pin No: (9 Pin Connector)	Description
Power		1	8	Unregulated DC supply. 24 to 30V
Signal Ground		2	6	Ground connection
TPB-	1	3	1	Differential Signal line for Signal Line B
TPB+	2	4	2	Differential Signal line for Signal Line B
TPA-	3	5	3	Differential Signal line for Signal Line A
TPA+	4	6	4	Differential Signal line for Signal Line A
TPA(S)			5	Shield for the differential signal line A. Normally grounded
TPB(S)			9	Shield for the differential signal line B. Normally grounded
NC			7	No connection

The Typical Embedded System

External Communication Interface – IEEE 1394 (Firewire)

- ✓ The IEEE 1394 connector contains two differential data transfer lines namely A and B
- ✓ The differential lines of A are connected to B (TPA+ to TPB+ and TPA- to TPB-) and vice versa
- ✓ Unlike USB interface (Except USB OTG), IEEE 1394 doesn't require a host for communicating between devices. Example, a scanner can be directly connected to a printer for printing
- ✓ The data rate supported by 1394 is far higher than the one supported by USB2.0 interface
- ✓ 1394 is a popular communication interface for connecting embedded devices like ‘Digital Camera’, ‘Camcorder’, ‘Scanners’ with desktop Computers for data transfer and storage

The Typical Embedded System

External Communication Interface – Infrared (IrDA)

- ✓ A serial, half duplex, line of sight based wireless technology for data communication between devices
- ✓ Infrared communication technique makes use of Infrared waves of the electromagnetic spectrum for transmitting the data
- ✓ IrDA supports point-point and point-to-multipoint communication, provided all devices involved in the communication are within the line of sight
- ✓ The typical communication range for IrDA lies in the range 10cm to 1 m
- ✓ IR supports data rates ranging from 9600bits/second to 16Mbps. Depending on the speed of data transmission IR is classified into Serial IR (SIR), Medium IR (MIR), Fast IR (FIR), Very Fast IR (VFIR) and Ultra Fast IR (UFIR)
- ✓ SIR supports transmission rates ranging from 9600bps to 115.2kbps. MIR supports data rates of 0.576Mbps and 1.152Mbps. FIR supports data rates up to 4Mbps. VFIR is designed to support high data rates up to 16Mbps. The UFIR specs are under development and it is targeting a data rate up to 100Mbps
- ✓ IrDA communication involves a transmitter unit for transmitting the data over IR and a receiver for receiving the data. Infrared Light Emitting Diode (LED) is used as the IR source for transmitter and at the receiving end a photodiode is used as the receiver

The Typical Embedded System

External Communication Interface – Bluetooth

- ✓ Low cost, low power, short range wireless technology for data and voice communication
- ✓ Bluetooth operates at 2.4GHz of the Radio Frequency spectrum and uses the Frequency Hopping Spread Spectrum (FHSS) technique for communication.
- ✓ Bluetooth supports a theoretical maximum data rate of up to 1Mbps and a range of approximately 30 feet for data communication
- ✓ Bluetooth communication has two essential parts; a physical link part and a protocol part. The physical link is responsible for the physical transmission of data between devices supporting Bluetooth communication and protocol part is responsible for defining the rules of communication
- ✓ The physical link works on the Wireless principle making use of RF waves for communication
- ✓ Bluetooth enabled devices essentially contain a Bluetooth wireless radio for the transmission and reception of data
- ✓ The rules governing the Bluetooth communication is implemented in the ‘Bluetooth protocol stack’. The Bluetooth communication IC holds the stack
- ✓ Each Bluetooth device will have a 48 bit unique identification number. Bluetooth communication follows packet based data transfer
- ✓ Bluetooth supports point-to-point (device to device) and point-to-multipoint (device to multiple device broadcasting) wireless communication. The point-to-point communication follows the master-slave relationship. A Bluetooth device can function as either master or slave
- ✓ A network formed with one Bluetooth device as master and more than one device as slaves is known as Piconet

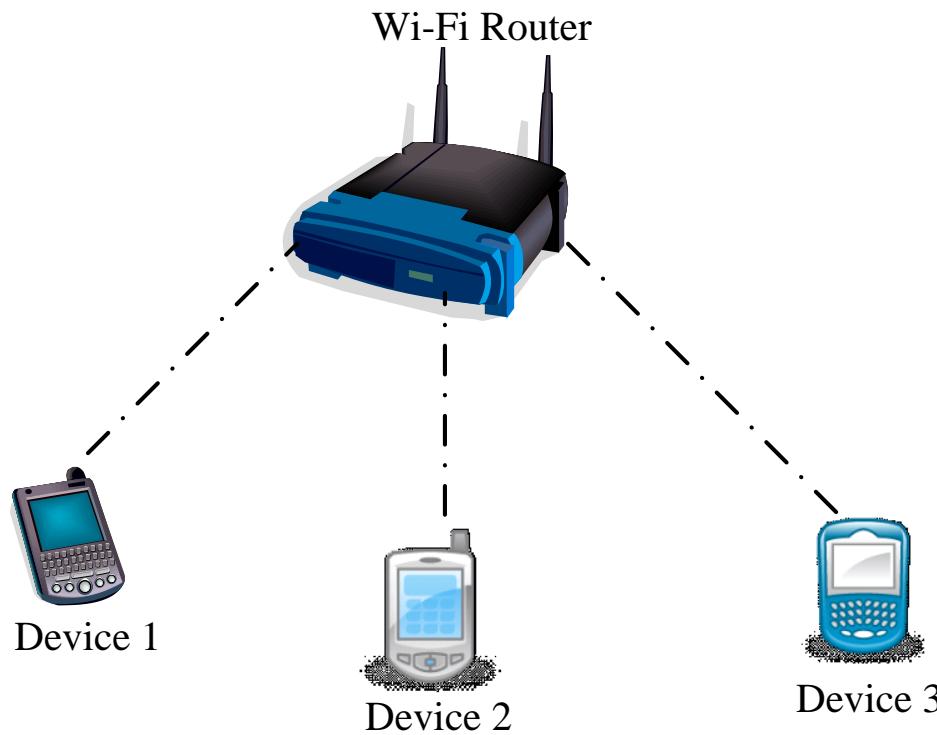
The Typical Embedded System

External Communication Interface – Wi-Fi

- ✓ The popular wireless communication technique for networked communication of devices
- ✓ Wi-Fi follows the IEEE 802.11 standard
- ✓ Wi-Fi is intended for network communication and it supports Internet Protocol (IP) based communication
- ✓ Wi-Fi based communications require an intermediate agent called Wi-Fi router/Wireless Access point to manage the communications
- ✓ The Wi-Fi router is responsible for restricting the access to a network, assigning IP address to devices on the network, routing data packets to the intended devices on the network
- ✓ Wi-Fi enabled devices contain a wireless adaptor for transmitting and receiving data in the form of radio signals through an antenna
- ✓ Wi-Fi operates at 2.4GHZ or 5GHZ of radio spectrum and they co-exist with other ISM band devices like Bluetooth
- ✓ A Wi-Fi network is identified with a Service Set Identifier (SSID). A Wi-Fi device can connect to a network by selecting the SSID of the network and by providing the credentials if the network is security enabled
- ✓ Wi-Fi networks implements different security mechanisms for authentication and data transfer
- ✓ Wireless Equivalency Protocol (WEP), Wireless Protected Access (WPA) etc are some of the security mechanisms supported by Wi-Fi networks in data communication

The Typical Embedded System

External Communication Interface – Wi-Fi



The Typical Embedded System

External Communication Interface – ZigBee

- ✓ Low power, low cost, wireless network communication protocol based on the IEEE 802.15.4-2006 standard
- ✓ ZigBee is targeted for low power, low data rate and secure applications for Wireless Personal Area Networking (WPAN)
- ✓ The ZigBee specifications support a robust mesh network containing multiple nodes. This networking strategy makes the network reliable by permitting messages to travel through a number of different paths to get from one node to another.
- ✓ ZigBee operates worldwide at the unlicensed bands of Radio spectrum, mainly at 2.400 to 2.484 GHz, 902 to 928 MHz and 868.0 to 868.6 MHz
- ✓ ZigBee Supports an operating distance of up to 100 meters and a data rate of 20 to 250Kbps
- ✓ ZigBee is primarily targeting application areas like Home & Industrial Automation, Energy Management, Home control/security, Medical/Patient tracking, Logistics & Asset tracking and sensor networks & active RFID
- ✓ Automatic Meter Reading (AMR), smoke and detectors, wireless telemetry, HVAC control, heating control, Lighting controls, Environmental controls, etc are examples for applications which can make use of the ZigBee technology

The Typical Embedded System

External Communication Interface – ZigBee

In the ZigBee terminology, each ZigBee device falls under any one of the following ZigBee device category

ZigBee Coordinator (ZC)/Network Coordinator:

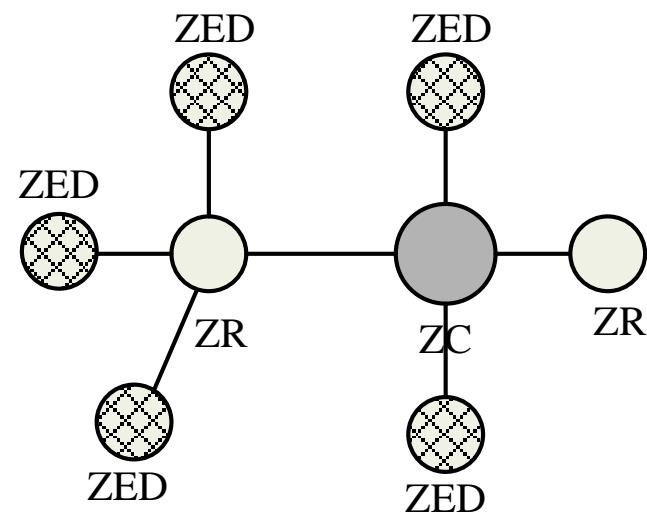
The ZigBee coordinator acts as the root of the ZigBee network. The ZC is responsible for initiating the ZigBee network and it has the capability to store information about the network

ZigBee Router (ZR)/Full function Device (FFD):

Responsible for passing information from device to another device or to another ZR

ZigBee End Device (ZED)/Reduced Function Device (RFD):

End device containing ZigBee functionality for data communication. It can talk only with a ZR or ZC and doesn't have the capability to act as a mediator for transferring data from one device to another.



The Typical Embedded System

External Communication Interface – General Packet Radio Service (GPRS)

- ✓ A communication technique for transferring data over a mobile communication network like GSM
- ✓ Data is sent as packets. The transmitting device splits the data into several related packets. At the receiving end the data is re-constructed by combining the received data packets
- ✓ GPRS supports a theoretical maximum transfer rate of 171.2kbps
- ✓ In GPRS communication, the radio channel is concurrently shared between several users instead of dedicating a radio channel to a cell phone user. The GPRS communication divides the channel into 8 timeslots and transmits data over the available channel
- ✓ GPRS supports Internet Protocol (IP), Point to Point Protocol (PPP) and X.25 protocols for communication.
- ✓ GPRS is mainly used by mobile enabled embedded devices for data communication. The device should support the necessary GPRS hardware like GPRS modem and GPRS radio
- ✓ GPRS is an old technology and it is being replaced by new generation data communication techniques like EDGE, High Speed Downlink Packet Access (HSDPA) etc which offers higher bandwidths for communication

The Typical Embedded System

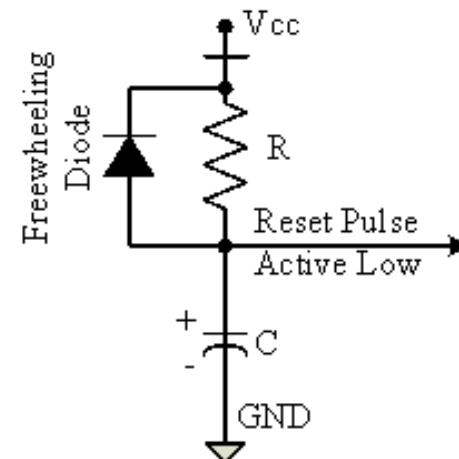
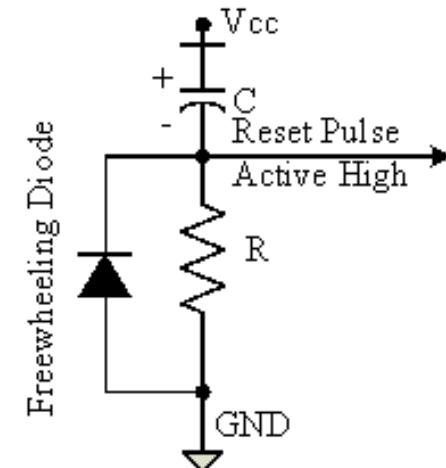
Embedded Firmware

- ✓ The control algorithm (Program instructions) and or the configuration settings that an embedded system developer dumps into the code (Program) memory of the embedded system
- ✓ The embedded firmware can be developed in various methods like
 - ✓ Write the program in high level languages like Embedded C/C++ using an Integrated Development Environment (The IDE will contain an editor, compiler, linker, debugger, simulator etc. IDEs are different for different family of processors/controllers.
 - ✓ Write the program in Assembly Language using the Instructions Supported by your application's target processor/controller

The Typical Embedded System

Other System Components – Reset Circuit

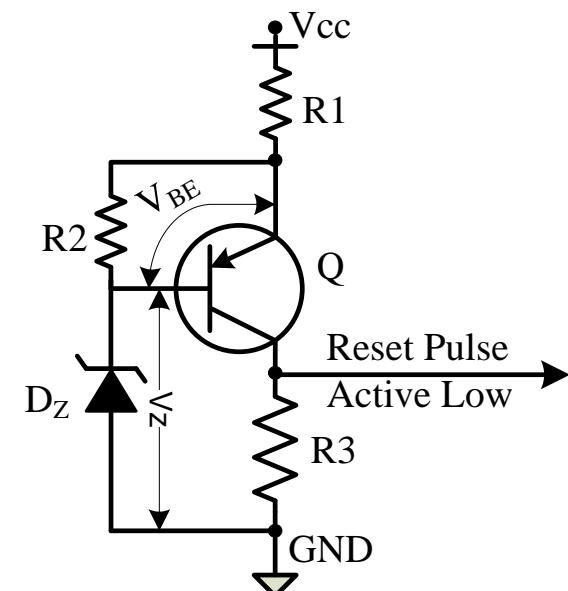
- ✓ The Reset circuit is essential to ensure that the device is not operating at a voltage level where the device is not guaranteed to operate, during system power ON
- ✓ The Reset signal brings the internal registers and the different hardware systems of the processor/controller to a known state and starts the firmware execution from the reset vector (Normally from vector address 0x0000 for conventional processors/controllers)
- ✓ The reset vector can be relocated to an address for processors/controllers supporting bootloader
- ✓ The reset signal can be either active high (The processor undergoes reset when the reset pin of the processor is at logic high) or active low (The processor undergoes reset when the reset pin of the processor is at logic low).



The Typical Embedded System

Other System Components – Brown-out Protection Circuit

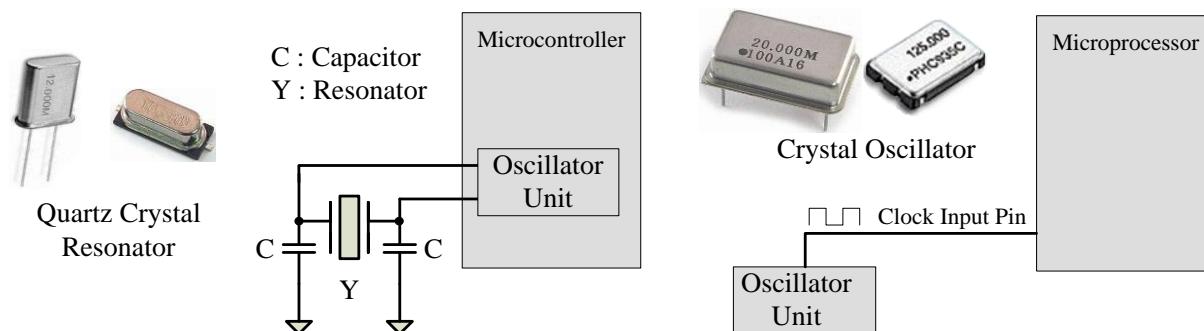
- ✓ Brown-out protection circuit prevents the processor/controller from unexpected program execution behavior when the supply voltage to the processor/controller falls below a specified voltage
- ✓ The processor behavior may not be predictable if the supply voltage falls below the recommended operating voltage. It may lead to situations like data corruption
- ✓ A brown-out protection circuit holds the processor/controller in reset state, when the operating voltage falls below the threshold, until it rises above the threshold voltage
- ✓ Certain processors/controllers support built in brown-out protection circuit which monitors the supply voltage internally
- ✓ If the processor/controller doesn't integrate a built-in brown-out protection circuit, the same can be implemented using external passive circuits or supervisor ICs



The Typical Embedded System

Other System Components – Oscillator Unit

- ✓ A microprocessor/microcontroller is a digital device made up of digital combinational and sequential circuits
- ✓ The instruction execution of a microprocessor/controller occurs in sync with a clock signal
- ✓ The oscillator unit of the embedded system is responsible for generating the precise clock for the processor
- ✓ Certain processors/controllers integrate a built-in oscillator unit and simply require an external ceramic resonator/quartz crystal for producing the necessary clock signals
- ✓ Certain processor/controller chips may not contain a built-in oscillator unit and require the clock pulses to be generated and supplied externally
- ✓ Quartz crystal Oscillators are example for clock pulse generating devices



The Typical Embedded System

Other System Components – Real Time Clock (RTC)

- ✓ The system component responsible for keeping track of time. RTC holds information like current time (In hour, minutes and seconds) in 12 hour /24 hour format, date, month, year, day of the week etc and supplies timing reference to the system
- ✓ RTC is intended to function even in the absence of power. RTCs are available in the form of Integrated Circuits from different semiconductor manufacturers like Maxim/Dallas, ST Microelectronics etc
- ✓ The RTC chip contains a microchip for holding the time and date related information and backup battery cell for functioning in the absence of power, in a single IC package
- ✓ The RTC chip is interfaced to the processor or controller of the embedded system
- ✓ For Operating System based embedded devices, a timing reference is essential for synchronizing the operations of the OS kernel. The RTC can interrupt the OS kernel by asserting the interrupt line of the processor/controller to which the RTC interrupt line is connected
- ✓ The OS kernel identifies the interrupt in terms of the Interrupt Request (IRQ) number generated by an interrupt controller
- ✓ One IRQ can be assigned to the RTC interrupt and the kernel can perform necessary operations like system date time updation, managing software timers etc when an RTC timer tick interrupt occurs

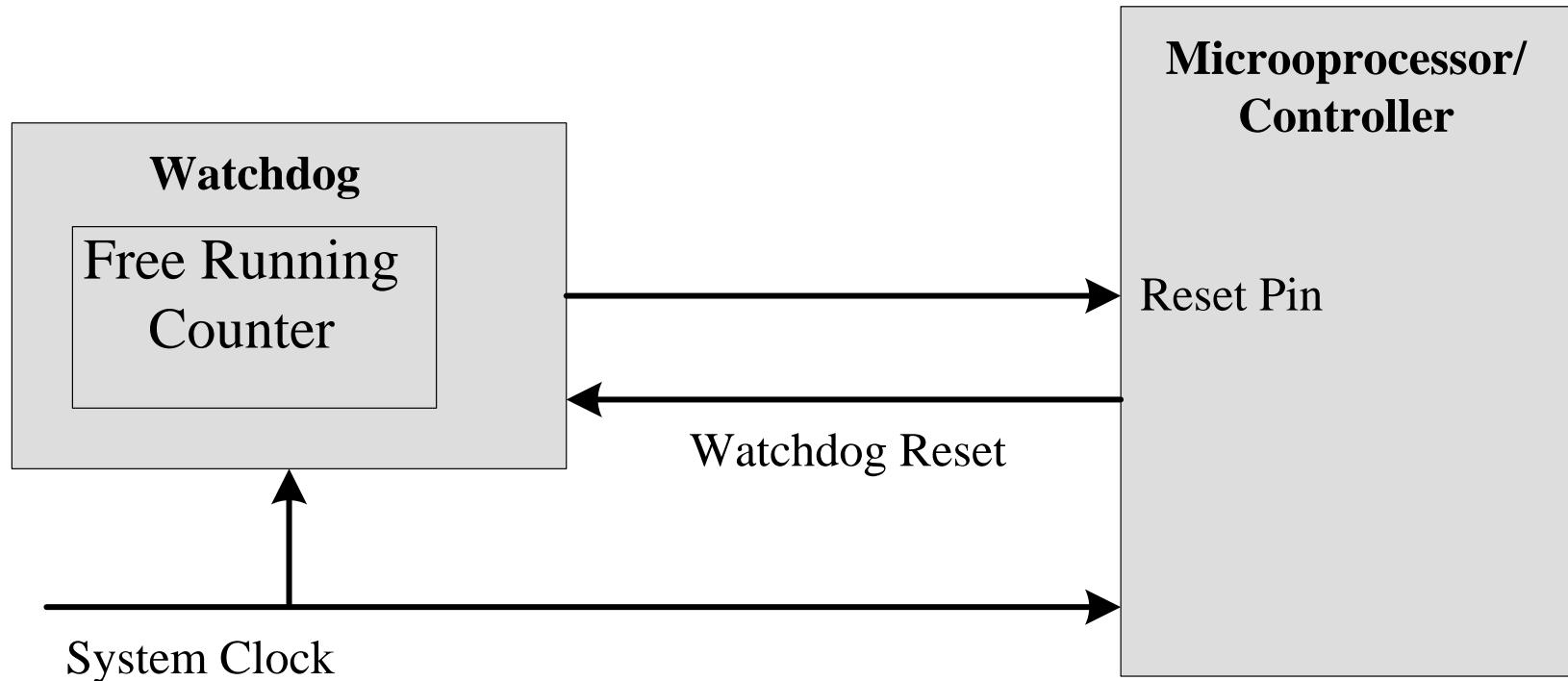
The Typical Embedded System

Other System Components – Watch Dog Timer (WDT)

- ✓ A timer unit for monitoring the firmware execution
- ✓ Depending on the internal implementation, the watchdog timer increments or decrements a free running counter with each clock pulse and generates a reset signal to reset the processor if the count reaches zero for a down counting watchdog, or the highest count value for an up counting watchdog
- ✓ If the watchdog counter is in the enabled state, the firmware can write a zero (for up counting watchdog implementation) to it before starting the execution of a piece of code (subroutine or portion of code which is susceptible to execution hang up) and the watchdog will start counting. If the firmware execution doesn't complete due to malfunctioning, within the time required by the watchdog to reach the maximum count, the counter will generate a reset pulse and this will reset the processor
- ✓ If the firmware execution completes before the expiration of the watchdog timer the WDT can be stopped from action
- ✓ Most of the processors implement watchdog as a built-in component and provides status register to control the watchdog timer (like enabling and disabling watchdog functioning) and watchdog timer register for writing the count value. If the processor/controller doesn't contain a built in watchdog timer, the same can be implemented using an external watchdog timer IC circuit.

The Typical Embedded System

Other System Components – Watch Dog Timer (WDT)



External Watch Dog Timer Unit Interfacing with Processor

Image Sensing and Acquisition

- Image sensing and acquisition are used for processing the analog images of physical scenes or the structure of an object, and converting it into digital.
- Image sensing refers to sensing an analog image and giving it as input to the machine.
- Image acquisition includes,
 - Processing of image.
 - Compression of image.
 - finally storing of image into digital form.

IMAGE SENSING

- An image sensing is a process to detect or sense the information that constitutes an image.
- As per nature of the object, the images can be generated in two ways
 1. The illumination generated by the object
 2. With the combination of an illumination source and the reflection or absorption of energy from the source by the object.
- Sensor arrangement is used to sense the illumination energy from the analog scene of object.

IMAGE ACQUISITION

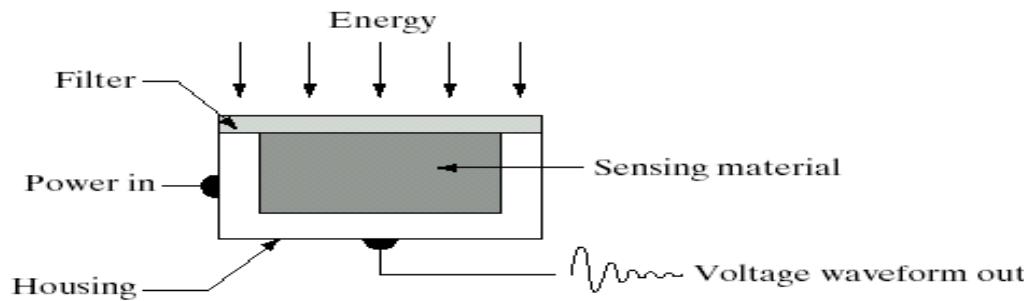
- Image acquisition can be defined as the action of retrieving an image from some source, usually it may be hardware based source.
- Performing image acquisition is the first image processing work flow sequence.
- Photo diode is mostly used sensor in this category for image acquisition.
- There are various method of image acquisition
 - (i) Single imaging Sensor
 - (ii) Line sensor
 - (iii) Array sensor

Image Sensing and Acquisition

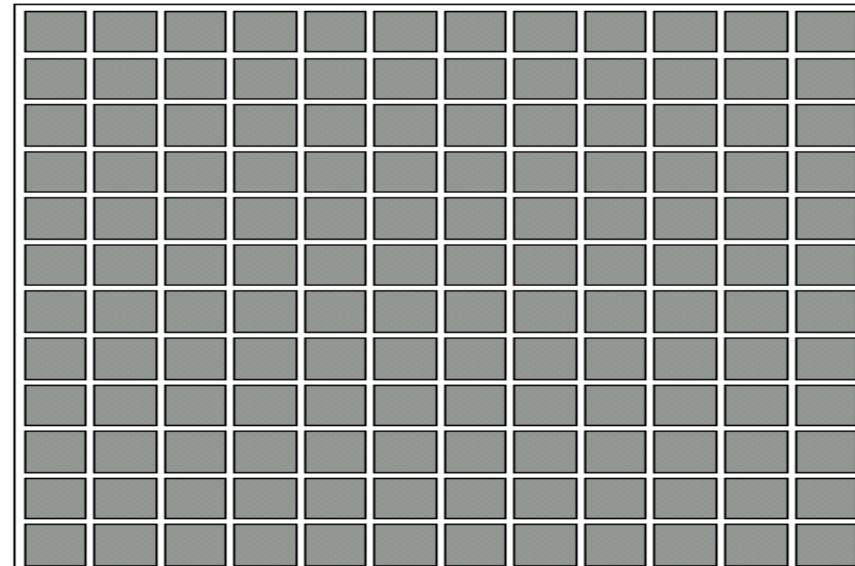
- There are 3 principal sensor arrangements (produce an electrical output proportional to light intensity).
(i) Single imaging Sensor (ii) Line sensor (iii) Array sensor

a
b
c

FIGURE 2.12
(a) Single imaging
sensor.
(b) Line sensor.
(c) Array sensor.



Transform
illumination
energy into
digital images



i) Image acquisition using a single sensor

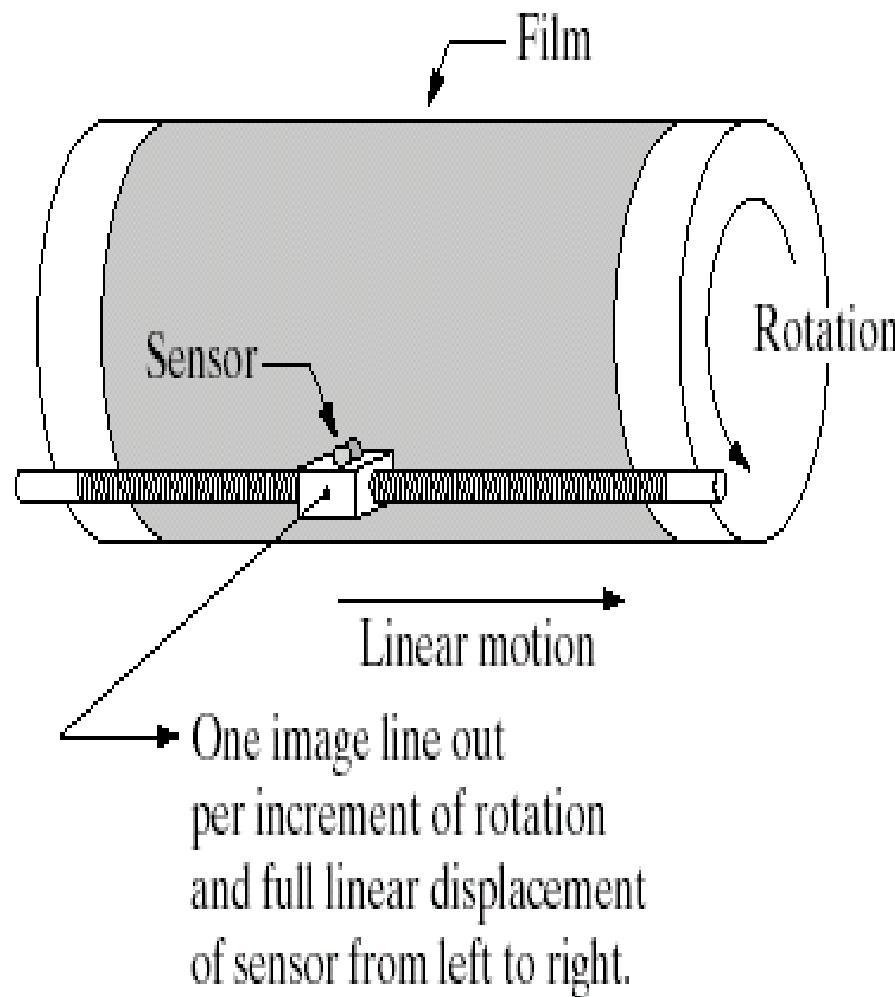


FIGURE 2.13 Combining a single sensor with motion to generate a 2-D image.

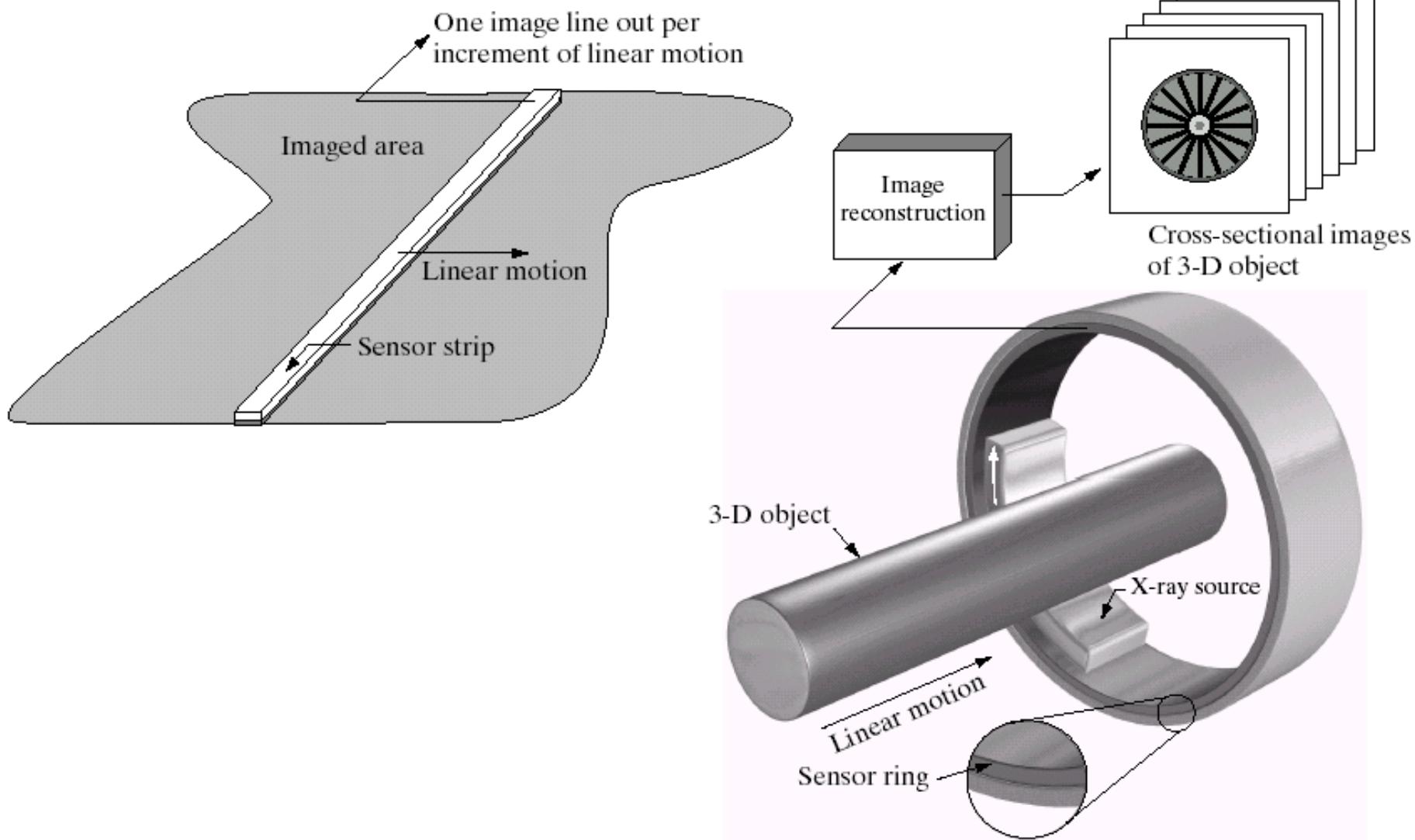
i) Image acquisition using a single sensor

- The most common sensor of this type is the photodiode.
- photodiode is constructed of silicon materials and whose output voltage waveform is proportional to light.
- The use of a filter in front of a sensor improves selectivity.

- **For example:** A green (pass) filter in front of a light sensor favours light in the green band of the color spectrum.
- As a consequence, the sensor output will be stronger for green light than for other components in the visible spectrum.

- To generate a 2-D image using a single sensor, there have to be relative displacements in both the x- and y-directions between the sensor and the area to be imaged.
- An arrangement used in high precision scanning.
- A film negative is mounted onto a drum.
- Drum mechanical rotation provides displacement in one dimension.
- The single sensor is mounted on a lead screw that provides motion in the perpendicular direction.
- Since mechanical motion can be controlled with high precision, this method is an inexpensive (but slow) way to obtain high-resolution images.

ii) Image Acquisition Using Sensor Strips



a | b

FIGURE 2.14 (a) Image acquisition using a linear sensor strip. (b) Image acquisition using a circular sensor ring.

ii)Image Acquisition Using Sensor Strips

- The sensor strip provides imaging elements in one direction.
- Motion perpendicular to the sensor strip provides imaging in the other direction.
- This is the most common type of arrangement used in most flatbed scanners.
- Sensing devices with 4000 or more in-line sensors are possible in linear sensor strips.
- In-line sensors are used routinely in airborne imaging applications.
- Imaging system is mounted on an aircraft that flies at a constant altitude and speed over the geographical area to be imaged.
- One-dimensional imaging sensor strips that respond to various bands of the electromagnetic spectrum.

- Sensor strips mounted in a ring configuration are used in medical and industrial imaging.
- Ring configuration sensors helps to obtain crosssectional (“slice”) images of 3-D objects.
- A rotating X-ray source provides illumination and the portion of the sensors opposite the source collect the X-ray energy that pass through the object.
- This is the basis for medical and industrial computerized axial tomography (CAT) imaging.

iii) Image Acquisition using Sensor Arrays

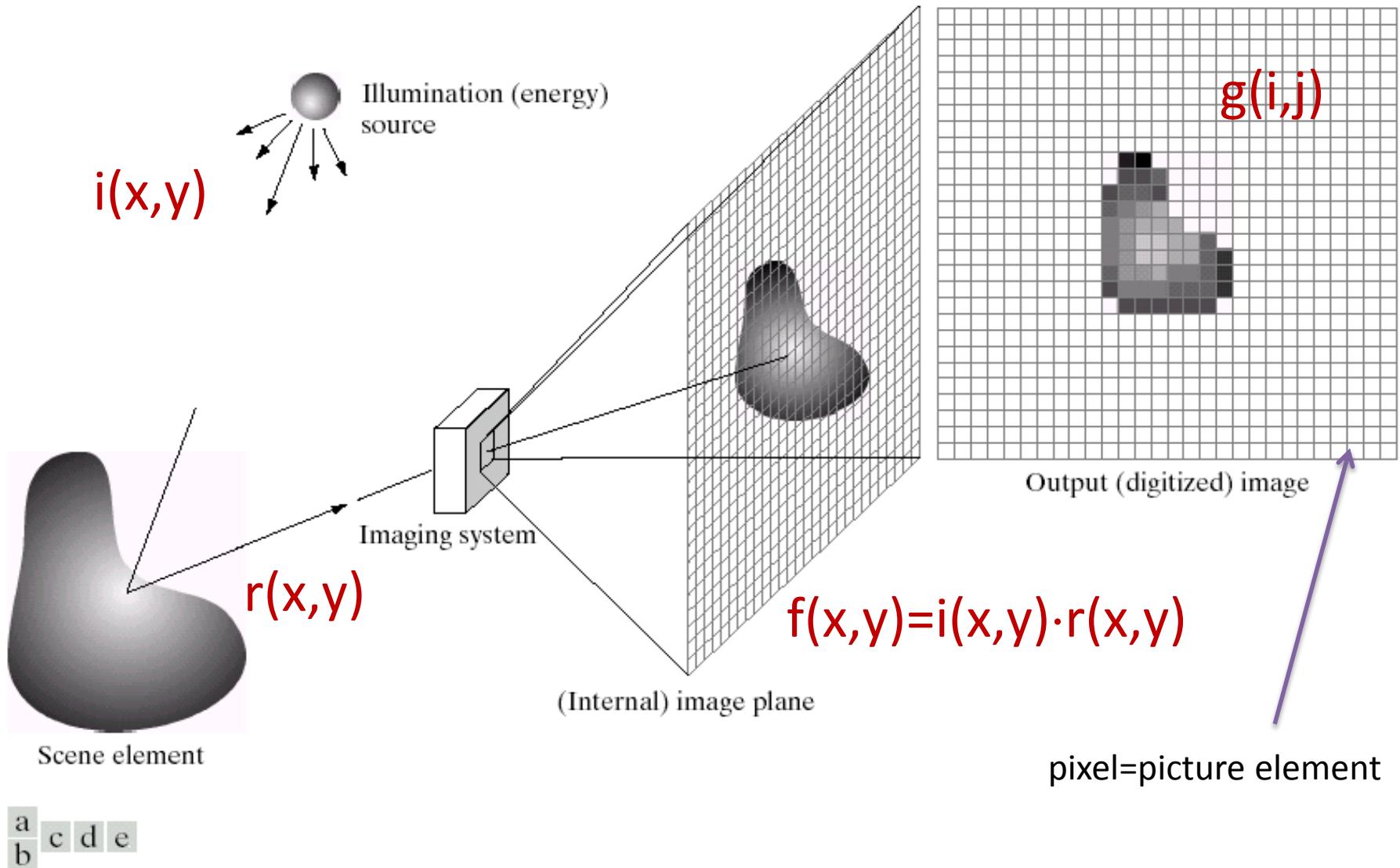


FIGURE 2.15 An example of the digital image acquisition process. (a) Energy (“illumination”) source. (b) An element of a scene. (c) Imaging system. (d) Projection of the scene onto the image plane. (e) Digitized image.

iii) Image Acquisition using Sensor Arrays

- This type of arrangement (Sensor arrays) is found in digital cameras.
- A typical sensor for these cameras is a CCD(Charge Coupled Devices) array
- CCD of cameras can be manufactured with a broad range of sensing properties.
- These can be packaged in rugged arrays of 4000 * 4000 elements or more sensors.
- CCD sensors are used widely in digital cameras and other light sensing instruments.
- The response of each sensor is proportional to the integral of the light energy projected onto the surface of the sensor.

- The above property of sensors that is used in astronomical and other applications requiring low noise images.
- The first function performed by the imaging system is to collect the incoming energy and focus it onto an image plane.
- If the illumination is light, the front end of the imaging system is a lens, which projects the viewed scene onto the lens focal plane.
- The sensor array, which is coincident with the focal plane, produces outputs proportional to the integral of the light received at each sensor.

A Simple Image Formation Model

- An image is defined by two dimensional function $f(x,y)$.
- The value or amplitude of ‘ f ’ at spatial coordinates (x,y) is a positive scalar quantity.
- When an image is generated from a physical process, its value are proportional to energy radiated by physical source.
- As a consequence, $f(x,y)$ must be nonzero and finite; that is,

$$0 < f(x,y) < \infty$$

- The function $f(x,y)$ may be characterized by two components:
 - (1) the amount of source illumination incident on the scene being viewed and
 - (2) the amount of illumination reflected by the objects in the scene.

$$f(x, y) = i(x, y)r(x, y)$$

$f(x, y)$: intensity at the point (x, y)

$i(x, y)$: illumination at the point (x, y)

(the amount of source illumination incident on the scene)

$r(x, y)$: reflectance/transmissivity at the point (x, y)

(the amount of illumination reflected/transmitted by the object)

where $0 < i(x, y) < \infty$ and $0 < r(x, y) < 1$

$r(x, y) = 0$ means total absorption

$r(x, y) = 1$ means total reflectance.

Some Typical Ranges of illumination

► Illumination

Lumen — A unit of light flow or luminous flux

Lumen per square meter (lm/m^2) — The metric unit of measure for illuminance of a surface

- On a clear day, the sun may produce in excess of 90,000 lm/m^2 of illumination on the surface of the Earth
- On a cloudy day, the sun may produce less than 10,000 lm/m^2 of illumination on the surface of the Earth
- On a clear evening, the moon yields about 0.1 lm/m^2 of illumination
- The typical illumination level in a commercial office is about 1000 lm/m^2

Some Typical Ranges of Reflectance

► Reflectance

- 0.01 for black velvet
- 0.65 for stainless steel
- 0.80 for flat-white wall paint
- 0.90 for silver-plated metal
- 0.93 for snow

Image Sampling and Quantization

- To create a digital image, we need to convert the continuous sensed data into digital form.
- Sampling and quantization are the two important processes used to convert continuous analog image into digital image.

Sampling:

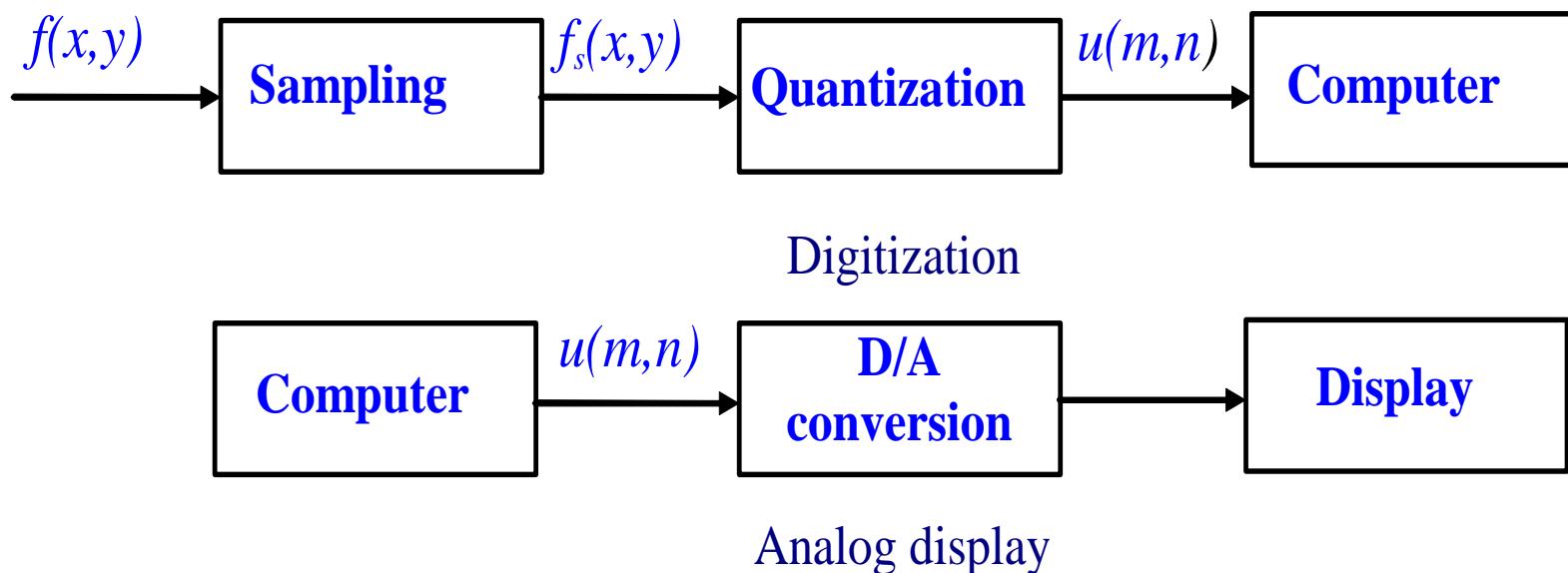
- Image sampling refers to discretization of spatial coordinates (along x axis) of Image.
- Its related to coordinates values of Image.

Quantization:

- It refers to discretization of gray level values of Image (amplitude (along y axis)).
- It related to intensity values of Image.

Sampling & Quantization

- The spatial and amplitude digitization of $f(x,y)$ is called:
 - Image sampling when it refers to spatial coordinates (x,y) and
 - gray-level quantization when it refers to the amplitude.



Basic Concepts in Sampling and Quantization

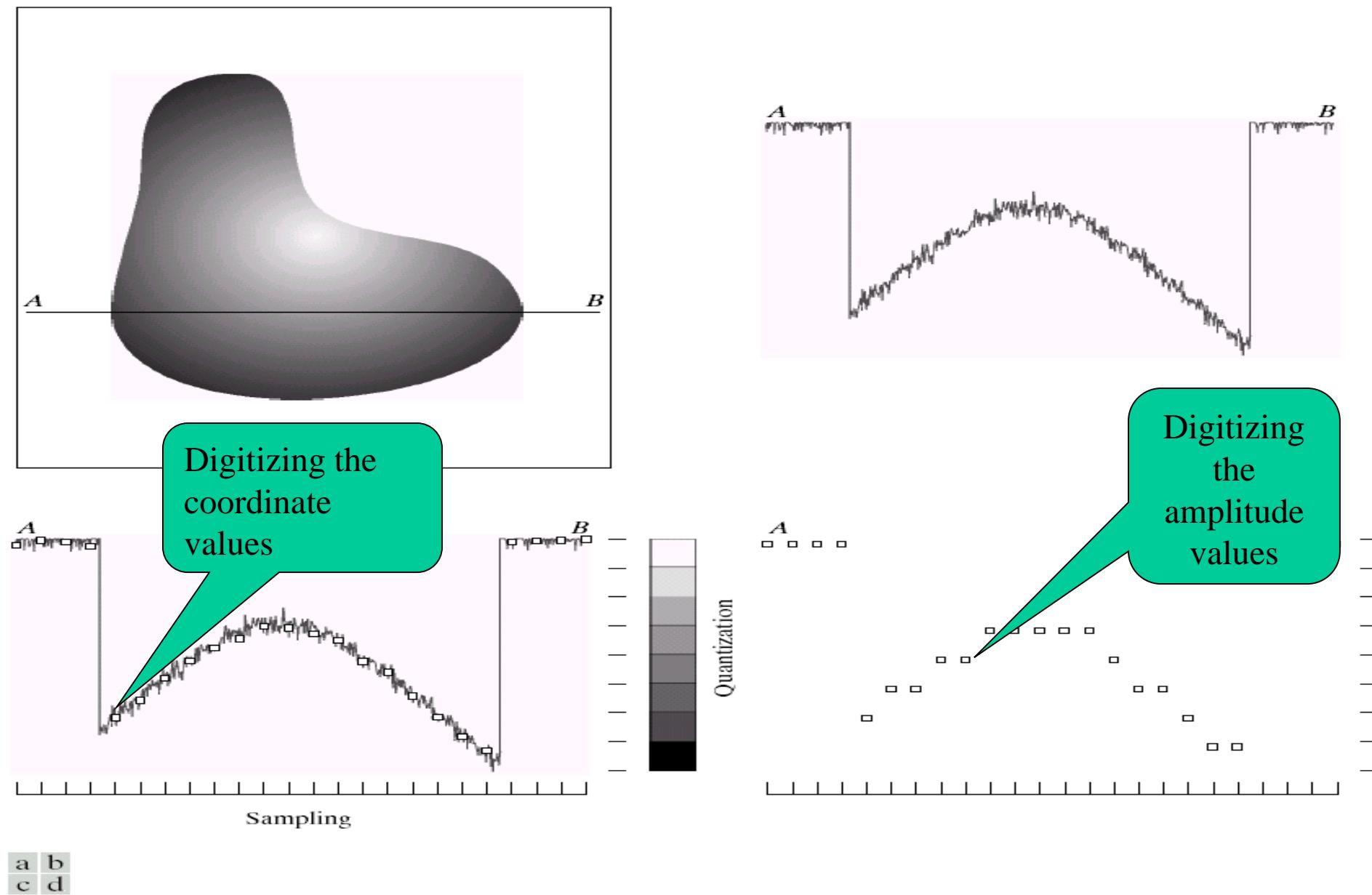
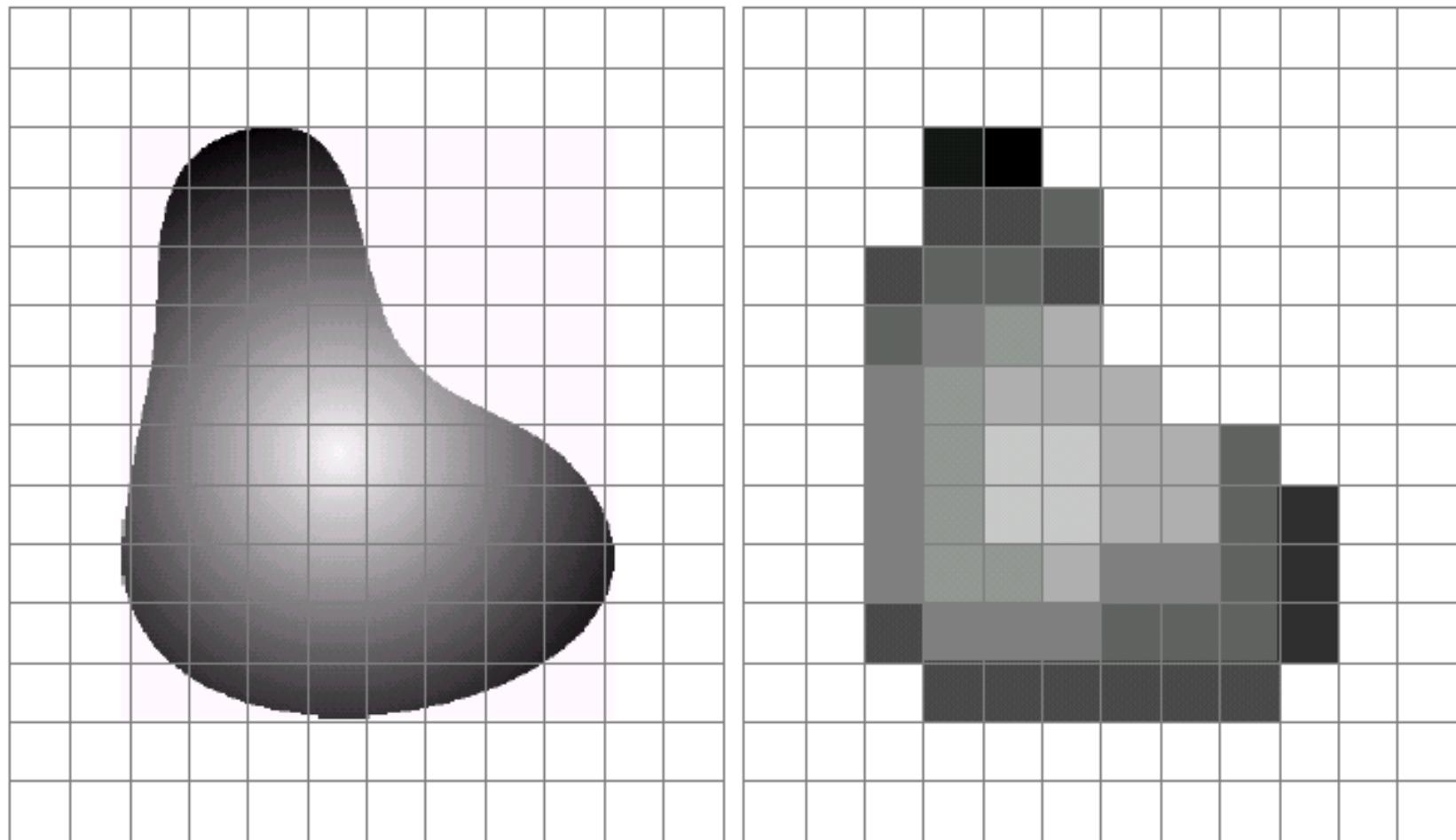


FIGURE 2.16 Generating a digital image. (a) Continuous image. (b) A scan line from **A** to **B** in the continuous image, used to illustrate the concepts of sampling and quantization. (c) Sampling and quantization. (d) Digital scan line.

- The basic idea behind sampling and quantization is illustrated fig (a)
- An image may be continuous with respect to the x- and y-coordinates, and also in amplitude.
- To convert it to digital form, we have to sample the function in both coordinates and in amplitude.
- The one-dimensional function shown in Fig(b).
- The random variations are due to image noise.
- To sample this function ,we take equally spaced samples along line AB, as shown in Fig(c).
- The set of these discrete locations gives the sampled function.

- In order to form a digital function, the gray-level values also must be converted(*quantized*) *into discrete quantities*.
- Fig. (c) shows the gray-level scale divided into eight discrete levels, ranging from black to white.
- The vertical tick marks indicate the specific value assigned to each of the eight gray levels.
- The continuous gray levels are quantized simply by assigning one of the eight discrete gray levels to each sample.
- The digital samples resulting from both sampling and quantization are shown in Fig. (d).

Sampling and Quantization



a b

FIGURE 2.17 (a) Continuous image projected onto a sensor array. (b) Result of image sampling and quantization.

Representing Digital Images

- Let $f(s,t)$ represent a continuous image function of two continuous variables, s and t .
- To convert this function into a *digital image by sampling and quantization*.
- Suppose that we sample the continuous image into a 2-D array, $f(x,y)$, containing M rows and N columns, where (x,y) are discrete coordinates.
- we use integer values for these discrete coordinates:
 $x = 0, 1, 2, \dots, M - 1$ and $y = 0, 1, 2, \dots, N - 1$.
- The value of the digital image at the origin is $f(0,0)$, and the next coordinate value along the first row is $f(0,1)$.
- Here, the notation $(0, 1)$ is used to signify the second sample along the first row.
- In general, the value of the image at any coordinates (x,y) is denoted $f(x,y)$, where x and y are integers.

- The representation of an $M \times N$ numerical array as

$$f(x, y) = \begin{bmatrix} f(0,0) & f(0,1) & \dots & f(0,N-1) \\ f(1,0) & f(1,1) & \dots & f(1,N-1) \\ \dots & \dots & \dots & \dots \\ f(M-1,0) & f(M-1,1) & \dots & f(M-1,N-1) \end{bmatrix}$$

↓ ↓

Digital Image Image Elements
(Pixels)

Matrix of Real Numbers

A Digital Image

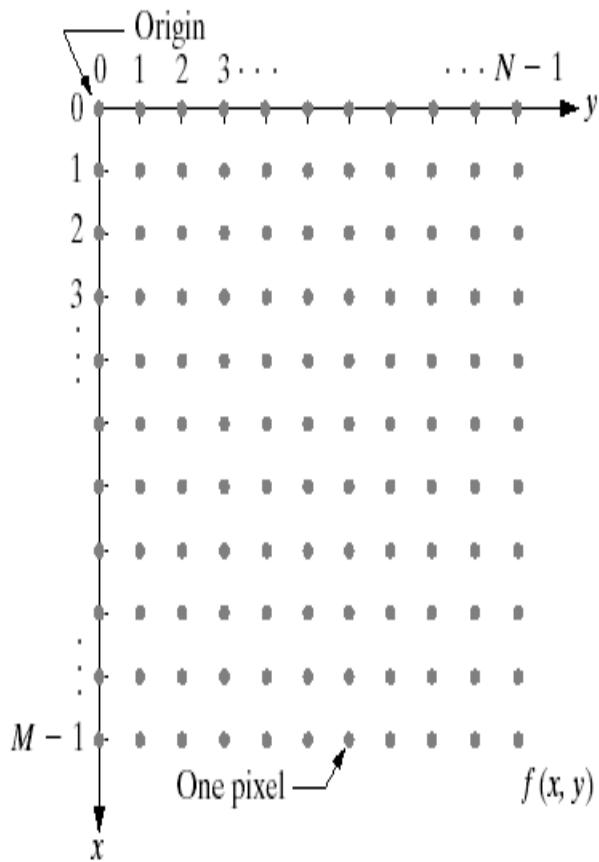
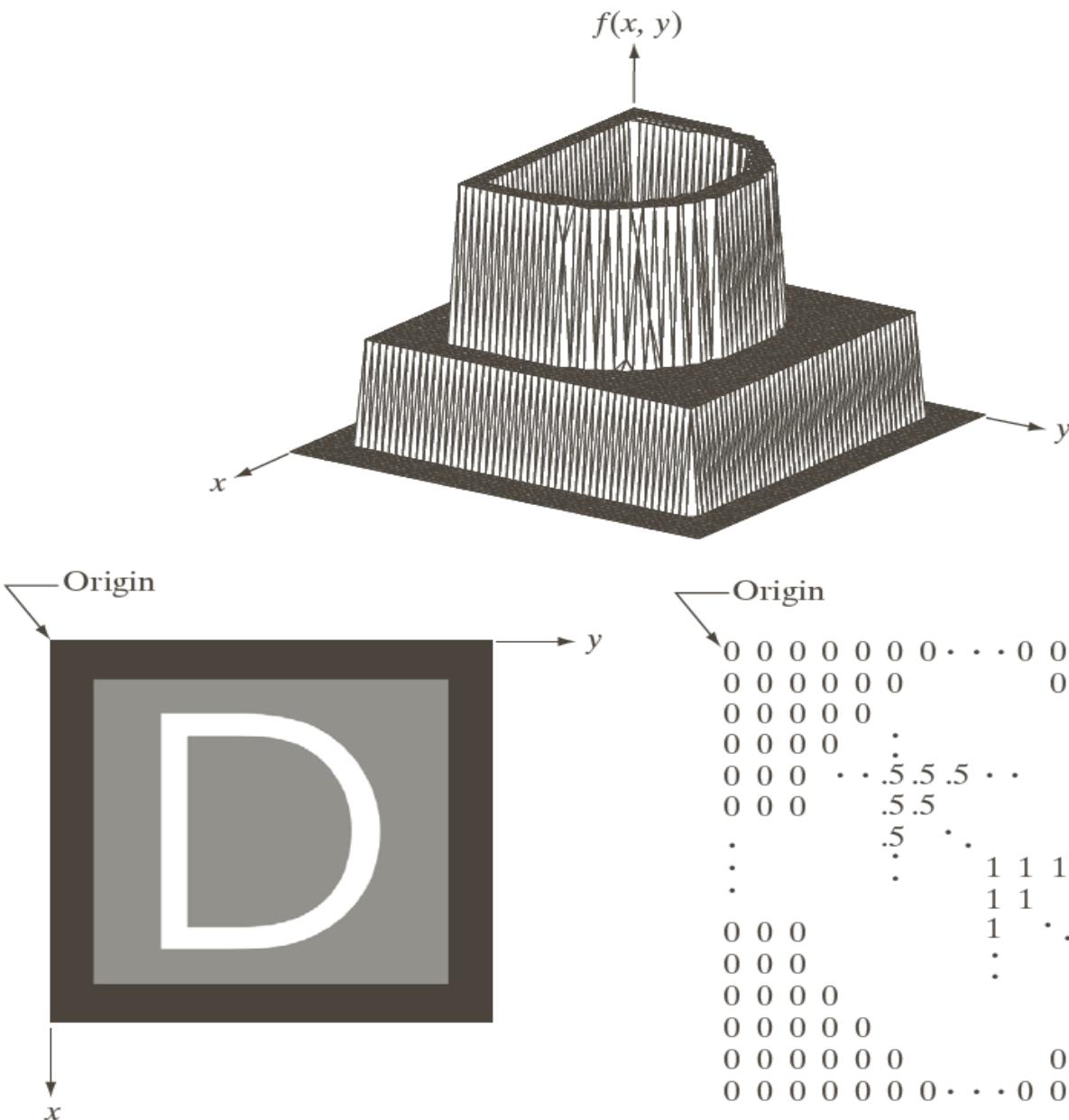


FIGURE 2.18
Coordinate
convention used
in this book to
represent digital
images.

Representing Digital Images



a
b c

FIGURE 2.18

(a) Image plotted as a surface.

(b) Image displayed as a visual intensity array.

(c) Image shown as a 2-D numerical array (0, .5, and 1 represent black, gray, and white, respectively).

- As Fig. 2.18 shows, there are three basic ways to represent .
- Figure 2.18(a) is a plot of the function, with two axes determining spatial location $f(x, y)$ and the third axis being the values of f (*intensities*) as a function of the two spatial variables x and y .
- The representation in Fig. 2.18(b) is much more common. It shows $f(x, y)$ as it would appear on a monitor or photograph.
- Here, the intensity of each point is proportional to the value of f at that point.
- In figure, 2.18(c) there are only three equally spaced intensity values. If the intensity is normalized to the interval $[0, 1]$, then each point in the image has the value 0, 0.5, or 1.
- The third representation is simply to display the numerical values of $f(x, y)$ as an array (matrix). In this example, f is of size 600×600 elements, or 360,000 numbers.

- Clearly, printing the complete array would be cumbersome and convey little information.
- When developing algorithms, however, this representation is quite useful when only parts of the image are printed and analyzed as numerical values.
- We can conclude from the figures that the representations In Figs. (b) and (c) are the most useful.
- Image displays allow us to view results at a glance.
- Numerical arrays are used for processing and algorithm development.

Representing Digital Images

Clearly, $a_{ij} = f(x = i, y = j) = f(i, j)$,

$$A = \begin{bmatrix} a_{0,0} & a_{0,1} & \dots & a_{0,N-1} \\ a_{1,0} & a_{1,1} & \dots & a_{1,N-1} \\ \dots & \dots & \dots & \dots \\ a_{M-1,0} & a_{M-1,1} & \dots & a_{M-1,N-1} \end{bmatrix}$$

Representing Digital Images

- The representation of an $M \times N$ numerical array in MATLAB

$$f(x, y) = \begin{bmatrix} f(1,1) & f(1,2) & \dots & f(1,N) \\ f(2,1) & f(2,2) & \dots & f(2,N) \\ \dots & \dots & \dots & \dots \\ f(M,1) & f(M,2) & \dots & f(M,N) \end{bmatrix}$$

Representing Digital Images

- This digitization process requires that decisions be made regarding the values for M , N , and *for the number, L , of discrete intensity levels*.
- *There are no restrictions placed on M and N , other than they have to be positive integers.*
- However, due to storage and quantizing hardware considerations, the number of intensity levels typically is an integer power of 2:
- Discrete intensity interval $[0, L-1]$, $L=2^k$
- The number b of bits required to store a $M \times N$ digitized image

$$b = M \times N \times k$$

- When $M=N$, this equation becomes $b=N^2$

Representing Digital Images

TABLE 2.1

Number of storage bits for various values of N and k .

N/k	1 ($L = 2$)	2 ($L = 4$)	3 ($L = 8$)	4 ($L = 16$)	5 ($L = 32$)	6 ($L = 64$)	7 ($L = 128$)	8 ($L = 256$)
32	1,024	2,048	3,072	4,096	5,120	6,144	7,168	8,192
64	4,096	8,192	12,288	16,384	20,480	24,576	28,672	32,768
128	16,384	32,768	49,152	65,536	81,920	98,304	114,688	131,072
256	65,536	131,072	196,608	262,144	327,680	393,216	458,752	524,288
512	262,144	524,288	786,432	1,048,576	1,310,720	1,572,864	1,835,008	2,097,152
1024	1,048,576	2,097,152	3,145,728	4,194,304	5,242,880	6,291,456	7,340,032	8,388,608
2048	4,194,304	8,388,608	12,582,912	16,777,216	20,971,520	25,165,824	29,369,128	33,554,432
4096	16,777,216	33,554,432	50,331,648	67,108,864	83,886,080	100,663,296	117,440,512	134,217,728
8192	67,108,864	134,217,728	201,326,592	268,435,456	335,544,320	402,653,184	469,762,048	536,870,912

Spatial and Intensity Resolution

- Spatial resolution
 - A measure of the smallest discernible detail in an image
 - stated with *line pairs per unit distance, dots (pixels) per unit distance, dots per inch (dpi)*
- Intensity resolution
 - The smallest discernible change in intensity level
 - stated with *8 bits, 12 bits, 16 bits, etc.*

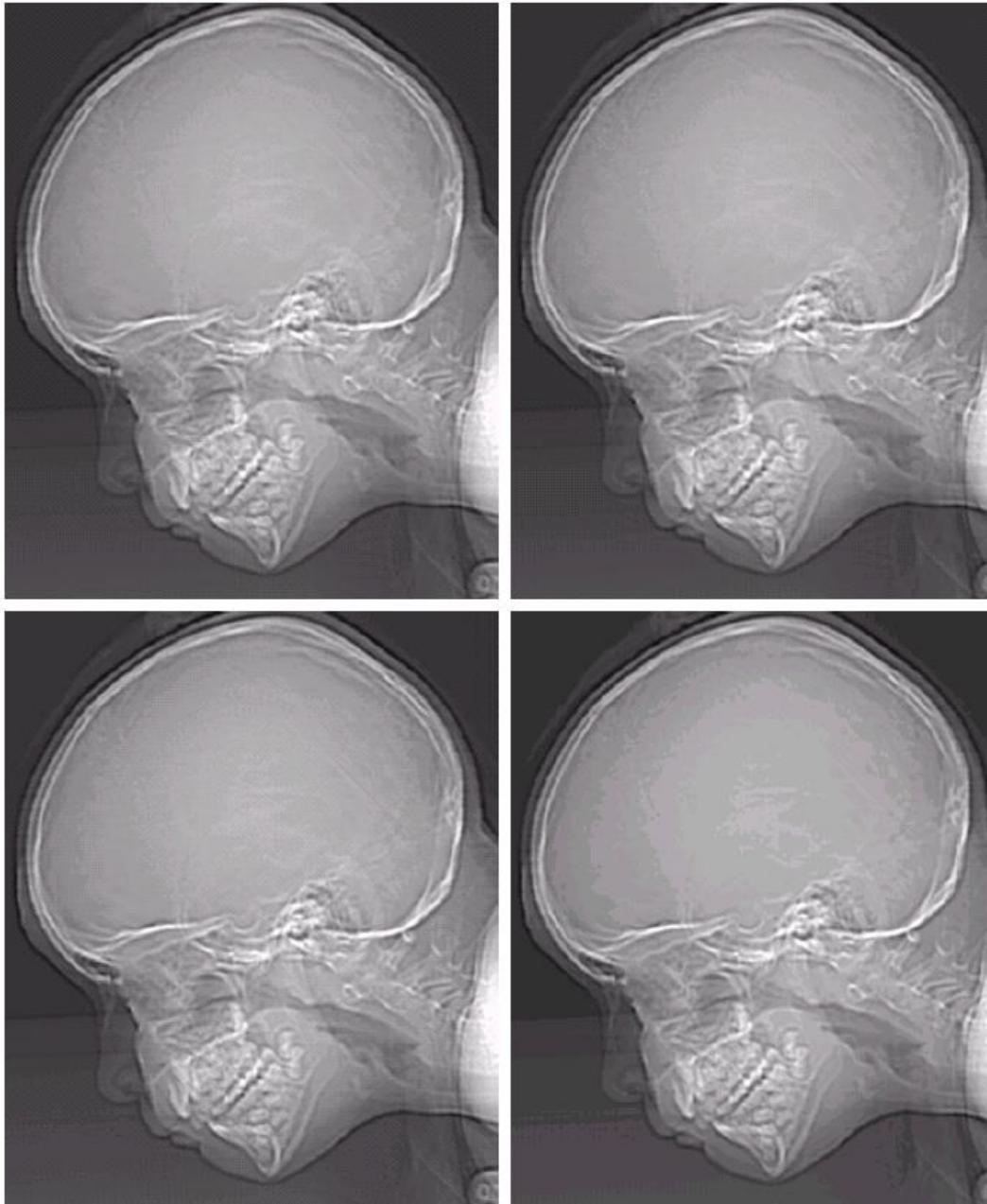
Spatial and Intensity Resolution



Weeks 1 &

FIGURE 2.20 Typical effects of reducing spatial resolution. Images shown at: (a) 1250 dpi, (b) 300 dpi, (c) 150 dpi, and (d) 72 dpi. The thin black borders were added for clarity. They are not part of the data.

Spatial and Intensity Resolution



a
b
c
d

FIGURE 2.21
(a) 452×374 ,
256-level image.
(b)–(d) Image
displayed in 128,
64, and 32 gray
levels, while
keeping the
spatial resolution
constant.

Spatial and Intensity Resolution

e f
g h

FIGURE 2.21
(Continued)
(e)–(h) Image displayed in 16, 8, 4, and 2 gray levels. (Original courtesy of Dr. David R. Pickens, Department of Radiology & Radiological Sciences, Vanderbilt University Medical Center.)

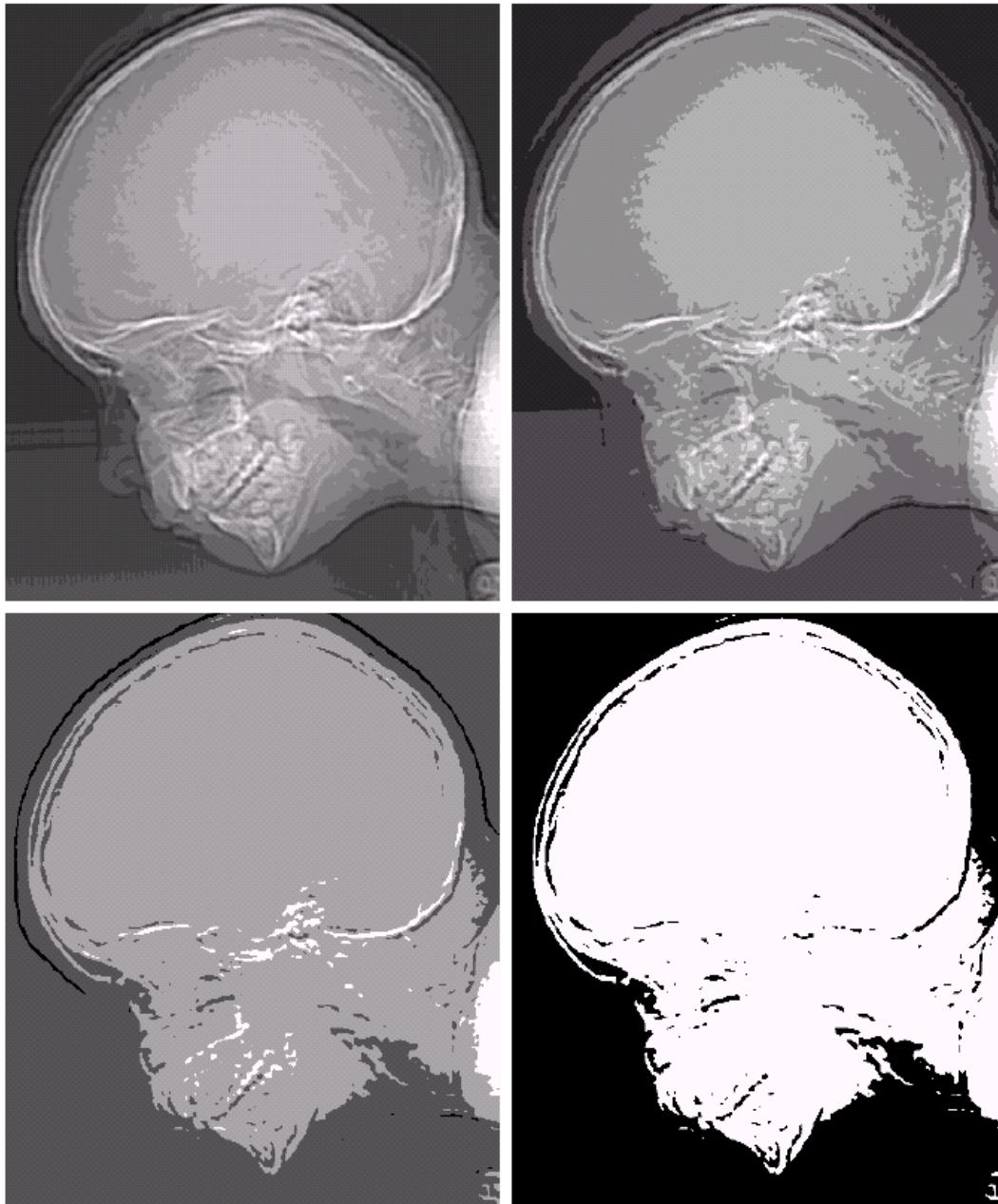


Image Interpolation

- **Interpolation** – Process of using known data to estimate unknown values
 - e.g., zooming, shrinking, rotating, and geometric correction
 - **Interpolation (sometimes called *resampling*)** – an imaging method to increase (or decrease) the number of pixels in a digital image.
 - Some digital cameras use interpolation to produce a larger image than the sensor captured or to create digital zoom
- 1) Nearest Neighbor Interpolation
 - 2) Bilinear Interpolation
 - 3) Bicubic Interpolation

Image Interpolation:

1) Nearest Neighbor Interpolation

$$f_1(x_2, y_2) = f(\text{round}(x_2), \text{round}(y_2)) \\ = f(x_1, y_1)$$

$$f_1(x_3, y_3) = f(\text{round}(x_3), \text{round}(y_3)) \\ = f(x_1, y_1)$$

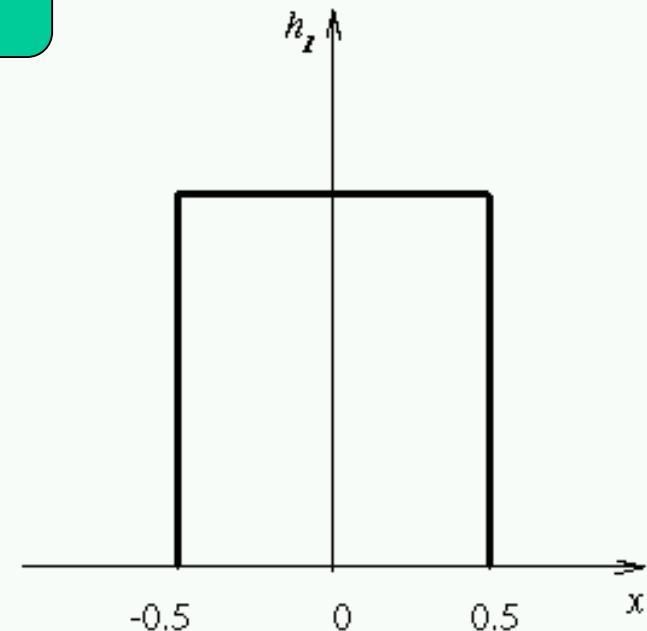
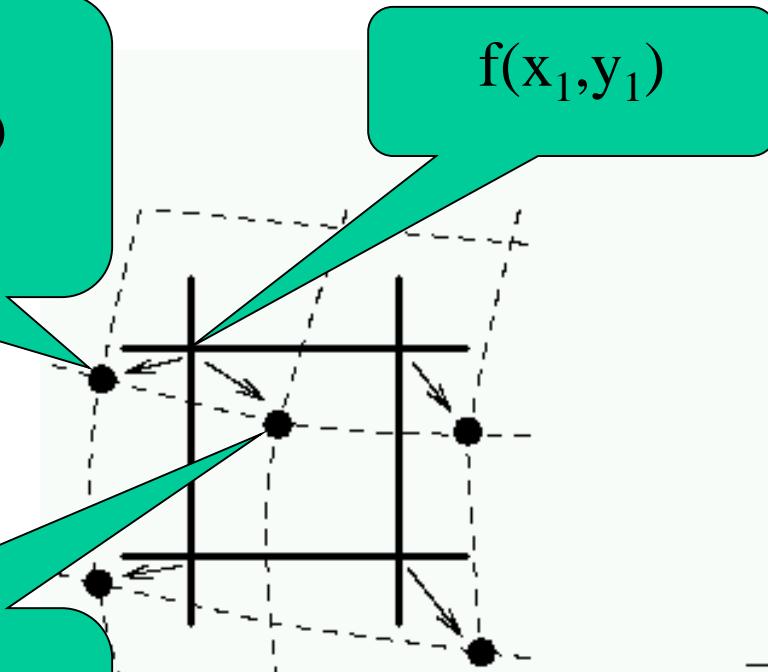
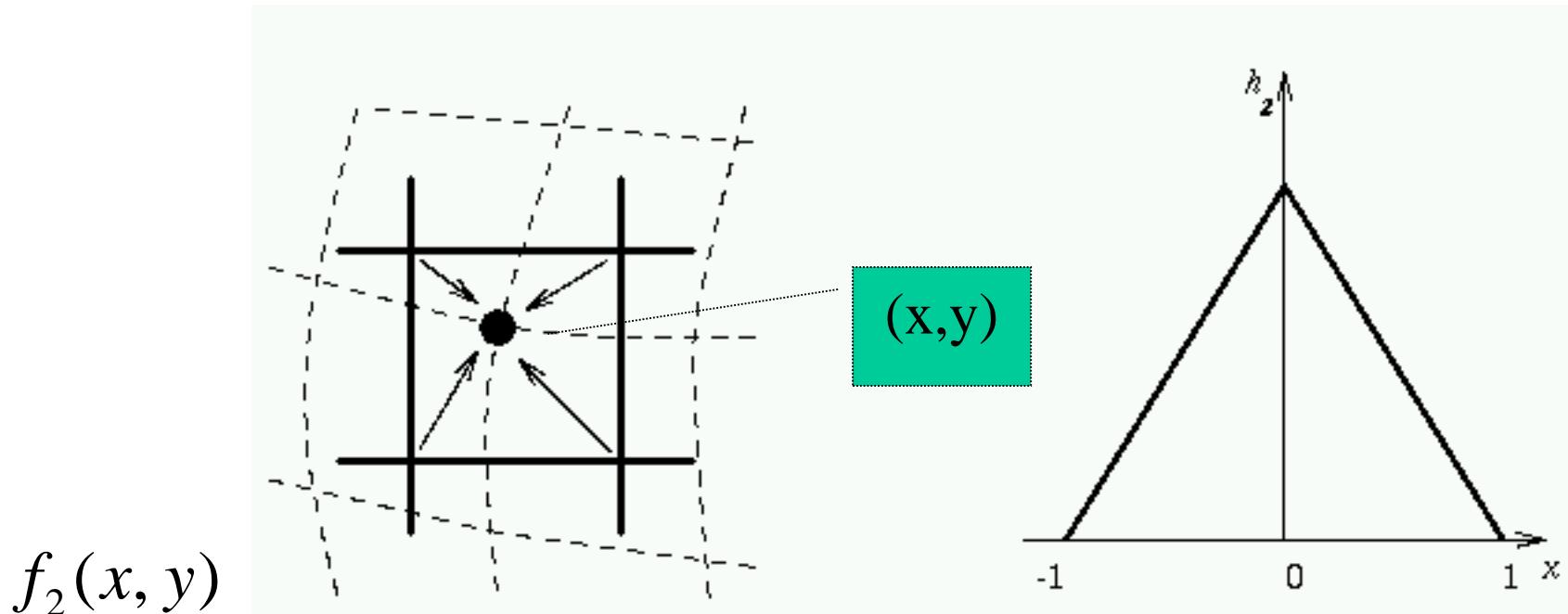


Image Interpolation:

2) Bilinear Interpolation



$$f_2(x, y)$$

$$= (1-a)(1-b)f(l, k) + a(1-b)f(l+1, k)$$

$$+ (1-a)b f(l, k+1) + a b f(l+1, k+1)$$

$$l = \text{floor}(x), k = \text{floor}(y), a = x - l, b = y - k.$$

Image Interpolation:

3) Bicubic Interpolation

- The intensity value assigned to point (x,y) is obtained by the following equation

$$f_3(x, y) = \sum_{i=0}^3 \sum_{j=0}^3 a_{ij} x^i y^j$$

- The sixteen coefficients are determined by using the sixteen nearest neighbors.

Examples: Interpolation



Nearest Neighbor Interpolation



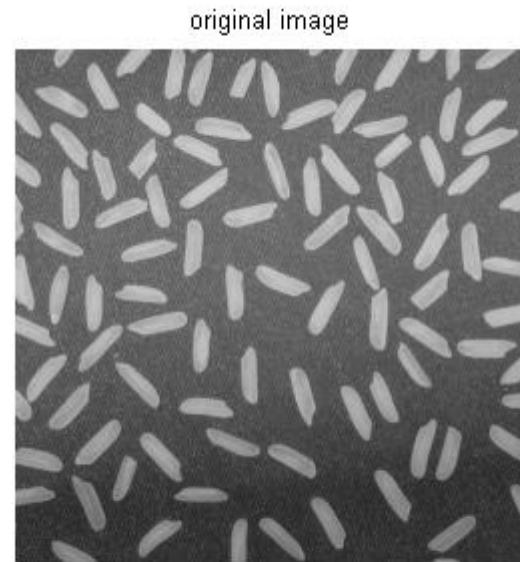
Bilinear Interpolation



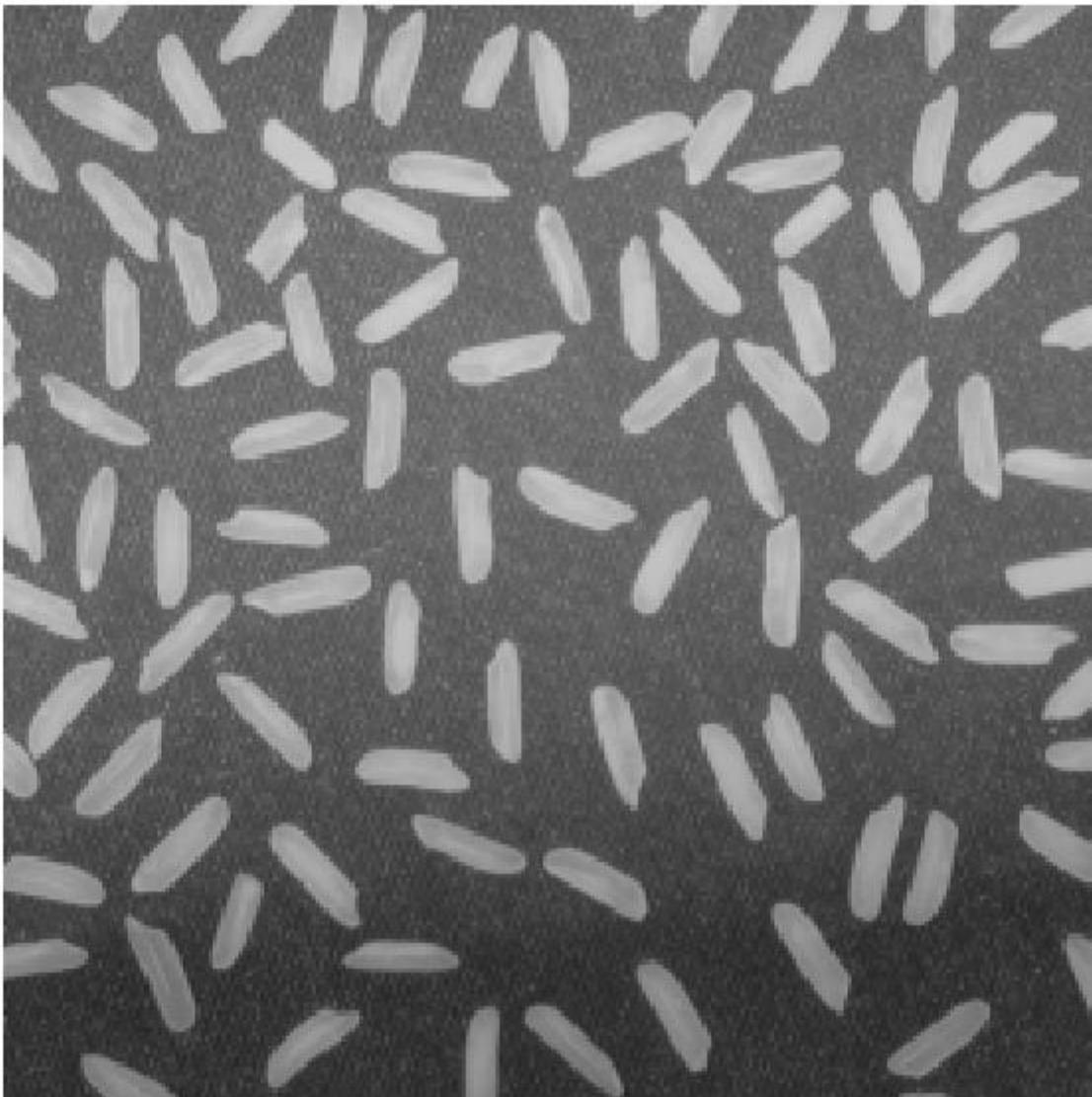
Bicubic Interpolation



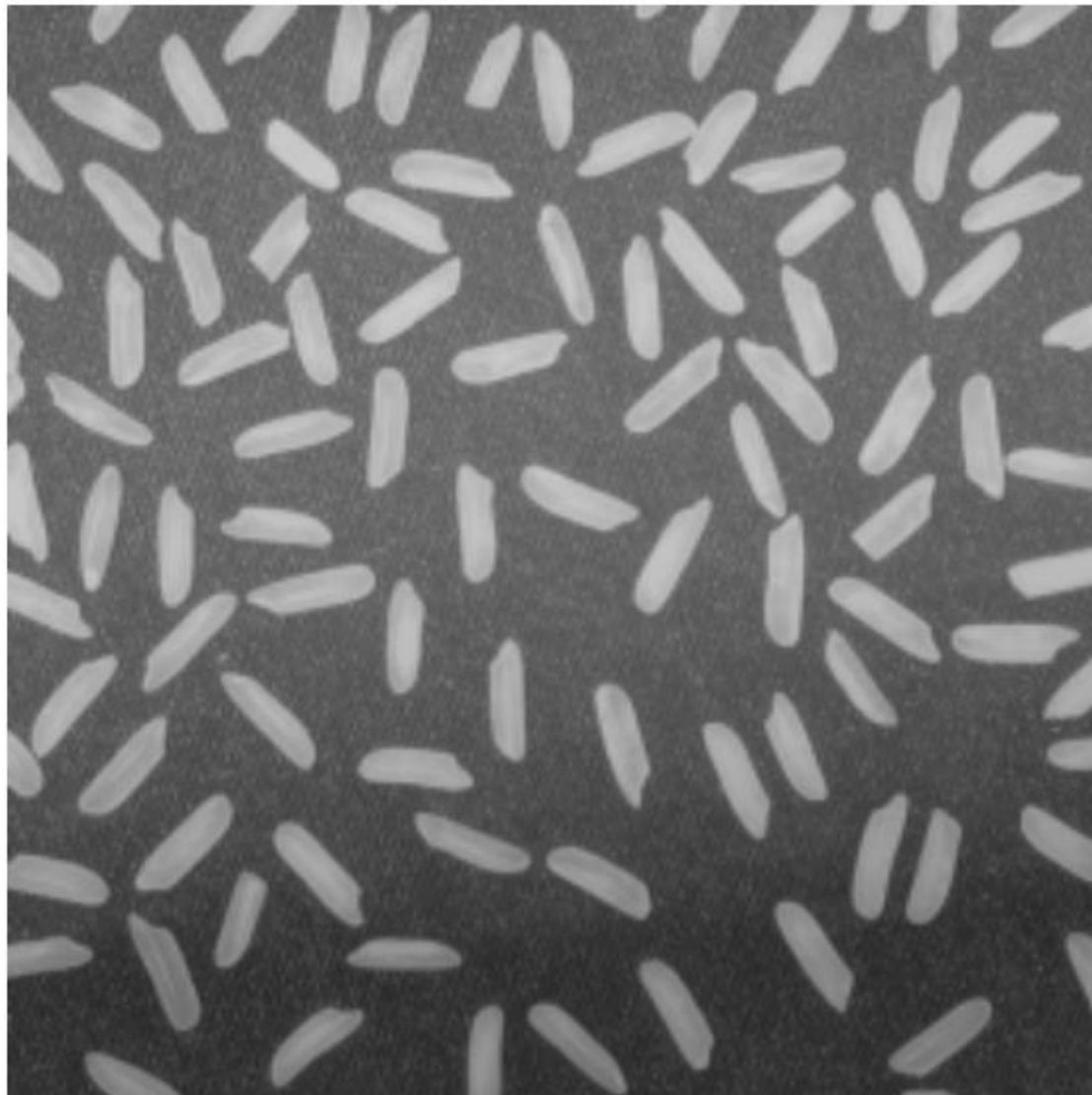
Examples: Interpolation



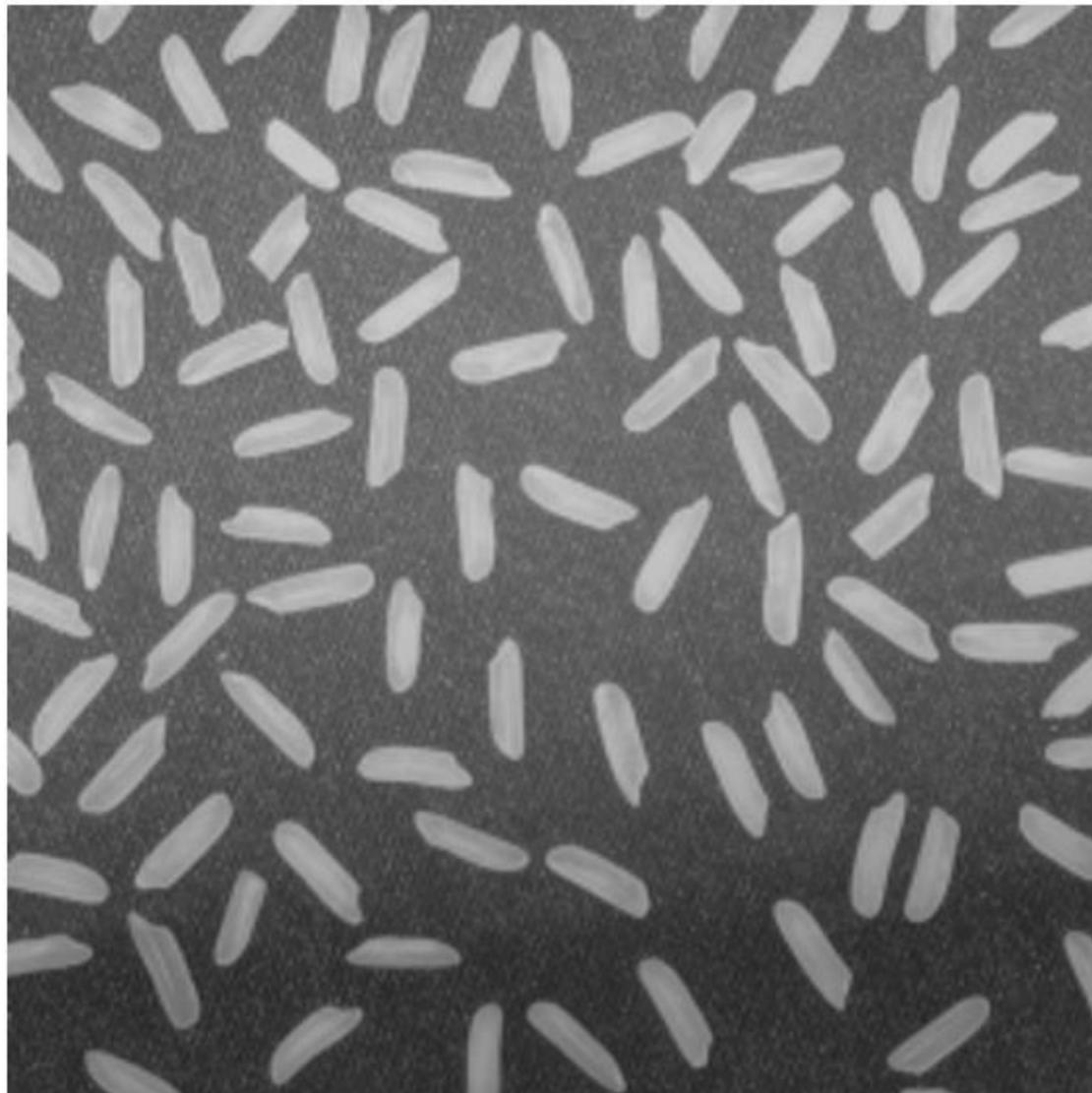
nearest



bilinear



bicubic

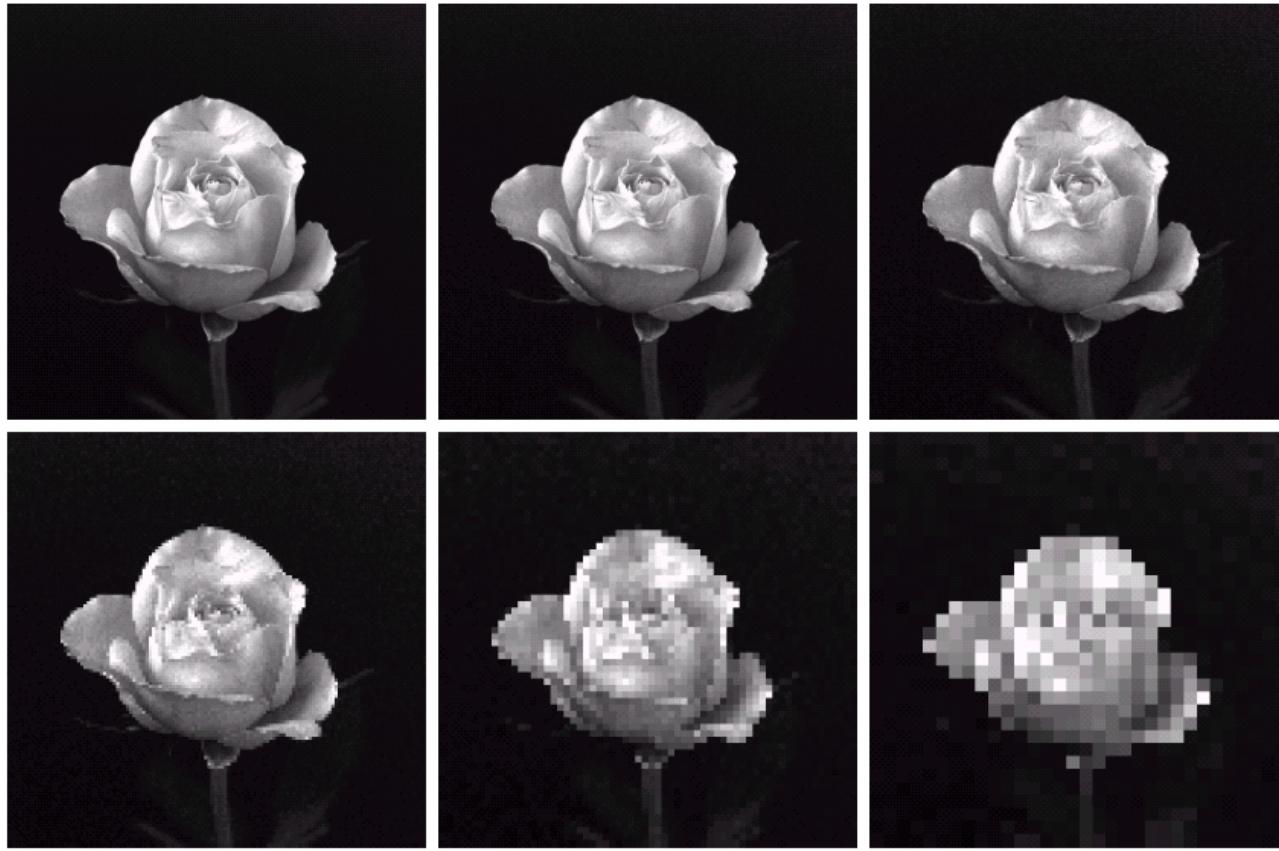


Examples



FIGURE 2.19 A 1024×1024 , 8-bit image subsampled down to size 32×32 pixels. The number of allowable gray levels was kept at 256.

Examples



a b c
d e f

FIGURE 2.20 (a) 1024×1024 , 8-bit image. (b) 512×512 image resampled into 1024×1024 pixels by row and column duplication. (c) through (f) 256×256 , 128×128 , 64×64 , and 32×32 images resampled into 1024×1024 pixels.

Basic Relationship between pixels read from below link

<https://www.slideshare.net/pareshkamble/pixel-relationships>

Region: Let R be a subset of pixels in an image. Two regions R_i and R_j are said to be adjacent if their union form a connected set.

Regions that are not adjacent are said to be disjoint.

We consider 4- and 8- adjacency when referring to regions.

Below regions are adjacent only if 8-adjacency is used.

1	1	1	
1	0	1	R_i
0	1	0	
0	0	1	
1	1	1	R_j
1	1	1	

Boundary (or border)

- The *boundary* of the region R is the set of pixels in the region that have one or more neighbors that are not in R.
- If R happens to be an entire image, then its boundary is defined as the set of pixels in the first and last rows and columns of the image.

Regions & Boundaries

Boundaries (border or contour): The boundary of a region R is the set of points that are adjacent to points in the complement of R.

o	o	o	o	o
o	1	1	o	o
o	1	1	o	o
o	1	1	1	o
o	1	1	1	o
o	o	o	o	o

RED colored 1 is NOT a member of border if 4-connectivity is used between region and background. It is if 8-connectivity is used.

Distance Measures

Distance Measures: Distance between pixels p, q & z with coordinates (x, y), (s, t) & (v, w) resp. is given by:

- a) $D(p, q) \geq 0$ [$D(p, q) = 0$ if $p = q$] called reflexivity
- b) $D(p, q) = D(q, p)$ called symmetry
- c) $D(p, z) \leq D(p, q) + D(q, z)$ called transitivity

Euclidean distance between p & q is defined as-

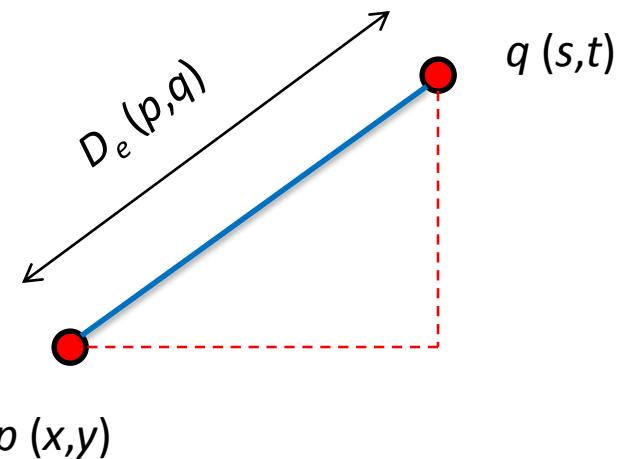
$$D_e(p, q) = [(x - s)^2 + (y - t)^2]^{1/2}$$

Distance Measures

- The *Euclidean Distance* between p and q is defined as:

$$D_e(p,q) = [(x - s)^2 + (y - t)^2]^{1/2}$$

Pixels having a distance less than or equal to some value r from (x,y) are the points contained in a disk of radius r centered at (x,y)



Distance Measures

City Block Distance: The D_4 distance between p & q is defined as

$$D_4(p, q) = |x - s| + |y - t|$$

In this case, pixels having D_4 distance from (x, y) less than or equal to some value r form a diamond centered at (x, y) .



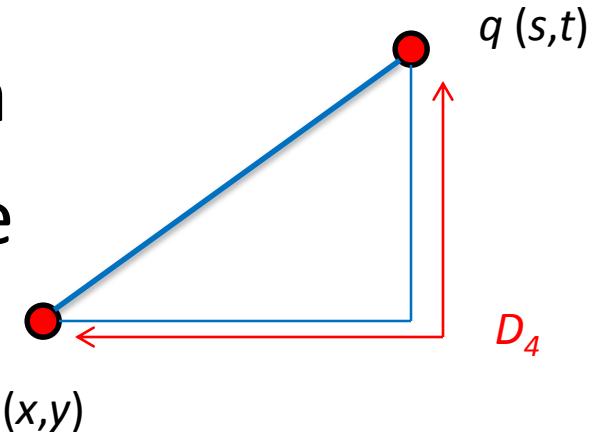
Pixels with D_4 distance ≤ 2 forms the following contour of constant distance.

Distance Measures

- The D_4 distance (also called *city-block distance*) between p and q is defined as:

$$D_4(p,q) = |x - s| + |y - t|$$

Pixels having a D_4 distance from (x,y) , less than or equal to some value r form a Diamond centered at (x,y)



Distance Measures

Example:

The pixels with distance $D_4 \leq 2$ from (x,y) form the following contours of constant distance.

The pixels with $D_4 = 1$ are the 4-neighbors of (x,y)

			2	
	2	1	2	
2	1	0	1	2
	2	1	2	
			2	

Distance Measures

Chess-Board Distance: The D_8 distance between p & q is defined as

$$D_8(p, q) = \max(|x - s|, |y - t|)$$

In this case, pixels having D_8 distance from (x, y) less than or equal to some value r form a square centered at (x, y) .

2	2	2	2	2
2	1	1	1	2
2	1	0	1	2
2	1	1	1	2
2	2	2	2	2

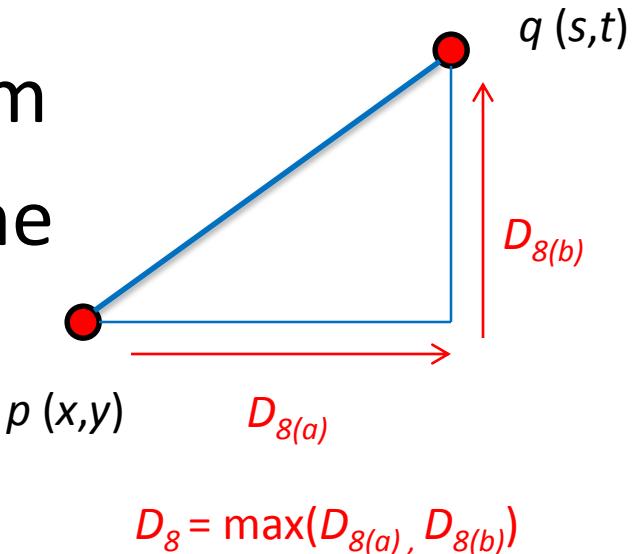
Pixels with D_8 distance ≤ 2 forms the following contour of constant distance.

Distance Measures

- The D_8 distance (also called *chessboard distance*) between p and q is defined as:

$$D_8(p,q) = \max(|x - s|, |y - t|)$$

Pixels having a D_8 distance from (x,y) , less than or equal to some value r form a square centered at (x,y)



Distance Measures

- **D_m distance:**

is defined as the shortest m-path between the points.

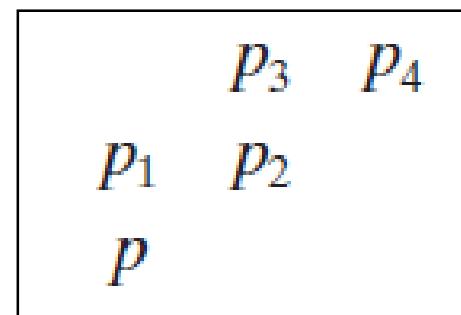
In this case, the distance between two pixels will depend on the values of the pixels along the path, as well as the values of their neighbors.

Distance Measures

- Example:

Consider the following arrangement of pixels and assume that p , p_2 , and p_4 have value 1 and that p_1 and p_3 can have can have a value of 0 or 1

Suppose that we consider the adjacency of pixels values 1 (i.e. $V = \{1\}$)



Distance Measures

- Cont. Example:

Now, to compute the D_m between points p and p_4

Here we have 4 cases:

Case1: If $p_1 = 0$ and $p_3 = 0$

The length of the shortest m-path
(the D_m distance) is 2 (p, p_2, p_4)

	0	1
0	1	
1		

Distance Measures

- Cont. Example:

Case2: If $p_1 = 1$ and $p_3 = 0$

now, p_1 and p_3 will no longer be adjacent (see m -adjacency definition)

then, the length of the shortest path will be 3 (p, p_1, p_2, p_4)

	0	1
1	1	
1		

Distance Measures

- Cont. Example:

Case3: If $p_1 = 0$ and $p_3 = 1$

The same applies here, and the shortest –m-path will be 3 (p, p_2, p_3, p_4)

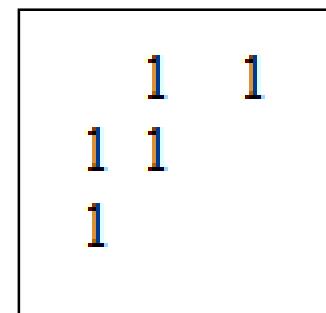
		1	1
0	1		
1			

Distance Measures

- Cont. Example:

Case4: If $p_1 = 1$ and $p_3 = 1$

The length of the shortest m-path will be 4 (p, p_1, p_2, p_3, p_4)



Relationship between pixels (Contd..)

Arithmetic/Logic Operations:

- **Addition :** $p + q$
- **Subtraction:** $p - q$
- **Multiplication:** $p * q$
- **Division:** p/q
- **AND:** $p \text{ AND } q$
- **OR :** $p \text{ OR } q$
- **Complement:** $\text{NOT}(q)$

Arithmetic/Logic Operations

- Tasks done using neighborhood processing:
 - Smoothing / averaging
 - Noise removal / filtering
 - Edge detection
 - Contrast enhancement

Linear and nonLinear operations

- Important classification of an image-processing method

- Is it a linear or a nonlinear method ?

- Let H be a general operator

$$H[f(x, y)] = g(x, y)$$

- H is said to be a ***linear operator*** if

$$H[a_i f_i(x, y) + a_j f_j(x, y)] = a_i H[f_i(x, y)] + a_j H[f_j(x, y)]$$

$$= a_i g_i(x, y) + a_j g_j(x, y)$$

► Linear vs. Nonlinear Operation

$$H[f(x, y)] = g(x, y)$$

$$H[a_i f_i(x, y) + a_j f_j(x, y)]$$

Additivity

$$= H[a_i f_i(x, y)] + H[a_j f_j(x, y)]$$

Homogeneity

$$= a_i H[f_i(x, y)] + a_j H[f_j(x, y)]$$

$$= a_i g_i(x, y) + a_j g_j(x, y)$$

H is said to be a **linear operator**;

H is said to be a **nonlinear operator** if it does not meet the above qualification.

As a simple example, suppose that H is the sum operator, Σ ; that is, the function of this operator is simply to sum its inputs. To test for linearity, we start with the left side of Eq. (2.6-2) and attempt to prove that it is equal to the right side:

$$\begin{aligned}\sum[a_i f_i(x, y) + a_j f_j(x, y)] &= \sum a_i f_i(x, y) + \sum a_j f_j(x, y) \\ &= a_i \sum f_i(x, y) + a_j \sum f_j(x, y) \\ &= a_i g_i(x, y) + a_j g_j(x, y)\end{aligned}$$

where the first step follows from the fact that summation is distributive. So, an expansion of the left side is equal to the right side of Eq. (2.6-2), and we conclude that the sum operator is linear.

On the other hand, consider the max operation, whose function is to find the maximum value of the pixels in an image. For our purposes here, the simplest way to prove that this operator is nonlinear, is to find an example that fails the test in Eq. (2.6-2). Consider the following two images

$$f_1 = \begin{bmatrix} 0 & 2 \\ 2 & 3 \end{bmatrix} \quad \text{and} \quad f_2 = \begin{bmatrix} 6 & 5 \\ 4 & 7 \end{bmatrix}$$

and suppose that we let $a_1 = 1$ and $a_2 = -1$. To test for linearity, we again start with the left side of Eq. (2.6-2):

$$\begin{aligned} \max \left\{ (1) \begin{bmatrix} 0 & 2 \\ 2 & 3 \end{bmatrix} + (-1) \begin{bmatrix} 6 & 5 \\ 4 & 7 \end{bmatrix} \right\} &= \max \left\{ \begin{bmatrix} -6 & -3 \\ -2 & -4 \end{bmatrix} \right\} \\ &= -2 \end{aligned}$$

Working next with the right side, we obtain

$$\begin{aligned} (1) \max \left\{ \begin{bmatrix} 0 & 2 \\ 2 & 3 \end{bmatrix} \right\} + (-1) \max \left\{ \begin{bmatrix} 6 & 5 \\ 4 & 7 \end{bmatrix} \right\} &= 3 + (-1)7 \\ &= -4 \end{aligned}$$

The left and right sides of Eq. (2.6-2) are not equal in this case, so we have proved that in general the max operator is nonlinear.

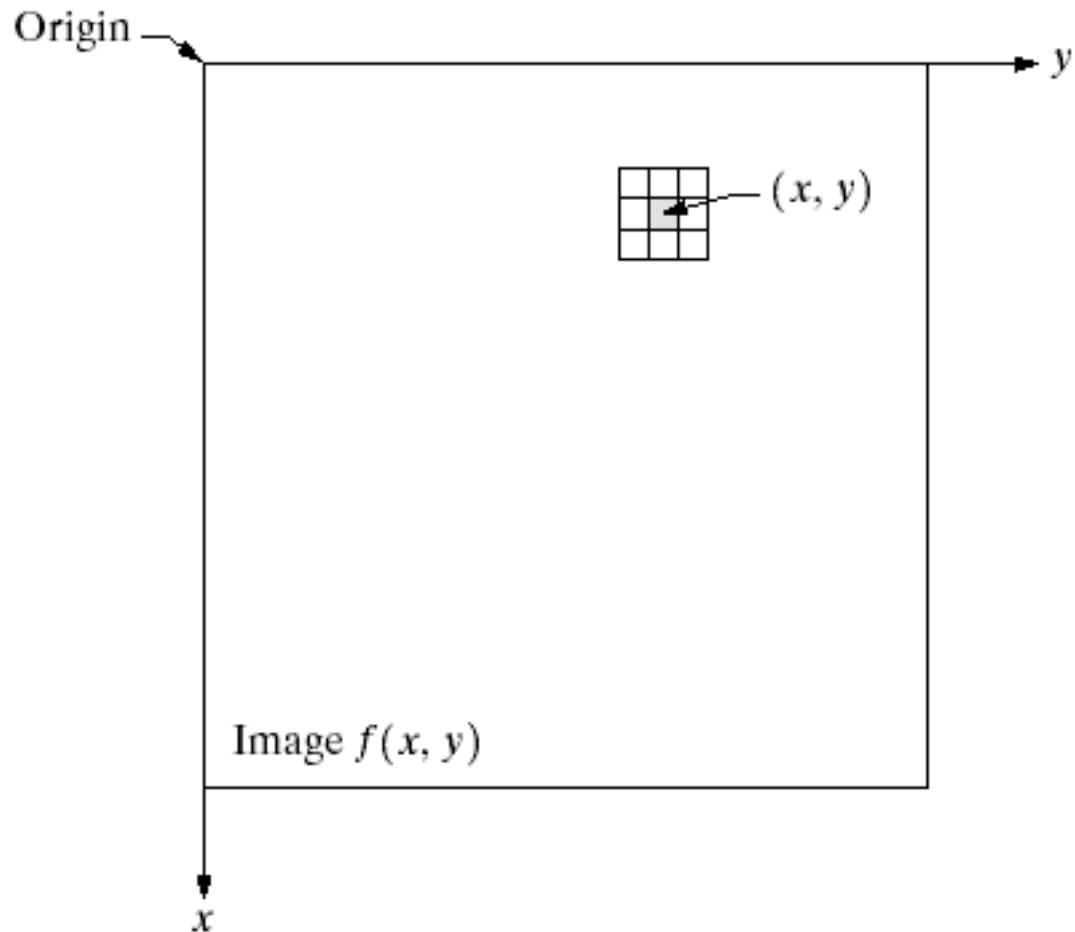
Fundamentals of Spatial Filtering

- Spatial filtering is one of the principal tools used in image processing for image enhancement.
- Filtering refers to accepting(passing) or rejecting certain frequency components. This effectively smoothens or sharpens the image.
- E.g. Low pass filter, high pass filter, etc.
- Such operations can be directly carried out on image in spatial domain also by using spatial filters (kernels, spatial masks, templates, & windows).
- Spatial filters are more versatile as they are used in linear as well as non-linear filtering (Difficult in frequency domain).

- Types of Spatial Filtering
- 1) Point to point (pixel to pixel) operation (discussed so far)
- 2) Mask based (Neighborhood) operations
 - i) Operation with 3x3 filter (E.g. Mean, max, min, etc)
 - ii) Correlation or Convolution
- Linear vs Non-Linear Filter
 - If the operation performed on the image pixels is linear, then the filter is called a linear spatial filter, otherwise nonlinear.

Basics of spatial filtering

- $g(x,y) = T[f(x,y)]$
- T operates on a neighborhood



The Mechanics of Spatial Filtering

- The **mechanics** of spatial filtering spatial filters consists of:
 1. Neighbourhood (small rectangle).
 2. Predefined operation that is performed on the image pixel.

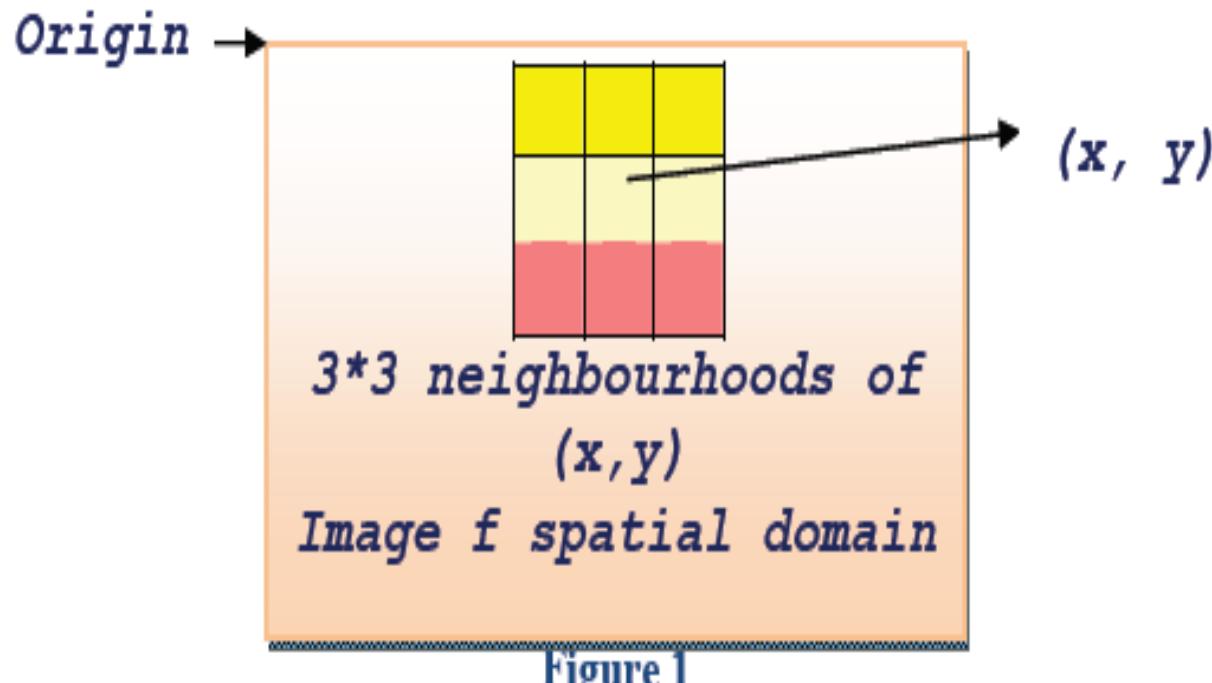
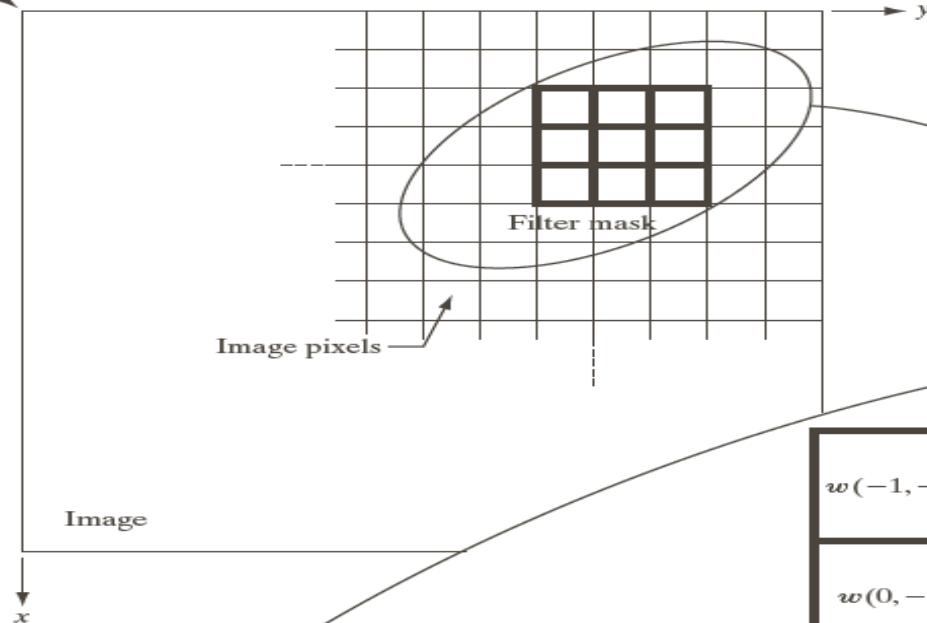


Figure 1

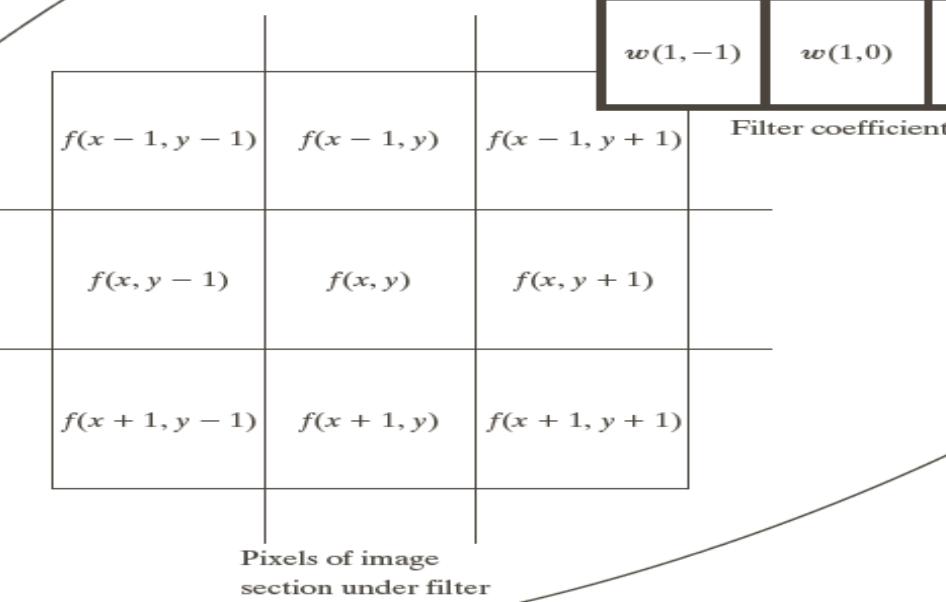
- **Filtering** creates new pixel with coordinates equal to the coordinates of the centre of the neighbourhood, and whose value is the result of the filtering operation.

Image origin



Image

x



Pixels of image
section under filter

Image origin

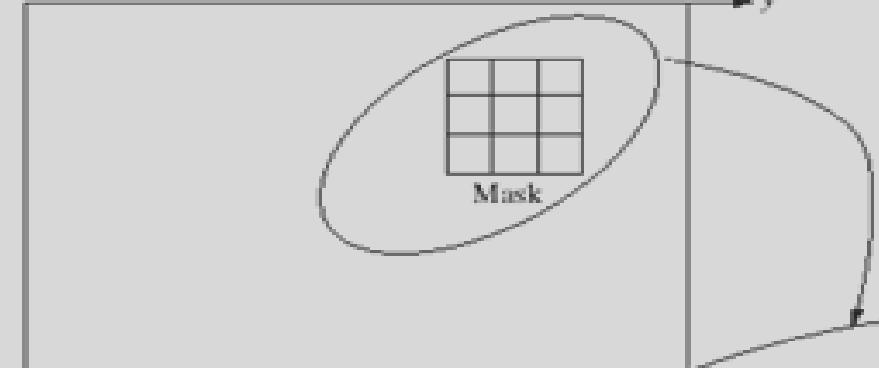


Image $f(x, y)$

$w(-1, -1)$	$w(-1, 0)$	$w(-1, 1)$
$w(0, -1)$	$w(0, 0)$	$w(0, 1)$
$w(1, -1)$	$w(1, 0)$	$w(1, 1)$

Mask coefficients, showing coordinate arrangement

$f(x - 1, y - 1)$	$f(x - 1, y)$	$f(x - 1, y + 1)$
$f(x, y - 1)$	$f(x, y)$	$f(x, y + 1)$
$f(x + 1, y - 1)$	$f(x + 1, y)$	$f(x + 1, y + 1)$

Pixels of image section under mask

FIGURE 3.32 The mechanics of spatial filtering. The magnified drawing shows a 3×3 mask and the image section directly under it; the image section is shown displaced out from under the mask for ease of readability.

mask coefficients

X (product) → output

underlying neighborhood

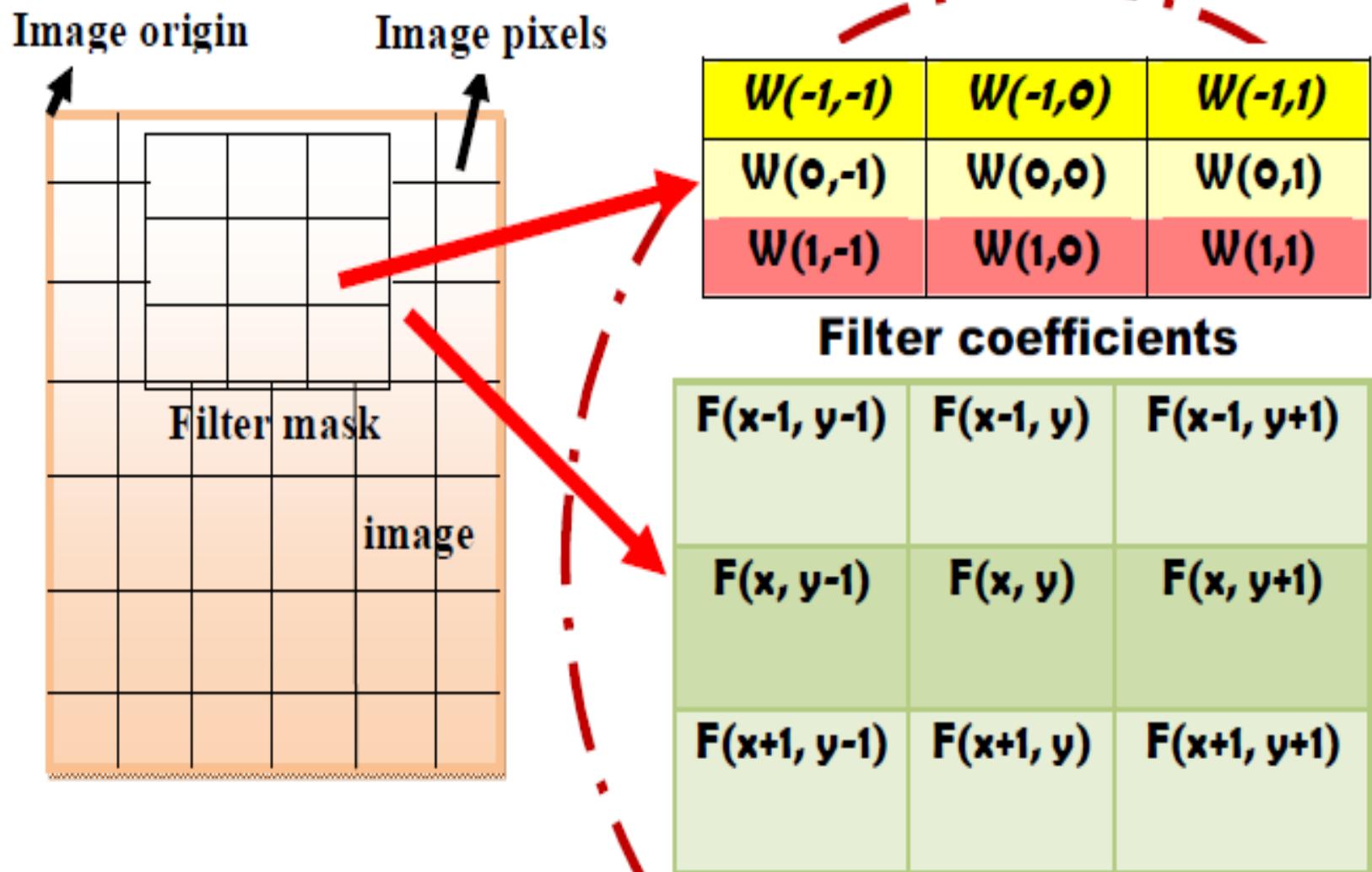


Figure 1

Operation with 3x3 Filter

- At any point (x, y) in the image, the response, $g(x, y)$, of the filter is the sum of products of the filter coefficients and the image pixels encompassed by the filter:

$$\begin{aligned} g(x, y) = & f(x-1, y-1).w_1 + f(x-1, y).w_2 + f(x-1, y+1).w_3 \\ & + f(x, y-1).w_4 + f(x, y).w_5 + f(x, y+1).w_6 \\ & + f(x+1, y-1).w_7 + f(x+1, y).w_8 + f(x+1, y+1).w_9 \end{aligned}$$

- For the mask of size $m \times n$, we assume

$$m = 2a + 1;$$

$$n = 2b + 1;$$

where a & b are positive integers.

- 3x3 is the smallest filter.

(Odd filters)

In general, linear spatial filtering of an image of size $M * N$ with a filter of size $m * n$ is given by the expression:

$$g(x, y) = \sum_{s=-a}^a \sum_{t=-b}^b w(s, t) \cdot f(x + s, y + t)$$

Where x and y are varied so that each pixel in w visits every pixel in f .

Operation with 3x3 Filter

- **3 x 3 Neighborhood / Mask / Window / Template:**

	(y - 1)	y	(y + 1)	Y
(x - 1)	$w(-1, -1)$ $f(x-1, y-1)$	$w(-1, 0)$ $f(x-1, y)$	$w(-1, 1)$ $f(x-1, y+1)$	
x	$w(0, -1)$ $f(x, y-1)$	$w(0, 0)$ $f(x, y)$	$w(0, 1)$ $f(x, y+1)$	
(x + 1)	$w(1, -1)$ $f(x+1, y-1)$	$w(1, 0)$ $f(x+1, y)$	$w(1, 1)$ $f(x+1, y+1)$	

Spatial Correlation and convolution

- Correlation & Convolution are two closely related concepts used in linear spatial filtering.
- *Correlation*: It is a process of moving a filter mask over an image & computing the sum of products at each location.
- *Convolution*: Here, the mechanics are same, except that the filter is first rotated by 180° .
- Correlation & Convolution are function of displacement. Correlation & Convolution are exactly same if the filter mask is symmetric.
- 1D correlation and convolution of a filter with a discrete unit impulse is shown below.

Correlation

Convolution

Origin	<i>f</i>	<i>w rotated 180°</i>	
0 0 0 1 0 0 0		8 2 3 2 1	(i)
	0 0 0 1 0 0 0		(j)
8 2 3 2 1			

(c) 

$$\begin{matrix} 1 & 2 & 3 & 2 & 8 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{matrix}$$

(d) 

$$\begin{matrix} 1 & 2 & 3 & 2 & 8 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{matrix}$$

↑ Position after one shift

(e) 

$$\begin{matrix} 1 & 2 & 3 & 2 & 8 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{matrix}$$

↑ Position after four shifts

(f) 

$$\begin{matrix} 1 & 2 & 3 & 2 & 8 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{matrix}$$

↑ Final position

(e) 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0
 1 2 3 2 8
 ↑ Position after four shifts

0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 (m)
8 2 3 2 1

0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 (n)
 8 2 3 2 1

Full correlation result

(g) 0 0 0 8 2 3 2 1 0 0 0 0

Full convolution result

0 0 0 1 2 3 2 8 0 0 0 0 (o)

Cropped correlation result

(h) 0 8 2 3 2 1 0 0

Cropped convolution result

$$0 \ 1 \ 2 \ 3 \ 2 \ 8 \ 0 \ 0 \quad (\text{B})$$

Correlation & Convolution

- Correlation is a function of displacement of the filter.
- Correlating a filter w with a function that contains all '0' & single '1' yields a 180° rotated copy of w .
- Correlating a function with discrete unit impulse yields a rotated (time inverted) version of the function.
- Convolving a function with a unit impulse yields the same function.
- Thus, to perform convolution all we have to do is rotate one function by 180° & perform same operation as in correlation.

FIGURE 3.30
 Correlation
 (middle row) and
 convolution (last
 row) of a 2-D
 filter with a 2-D
 discrete, unit
 impulse. The 0s
 are shown in gray
 to simplify visual
 analysis.

Initial position for w						
1	2	3	0	0	0	0
4	5	6	0	0	0	0
7	8	9	0	0	0	0

Rotated <i>w</i>		
9	8	7
6	5	4
3	2	1

Full correlation result

Full convolution result

(c) (d)

Estimated as **Full communication content**

→ Rotated 10° Full convolution result

9 8 7 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0

3 **2** **1** 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0

0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 4 5 6 0 0 0 0

• **W**hen you're writing a research paper, it's important to cite your sources correctly. This will help you avoid plagiarism and give credit where it's due.

Cropped correlation result				
0	0	0	0	0
0	9	8	7	0
0	6	5	4	0
0	3	2	1	0
0	0	0	0	0

(e)

Cropped convolution result

0	0	0	0	0
0	1	2	3	0
0	4	5	6	0
0	7	8	9	0
0	0	0	0	0

10

Summary:

- **Correlation** of a filter $w(x, y)$ of size $m * n$ with an image $f(x, y)$ denoted as

$$W(x, y) \circ f(x, y) = \sum_{s=-a}^a \sum_{t=-b}^b w(s, t)f(x + s, y + t)$$

- In similar manner, the **convolution** of $w(x, y)$ and $f(x, y)$ denoted by $w(x, y) * f(x, y)$ is given by:

$$W(x, y) * f(x, y) = \sum_{s=-a}^a \sum_{t=-b}^b w(s, t)f(x - s, y - t)$$

Where the minus sign on the right flip (rotate by 180°)

(We can flip and shift either f or w)

Vector representation of Linear Filtering

w_1	w_2	w_3
w_4	w_5	w_6
w_7	w_8	w_9

$$R = w_1 z_1 + w_2 z_2 + \dots + w_9 z_9$$

$$= \sum_{k=1}^9 w_k z_k \\ = w^T z$$

Where, w & z are 9-dimensional vectors formed from coefficients of the mask & image intensities encompassed by the mask, resp.

Generating Spatial Filter Masks

General implementation for filtering an $M \times N$ image with a weighted average filter of size $m \times n$ is given by:

$$g(x, y) = \frac{\sum_{s=-a}^a \sum_{t=-b}^b w(s, t) f(x + s, y + t)}{\sum_{s=-a}^a \sum_{t=-b}^b w(s, t)}$$

2) Exponential Filter

- Some applications have a continuous function of 2 variables.
E.g. Gaussian function Spatial filter mask has the basic form:

$$h(x, y) = e^{-\frac{x^2 + y^2}{2\sigma^2}}$$

where, σ is standard deviation

Smoothing Spatial Filter

- Smoothing filters are used for
 - ❖ blurring
 - ❖ noise reduction.
- **Blurring** is used in preprocessing steps to removal of small details from an image prior to object extraction and bridging of small gaps in lines or curves
- **Noise reduction** can be accomplished by blurring

Types of Smoothing Filter

There are 2 way of smoothing spatial filters

- **Linear Filters** – operations performed on image pixel
- **Order-Statistics (non-linear) Filters** - based on ranking the pixels

Linear Filter

- Linear spatial filter is simply the **average** of the pixels contained in the **neighborhood** of the filter mask.
- The idea is **replacing** the value of **every pixel** in an image by the **average** of the gray levels in the **neighborhood** defined by the filter mask.

Linear Filter (cont..)

- This process result in an image **reduce the sharp transitions in intensities.**
- Two mask
 - **Averaging filter**
 - **Weighted averaging filter**

Averaging Filter

- A major use of averaging filters is in the reduction of irrelevant detail in image.
- $m \times n$ mask would have a normalizing constant equal to $1/mn$.
- Its also known as low pass filter.
- A spatial averaging filter in which all coefficients are equal is called a box filter.

Averaging Filter - Example

$$\frac{1}{9} \times \begin{array}{|c|c|c|} \hline 1 & 1 & 1 \\ \hline 1 & 1 & 1 \\ \hline 1 & 1 & 1 \\ \hline \end{array}$$

Generating Spatial Filter Masks

1) Average Mean Filter

- The average value at any location (x, y) in the image is the sum of the nine intensity values in the 3×3 neighborhood centered on (x, y) divided by 9.
- If z_i , $i = 1, 2, \dots, 9$ denote these intensities, then the average is:

$$R = \frac{1}{9} \sum_{i=1}^9 z_i$$

$$\frac{1}{9} \times \begin{array}{|c|c|c|} \hline 1 & 1 & 1 \\ \hline 1 & 1 & 1 \\ \hline 1 & 1 & 1 \\ \hline \end{array}$$
$$\frac{1}{16} \times \begin{array}{|c|c|c|} \hline 1 & 2 & 1 \\ \hline 2 & 4 & 2 \\ \hline 1 & 2 & 1 \\ \hline \end{array}$$

Weighted Averaging Filter

- Pixels are multiplied by different coefficients, thus giving more weight to some pixel at the expense of others.
- The center pixel is multiplied by a higher value than any other, thus giving the pixel more importance in the calculation of average.
- The other pixels are inversely weighted as a function of their distance from center of mask

Weighted Averaging Filter

- The general implementation for filtering an MxN image with a weighted averaging filter of size m x n is given by the expression

$$g(x, y) = \frac{\sum_{s=-a}^a \sum_{t=-b}^b w(s, t) f(x + s, y + t)}{\sum_{s=-a}^a \sum_{t=-b}^b w(s, t)}$$

- For complete filtered image apply $x = 0, 1, 2, 3, \dots, m-1$ and $y = 0, 1, 2, 3, \dots, n-1$ in the above equation.

Weighted Average Filter - Example

$$\frac{1}{16} \times \begin{array}{|c|c|c|} \hline 1 & 2 & 1 \\ \hline 2 & 4 & 2 \\ \hline 1 & 2 & 1 \\ \hline \end{array}$$

Order-Statistics Filter

- Order-statistics filters are **nonlinear spatial filters**.
- It is based on **ordering (ranking) the pixels**
(increasing / decreasing)
contained in the image area encompassed by the filter,
- It **replacing** the value of the **center pixel** with the value determined by the **ranking result**.
- The filter selects a sample from the window, **does not average**
- **Edges** are better **preserved** than with liner filters
- Best suited for “salt and pepper” noise

Order-Statistic Filters

Ex. 2) 8x8 Pseudo image with a single edge (High Frequency) of 10 & 50. Remove using a 3x3 size median filter mask.

10	10	10	10	10	10	10	10
10	10	10	10	10	10	10	10
10	250	10	10	10	10	10	10
10	10	10	10	10	10	10	10
50	50	50	50	250	50	50	50
50	50	50	50	50	50	50	50
50	50	50	50	50	50	50	50
50	50	50	50	50	50	50	50

8x8 Image

10	10	10	10	10	10	10	10
10	10	10	10	10	10	10	10
10	250	10	10	10	10	10	10
10	10	10	10	10	10	10	10
50	50	50	50	250	50	50	50
50	50	50	50	50	50	50	50
50	50	50	50	50	50	50	50
50	50	50	50	50	50	50	50

10	10	10	10	10	10	10	10
10	10	10	10	10	10	10	10
10	250	10	10	10	10	10	10
10	10	10	10	10	10	10	10
50	50	50	50	250	50	50	50
50	50	50	50	50	50	50	50
50	50	50	50	50	50	50	50
50	50	50	50	50	50	50	50

10	10	10	10	10	10	10	10
10	10	10	10	10	10	10	10
10	10	10	10	10	10	10	10
10	10	10	10	10	10	10	10
50	50	50	50	250	50	50	50
50	50	50	50	50	50	50	50
50	50	50	50	50	50	50	50
50	50	50	50	50	50	50	50

10	10	10	10	10	10	10	10
10	10	10	10	10	10	10	10
10	10	10	10	10	10	10	10
10	10	10	10	10	10	10	10
50	50	50	50	250	50	50	50
50	50	50	50	50	50	50	50
50	50	50	50	50	50	50	50
50	50	50	50	50	50	50	50

10	10	10	10	10	10	10	10
10	10	10	10	10	10	10	10
10	10	10	10	10	10	10	10
10	10	10	10	10	10	10	10
50	50	50	50	250	50	50	50
50	50	50	50	50	50	50	50
50	50	50	50	50	50	50	50
50	50	50	50	50	50	50	50

10	10	10	10	10	10	10	10
10	10	10	10	10	10	10	10
10	10	10	10	10	10	10	10
10	10	10	10	10	10	10	10
50	50	50	50	50	50	50	50
50	50	50	50	50	50	50	50
50	50	50	50	50	50	50	50
50	50	50	50	50	50	50	50

Types of order-statics filter

Different types of order-statics filters are

- Minimum filter
- Maximum filter
- Median filter

Minimum Filter

- The 0th percentile filter is the min filter.
- Minimum filter selects the smallest value in the window and replace the center by the smallest value
- Using comparison the minimum value can be obtained fast.(not necessary to sort)
- It enhances the dark areas of image

Minimum Filter - Example



(mask size = 3×3)



(mask size = 7×7)

Maximum Filter

- The maximum filter selects the largest value within of pixel values, and replace the center by the largest value.
 - Using comparison the maximum value can be obtained fast.(not necessary to sort)
 - Using the 100th percentile results in the so-called *max filter*
 - it enhances bright areas of image
- It is used to find the brightest points in an image.
– Response of a 3×3 max filter is given by

$$R = \max\{z_k | k = 1, 2, \dots, 9\}$$

Maximum Filter



mask (3 × 3)



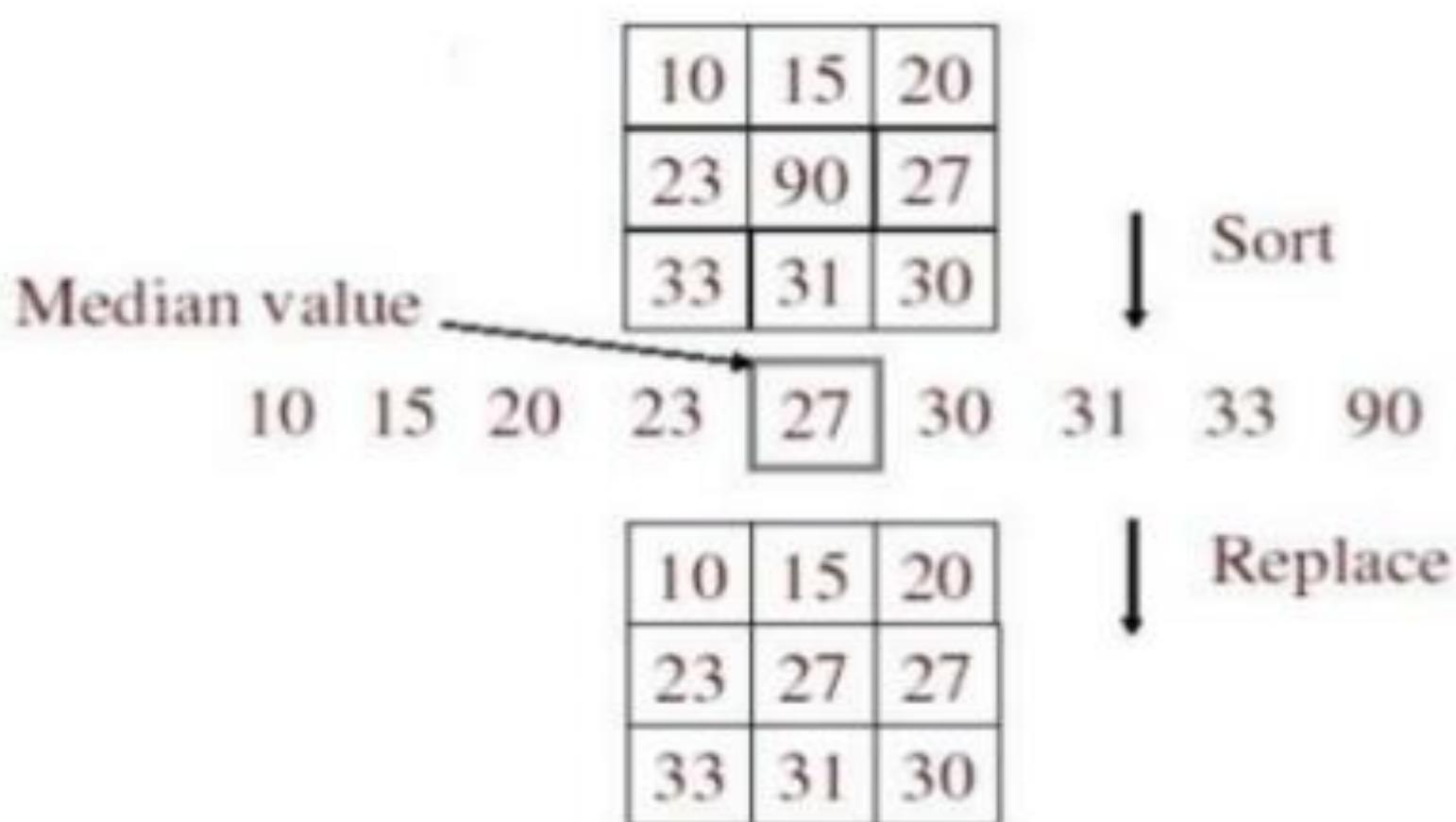
mask (7 × 7)

Median Filter

Three steps to be followed to run a median filter:

1. Consider each pixel in the image
 2. Sort the neighboring pixels into order based upon their intensities
 3. Replace the original value of the pixel with the median value from the list.
- Popular with certain random noise and impulse noise (Salt & Pepper noise).
 - They provide excellent noise reduction
 - Comparatively less blurring than linear smoothing filter of same size.

Median Filter - Process



Median Filter - Example

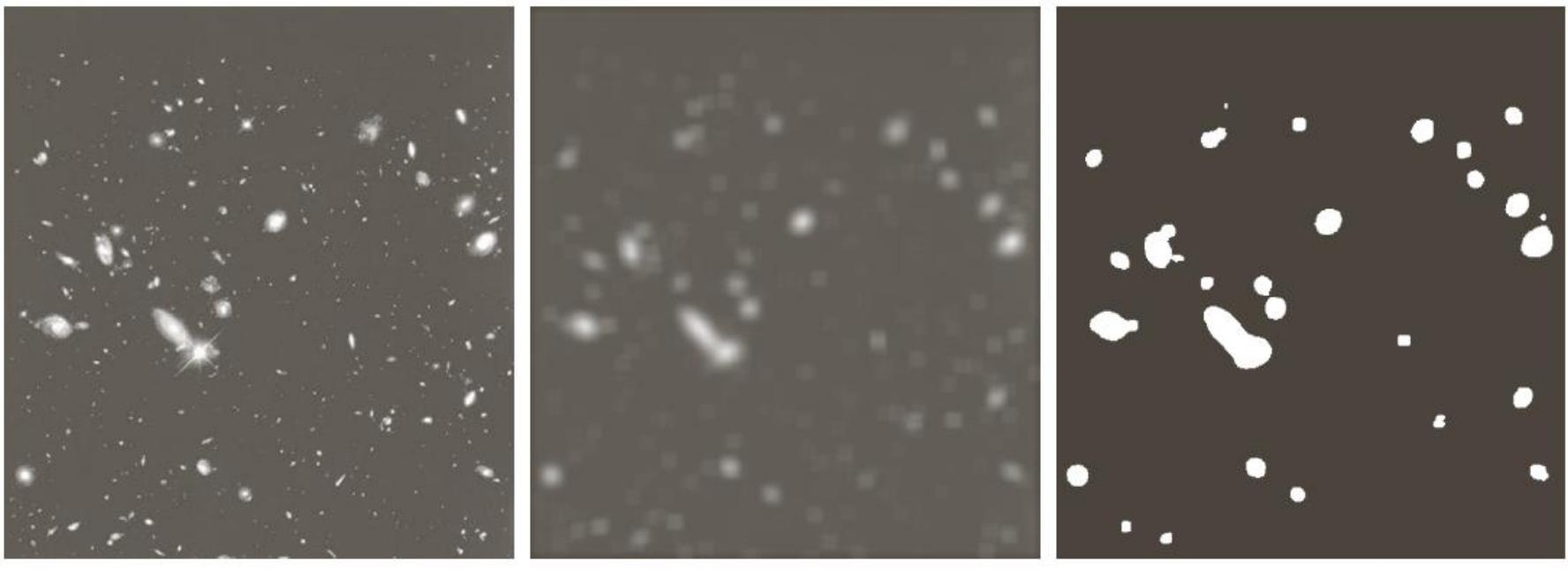


Median Filter size = 3×3



Median Filter size = 7×7

Example: Gross Representation of Objects



a b c

FIGURE 3.34 (a) Image of size 528×485 pixels from the Hubble Space Telescope. (b) Image filtered with a 15×15 averaging mask. (c) Result of thresholding (b). (Original image courtesy of NASA.)

Example: Use of Median Filtering for Noise Reduction

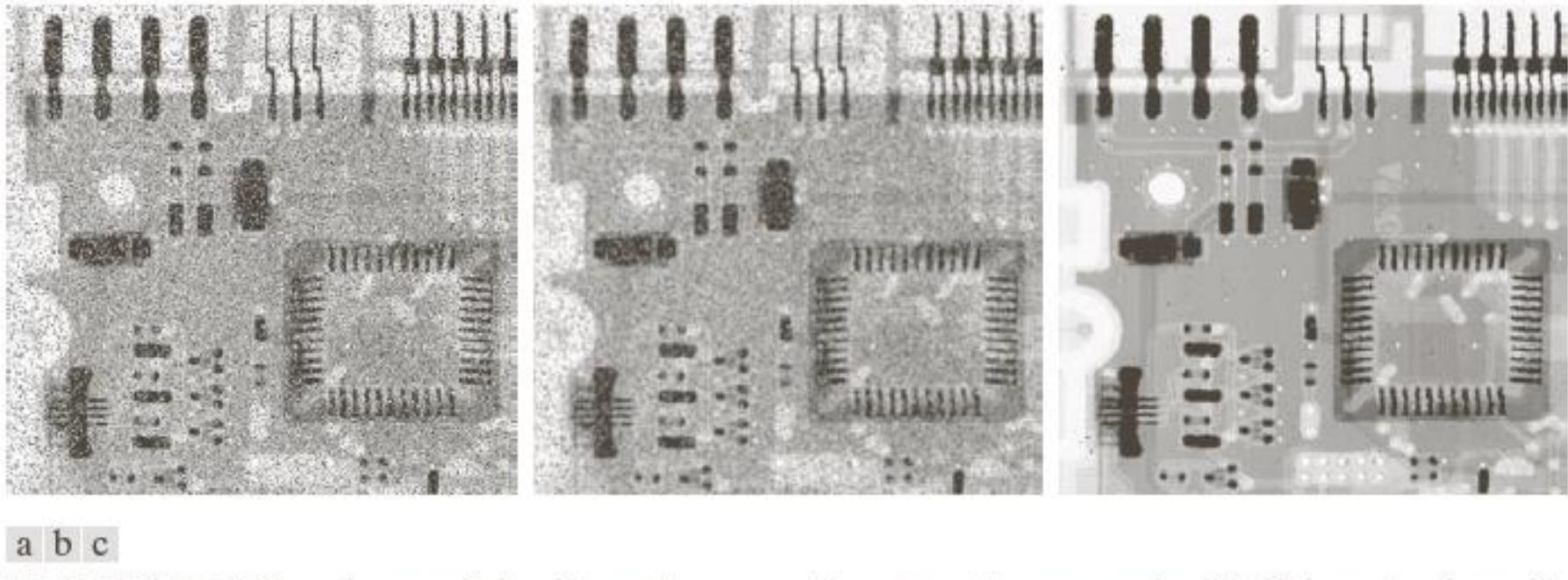


FIGURE 3.35 (a) X-ray image of circuit board corrupted by salt-and-pepper noise. (b) Noise reduction with a 3×3 averaging mask. (c) Noise reduction with a 3×3 median filter. (Original image courtesy of Mr. Joseph E. Pascente, Lixi, Inc.)

Conclusion of Smoothing Filter

- A linear filter cannot totally eliminate impulse noise, as a single pixel which acts as an intensity spike can contribute significantly to the weighted average of the filter.
- Non-linear filters can be robust to this type of noise because single outlier pixel intensities can be eliminated entirely.

Sharpening Spatial Filters

Sharpening:

➤ The term sharpening is referred to the techniques suited for enhancing the intensity transitions.

Blurring vs Sharpening:

- Blurring/smooth is done in spatial domain by pixel averaging in a neighbors, it is a process of integration
- Sharpening is an inverse process, to find the difference by the neighborhood, done by spatial differentiation.

➤ The intensity transitions between adjacent pixels are related to the derivatives of the image.

➤ Hence, operators (possibly expressed as linear filters) able to compute the derivatives of a digital image are very interesting.

Some Applications

- ❑ Photo Enhancement
- ❑ Medical image visualization
- ❑ Industrial defect detection
- ❑ Electronic printing
- ❑ Autonomous guidance in military systems

Sharpening Spatial Filters

- ▶ Foundation
- ▶ Laplacian Operator
- ▶ Unsharp Masking and Highboost Filtering
- ▶ Using First-Order Derivatives for Nonlinear Image Sharpening — The Gradient

Foundation

- ❑ Sharpening Filters to find details about
 - ❑ Remove blurring from images.
 - ❑ Highlight edges
- ❑ We are interested in the behavior of these derivatives in areas of constant gray level(flat segments), at the onset and end of discontinuities(step and ramp discontinuities), and along gray-level ramps.
- ❑ These types of discontinuities can be noise points, lines, and edges.

Laplacian Operator: First and Second Derivative

First derivative of an image

- ▶ Since the image is a discrete function, the traditional definition of derivative cannot be applied.
- ▶ Hence, a suitable operator have to be defined such that it satisfies the main properties of the first derivative:
 1. it is equal to zero in the regions where the intensity is constant;
 2. it is different from zero for an intensity transition;
 3. it is constant on ramps where the intensity transition is constant.
- ▶ The natural derivative operator is the difference between the intensity of neighboring pixels (spatial differentiation).
- ▶ For simplicity, the monodimensional case can be considered:

$$\frac{\partial f}{\partial x} = f(x + 1) - f(x)$$

- ▶ Since $\frac{\partial f}{\partial x}$ is defined using the next pixel:
 - ▶ it cannot be computed for the last pixel of each row (and column);
 - ▶ it is different from zero in the pixel before a step.

Definition for a first derivative

- Must be zero in flat segments
- Must be nonzero at the onset of a gray-level step or ramp
- Must be nonzero along ramps
- A basic definition of the first-order derivative of a one-dimensional function $f(x)$ is

$$\frac{\partial f}{\partial x} = f(x+1) - f(x)$$

(Diff b/w subsequent values & measures the rate of change of the function)

Second derivative of an image

- ▶ Similarly, the second derivative operator can be defined as:

$$\begin{aligned}\frac{\partial^2 f}{\partial x^2} &= f(x+1) - f(x) - (f(x) - f(x-1)) \\ &= f(x+1) - 2f(x) + f(x-1)\end{aligned}$$

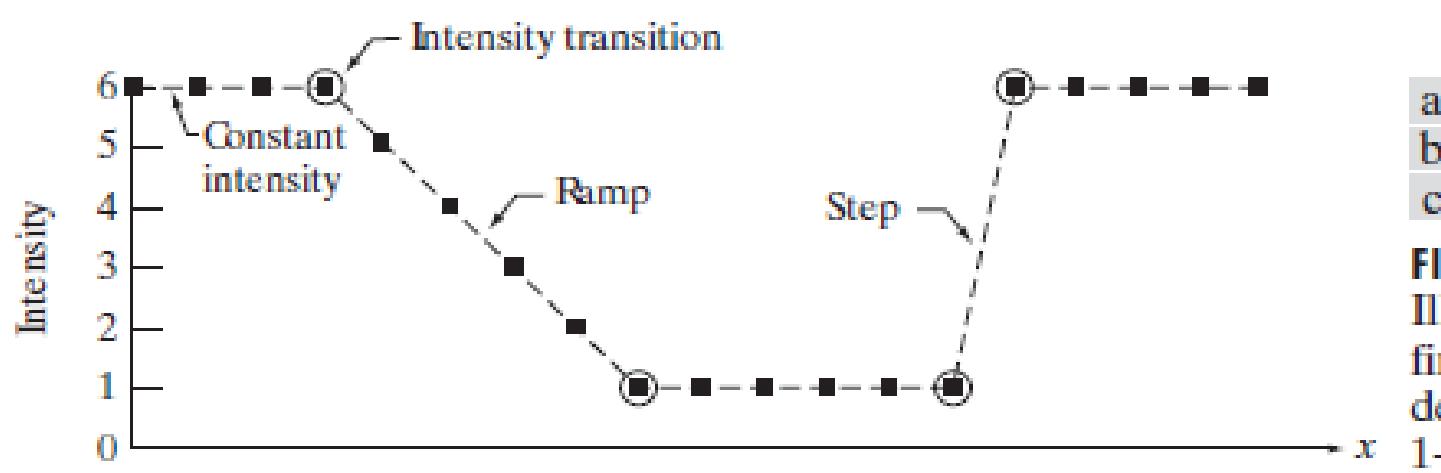
- ▶ This operator satisfies the following properties:
 1. it is equal to zero where the intensity is constant;
 2. it is different from zero at the begin of a step (or a ramp) of the intensity;
 3. it is equal to zero on the constant slope ramps.
- ▶ Since $\frac{\partial^2 f}{\partial x^2}$ is defined using the previous and the next pixels:
 - ▶ it cannot be computed with respect to the first and the last pixels of each row (and column);
 - ▶ it is different from zero in the pixel that precedes and in the one that follows a step.

Definition for a second derivative

- Must be zero in flat areas
- Must be non zero at the onset and end of a gray-level step or ramp
- Must be zero along ramps of constant slope
- We define a second-order derivative as the difference

$$\frac{\partial^2 f}{\partial x^2} = f(x+1) + f(x-1) - 2f(x)$$

(the values both before & after the current value)



Scan line

6	6	6	6	5	4	3	2	1	1	1	1	1	6	6	6	6	6

 $\rightarrow x$

1st derivative 0 0 -1 -1 -1 -1 0 0 0 0 0 0 5 0 0 0 0

2nd derivative 0 0 -1 0 0 0 0 1 0 0 0 0 5 -5 0 0 0

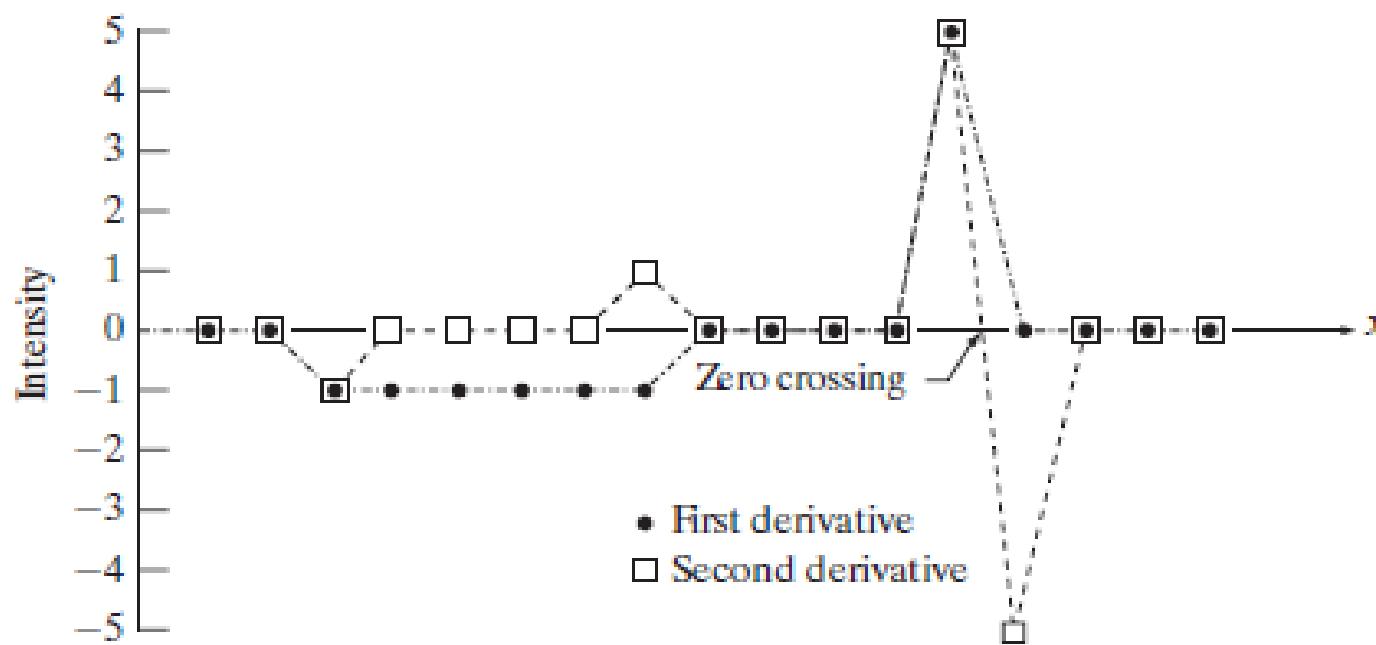


FIGURE 3.36
Illustration of the first and second derivatives of a 1-D digital function representing a section of a horizontal intensity profile from an image. In (a) and (c) data points are joined by dashed lines as a visualization aid.

- ❑ First and second-order derivatives in digital form => difference

$$\left\{ \begin{array}{l} \frac{\partial f}{\partial x} = f(x+1) - f(x) \\ \frac{\partial^2 f}{\partial x^2} = [f(x+1) - f(x)] - [f(x) - f(x-1)] \\ \qquad\qquad\qquad = f(x+1) + f(x-1) - 2f(x) \end{array} \right.$$

- ❑ The 1st-order derivative is nonzero along the entire ramp, while the 2nd-order derivative is nonzero only at the onset and end of the ramp.
- ❑ The response at and around the point is much stronger for the 2nd- than for the 1st-order derivative.

Using the Second Derivative for Image Sharpening—The Laplacian

- ▶ Usually the sharpening filters make use of the second order operators.
 - ▶ A second order operator is more sensitive to intensity variations than a first order operator.
- ▶ Besides, partial derivatives has to be considered for images.
 - ▶ The derivative in a point depends on the direction along which it is computed.
- ▶ Operators that are invariant to rotation are called *isotropic*.
 - ▶ Rotate and differentiate (or filtering) has the same effects of differentiate and rotate.
- ▶ The *Laplacian* is the simpler isotropic derivative operator (wrt. the principal directions):

$$\nabla^2 f = \frac{\partial^2 f}{\partial x^2} + \frac{\partial^2 f}{\partial y^2}$$

Laplacian filter

- ▶ In a digital image, the second derivatives wrt. x and y are computed as:

$$\frac{\partial^2 f}{\partial x^2} = f(x+1, y) - 2f(x, y) + f(x-1, y)$$

$$\frac{\partial^2 f}{\partial y^2} = f(x, y+1) - 2f(x, y) + f(x, y-1)$$

- ▶ Hence, the Laplacian results:

$$\begin{aligned}\nabla^2 f(x, y) &= f(x+1, y) + f(x-1, y) + f(x, y+1) \\ &\quad + f(x, y-1) - 4f(x, y)\end{aligned}$$

- ▶ Also the derivatives along to the diagonals can be considered:

$$\begin{aligned}\nabla^2 f(x, y) &+ f(x-1, y-1) + f(x+1, y+1) \\ &+ f(x-1, y+1) + f(x+1, y-1) - 4f(x, y)\end{aligned}$$

Sharpening Spatial Filters: Laplace Operator

0	1	0
1	-4	1
0	1	0

Laplacian filter invariant to 90° rotations

(a) Filter mask used
to implement
Eq. (3.6-6).

(b) Mask used to
implement an
extension of this
equation that
includes the
diagonal terms.

1	1	1
1	-8	1
1	1	1

Laplacian filter invariant to 45° rotations

Sharpening Spatial Filters: Laplace Operator

0	1	0
1	-4	1
0	1	0

1	1	1
1	-8	1
1	1	1

0	-1	0
-1	4	-1
0	-1	0

-1	-1	-1
-1	8	-1
-1	-1	-1

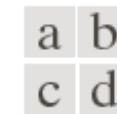


FIGURE 3.37

(a) Filter mask used to implement Eq. (3.6-6).

(b) Mask used to implement an extension of this equation that includes the diagonal terms.

(c) and (d) Two other implementations of the Laplacian found frequently in practice.

Implementation

$$g(x, y) = \begin{cases} f(x, y) - \nabla^2 f(x, y) & \text{If the center coefficient is negative} \\ f(x, y) + \nabla^2 f(x, y) & \text{If the center coefficient is positive} \end{cases}$$

Where $f(x, y)$ is the original image

$\nabla^2 f(x, y)$ is Laplacian filtered image

$g(x, y)$ is the sharpen image

Algorithm

1. Using Laplacian filter to original image
2. And then add the image result from step 1 and the original image
3. We will apply two step to be one mask

$$g(x, y) = f(x, y) - f(x+1, y) - f(x-1, y) - f(x, y+1) - f(x, y-1) + 4f(x, y)$$

$$g(x, y) = 5f(x, y) - f(x+1, y) - f(x-1, y) - f(x, y+1) - f(x, y-1)$$

Result of Algorithm

0	-1	0
-1	5	-1
0	-1	0

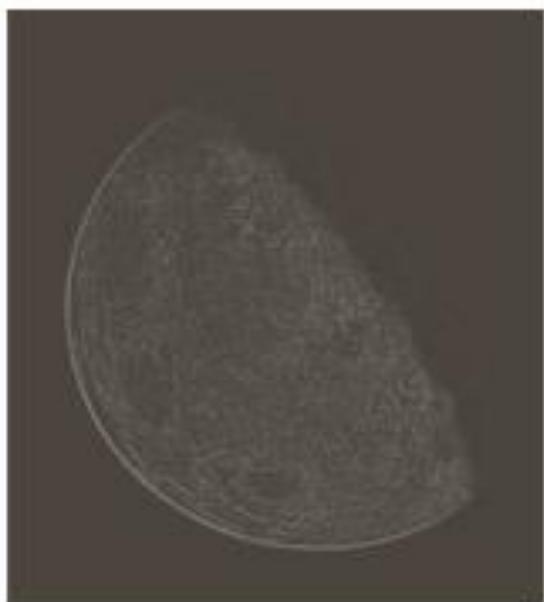
-1	-1	-1
-1	9	-1
-1	-1	-1

Laplacian filter: example

- ▶ The Laplacian has often negative values.
- ▶ In order to be visualized, it must be properly scaled to the representation interval $[0, \dots, L - 1]$.



(a)



(b)



(c)

(a) Original image, (b) its Laplacian, (c) its Laplacian scaled such that zero is displayed as the intermediate gray level.

Laplacian filter: example (2)

- ▶ The Laplacian is positive at the onset of a step and negative at the end.
- ▶ Subtracting the Laplacian (or a fraction of it) from the image, the height of the step is increased.



(a)



(b)



(c)

(a) Original image, (b) Laplacian filtered, (c) Laplacian with diagonals filtered.

Unsharp Masking and Highboost Filtering

► Unsharp masking

Sharpen images consists of subtracting an unsharp (smoothed) version of an image from the original image

e.g., printing and publishing industry

► Steps

1. Blur the original image
2. Subtract the blurred image from the original
3. Add the mask to the original

Letting $\bar{f}(x, y)$ denote the blurred image, unsharp masking is expressed in equation form as follows. First we obtain the mask:

$$g_{\text{mask}}(x, y) = f(x, y) - \bar{f}(x, y)$$

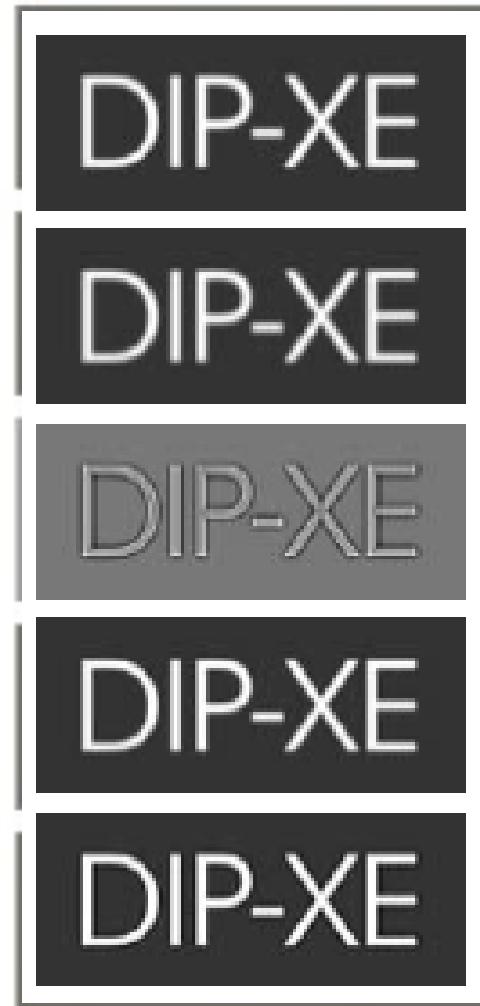
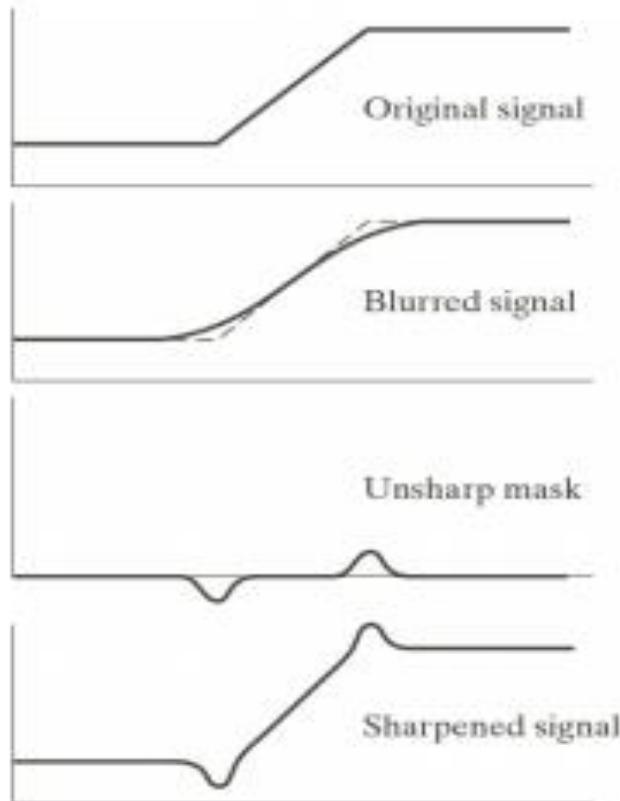
Then we add a weighted portion of the mask back to the original image:

$$g(x, y) = f(x, y) + k * g_{\text{mask}}(x, y)$$

where we included a weight, k ($k \geq 0$), for generality. When $k = 1$, we have unsharp masking, as defined above. When $k > 1$, the process is referred to as *highboost filtering*. Choosing $k < 1$ de-emphasizes the contribution of the unsharp mask.

Unsharp Masking and Highboost Filtering: Example

Unsharp masking (2)



Using First-Order Derivatives for (Nonlinear) Image Sharpening—The Gradient

- First Derivatives in image processing are implemented using the magnitude of the gradient.
- ▶ The *gradient* of a function is the vector formed by its partial derivatives.
- ▶ For a bidimensional function, $f(x, y)$:

$$\nabla f \equiv \text{grad}(f) \equiv \begin{bmatrix} g_x \\ g_y \end{bmatrix} = \begin{bmatrix} \frac{\partial f}{\partial x} \\ \frac{\partial f}{\partial y} \end{bmatrix}$$

- ▶ The gradient vector points toward the direction of maximum variation.
- ▶ The gradient *magnitude*, $M(x, y)$ is:

$$M(x, y) = \text{mag}(\nabla f) = \sqrt{g_x^2 + g_y^2}$$

- ▶ It is also called *gradient image*.
- ▶ Often approximated as $M(x, y) \approx |g_x| + |g_y|$.

- The magnitude of this vector is given by

$$mag(\nabla f) = \sqrt{G_x^2 + G_y^2} \approx |G_x| + |G_y|$$

G_x

-1	1
----	---

This mask is simple, and no isotropic. Its result only horizontal and vertical.

G_y

1
-1

Image Sharpening based on First-Order Derivatives

The *magnitude* of vector ∇f , denoted as $M(x, y)$

$$M(x, y) = \text{mag}(\nabla f) = \sqrt{{g_x}^2 + {g_y}^2}$$

$$M(x, y) \approx |g_x| + |g_y|$$

z_1	z_2	z_3
z_4	z_5	z_6
z_7	z_8	z_9

$$M(x, y) = |z_8 - z_5| + |z_6 - z_5|$$

Robert's Method

- The simplest approximations to a first-order derivative that satisfy the conditions stated in that section are

z_1	z_2	z_3
z_4	z_5	z_6
z_7	z_8	z_9

$$G_x = (z_9 - z_5) \text{ and } G_y = (z_8 - z_6)$$

$$\nabla f = \sqrt{(z_9 - z_5)^2 + (z_8 - z_6)^2}$$

$$\nabla f \approx |z_9 - z_5| + |z_8 - z_6|$$

- These mask are referred to as the Roberts cross-gradient operators.

-1	0
0	1

0	-1
1	0

Image Sharpening based on First-Order Derivatives

Roberts Cross-gradient Operators

$$M(x, y) \approx |z_9 - z_5| + |z_8 - z_6|$$

Sobel Operators

$$M(x, y) \approx |(z_7 + 2z_8 + z_9) - (z_1 + 2z_2 + z_3)| + |(z_3 + 2z_6 + z_9) - (z_1 + 2z_4 + z_7)|$$

z_1	z_2	z_3
z_4	z_5	z_6
z_7	z_8	z_9

3/6/2019

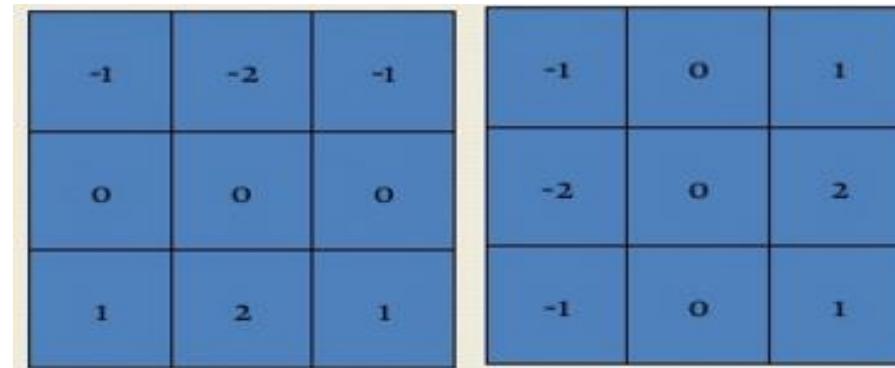


Image Sharpening based on First-Order Derivatives

z_1	z_2	z_3
z_4	z_5	z_6
z_7	z_8	z_9

-1	0
0	1

0	-1
1	0

-1	-2	-1
0	0	0
1	2	1

-1	0	1
-2	0	2
-1	0	1

a
b c
d e

FIGURE 3.41

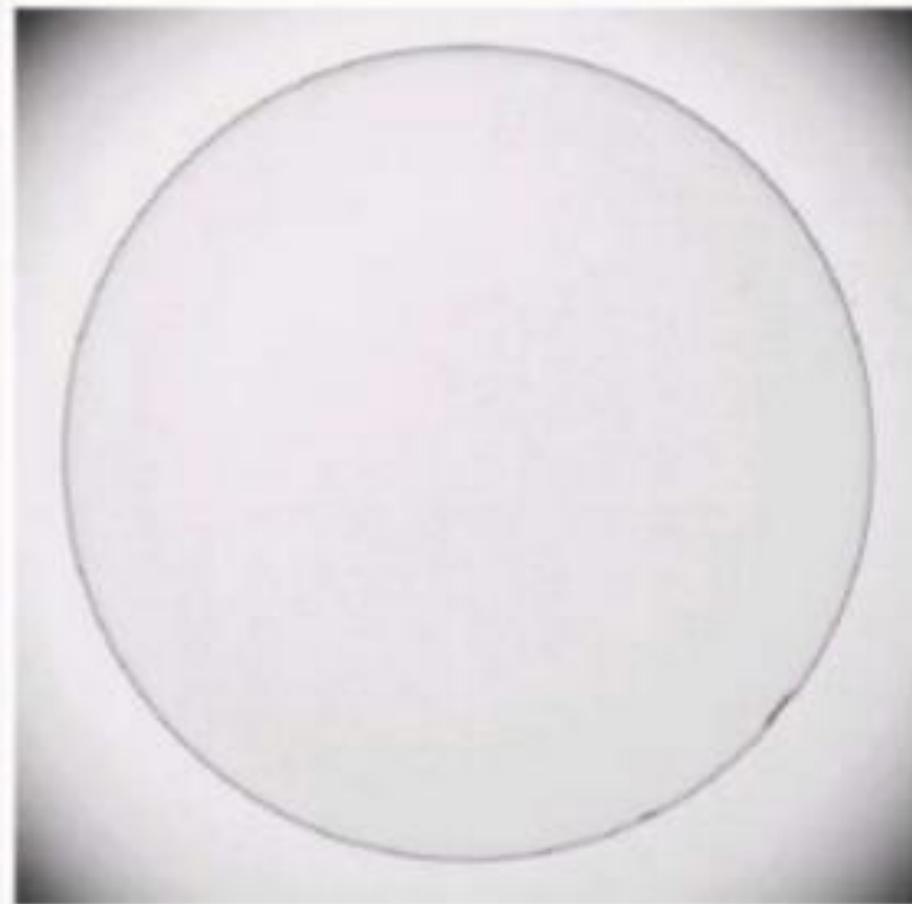
A 3×3 region of an image (the z s are intensity values).

(b)–(c) Roberts cross gradient operators.

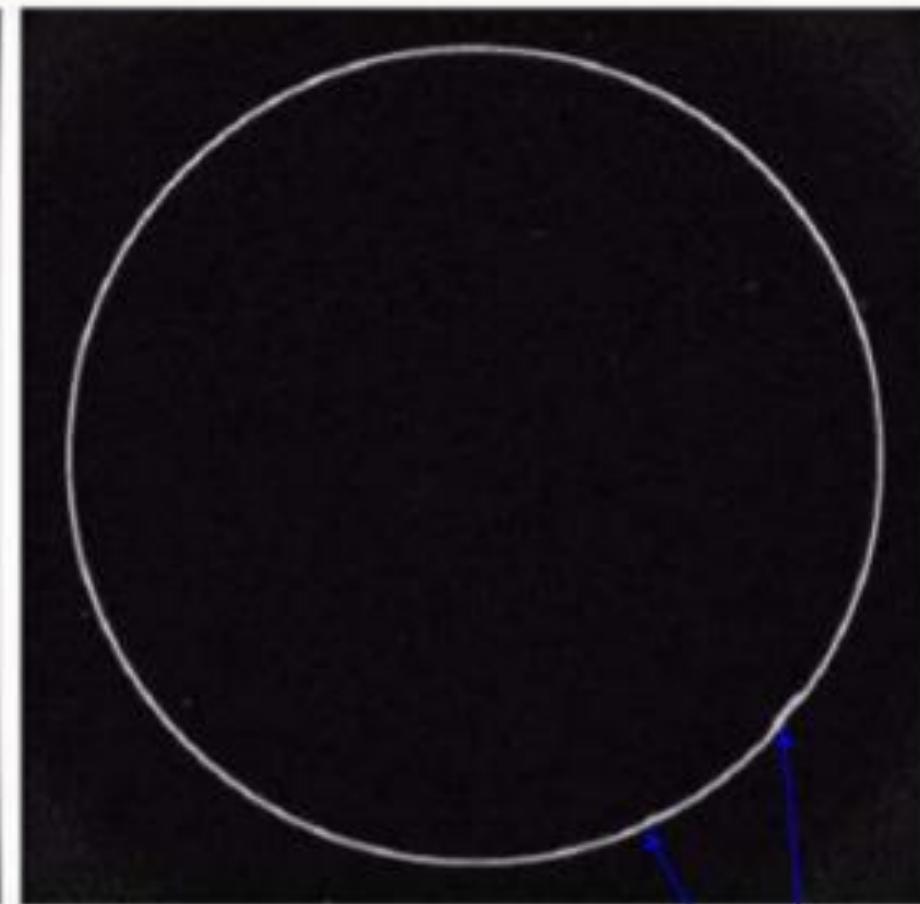
(d)–(e) Sobel operators. All the mask coefficients sum to zero, as expected of a derivative operator.

Gradient: example

- Enhance defects and eliminate slowly changing background



original(contact lens)



Sobel gradient

defects