

Vol.1

FE Exam

Preparation Book

Preparation Book for Fundamental Information Technology Engineer Examination

Part1: Preparation for Morning Exam

Part2: Trial Exam Set

IPA

Information-Technology Promotion Agency, Japan

Table of Contents

Chapter 1

Computer Science Fundamentals	2
1.1 Basic Theory of Information	3
1.1.1 Radix Conversion	3
1.1.2 Numerical Representations	7
1.1.3 Non-Numerical Representations	10
1.1.4 Operations and Accuracy	11
Quiz	14
1.2 Information and Logic	15
1.2.1 Logical Operations	15
1.2.2 BNF	18
1.2.3 Reverse Polish Notation	21
Quiz	24
1.3 Data Structures	25
1.3.1 Arrays	25
1.3.2 Lists	27
1.3.3 Stacks	29
1.3.4 Queues (Waiting lists)	30
1.3.5 Trees	32
1.3.6 Hash	34
Quiz	37
1.4 Algorithms	38
1.4.1 Search Algorithms	38
1.4.2 Sorting Algorithms	41
1.4.3 String Search Algorithms	45
1.4.4 Graph Algorithms	48
Quiz	50
Questions and Answers	51

Chapter 2	
Computer Systems	62
2.1 Hardware.....	63
2.1.1 Information Elements (Memory).....	63
2.1.2 Processor Architecture.....	65
2.1.3 Memory Architecture.....	68
2.1.4 Magnetic Tape Units.....	70
2.1.5 Hard Disks.....	73
2.1.6 Terms Related to Performance/ RAID.....	77
2.1.7 Auxiliary Storage / Input and Output Units.....	79
2.1.8 Input and Output Interfaces.....	81
Quiz.....	83
2.2 Operating Systems.....	85
2.2.1 Configuration and Objectives of OS.....	85
2.2.2 Job Management.....	87
2.2.3 Task Management.....	89
2.2.4 Data Management and File Organization.....	90
2.2.5 Memory Management.....	95
Quiz.....	99
2.3 System Configuration Technology.....	100
2.3.1 Client Server Systems.....	100
2.3.2 System Configurations.....	102
2.3.3 Centralized Processing and Distributed Processing.....	104
2.3.4 Classification by Processing Mode.....	106
Quiz.....	108
2.4 Performance and Reliability of Systems.....	109
2.4.1 Performance Indexes.....	109
2.4.2 Reliability.....	111
2.4.3 Availability.....	113
Quiz.....	116
2.5 System Applications.....	118
2.5.1 Network Applications.....	118
2.5.2 Database Applications.....	121
2.5.3 Multimedia Systems.....	123
Quiz.....	125
Questions and Answers.....	126

Chapter 3	
System Development	138
3.1 Methods of System Development	139
3.1.1 Programming Languages.....	139
3.1.2 Program Structures and Subroutines.....	141
3.1.3 Language Processors.....	143
3.1.4 Development Environments and Software Packages.....	144
3.1.5 Development Methods.....	147
3.1.6 Requirement Analysis Methods.....	149
3.1.7 Software Quality Management.....	151
Quiz.....	154
3.2 Tasks of System Development Processes	155
3.2.1 External Design.....	155
3.2.2 Internal Design.....	157
3.2.3 Software Design Methods.....	159
3.2.4 Module Partitioning Criteria.....	162
3.2.5 Programming.....	163
3.2.6 Types and Procedures of Tests.....	165
3.2.7 Test Techniques.....	167
Quiz.....	170
Questions and Answers	172
Chapter 4	
Network Technology	181
4.1 Protocols and Transmission Control	182
4.1.1 Network Architectures.....	182
4.1.2 Transmission Control.....	184
Quiz.....	187
4.2 Transmission Technology	188
4.2.1 Error Control.....	188
4.2.2 Synchronization Control.....	190
4.2.3 Multiplexing and Communications.....	192
4.2.4 Switching.....	194
Quiz.....	195
4.3 Networks	196
4.3.1 LANs.....	196
4.3.2 The Internet.....	198
4.3.3 Various Communication Units.....	200
4.3.4 Telecommunications Services.....	202
Quiz.....	204
Questions and Answers	205

Chapter 5	
Database Technology	212
5.1 Data Models	213
5.1.1 3-layer Schemata	213
5.1.2 Logical Data Models	215
5.1.3 E-R Model and E-R Diagrams	217
5.1.4 Normalization and Reference Constraints	218
5.1.5 Data Manipulation in Relational Database	221
Quiz	223
5.2 Database Languages	224
5.2.1 DDL and DML	224
5.2.2 SQL	226
Quiz	231
5.3 Control of Databases	232
5.3.1 Database Control Functions	232
5.3.2 Distributed Databases	234
Quiz	236
Questions and Answers	237

Chapter 6	
Security and Standardization	244
6.1 Security	245
6.1.1 Security Protection	245
6.1.2 Computer Viruses	247
6.1.3 Computer Crime	249
Quiz	251
6.2 Standardization	252
6.2.1 Standardization Organizations and Standardization of Development and Environment	252
6.2.2 Standardization of Data	254
6.2.3 Standardization of Data Exchange and Software	256
Quiz	258
Questions and Answers	259

Chapter 7

Computerization and Management	262
7.1 Information Strategies	263
7.1.1 Management Control	263
7.1.2 Computerization Strategies	265
Quiz	267
7.2 Corporate Accounting	268
7.2.1 Financial Accounting	268
7.2.2 Management Accounting	270
Quiz	274
7.3 Management Engineering	275
7.3.1 IE	275
7.3.2 Schedule Control (OR)	278
7.3.3 Linear Programming	282
7.3.4 Inventory Control (OR)	284
7.3.5 Probability and Statistics	286
Quiz	290
7.4 Use of Information Systems	291
7.4.1 Engineering Systems	291
7.4.2 Business Systems	293
Quiz	296
Questions and Answers	297

PREPARATION FOR MORNING EXAM

The Morning Exam questions are formulated from the following seven fields: Computer Science Fundamentals, Computer Systems, System Development, Network Technology, Database Technology, Security and Standardization, and Computerization and Management.

Here, detailed explanations of each field are provided in the beginning of each chapter, followed by the actual questions used in the past exams, as well as answers and comments that are included in the end of each chapter.

1 Computer Science Fundamentals

Chapter Objectives

In order to become an information technology engineer, it is necessary to understand the structures of information processed by computers and the meaning of information processing. All information is stored as binary numbers in computers; therefore, in Section 1, we will learn the form in which decimal numbers and characters we use in daily life are stored in computers. In Section 2, we will study logical operations as a specific example of information processing. In Section 3, we will learn data structures, of which modification is necessary to increase the ease of data processing. Lastly, in Section 4, we will study specific data processing methods.

- 1.1 Basic Theory of Information**
- 1.2 Information and Logic**
- 1.3 Data Structures**
- 1.4 Algorithms**

[Terms and Concepts to Understand]

Radix, binary, hexadecimal, fixed point, floating point, logical sum, logical product, exclusive logical sum, adder, list, stack, queue, linear search, binary search, bubble sort

1.1 Basic Theory of Information

Introduction

All information (such as characters and numerals) is expressed by combinations of 1s and 0s inside computers. An expression using only 1s and 0s is called a binary number. Here, we will learn expressive forms for information.

1.1.1 Radix Conversion

Points

- In computers, all data is expressed by using binary numbers.
- Hexadecimal numbers are expressed by separating binary numbers into 4-bit groups.

The term “Radix¹ conversion” means, for instance, converting a decimal number to a binary number. Here, “10” in decimal numbers and “2” in binary numbers are called the radices. Inside a computer, all data is expressed as **binary numbers** since the two conditions of electricity, ON and OFF, correspond to the binary numbers. Each digit of a binary number is either a “0” or a “1,” so all numbers are expressed by two symbols—0 and 1.

However, binary numbers, expressed as combinations of 0s and 1s, tend to be long and hard to understand, so the concept of **hexadecimal notation** was introduced. In hexadecimal notation, 4 bits² (corresponding to numbers 0 through 15 in decimal notation) are represented by one digit (0 through F).

The table below shows the correspondence among the decimal, binary, and hexadecimal notations.

Decimal	Binary	Hexadecimal	Decimal	Binary	Hexadecimal
0	0000	0	8	1000	8
1	0001	1	9	1001	9
2	0010	2	10	1010	A
3	0011	3	11	1011	B
4	0100	4	12	1100	C
5	0101	5	13	1101	D
6	0110	6	14	1110	E
7	0111	7	15	1111	F
			16	10000	10

¹ **Radix:** It is the number that forms a unit of weight for each digit in a numeration system such as binary, octal, decimal, and hexadecimal notations. The radix in each of these notations is 2, 8, 10, and 16, respectively.

Binary system: uses 0 and 1.

Octal system: uses 0 through 7.

Decimal system: uses 0 through 9.

Hexadecimal system: uses 0 through F.

² **Bit:** It means the smallest unit of information inside a computer, expressed by a “0” or a “1.” Data inside a computer is expressed in binary, so a bit represents one digit in binary notation. For the purpose of convenience, the hexadecimal and octal notations are represented by partitioning binary numbers as follows:

Quaternary: 2 bits (0 through 3)

Octal: 3 bits (0 through 7)

Hexadecimal: 4 bits (0 through F)

◆ Conversion of Binary or Hexadecimal Numbers into Decimal Numbers

In general, when a value is given in the numeration system with radix r (r -ary system), we multiply each digit value with its corresponding weight³ and adds up the products in order to check what the value is in decimal. For digits to the left of the radix point, the weights are r^0, r^1, r^2, \dots from the lowest digit. Thus, the conversion is shown below. (In these examples, (a) is shown in hexadecimal, and (b) is in binary.)

$$\begin{aligned}(12A)_{16} &= 1 \times 16^2 + 2 \times 16^1 + A \times 16^0 \\ &= 256 + 32 + 10 \\ &= (298)_{10}\end{aligned}\quad \dots\dots (a)$$

$$\begin{aligned}(1100100)_2 &= 1 \times 2^6 + 1 \times 2^5 + 0 \times 2^4 + 0 \times 2^3 + 1 \times 2^2 + 0 \times 2^1 + 0 \times 2^0 \\ &= 64 + 32 + 4 \\ &= (100)_{10}\end{aligned}\quad \dots\dots (b)$$

For those digits to the right of the radix point, the weights are $r^{-1}, r^{-2}, r^{-3}, \dots$ in order. Thus, the conversion is shown below. (In these examples, (c) is shown in hexadecimal, and (d) is in binary.)

$$\begin{aligned}(0.4B)_{16} &= 4 \times 16^{-1} + B \times 16^{-2} \\ &= 4 / 16 + 11 / 16^2 \\ &= 0.25 + 0.04296875 \\ &= (0.29296875)_{10}\end{aligned}\quad \dots\dots (c)$$

$$\begin{aligned}(0.01011)_2 &= 0 \times 2^{-1} + 1 \times 2^{-2} + 0 \times 2^{-3} + 1 \times 2^{-4} + 1 \times 2^{-5} \\ &= 0.25 + 0.0625 + 0.03125 \\ &= (0.34375)_{10}\end{aligned}\quad \dots\dots (d)$$

◆ Conversion of Decimal Integers to Binary Numbers

Mathematically, using the fact that the n -th digit from the right (lowest) represents the place value of 2^{n-1} in binary, we can decompose a decimal number into a sum of powers of 2 (values 2^n for some n).

$$\begin{aligned}(59)_{10} &= 32 + 16 + 8 + 2 + 1 = 2^5 + 2^4 + 2^3 + 2^1 + 2^0 \\ &= 1 \times 2^5 + 1 \times 2^4 + 1 \times 2^3 + 0 \times 2^2 + 1 \times 2^1 + 1 \times 2^0 \\ &\quad \downarrow \quad \downarrow \quad \downarrow \quad \downarrow \quad \downarrow \quad \downarrow \\ &\quad (1 \quad 1 \quad 1 \quad 0 \quad 1 \quad 1)_2\end{aligned}$$

³ **Weight:** the value that indicates each scaling position in numerical expressions such as binary, octal, decimal, and hexadecimal.

However, we can also divide the given number by 2 sequentially and repeat it until the quotient becomes 0. This is a mechanical conversion method, so calculation errors can be reduced.⁴

Remainder

$$\begin{array}{r}
 2 \overline{)59} \cdots 1 \rightarrow (1) 59 / 2 = 29 \text{ remainder } 1 \\
 2 \overline{)29} \cdots 1 \rightarrow (2) 29 / 2 = 14 \text{ remainder } 1 \\
 2 \overline{)14} \cdots 0 \rightarrow (3) 14 / 2 = 7 \text{ remainder } 0 \\
 2 \overline{)7} \cdots 1 \rightarrow (4) 7 / 2 = 3 \text{ remainder } 1 \\
 2 \overline{)3} \cdots 1 \rightarrow (5) 3 / 2 = 1 \text{ remainder } 1 \\
 2 \overline{)1} \cdots 1 \rightarrow (6) 1 / 2 = 0 \text{ remainder } 1
 \end{array}$$

$0 \leftarrow$ "The process ends when the quotient is 0." (7) List the remainders from the bottom. $\rightarrow (59)_{10} = (111011)_2$

In addition, in order to convert a decimal number to a hexadecimal number, we can use 16 instead of 2 here. In general, to convert a decimal number to an n -ary number, use n instead of 2.

◆ Conversion of Decimal Numbers into Binary Numbers

Mathematically, using the fact that the n -th digit after the radix point in binary represents the place value of 2^n , we can decompose a decimal number into a sum of powers of 2 (values 2^n for some n).

$$\begin{aligned}
 (0.59375)_{10} &= 0.5 + 0.0625 + 0.03125 \\
 &= 2^{-1} + 2^{-4} + 2^{-5} \\
 &= 1 \times 2^{-1} + 0 \times 2^{-2} + 0 \times 2^{-3} + 1 \times 2^{-4} + 1 \times 2^{-5} \\
 &\quad \parallel \quad \parallel \quad \parallel \quad \parallel \quad \parallel \\
 (0.1 &\quad 0 \quad 0 \quad 1 \quad 1)_2
 \end{aligned}$$

However, we can also multiply the fractional part (the part to the right of the decimal (or radix) point) by 2 sequentially and repeat it until the fractional part becomes 0. This is a mechanical conversion method, so calculation errors can be reduced.

(5) List the integer-part values from the top. $\rightarrow (0.59375)_{10} = (0.10011)_2$

$$\begin{array}{r}
 0.59375 \times 2 = 1 .1875 \rightarrow (1) \text{ Write down only the fractional part.} \\
 \hline
 0.1875 \times 2 = 0 .375 \rightarrow (2) \text{ Write down only the fractional part.} \\
 \hline
 0.375 \times 2 = 0 .75 \rightarrow (3) \text{ Write down only the fractional part.} \\
 \hline
 0.75 \times 2 = 1 .5 \rightarrow (4) \text{ Write down only the fractional part.} \\
 \hline
 0.5 \times 2 = 1 .0 \leftarrow \text{The process ends when the fractional part becomes } 0. \text{ }^5
 \end{array}$$

In addition, in order to convert a decimal number to a hexadecimal number, use 16 instead of 2. In general, to convert a decimal number to an n -ary number, use n instead of 2.

⁴ (Note) There is no guarantee that multiplying the fractional part by 2 always produces 0. We can verify this fact by converting 0.110 into the binary number; it becomes a repeating binary fraction. It is always possible to convert a binary fraction to a decimal fraction, but not vice versa. In such a case, we can stop the conversion at an appropriate place.

⁵ **Repeating fraction:** a number with a radix point where a sequence of digits is repeated indefinitely. For instance, $1 / 3 = 0.333\dots$, and $1 / 7 = 0.142857142857\dots$, wherein the patterns "3" and "142857" are repeated, respectively.

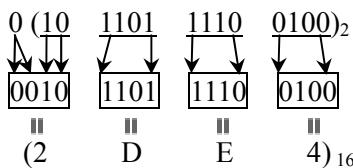
◆ Conversion between Hexadecimal and Binary Numbers

We can use the fact that each digit of a hexadecimal number corresponds to 4 bits in binary.

FROM BINARY TO HEXADECIMAL

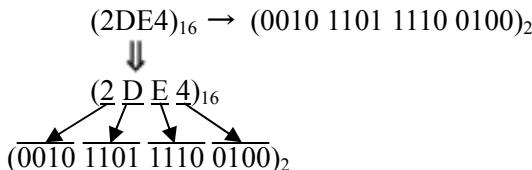
As shown below, we can group the binary number into blocks of 4 bits, starting from the lowest bit (rightmost bit), and then assign the corresponding hexadecimal digit for each block. If the last (leftmost) block is less than 4 bits, it is padded with leading 0s.

$$(10110111100100)_2 \rightarrow (10\ 1101\ 1110\ 0100)_2 \rightarrow (2DE4)_{16}$$



FROM HEXADECIMAL TO BINARY

As shown below, we can assign the corresponding 4-bit binary number to each digit of the given hexadecimal number.

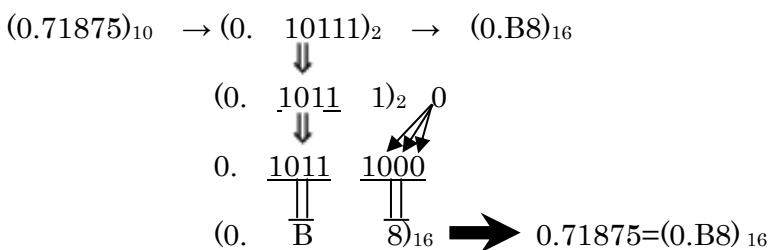


◆ Conversion between Hexadecimal Fractions and Decimal Fractions

To convert between hexadecimal fractions and decimal fractions, we can combine the conversion between decimal and binary numbers together with the conversion between binary and hexadecimal numbers to reduce errors.

FROM DECIMAL TO HEXADECIMAL FRACTION

We can convert the given decimal number to binary first, and then convert the binary number to the corresponding hexadecimal number. In converting binary to hexadecimal, we can group the bits into 4-bit blocks, starting from the highest (leftmost) bit of the fractional part, and convert each block to the corresponding hexadecimal digit. If the last (rightmost) block is fewer than 4 bits, it is padded with trailing 0s.



FROM HEXADECIMAL TO DECIMAL⁶

First, we can convert the given hexadecimal number to the corresponding binary number, and then convert the binary number to the corresponding decimal number.

$$(0.B8)_{16} \rightarrow (0.10111000)_2 \rightarrow 0.71875$$

$$\downarrow$$

$$(0.\underline{1011} \underline{1000})_2 \quad 0$$

$$\downarrow$$

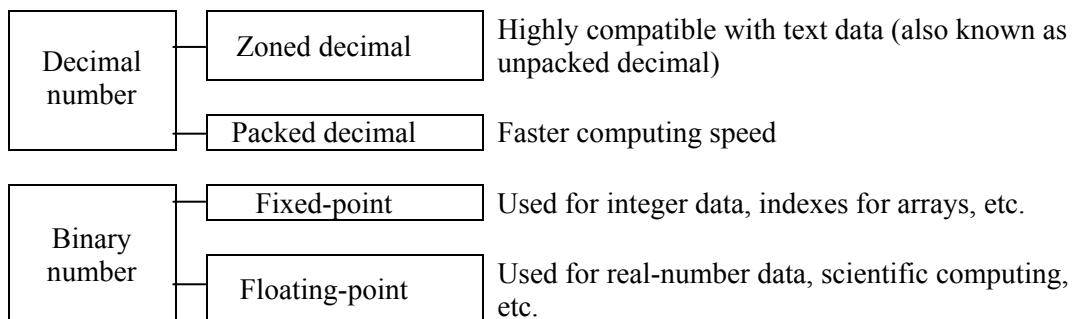
$$(0.10111000)_2 = 2^{-1} + 2^{-3} + 2^{-4} + 2^{-5} = (0.71875)_{10}$$

1.1.2 Numerical Representations

Points

- Decimal numbers are represented in packed or zoned format.
- Binary numbers are represented in fixed-point or floating-point format.

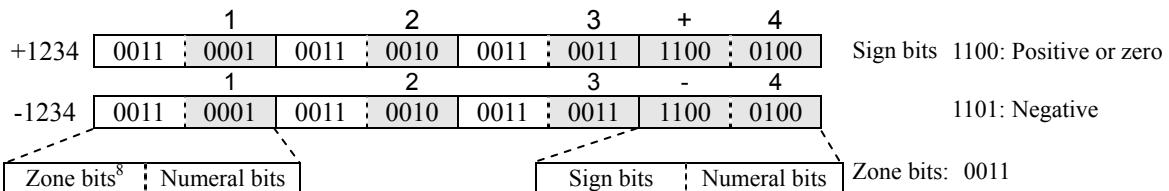
Decimal numbers used in our daily life need converting to a format which is convenient for computer processing, so there are various formats available to represent numerical values. Some of the formats that represent numerical values in a computer are shown below.



⁶ (FAQ) There are many questions mixing multiple radices (bases) such as “Which of the following is the correct result (in decimal) of adding the hexadecimal and binary numbers?” If the final result is to be represented in decimal, it is better that you convert the original numbers to decimal first and then calculate it. If the final result is to be represented in a radix other than 10 (binary, octal, hexadecimal, etc.), it is better that you convert the original numbers to binary first and then carry on the calculation.

◆ Decimal Number Representation

In the zoned decimal format, each digit of the given decimal number is represented by 8 bits, and the highest 4 bits of the last digit are used for the sign information.⁷ The numeral bits of each byte contain the corresponding numerical value in binary

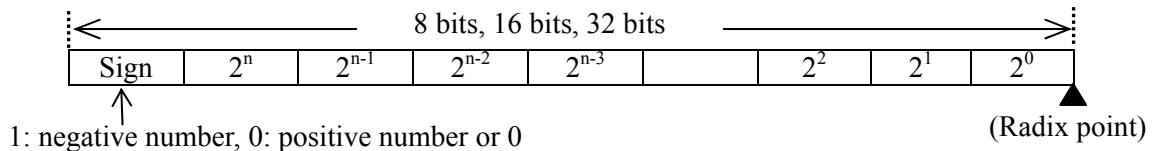


In the packed decimal format, each digit of the decimal number is represented with 4 bits, and the last four bits indicate the sign. The leading space of the highest byte is padded with 0s. The bit pattern of the sign bits is the same as that of the zoned decimal format. In the examples shown below, 2 bytes and 4 bits are sufficient to represent the numbers, but in both cases 3 bytes are used by appending four leading 0s since computers reserve areas in byte⁹ units.

0	1	2	3	4	+	0	1	2	3	4	-		
+1234	0000	0001	0010	0011	0100	1100	-1234	0000	0001	0010	0011	0100	1101

◆ Fixed-Point Number Representation

In fixed-point number, binary integers are represented in fixed-length binary. Two's complement is used to represent negative numbers, so the leading bit (sign bit) of a negative number is always a "1."



⁷ (Hints and Tips) If the sign (positive or negative) is not used in the zoned decimal format, the sign bits are identical to the zone bits.

⁸ (Note) The bit patterns in the zone bits are different depending on the computer. The examples shown here have "0011," but some computers use "1111." The numeral bits, however, are identical.

⁹ **Byte:** A byte is a unit of 8 bits. It is the unit for representing characters.

Let us represent the decimal number “-20” in two's complement. First, we can represent the decimal number “+20” in binary as shown below.

$(+20)_{10} =$	<table border="1" style="display: inline-table; vertical-align: middle;"> <tr><td>0</td><td>0</td><td>0</td><td>1</td><td>0</td><td>1</td><td>0</td><td>0</td></tr> </table>	0	0	0	1	0	1	0	0	\Downarrow	Reverse each bit.
0	0	0	1	0	1	0	0				
			<table border="1" style="display: inline-table; vertical-align: middle;"> <tr><td>1</td><td>1</td><td>1</td><td>0</td><td>1</td><td>0</td><td>1</td><td>1</td></tr> </table>	1	1	1	0	1	0	1	1
1	1	1	0	1	0	1	1				
(+)			One's complement ¹⁰								
$(-20)_{10} =$	<table border="1" style="display: inline-table; vertical-align: middle;"> <tr><td>1</td><td>1</td><td>1</td><td>0</td><td>1</td><td>1</td><td>0</td><td>0</td></tr> </table>	1	1	1	0	1	1	0	0	1 Add 1	Two's complement
1	1	1	0	1	1	0	0				

Hence, $(-20)_{10}$ is represented as $(11101100)_2$. The bit length varies from computer to computer.

In general, the numbers from -2^{n-1} through $2^n - 1$, a total of 2^n numbers, can be represented by using n bits. Note that, considering only the absolute values, one more negative number can be represented in comparison with positive numbers.

◆ Floating-Point Number Representation

In floating-point number, a real number is represented in exponential form ($a = \pm m \times r^e$) using a fixed-length binary number, so it is possible to represent very large (and very small) numbers, such as those used in scientific computing. However, since the computer register¹¹ has a limited number of digits, an error may occur in representing the value of repeating fraction.

1 bit	8 bits	23 bits (single precision)
0	10000100	1101000000000000000000000
↑	↑	▲
Mantissa sign \pm^{12}	Exponent e	Radix point
		↑
		Mantissa m

This is the International Standard Form known as IEEE754.

¹⁰ **Complement:** The complement of a number is the value obtained by subtracting the given number from a certain fixed number, which is a power of the radix or a power of the radix minus 1. For instance, in decimal, there are ten's complements and nine's complements. In binary, there are two's complements and one's complements. In general, in the r -ary system, there are r 's complements and $(r-1)$'s complements. If x is an n -digit number in the r -ary system, r 's complement of x is $(r^n - x)$, and $(r-1)$'s complement of x is $((r^n - 1) - x)$. For example, the three-digit number “123” in decimal has the following complements: ten's complement is “ $1000 - 123 = 877$,” and nine's complement is “ $999 - 123 = 876$.” The 4-bit number “0101” in binary has the following complements: two's complement is “ $10000 - 0101 = 1011$,” and one's complement is “ $1111 - 0101 = 1010$.”

ten's complement	nine's complement
$10^3 = 1000$	$10^3 - 1 = 999$
$\underline{-} \quad 123$	$\underline{-} \quad 123$
877	876
two's complement	one's complement
$2^4 = 10000$	$2^4 - 1 = 1111$
$\underline{-} \quad 0101$	$\underline{-} \quad 0101$
1011	1010

Note that one's complement in binary is just the reverse of each bit (0 becomes 1 and vice-versa). Two's complement is one's complement plus 1.

¹¹ **Register:** It is low-capacity, high-speed memory placed in the CPU for temporary storage of data.

¹² (FAQ) There are many questions on converting a given binary number into the corresponding negative number and converting a given negative number into the corresponding positive number.

1.1.3 Non-Numerical Representations

Points

- In general, each character is represented by 8 bits.
- In multimedia, data associated with still image data, moving picture data, and sound data is handled.

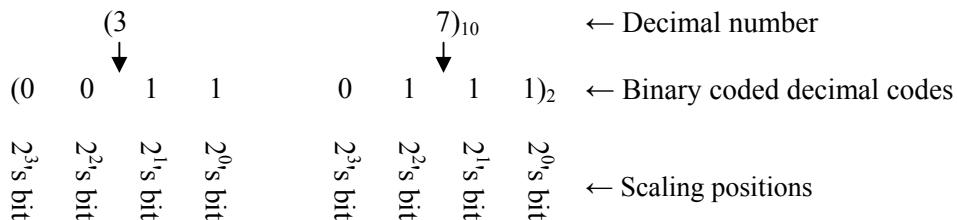
A non-numerical representation refers to a representation of data other than numerical values. In other words, it refers to a representation of character, sound, or image. The way in which the data is internally represented differs from computer to computer. Hence, in order to ensure smooth data exchange between computers, it is necessary to establish some standardized representations.

◆ Character Representations

Using n -bit binary numbers, there are 2^n types of codes available, and one-to-one correspondence to those codes allows us to represent 2^n types of characters (alphabet characters, numeric characters, special characters, and various symbols).

BCD Code (Binary Coded Decimal Code)

Each digit of decimal number can be represented by using 4 bits. The following shows such an example.



Standardizations of Character Codes

Code Name	Explanation
EBCDIC	Computer code defined by IBM for general purpose computers 8 bits represent one character.
ASCII	7-bit code established by ANSI (American National Standards Institute) Used in PCs, etc.
ISO code	ISO646 published as a recommendation by the International Organization for Standardization (ISO), based on ASCII 7-bit code for information exchange
Unicode	An industry standard allowing computers to consistently represent characters used in most of the countries Every character is represented with 2 bytes.
EUC	2-byte and 1-byte characters can be used together on UNIX (extended UNIX code). Chinese and Hangul characters are also handled.

◆ Image and Sound Representations

The amount of information, such as images, sound, and characters, processed by multimedia systems is enormous. Hence, data compression technologies are crucial in constructing a multimedia system. Their representation technologies are also important. On the other hand, data for representing multimedia such as still images and sound are readily available on PCs since the technology for digitizing analog data has been advancing.

Still images	GIF	Format to save graphics, 256 colors displayable
	JPEG ¹³	Compression format for color still images, or the name of the joint organization of ISO and ITU-T establishing this standard
Moving Pictures	MPEG	Compression format for color moving pictures, or the name of the joint organization of ISO and IEC which established this standard
		MPEG-1 Data stored mainly on CD-ROM
		MPEG-2 Stored images like video; real-time images
		MPEG-4 Standardization for mobile terminals
Sound	PCM	Converting analog signals (sound, etc.) into digital signals
	MIDI	Interface to connect a musical instrument with a computer

1.1.4 Operations and Accuracy

Points

- There are two types of shift operations: arithmetic shift and logical shift.
- Operations in computer depend on the number of significant digits, so the result could have a margin of error.

Computers are equipped with circuits to perform the four fundamental arithmetic operations and shift operations. For operations such as computing 2^n , the operation speed improves by using shift operations (or moving digits). All computer operations are executed in the register. This register¹⁴ has only the limited number of significant digits, so an operation result may contain a margin of error.

¹³ (FAQ) There have been many exam questions that require some knowledge of organizations which have established functions and standards regarding JPEG and MPEG. Several keywords, such as JPEG, ISO, and ITU-T for still images, MPEG, ISO, and IEC for motion pictures, should be checked prior to the exam.

¹⁴ **Register:** It is the low-capacity, high-speed memory placed in the CPU for temporary storage of data; this includes general-purpose registers used by the CPU to carry out operations.

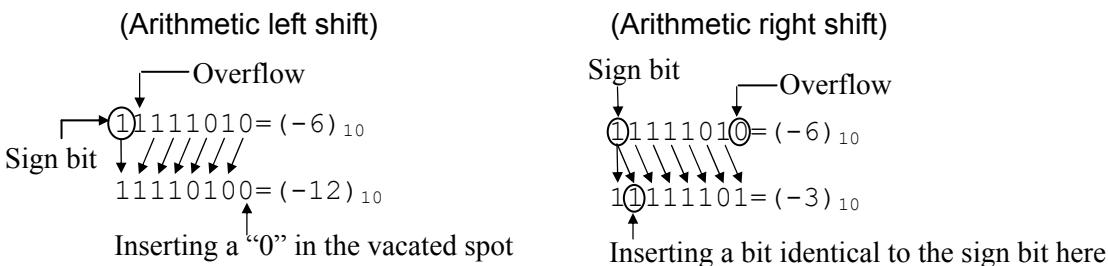
◆ Shift Operations

A **shift operation** is the operation of shifting (moving) a bit string to the right or to the left. Shift methods can be classified as shown below.

	Arithmetic shift	Logical shift
Left shift	Arithmetic left shift	Logical left shift
Right shift	Arithmetic right shift	Logical right shift

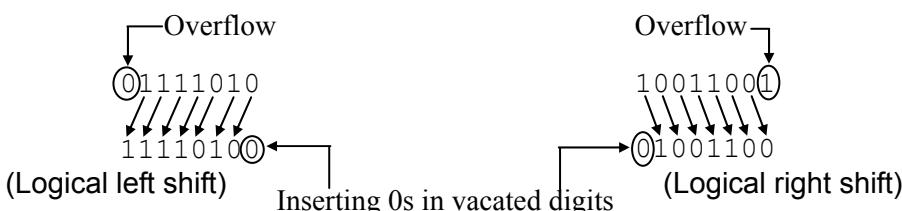
Arithmetic shift

An arithmetic shift is used when data is handled as numeric data with a positive or negative sign; it is an operation of shifting a bit string, except for the sign bit, representing a fixed-point number. The **arithmetic left shift** inserts a “0” in the rightmost place that has been made empty by the shift. In general, shifting left by n bits increases the number by 2^n times. The **arithmetic right shift**, on the other hand, inserts a value identical to the sign bit into the leftmost place that has been made empty by the shift. In general, shifting right by n bits reduces the number by 2^n times ($1/2^n$). Examples of 1-bit arithmetic shifts are illustrated below. Shifting 1 bit to the left doubles the value while shifting 1 bit to the right reduces the value to half.



Logical shift

Unlike an arithmetic shift, a logical shift does not handle the data as numeric data; rather, it handles the data merely as bit strings. It shifts an entire bit string of data and inserts 0s in places vacated by the shift. In **logical shifts**¹⁵, there is no such relation as a change by 2^n or 2^{-n} times in arithmetic shifts. Examples of 1-bit logical shifts are illustrated below:



¹⁵ (Note) In a logical shift, the figure indicates that the sign bit of 0 may become 1 after the shift. If the data is numeric, this means that a positive number changes to a negative number by the shift operation.

◆ Errors

Since operations are executed by a computer register with a limited number of digits, numerical values that cannot be contained in the register are ignored, resulting in differences between the operation results and true values. Such a difference is called an **error**.

Rounding errors

Since computers cannot handle an infinite (non-terminating) fraction, bits smaller than a certain bit are rounded off, rounded down, or rounded up to the value with the limited number of significant digits. The difference between the true value and the result of such rounding is called the **rounding¹⁶ error** (or round-off error).

Cancellation of significant digits

When one number is subtracted from another number almost identical to it, or when two numbers, one positive and the other negative, with almost identical absolute values are added together, the number of significant digits could drop drastically. This is called a cancellation of significant digits (or cancellation error).

$$\begin{array}{r}
)356.3622 \\
 -356.3579 \\
 \hline
 0.0043
 \end{array}$$

↑ Since the higher digits become 0, the number of significant digits decreases drastically.

Loss of trailing digits

When a very large number and a very small number are added together, or when one is subtracted from the other, some information (or a part thereof) in the lower digits, which cannot be contained in the mantissa, can be lost due to the alignment of the numbers. This is called a **loss of trailing digits**. In order to keep the error by loss of trailing digits small, it is necessary to do addition and subtraction in an order starting with numbers with small absolute values.

$$\begin{array}{r}
 356.3622 \\
 -0.000015 \leftarrow \text{Digits in extremely small place values get omitted.} \\
 \hline
 356.3622
 \end{array}$$

¹⁶ **Rounding:** It is a way to approximate a number by rounding off, rounding down, or rounding up so that it can be easily handled by people. For instance, if 2.15 is rounded to the nearest integer, it is rounded to 2, with an error of 0.15.

Quiz

- Q1** Express the decimal number 100 in the binary, octal, and hexadecimal notations.
- Q2** Perform arithmetic right and logical right shifts by 3 bits on the 8-bit binary number 11001100.
- Q3** Explain “cancellation of significant digits” and “loss of trailing digits.”

Information and Logic

Introduction

To make a computer perform a task, a program written according to rules is needed. Here, we will learn about logical operations, BNF, and reverse Polish notation. Logic operations are fundamental to the mechanism of operations. BNF is syntax rules for writing programs. Reverse Polish notation is used to interpret mathematical formulas written in programs.

1.2.1 Logical Operations

Points

- Logical sum, logical product, logical negation, and exclusive logical sum are the basic logic operations.
- The grammar of a programming language is written in BNF.

Basic logical operations include **logical product (AND)**, **logical sum (OR)**, **logical negation (NOT)**, and **exclusive logical sum (EOR, XOR)**. Logical negation is sometimes referred to simply as negation.

◆ Definitions of Logical Operations

The table below shows the notation of logical variables A and B , along with the meanings of their operations.¹⁷ Each logical variable is a 1-bit binary number, either a “1” or a “0.”

Logical operation	Notation	Meaning
Logical product (AND)	$A \cdot B$	The result is 1 only when both bits are 1s.
Logical sum (OR)	$A + B$	The result is 1 when at least one of the bits is 1.
Logical negation (NOT)	\bar{A}	Reversal of the bit (0 for 1; 1 for 0)
Exclusive logical sum (EOR, XOR)	$A \oplus B$	The result is 0 if the bits are the same, and 1 if the bits are not equal to each other.

¹⁷ (Note) The inside of a computer is equipped with circuits corresponding to logical product, logical sum, and logical negation. All operations are executed using combinations of these circuits.

◆ Truth Tables / Logical Operations¹⁸

A table summarizing results of logical operations is called a truth table.¹⁹ The following table shows logical product, logical sum, exclusive logical sum, and logical negation.

A	B	Logical product	Logical sum	Exclusive logical sum	Logical negation	
		$A \cdot B$	$A + B$	$A \oplus B$	\bar{A}	\bar{B}
0	0	0	0	0	1	1
0	1	0	1	1	1	0
1	0	0	1	1	0	1
1	1	1	1	0	0	0

Exclusive logical sum can be expanded, as shown below. Many questions can be easily answered if you know the expanded form of exclusive logical sum, so be sure to know the expanded formula.

$$A \oplus B = A \cdot \bar{B} + \bar{A} \cdot B$$

A	B	\bar{A}	\bar{B}	$A \oplus B$	$A \cdot \bar{B}$	$\bar{A} \cdot B$	$A \cdot \bar{B} + \bar{A} \cdot B$
0	0	1	1	0	0	0	0
0	1	1	0	1	0	1	1
1	0	0	1	1	1	0	1
1	1	0	0	0	0	0	0

◆ De Morgan's Theorem

A well-known set of formulas concerning logical operations is **De Morgan's theorem**. These laws give the relations, as shown below. You can easily memorize them if you remember to exchange logical products and logical sums when removing parentheses. Many questions can be easily answered if you know De Morgan's theorem, so be sure to know these formulas.²⁰

$$\overline{(A \cdot B)} = \bar{A} + \bar{B}$$

$$\overline{(A + B)} = \bar{A} \cdot \bar{B}$$

¹⁸ Negation of logical sum: $(A + B) = \bar{A} \cdot \bar{B}$. Negation of logical product: $(A \cdot B) = \bar{A} + \bar{B}$.

¹⁹ (Note) Some truth tables represent 1 with “T” (or true) and 0 with “F” (or false).

²⁰ (FAQ) Many questions can be easily answered if you know De Morgan's theorem. There are also many questions that can be easily answered if you know the expanded form of exclusive logical sum.

◆ Adder

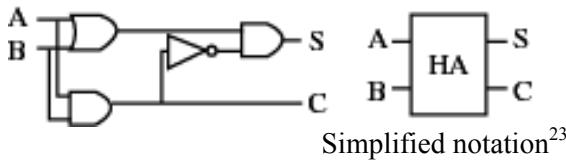
An **adder** is a circuit that performs addition of 1-bit binary numbers, consisting of AND, OR, and NOT logic circuits.²¹ There are half adders, which do not take into account carry-overs from lower bits, and full adders, which take into account carry-overs from lower bits.

Half adder (HA)

When a signal of a “0” or a “1” is sent to the inputs A and B of a circuit, the addition result appears as outputs C and S . Here, C indicates a carry-over, and S is the lower bit of the result of addition. The binary addition result is shown below. As seen here, C is the logical product, and S is the exclusive logical sum.²²

A	B	=	C	S
0	0	=	0	0
0	1	=	0	1
1	0	=	0	1
1	1	=	1	0

In the figure below, the circuit structure of a half adder is shown on the left. The figure on the right is simplified notation for a half adder, which is generally used.



²¹ (FAQ) There are many questions on the use of adders. As a shortcut, most of these questions can be answered if you know logical operations, but you can save time by knowing the operation results of adders.

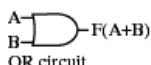
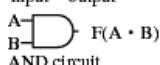
²² (Hints and Tips) Be sure to understand the binary 1-bit operations correctly. Be very careful since it is easy to make careless mistakes. The four additions of 1-bit binary numbers are shown below.

$$\begin{array}{r} A \quad 0 \quad 0 \quad 1 \quad 1 \\ B \quad + 0 \quad + 1 \quad + 0 \quad + 1 \\ \hline 0 \quad 1 \quad 1 \quad 10 \end{array}$$

If A and B are both 1s, simple addition gives the sum of 2, but in binary, in which only 0s and 1s are used, a carry-over takes place, resulting in the sum of “10.” If the adder circuit does not carry over, the output “0” is produced.

²³ You must have the circuit symbols memorized well. Be careful not to mix the AND and OR circuits.

input output

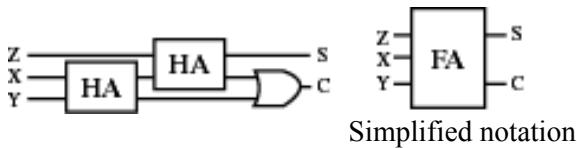


Full adder (FA)

For a full adder, there are three input values, one of which is the carry-over from the lower bit. Hence, a full adder adds three values X , Y , and Z . The addition results are as shown below. Unlike half adders, there are no general relations such as logical product and exclusive logical sum with a full adder.

X	Y	Z	=	C	S
0	0	0	=	0	0
0	0	1	=	0	1
0	1	0	=	0	1
0	1	1	=	1	0
1	0	0	=	0	1
1	0	1	=	1	0
1	1	0	=	1	0
1	1	1	=	1	1

In the figure below, the circuit structure of a full adder is shown on the left. As shown in the figure, a full adder consists of two half adders combined. The figure on the right is simplified notation for the full adder.



1.2.2 BNF

Points

- A means of strictly expressing the grammar of a programming language
- The terminal symbols cannot be further decomposed.

To define the grammar of a programming language (syntax definition), expressions free from any ambiguity are required. To express such a grammar, BNF (Backus-Naur Form) is often used.²⁴

BNF defines the rules of character orders by using characters; it also defines repetition and selection using appropriate character symbols. Since only characters are used in the definitions, the expressions are simple and close to the final descriptive style of the sentences. Furthermore, not only does BNF give unambiguous definitions, it is also considered to be easy to understand.

²⁴ (Note) BNF was first used to define ALGOL60, a programming language for technical calculations. BNF is a language to define syntax formally, not to stipulate any meanings. Hence, it cannot define every rule of a language, so today many extensions of BNF are used.

◆ Basic Forms of BNF

Expressions of BNF include sequence, repetition, and selection.

Sequence

$\langle x \rangle ::= \langle a \rangle \langle b \rangle$ ²⁵

This gives a definition which means “the syntax element x is a string of the character a and b .” The symbol “ $::=$ ” means “is defined to be.”

Repetition

$\langle x \rangle ::= \langle a \rangle \dots$

This gives a definition which means “the syntax element x is a repetition of the character a ,” It also means that the character a repeats once or more times.

Selection

$\langle x \rangle ::= \langle a \mid b \rangle$

This gives a definition which means “the syntax element x is either the character a or the character b .” If one of the options is missing, the following expression is used:

$\langle x \rangle ::= [\langle a \rangle]$

This gives a definition which means “the syntax element x is either the character a or the null character (blank).” The symbols “[]” means that it can be omitted.

◆ Terminal and Non-Terminal Symbols

A syntax element already defined can be used to define another element or even itself. These syntax elements are called **non-terminal symbols**.²⁶ Characters that are used directly in sentences are called **terminal symbols**.

In the following definitions, the underlined “ $\langle x \rangle$ ” is a non-terminal symbol whereas a , b , and c are terminal symbols.

$\langle y \rangle ::= \langle a \rangle \langle \underline{x} \rangle$
 $\langle x \rangle ::= \langle b \rangle \langle c \rangle$

²⁵ (Note) < >: These angle brackets are used when characters are consecutively placed or when the bounds are unclear; these do not have to be used.

²⁶ (Note) **Non-terminal symbols**: These are used to make the syntax definition easy to understand.

◆ Examples of BNF

For example, the syntax rules of “floating-point constant” are defined as follows:

```

<floating-point constant> ::= [<sign>]<radix constant>[<exponent>] | 
    [<sign>]<numeric string><exponent>
<radix constant> ::= [<numeric string>]<.><numeric string>|<numeric string><.>
<exponent> ::= <E>[<sign>]<numeric string>
<numeric string> ::= <numeral>|<numeric string><numeral>
<numeral> ::= 0|1|2|3|4|5|6|7|8|9
<sign> ::= +|-
  
```

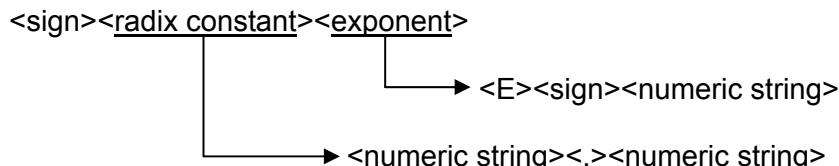
Let us follow the syntax rules above to see what `<floating-point constant>` looks like specifically.²⁷

For explanations, we number each element as follows:

```

<floating-point constant> ::= [<sign>]<radix constant>[<exponent>] | [<sign>]<numeric string><exponent>
    (1)           (2)           (3)           (4)           (5)           (6)
  
```

The definition of `<floating-point constant>` is separated by the line “|” which separates the group (1)~(3) from the group (4)~(6), so it has two possible forms. Let us take the first group (1)~(3) as our example to interpret what `<floating-point constant>` is. For clarity, we include in our example the part surrounded by [], which can be omitted. Each of the elements (1)~(3) can be further expanded as follows:



Here, there has risen a need to interpret `<numeric string>`.²⁸
`<numeric string>` is defined as follows:

```

<numeric string> ::= <numeral>|<numeric string><numeral>
  
```

Next, there is now a need to interpret `<numeral>`.

The numeral is defined as follows:

```

<numeral> ::= <0|1|2|3|4|5|6|7|8|9>
  
```

This means, for example, the following is possible:

```

<numeral> ::= 0
  
```

²⁷ (FAQ) An example of syntactical rules: often, questions follow the pattern of selecting the sentence that satisfies given syntactical rules.

²⁸ (Note) An expression with a character string combined with special symbols (\$, *, etc.) is called a regular expression. These designated characters are called meta-characters. Meta-characters have specific meanings. In UNIX, Windows, etc., if one searches for a file by entering “*.jpg,” then the system looks for all files with the extension “jpg.” Here, the symbol “*” is a meta-character.

Further, considering the definitions of <numerical string> and <numeral>, “0” itself is also a <numerical string>. Therefore, we can have the following:

<numerical string> ::= 0<numeral>

This allows “01” to be also a <numerical string>. We can go on.

<numerical string> ::= 01<numeral> ,

The allows “012” to be also <numerical string>. Hence, <numerical string> is any consecutive string of numerals. Hence, for example, <radix constant> can look as follows:

<radix constant> ::= 123.456

Thus, the interpretation of <floating-point constant> can give us the following example:

<sign><radix constant><exponent>
+ 123.456 E+123

Of course, the <sign> can be a negative sign “–” or can be omitted altogether, so the following strings can also be floating-point constants.

-123.456E-123
-123.456E123

As you can see, if all specific forms are to be expressed, that would result in an enormous amount of information. BNF is thus used to give general definitions to avoid such a situation.

1.2.3 Reverse Polish Notation

Points

- Reverse Polish Notation is a way to mechanically interpret mathematical formulas.
- It is characterized by two variables followed by an operator.

Reverse Polish Notation is a method of expressing mathematical formulas we use every day in a form more easily processed by computers. The basic concept of this notation is that the operators are written toward the end as opposed to the middle of a formula.

For example, $X = A + B * C$ means “Calculate the product of B and C, add A, and then move the result to X.” This is expressed by extracting the underlined parts as follows:

XABC*+=

◆ Conversion of Mathematical Formula to Reverse Polish Notation

For example, let us convert “ $e = a - b \div (c + d)$ ” into Reverse Polish Notation.²⁹ The order of operations is the usual order followed in performing mathematical operations. The underlined part is to be calculated first³⁰:

$$(1) \ e = a - b \div (c + d)$$

“(c+d)” is converted to Reverse Polish Notation. $\rightarrow cd+$

Let us call this string “P.”

$$(2) \ e = a - b \div P$$

“ $b \div P$ ” is converted to Reverse Polish Notation. $\rightarrow bP\div$

Let us call this string “Q.”

$$(3) \ e = a - Q$$

“ $a - Q$ ” is converted to Reverse Polish Notation. $\rightarrow aQ-$

Let us call this string “R.”

$$(4) \ e = R$$

“ $e = R$ ” is converted to Reverse Polish Notation. $\rightarrow eR=$

(5) Re-write P, Q, and R in Reverse Polish Notation (underlines indicate where replacement has occurred):

$$eR = \rightarrow eaQ- = \rightarrow eabP\div- = \rightarrow eabcd+\div- =$$

²⁹ (FAQ) Conversion into Reverse Polish Notation or into a mathematical formula is a very frequent theme on exams. It is best if you learn how to answer these questions intuitively.

³⁰ (Note) Intuitively, Reverse Polish Notation follows the order of operations in the formula when converting.

$e=a-b/(c+d)$
① $cd+$
② b ① \backslash
 $\rightarrow bcd+/-$
③ a ② $-$
 $\rightarrow abcd+/-$
④ e ③ $=$
 $\rightarrow eabcd+/-=$

◆ Conversion from Reverse Polish Notation into Mathematical Formula

A mathematical formula is converted into Reverse Polish Notation as follows:

- (i) Scan the Reverse Polish Notation from the beginning, looking for an operator.³¹
- (ii) Execute the operation indicated by the first operator, using the two variables immediately preceding the operator.
- (iii) Let the result of the operation of (ii) be a new variable, and repeat the first two steps (i) and (ii).

For example, consider the formula in Reverse Polish Notation “ $eabcd + \div - =$.” This is converted as follows:

Here, the underlined parts indicate the parts that can be converted.

- (1) Scan the Reverse Polish Notation “ $eabcd + \div - =$ ” from the beginning, searching for an operator. The first operator is “+,” so the focus is on that operator and the two variables preceding it, i.e., “ $cd+$.”

$$cd+ \rightarrow c+d \quad \text{Let this be } P. \rightarrow "eabP \div - ="$$

- (2) Scan the expression “ $eabP \div - =$ ” from the beginning, searching for an operator. The first operator is “ \div ,” so the focus is on that operator and the two variables preceding it, i.e., “ $bP\div$.”

$$bP\div \rightarrow b \div P \quad \text{Let this be } Q. \rightarrow "eaQ - ="$$

- (3) Scan the expression “ $eaQ - =$ ” from the beginning, searching for an operator. The first operator is “ $-$,” so the focus is on that operator and the two variables preceding it, i.e., “ $aQ-$.”

$$aQ- \rightarrow a-Q \quad \text{Let this be } R. \rightarrow "eR ="$$

- (4) Rewrite P , Q , and R as mathematical formulas (the underlined parts have been replaced).

$$eR = \rightarrow e = R \rightarrow e = a - Q \rightarrow e = a - b \div P \rightarrow e = a - (b \div (c + d))$$

Removing unnecessary parentheses, we get the following result:

$$e = a - b \div (c + d)$$

◆ Polish Notation

In **Polish Notation**, “ $a + b$ ” is expressed as “ $+ ab$ ”, for instance.³² Whereas the expression for this in Reverse Polish Notation is “ $ab +$,” Polish Notation places the operator in front of the variables. The fundamental concept is the same as that of Reverse Polish Notation. If “ $e = a - b \div (c + d)$ ” is converted to Polish Notation, we have the following:

$$e = a - b \div (c + d) \rightarrow = e - a \div b + cd$$

³¹ (Hints and Tips) In Reverse Polish Notation, once you find an operator, there will always be two variables that precede it immediately.

³² In Polish Notation, every operator is always followed by two variables. If there are not two variables, search for the next variable.

Quiz

- Q1** Given the values of logical variables x and y below, complete the table below by calculating the logical product, logical sum, and exclusive logical sum.

x	y	Logical product	Logical sum	Exclusive logical sum
0	0			
0	1			
1	0			
1	1			

- Q2** Explain “adders,” “half adders,” and “full adders.”

- Q3** Convert the formula “ $(a + b) \times (c - d)$ ” into Reverse Polish Notation.

1.3 Data Structures

Introduction

When considering procedures (algorithm) for a program, it is easier to create an algorithm if you put data in certain typical patterns. Such typical patterns are called data structures. Some popular data structures are arrays, lists, stacks, queues, and trees.

1.3.1 Arrays

Points

- Arrays can be used in every data structure.
- Arrays are referred to by index.

An **array** is a data structure consisting of multiple data of the same type. For example, imagine children lined up in a single row. This situation, in which objects with identical properties (here, the objects are “children”) are repeated, is similar to an array. Each child is identified as the “first child,” “second child,” etc. These numbers, “first, second, ...” are called index numbers. An array is used when multiple data of the same type are handled not individually but in relation to one another. The data is given an array name, and each data field (element) is identified by an index.

◆ 1-Dimensional Arrays

A **1-dimensional array** is conceptually shown below.³³

Index	1	2	3	4	...	25	26
Array <i>T</i>	<i>a</i>	<i>b</i>	<i>c</i>	<i>d</i>	...	<i>y</i>	<i>z</i>

Each array is given a name. In the example shown above, the name is “*T*.” To identify each element, an index is used. An index number represents the position of an element in the array.³⁴ For example, the fourth element “*d*” is designated by “*T*(4),” where the index number is in parentheses. In some languages, square brackets [] are used. In general, the *n*-th element of the array is denoted by “*T*(*n*).” By changing the value of *n*, we can indicate any element of the array.

³³ (FAQ) There are hardly any questions directly on arrays themselves. However, any question on a data structure or an algorithm always uses an array. Hence, you must understand arrays properly. More specifically, be sure that you understand how to use the index.

³⁴ (Hints and Tips) The index begins with 0 in some programming languages. Questions on algorithms on the exam may have indexes starting at 0 or 1, so care must be taken.

◆ 2-Dimensional Arrays

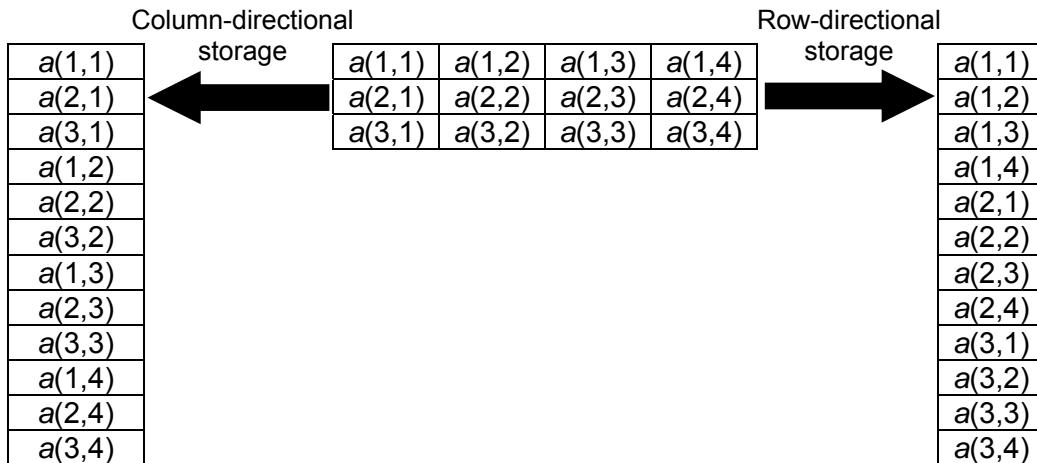
A **2-dimensional array** is conceptually shown below.³⁵

	Column 1	Column 2	Column 3	Column 4
Row 1	a (1,1)	a (1,2)	a (1,3)	a (1,4)
Row 2	a (2,1)	a (2,2)	a (2,3)	a (2,4)
Row 3	a (3,1)	a (3,2)	a (3,3)	a (3,4)

In general, the elements of a 2-dimensional array are identified using two sets of index numbers m and n . The notation is “ $a(m,n)$ ” or “ a_{mn} ,” where m is used for the row and n for the column. The array shown above is a 2-dimensional array with 3 rows and 4 columns, sometimes called a “3 by 4” array.

◆ Row-Directional Storage and Column-Directional Storage

When an array is stored in memory, it is stored as a 1-dimensional array. When the elements of a 2-dimensional array are stored as a 1-dimensional array, there are two methods that can be used, depending on the order in which the elements are stored: **row-directional storage** or **column-directional storage**.³⁶



In the figure above, take notice of the difference in the indexes. In column-directional storage, the x in “ $a(x,y)$ ” is changing first. In row-directional storage, y is changing first. When referring to an array, it is more efficient to look up the elements consecutively than to access skipping here and there. Hence, for efficient processing, the indexes are controlled as follows:

- Column-directional: x in “ $a(x,y)$ ” ($x = 1$ to m ; $y = 1$ to n) changes first.
- Row-directional: y in “ $a(x,y)$ ” ($x = 1$ to m ; $y = 1$ to n) changes first.

With this arrangement, referring to a 2-dimensional array is made more efficient when it is converted to a 1-dimensional array.

³⁵ (Hints and Tips) A 1-dimensional array is used when data is simply stored. A 2-dimensional array is used when storing objects like mathematical matrices.

³⁶ Among programming languages, Fortran uses column-directional storage whereas COBOL, PL/I, and C use row-directional storage.

1.3.2 Lists

Points

- Lists are characterized by being connected with pointers.
- Operations for a list are controlled by changing the values of pointers.

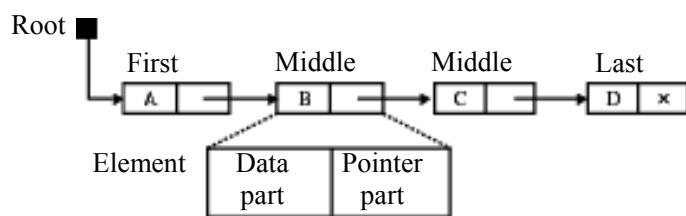
A **list** is a set of identical or similar data placed logically in one line (linear³⁷); its structure is similar to that of an array. The difference is that, whereas the elements of an array are placed physically right next to one another, the elements of a list can be placed at independent locations, and pointers establish connection between them. Because of this, sometimes arrays and lists are distinguished from each other by another pair of terms: an “array” to refer to a linear list, and a “list” to refer to a connected list because the elements are connected by pointers.

In general, the term “list” refers to “connected list”. In the explanations below, we refer to “connected list” simply as “list.”

◆ Structures of List

A list is a data structure in which the elements are connected by pointers. A pointer is information indicating the storage location (address) of the next element. Each element is connected by a pointer, so the elements need not be placed in order.

A list can have a variety of structures. The figure below is called a one-directional (unidirectional) list.³⁸



The pointer to the initial element is stored in the variable called the root. The last element (D) of the list has no element following it, so its pointer includes the symbol (X) indicating that the element is the last one in the list. In some programming languages, this symbol may be stored automatically; in others, any symbol can be given. The important thing is to assign a value that cannot exist as data.

³⁷ (Hints and Tips) The term “linear” here refers to a set of data placed in consecutive locations. An array is linear since the elements are placed in consecutive area. On the other hand, a (connected) list is a structure where the elements are linked by pointers, so they may not be placed in consecutive locations.

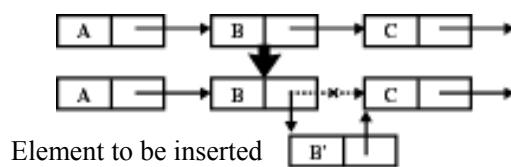
³⁸ Besides unidirectional lists, there are bidirectional lists and ring lists. A bidirectional list is one in which each element has a pointer indicating the previous element as well as a pointer indicating the next element. A ring list is one in which the last element has a pointer indicating the location of the first element.

◆ Basic Operations of List

There are some basic operations performed on lists; among them, insert and delete are particularly important operations.

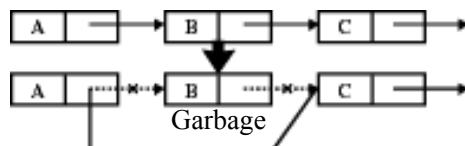
Insert³⁹

To insert an element into a given list, all we have to do is to change some pointers appropriately. First, in the pointer part of the element to be inserted to the list, enter the address of the element that is to immediately follow the element. Next, change the pointer part of the element immediately preceding the element to be inserted so that the pointer part can have the address of the element that is to be inserted.



Delete

To delete an element from a given list, just as in insertion, all we have to do is to change pointers. Change the pointer part of the element immediately preceding the element to be deleted so that the pointer can indicate the data immediately following the element to be deleted. The data to be deleted remains as garbage until the list is re-structured, so it is necessary to perform, in a timely manner, garbage collection⁴⁰ to delete unnecessary elements.⁴¹



³⁹ (FAQ) Many questions involve insertion into and deletion from a list. You must carefully consider which element it is whose pointer should be stored.

⁴⁰ **Garbage collection:** It is the procedure whereby small, fragmented unused memory and other areas not usable due to memory leak are combined together in order to increase usable memory space. If garbage is not collected, usable memory space continues to decrease and finally the system restart will be required.

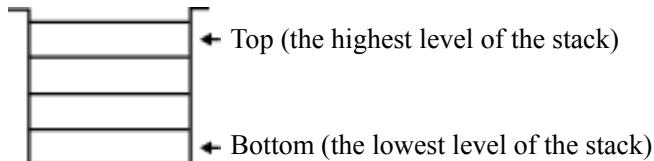
⁴¹ **Memory leak:** It means the situation wherein the main memory secured dynamically by an application is not released for some reason and remains in the main memory. To eliminate memory leak, garbage collection is necessary.

1.3.3 Stacks

Points

- Stacks are data structures of LIFO (Last-In First-Out).
- Stacks are used to manage the return addresses of subroutines.

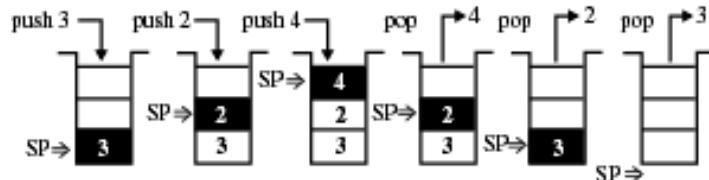
A **stack** is a data structure in which data insertion and deletion both take place on the same end of the list. Conceptually, it can be described as shown below.



The end where insertion (storage) and deletion (removal) of elements take place is called the top, and the other end is called the bottom. Insertion is called push-down while deletion is pop-up.

◆ Basic Operations of Stack

A **stack** is a data structure of **LIFO** (Last-In First-Out), meaning that the element stored last is first taken out. In the figure below, data is stored in the order of “3→2→4” and taken out in the order of “4→2→3.”⁴²



The pointer called stack pointer (SP) is used to keep track of where the top of the stack currently is; we can store an element into or remove an element from the position indicated by SP. The stack pointer sometimes points at the actual top element, and sometimes one place beyond it, depending on implementation.

◆ Use of Stack

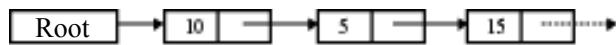
When a main program calls a subprogram (subroutine) or a function, often the return address of the program being executed is stored in a stack; when the subprogram is completed, the return address of the main program is taken from the stack to return the control. Further, if a subprogram calls other subprograms, the return addresses of the called programs are stored in the stack each time in sequence.⁴³

⁴² (FAQ) Many questions involve stacks. The pattern is that frequently there are questions asking what happens to the contents of a given stack when push and pop are repeated.

⁴³ Using a stack, a subprogram can be called from within another subprogram. Every time a subprogram is called sequentially, the return address is stored into the stack. Since taking out follows the order opposite the order in which storing took place, the subprograms are returned in the opposite order as well. A structure wherein a subprogram is called from within another subprogram like this is referred to as nested structure.

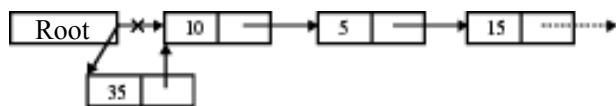
◆ Implementation of Stack using List

Using the structure of a list, we can implement a stack. In case of the following list, an element can be added to or deleted from the top of the list.



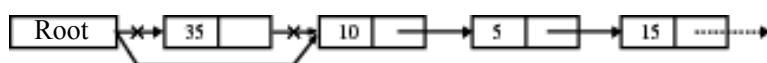
Insert to the list

To add the element “35” to the list, we place it as the first element, i.e., before the element “10.”



Delete from the list

We delete the first element “35” from the list, which is inserted in the above process. As a result, by combining insertion and deletion, we can implement the stack.

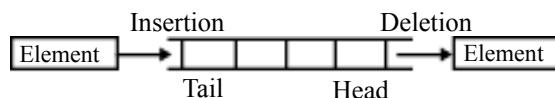


1.3.4 Queues (Waiting lists)

Points

- Queues are data structures of FIFO (First-In First-Out).
- Queues are used for online transaction processes.

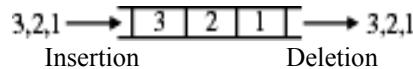
A **queue** is a data structure in which insertion takes place at one end while deletion (taking-out) occurs at the other end. Conceptually, it is described as shown below.



The first data in a queue is called the **head** while the last data is called the **tail**. A queue is sometimes referred to as a waiting list; this name came from the concept of processing sequentially.

◆ Operations of Queue

A queue is of the type referred to as FIFO (First-In First-Out), meaning that the element stored first is taken out first. In the figure below, the data is stored in the order of “1 → 2 → 3” and are taken out in the order of “1 → 2 → 3.”



In a queue, new data is always stored (enqueued) after the last data, and the first (oldest) data is always deleted (dequeued) first.⁴⁴

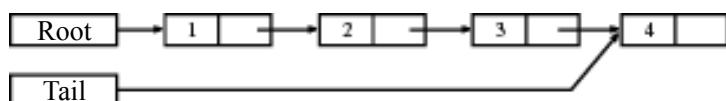
◆ Examples of Queues

In multiple programming, programs waiting to be executed are placed into the queue for execution as long as their priorities are equal, and they wait for the CPU to be available. In online transaction⁴⁵ processing, messages (electronic texts) are entered into a queue and processed in the order of entrance.

◆ Implementation of Queue using List

To implement a queue using a list, find the pointer that indicates the position of the last element of the list. Insertion is performed at the end of the list, and deletion is performed at the head.⁴⁶

Suppose there is a list as shown below. Here, the pointer to the last element is referred to as the “tail” for convenience.



Since we assume here that the element is to be added at the end of the list, the figure above indicates that the elements were added and stored in the order of “1 → 2 → 3 → 4.”

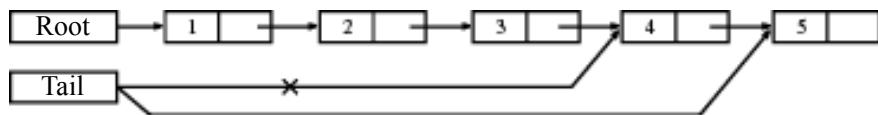
⁴⁴ (Note) Examples of queues are seen all around us in daily life. For example, a line of people waiting to purchase train tickets from a ticket vending machine is a queue, as those who joined the line first purchase tickets first. Because of this metaphor of people waiting in lines, sometimes a queue is called a waiting list.

⁴⁵ **Online transaction processing:** It is the processing mode in which a process request is immediately executed and the result is returned, such as seat-reservation systems of trains and airlines. For example, when ticketing is requested for a train ticket, the ticket is immediately printed. A request for processing is called a transaction.

⁴⁶ (Hints and Tips) A time-sharing system (TSS) appears on the surface as an online transaction process, but the method of processing is completely different. A queue processes tasks in the order in which they arrived; a TSS splits the processing time among the tasks. So even if a program (or a terminal) does not finish its processing, after a certain amount of time has elapsed, another program (or a terminal) begins its processing. A TSS is accomplished by multiprogramming.

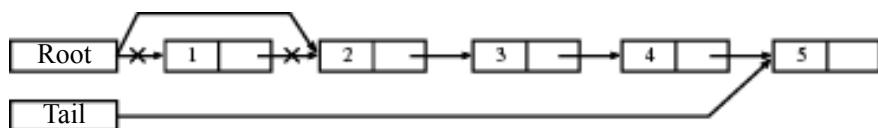
Insert into the list⁴⁷

The figure below shows how “5” was inserted. The pointer value that indicates the tail has been switched to point to “5.” Also, the pointer of the element “4,” which used to be the last element, has been changed so that it can point to the element just inserted.



Delete from the list

Since the first element is “1,” the root pointer is changed to point to the element “2.” Read the element “1” and see what the pointer says; that should be pointing to the position of the element “2.”

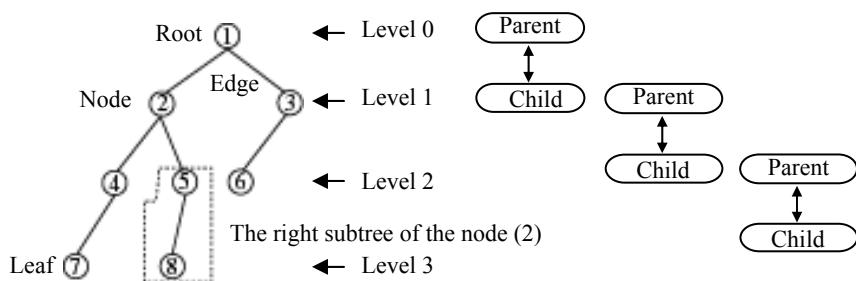


1.3.5 Trees

Points

- Trees clearly indicate a hierarchical structure.
- Among the various types of trees, binary trees are to be thoroughly understood.

A **tree** is a data structure that expresses the hierarchical structure between elements. It is used for the organizational chart of a company, system configuration, etc. It has a **root** at the top, and **nodes** are joined by **edges** (branches). A node directly above another node is called a **parent**, and a node directly below another is called a **child**.⁴⁸ Each node is placed at a level showing the degree of depth; the root is at level 0. A node without any children is called a **leaf**. A part of a tree is called a **subtree**. Given a node, the subtree to the left of it is called the **left subtree**; the one on the right is the **right subtree**.

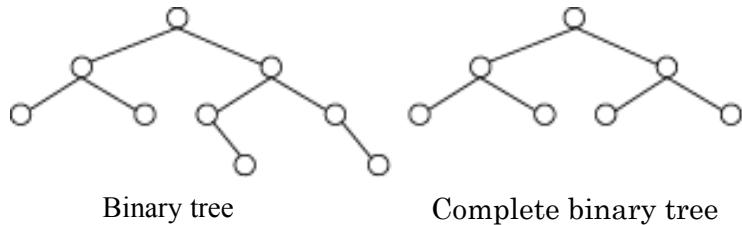


⁴⁷ **Multiprogramming:** It is a method where multiple programs appear to be running at the same time. No computers can actually execute multiple programs concurrently. Hence, the computer uses time-sharing to switch, at short time intervals, the program being executed so that it can appear as though multiple programs are being executed concurrently.

⁴⁸ (Hints and Tips) A pointer is used for a parent to indicate its child. Each parent thus has as many pointers as its children.

◆ Binary Trees and Complete Binary Trees

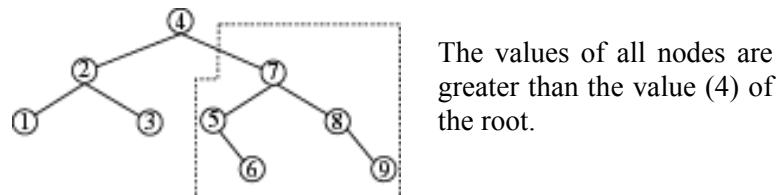
A tree in which each node has no more than two children is called a **binary tree**. If a binary tree is such that all leaves are at the same depth, or if the difference of depth between any two leaves is 1 or less and the leaves are laid out from the left, then such a tree is called a **complete binary tree**.⁴⁹



◆ Binary Search Trees

A **binary search tree** is a binary tree such that the value of an element is assigned to each node under the following restriction:⁵⁰

Value of the left child < Value of the parent element < Value of the right child



⁴⁹ (FAQ) On the Common FE Exam, questions involve only binary trees. Keep straight in your mind the characteristics of various binary trees, such as complete binary trees, binary search trees, and heaps.

⁵⁰ (Hints and Tips) In a binary search tree, note that the element with the minimum value is the leftmost leaf while the element with the maximum value is the rightmost leaf. This is a characteristic of a binary search tree.

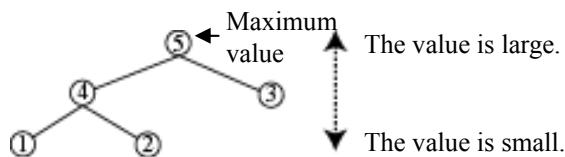
◆ Heaps

A binary tree is called a **heap** if the node values are assigned from the root level and from left to right on the same level with the following conditions:⁵¹

value of a parent element > value of a child element
(or value of a parent element < value of a child element)

The heap which meets the former condition is called the max-heap, and the min-heap for the latter condition.

As a result, elements with large (or small) values are close to the root whereas elements with small (or large) values are toward the leaves. It is a data structure suitable for retrieving a maximum (or minimum) value since the root is the element with the largest (or smallest) value.



1.3.6 Hash

Points

- Hash is the concept of using the key values directly as the index.
- Two methods to avoid collisions are the open-address method and the chain method.

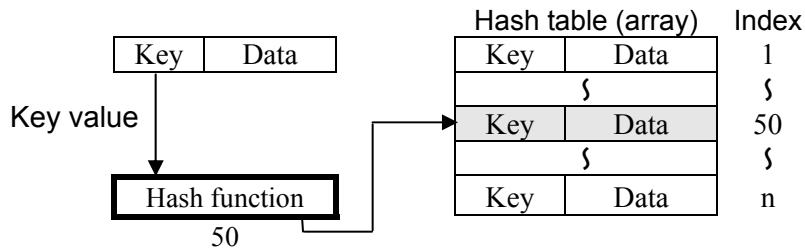
Hash is the concept of using key values directly as the storage locations of data. For example, suppose there is an array H of size 100. If the key values are two digits from 01 through 99 without duplication, these key values can be directly used as index numbers. This is called the **direct search method**.

However, it is rare that key values can be directly used as index numbers. Thus, to convert key values to index numbers, a hash function⁵² is used to calculate hash values, which are then used as index numbers. The array that stores elements using such a method is called a **hash table**.

Consider now the hash function that divides a given key value by the number of elements in the array and adds 1 to the remainder.

⁵¹ (Hints and Tips) Note that the heap shown here has the maximum value at the root. Take out the root, restructure the heap, and repeat this process; this way, you can take out the elements in the order of their values, from the largest to the smallest.

⁵² **Hash function:** A function that calculates data addresses (index numbers, etc.) from key values



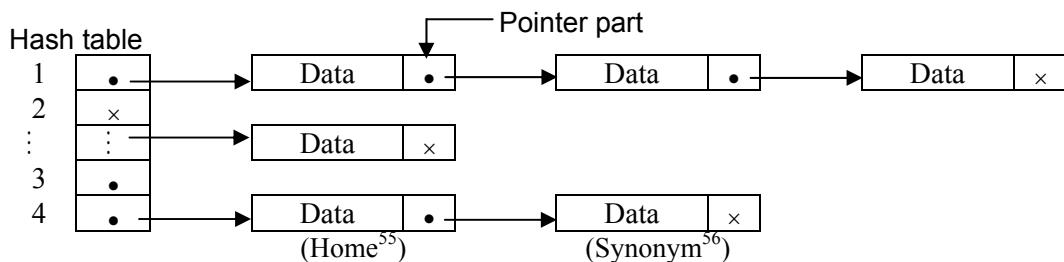
If there are n elements, then the remainder will be 0 through (n – 1), so adding 1 will give hash values of 1 through n. These can then be used as index numbers to be stored in the array.⁵³

However, the keys involve a variety of values, so the same index number can be produced from different key values by calculating the index number (hash value) 1 through n using the hash function. When the same hash value is generated in this way, it is called a **collision**.⁵⁴

◆ Chain Method (Open Hash Method)

This is the method of using a list to store elements with the same hash value when a collision occurs. The hash table contains in advance only the pointer indicating the first data of the list.

The figure below shows an example in which three pieces of data are stored in the position with index number 1 in the hash table. This hash table has a pointer indicating the first data. The position of the next data is found by looking up the pointer part when the first data is read.



⁵³ (FAQ) Many questions dealing with hash will ask you to calculate the storage position, and a "mod" function is often used as the hash function in such cases. "mod (a,b)" is the remainder of "a" divided by "b."

⁵⁴ **Collision:** When a hash function is used to calculate storage addresses, different key values could result in the same hash value. This is called a collision.

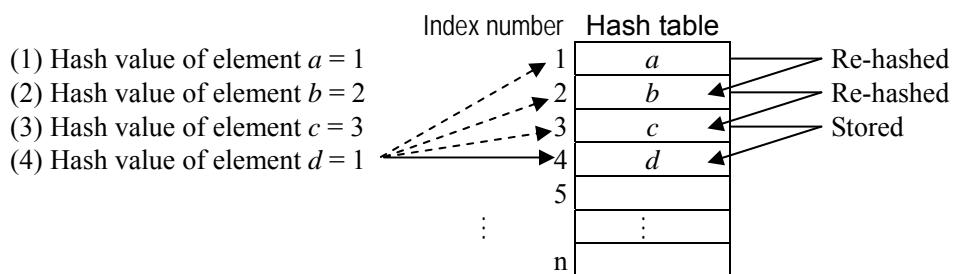
⁵⁵ **Home:** Data that had been stored first when a collision occurred

⁵⁶ **Synonym:** Data that came in later when a collision occurred

◆ Open Address Method (Closed Hash Method)

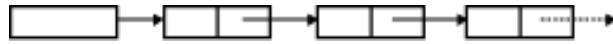
This is the method of dealing with collisions with re-hashing. Re-hashing refers to re-calculating the storage location when a collision occurs and storing the new data there if the location is empty.

For example, elements a , b , and c are stored in their respective positions (designated by the index numbers) according to the hash values calculated. Next, element d has the hash value 1, but position 1 is already taken by element a stored in that location. Here, for example, if the re-hashing method is determined in advance as “the original hash value + 1,” then the next position is the one designated by index number 2. But, that location is also taken, and the same goes for index number 3. Then, looking up position 4, that location is empty. As a result, element d is stored in the position with index number 4. If there happens to be no vacancy all the way to the end of the hash table, the search goes back to the first position of the table and looks for the first vacancy in a similar way.



Quiz

Q1 What do we call a data structure whose concept is shown in the following figure?



Q2 What do we call a data structure of “Last-In First-Out”?

Q3 What do we call a data structure of “First-In First-Out”?

Q4 Define “binary tree” and “complete binary tree.”

Q5 What do we call a binary tree with the following relation: “value of the left child < value of the parent element < value of the right child.”

1.4 Algorithms

Introduction

A set of procedures to solve a problem is called an algorithm. A figure expressing the set of procedures to obtain appropriate results is called a flowchart. Here, we have selected basic algorithms to study.⁵⁷

1.4.1 Search Algorithms

Points

- There are two types of search algorithms: linear search and binary search.
- In binary search, the elements are sorted in ascending or descending order.

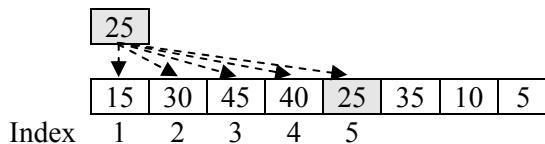
Search means finding an element in a table (1-dimensional array), and there are two types of search methods: **linear search** (sequential search) and **binary search**. Linear search can be performed regardless of how the elements are sorted, but binary search requires that the elements be sorted in ascending or descending order.

◆ Linear Search

This is the method of searching for the desired element in the table from the beginning of the table in order. It can be done regardless of how the elements are sorted, but it takes longer than binary search. If N is the number of elements, at least 1 (if the element to be sought is located at the beginning of the table) and at most N (if the element to be sought is at the end of the table or does not exist) comparisons are necessary.

In linear search, comparisons are made from index number 1 and continued as 1 is added to the previous index number until the index number reaches N .

For example, suppose “25” in the table is searched for by linear search. It is compared with the first value, the second, ..., the fifth. These numbers indicating the positions of elements are the index numbers.



⁵⁷ (FAQ) To express algorithms, questions on Morning Exams use flowcharts, while questions on Afternoon Exams use pseudo-language. Rules on the pseudo-language are not released, so it is a good idea to look through them in advance.

◆ Binary Search

This is an effective method when the elements in the table are sorted in ascending⁵⁸ or descending order.⁵⁹ Comparisons are made in sequence with the middle value of the table. After the first comparison, the right or left half of the table is discarded, and the middle value of the remaining part of the table is used for the next comparison. Since the range to be searched is reduced to half each time, this type of search is faster on average than linear search.

Let us explain a specific algorithm, using the following array as an example. Suppose that we are looking for the value “11.”

Index	1	2	3	4	5	6	7	8	9	10
Array T	0	1	3	5	7	9	11	13	15	17

First comparison

The range is the entire array. Let L be the lower bound and U be the upper bound of the range. Let M be the middle value (median).

Index	1	2	3	4	5	6	7	8	9	10
Array T	0	1	3	5	7	9	11	13	15	17

$L \leftarrow$ Search range $\rightarrow U$

$$M = (L + U) / 2 = (1 + 10) / 2 = 5.5 \rightarrow 5 \text{ (median, shaded value)}$$

The median can be obtained by rounding the quotient up or down; either is acceptable. Here, we round it down for our explanation.

$$T(M) = T(5) = 7$$

The value to be sought is “11,” so “11” cannot be found in the left half of the table, including the median value because the elements are sorted in ascending order and the desired value is larger than the median value.⁶⁰

Second comparison

Since the first comparison made it clear that the desired value is not in the left half of the table including the median, we change the search range. Here, the lower bound is changed to the value immediately to the right of the median. The value of L is then changed as follows:

$$L = M + 1 = 5 + 1 = 6$$

As a result, the search range changes as shown below.

Index	1	2	3	4	5	6	7	8	9	10
Array T	0	1	3	5	7	9	11	13	15	17

$L \leftarrow$ Search range $\rightarrow U$

In the same way as the first comparison, we find the new median as follows:

$$M = (L + U) / 2 = (6 + 10) / 2 = 8 \text{ (median, shaded value)}$$

$$T(M) = T(8) = 13$$

⁵⁸ **Ascending order:** Order in which data is sorted from the smallest key value to the largest key value

⁵⁹ **Descending order:** Order in which data is sorted from the largest key value to the smallest key value

⁶⁰ (Hints and Tips) When deleting the left half of a table, the new lower bound is the median plus 1; when deleting the right half, the new upper bound is the median minus 1.

Since we are comparing this to “11,” the desired value “11” cannot be in the right half of the search range including this new median. Since the elements are sorted in ascending order and the desired value is smaller than the value of the median.

Third comparison

Since the second comparison made it clear that the desired value is not in the right half of the search range including the median, we change the search range. Here, the upper bound is changed to the value immediately to the left of the median. The value of U is then changed as follows:

$$U = M - 1 = 8 - 1 = 7$$

As a result, the search range changes as shown below.

Index	1	2	3	4	5	6	7	8	9	10
Array T	0	1	3	5	7	9	11	13	15	17

$L \leftrightarrow U$
Search range

In the same way as the second comparison, we find the (new) median as follows:

$$M = (L + U) / 2 = (6 + 7) / 2 = 6 \text{ (median, shaded value)}$$

$$T(M) = T(6) = 9$$

Since we are comparing this to “11,” the desired value “11” cannot be in the left half of the search range including the median.

Fourth comparison

Since the third comparison made it clear that the desired value is not in the left half of the search range including the median, we change the lower bound in the same way as the second comparison.

$$L = M + 1 = 6 + 1 = 7$$

As a result, the search range is as shown below.

Index	1	2	3	4	5	6	7	8	9	10
Array T	0	1	3	5	7	9	11	13	15	17

$L=U$
Search range

$$M = (L + U) / 2 = (7 + 7) / 2 = 7 \text{ (median, shaded value)}$$

$$T(M) = T(7) = 11$$

Since we are comparing this to “11,” we can find the desired value.⁶¹

Procedure when search fails

Suppose, for example, that we search for “10.” At the fourth comparison, the formula “ $T(M) = 11 > 10$ ” holds true, so we have to change the upper bound of the search range. The new search

⁶¹ (Hints and Tips) The element “11,” which is $T(7)$ in array T , was found after 4 comparisons in binary search. Linear search requires 7 comparisons to find the value.

range is as follows:

$$L = 7 \text{ (remains unchanged)}$$

$$U = M - 1 = 7 - 1 = 6$$

Since L is the lower bound and U is the upper bound, we should have " $L \leq M$," but now we have " $L > M$." When this inequality holds, we determine that the desired element is not present.

◆ Comparison between Linear and Binary Search

When binary search is used to find "11," it is found after 4 comparisons. However, in case of linear search, since "11" has the index number "7," it takes 7 comparisons. Consequently, even though binary search is more complex, the number of comparisons is reduced.

However, consider searching for "0," the index of which is "1." Linear search can find it at the first comparison while binary search takes 3 comparisons. Here, linear search is faster.

To address this issue, there is a concept called the mean number of comparisons. When the number N of elements is very large, this value tells us how many comparisons are required on average. We omit detailed explanations here, but this is obtained by the following formulas:⁶²

$$\text{Mean number of comparisons for linear search} = N/2$$

$$\text{Maximum number of comparisons for linear search} = N$$

$$\text{Mean number of comparisons for binary search} = [\log_2 N]$$

$$\text{Maximum number of comparisons for binary search} =$$

$$\text{average number of comparisons for binary search} + 1$$

Let me add a word on the square brackets [] used in $[\log_2 N]$. In general, $\log_2 N$ is not an integer, but the number of comparisons must be an integer. Hence, [] denotes deleting, or truncating, the fractional part. For example, $[10.513]$ is 10.

1.4.2 Sorting Algorithms

Points	<ul style="list-style-type: none"> ➤ Be careful when manipulating index numbers in bubble sort, selection sort, and insertion sort. ➤ Recursive call is used in quick sort and merge sort.
---------------	--

Sort means rearranging elements and/or records of an array in a certain order according to a key. Arranging elements from the smallest key value to the largest is called **sorting in ascending order**, and ordering them from the largest key value to the smallest is called **sorting in descending order**.

Sorting the contents of an area in a program, such as an array, is called **internal sorting** whereas sorting data stored in an external device such as records in a file is called **external sorting** (file sorting).⁶³ Typical methods for internal sorting include bubble sort, selection sort,

⁶² (FAQ) The average number of comparisons and the maximum number of comparisons in binary search are frequently asked on exams, so it is a good idea to have these formulas memorized.

⁶³ (Hints and Tips) Questions involving internal sorting on the Common FE Exams are almost always about array manipulation. Be careful not to switch the index numbers when data is switched.

insertion sort, quick sort, merge sort, shell sort, and heap sort.⁶⁴

◆ Bubble Sort

In bubble sort, each adjacent pair of elements is sequentially compared and exchanged if necessary. In case of sorting in ascending order, the maximum value is put as the last element in the array. Next, going back to the beginning, the values are checked and exchanged when necessary. On the second run, the element at the end of the array is outside the sorting range. Continuing this process, the range gets smaller each time, and the sorting ends when the first and second elements are compared.

Below is an example of sorting in ascending order.

Before sorting	5	4	3	2	1:	
First run	5	↔	4	3	2	1: Exchanging 5 and 4
	4	5	↔	3	2	1: Exchanging 5 and 3
	4	3	5	↔	2	1: Exchanging 5 and 2
	4	3	2	5	↔	1: Exchanging 5 and 1
	4	3	2	1	5:	First run finished (the maximum value at the right end)
Second run	4	↔	3	2	1	5: Exchanging 4 and 3 (Values to the right of “ ” are sorted already.)
	3	4	↔	2	1	5: Exchanging 4 and 2
	3	2	4	↔	1	5: Exchanging 4 and 1
	3	2	1	4		5: Second run finished (second largest value at second from the right)
Third run	3	↔	2	1	4	5: Exchanging 3 and 2 (Values to the right of “ ” are sorted already.)
	2	3	↔	1	4	5: Exchanging 3 and 1
	2	1	3	4	5:	Third run finished (third largest value at third from the right)
Fourth run	2	↔	1	3	4	5: Exchanging 2 and 1 (Values to the right of “ ” are sorted already.)
	1	2		3	4	5: Fourth run finished (sorting complete)

◆ Selection Sort

Selection sort finds the maximum value (or the minimum value) from the array and exchanges it with the element at the end of the array. Next, it finds the maximum (or minimum) value from the array except for the last element and exchanges it with the second-to-the-last element of the array. Repeating this procedure, selection sort ends when it compares the first and second elements of the array.⁶⁵

Below is an example of sorting in ascending order.

First run	5	←	4	3	2	→	1: Since 5 is the maximum value, it is exchanged with the last element “1.”
Second run	1	←	4	3	2		5: Since 4 is the maximum value, it is exchanged to the last element in the second run.

⁶⁴ (FAQ) Questions of internal sorting appear on the Common FE Exams. Bubble sort and selection sort have appeared very frequently, so be sure to understand their algorithms well.

⁶⁵ (FAQ) Bubble sort and selection sort very frequently appear on the exams. The questions are given in a variety of ways, such as on the contents of an array at an intermediate stage and filling in blanks of a flowchart. Be sure that you understand the algorithms well.

Third run 1 2 3 | 4 5: Sorting complete

◆ Insertion Sort

Insertion sort starts with an already sorted array, compares the element to be inserted with the elements in the array, starting from the back, and inserts the element in the appropriate location.⁶⁶ Below, the elements to the left of “|” are already sorted. Here, since there is only one element on the first run, it is already considered to have been sorted.

Below is an example of sorting in ascending order.

First run	5 4 3 2	1:	Since 4 is the least value, it is inserted in the appropriate location (before 5).
Second run	4 5 3 2	1:	Since 3 is the least value, it is inserted in the appropriate location (before 4).
Third run	3 4 5 2	1:	Since 2 is the least value, it is inserted in the appropriate location (before 3).
Fourth run	2 3 4 5	1:	Since 1 is the least value, it is inserted in the appropriate location (before 2).
	1 2 3 4 5:		Sorting complete

◆ Quick Sort

Quick sort selects a random value from the array and uses its key value as the pivot; the elements are divided into two groups: the first group in which all elements are less than the pivot and the second group in which all elements are greater than the pivot (equal values can go either way). Then, the same procedure is repeated for each group. This is continued until there is only one element in each group. As a result, the array is sorted.⁶⁷

Below is an example of sorting in ascending order. The underlined values are the pivots. The line “|” indicates a block boundary.

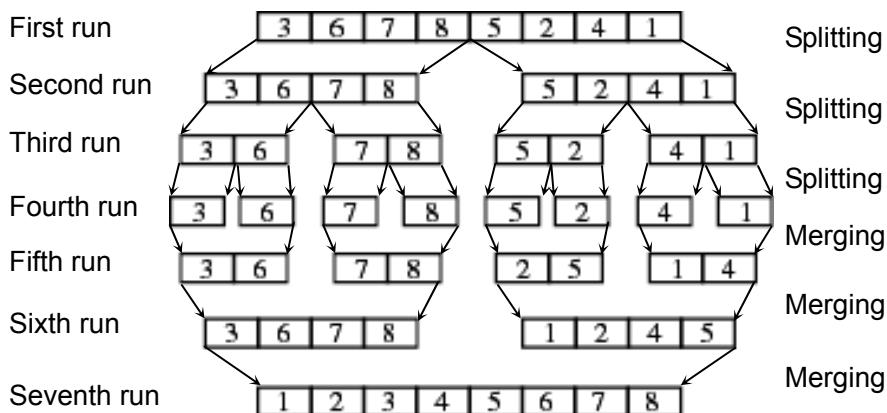
First run	2 5 6 4 1 <u>3</u> :	Divided into two blocks
Second run	2 1 <u>3</u> 5 6 <u>4</u> :	Divide each block into two
Third run	1 <u>2</u> <u>3</u> 4 5 <u>6</u> :	Divide each block into two (except for those groups with only 1 element)
Fourth run	1 2 3 4 5 6:	Sorting complete

⁶⁶ (Hints and Tips) When finding a location for insertion in insertion sort, the element to be inserted is compared from the back of an already sorted array. For example, on the third run here, the comparison will be “2 and 5,” “2 and 4,” and “2 and 3” in this order.

⁶⁷ (Note) Quick sort and merge sort differ from each other in the number of elements involved in splitting processes, but they use the same method. In such cases, a method known as “recursive call” is used.

◆ Merge Sort

In merge sort, two or more arrays, each of which is already sorted, are merged together to form one sorted array. In merge sort, splitting is repeated until each group has only one element. When each group has only one element, the elements are merged together in sequence.⁶⁸ Below is an example of sorting in ascending order.



◆ Shell Sort

This is an improved form of insertion sort; the sorting is made faster by increasing the moving distances of the elements.

First, the elements are sorted roughly by using insertion sort with gaps of a certain size. Then, insertion sort is used again to complete the sorting operation.

Below is an example of sorting in ascending order. Initially the gap is set to size 2, i.e., sorting every other element only. Then, the gap is made 1, and insertion sort is used.

Unsorted	2	4	5	3	1:	
First run	<u>2</u>	4	<u>5</u>	3	<u>1</u> :	Every other element is sorted (the underlined elements are sorted).
	<u>1</u>	4	<u>2</u>	3	<u>5</u> :	First run complete
Second run	1	<u>4</u>	2	<u>3</u>	5:	Every other remaining element is sorted (the underlined elements are sorted).
	1	<u>3</u>	2	<u>4</u>	5:	Second run complete
Third run	1	2	3	4	5:	Third run complete (sorting complete)

The reason that such a complicated method is used is the insert sort does not necessarily require exchanging of elements. For example, consider the following situation.

Case A: 2 4 6 | 1 ...
 Case B: 2 4 6 | 8 ...

⁶⁸ **Recursion:** It is a process in which a function calls itself from within itself. In Pascal and C, “recursive call” is allowed, but COBOL and Fortran do not allow this.

In Case A, in order to decide where to insert the “1,” comparisons are made $6 \rightarrow 4 \rightarrow 2$. Then, all the elements need to be moved over to make space to insert the “1.” In contrast, in Case B, as soon as the value is compared to “6,” the insertion location is obtained, without any sliding.

Hence, the amount of processing in insertion sort depends on how the elements are originally ordered. Shell sort reduces the work of sliding/moving elements by roughly sorting first.

◆ Heap Sort

A heap is a binary tree in which every subtree has the property that a parent has a value larger than its children. If the root element is picked, we can obtain the maximum value while the remaining elements can be re-structured to form a heap. We can again pick the root, which gives us the element with the second largest value. In other words, by repeating root extraction and re-structuring the heap, sorting can be achieved. This sorting method using a heap is called heap sort.⁶⁹

1.4.3 String Search Algorithms

Points	<ul style="list-style-type: none"> ➤ In general, string search compares one character at a time. ➤ Methods for string search include the brute-force (naïve) method, the Boyer-Moore method, etc.
---------------	---

String search means the process of looking for a designated sequence of characters in a text (character string). In most cases, strings are in arrays where each cell stores one character and is referenced by index. Two arrays are then given: the text and the designated string (pattern). The algorithm then searches for the pattern string in the string of the text.

In the example below, we want to check that string S , which is “XYZ,” is present in cells 6~8 and cells 10~12 in string R . It is obvious by visual inspection, but it is actually rather difficult to create an algorithm to check this.

String S	X	Y	Z	← Pattern								
String R	P	Q	A	C	Z	X	Y	Z	← Text			
Position	1	2	3	4	5	6	7	8	9	10	11	12

⁶⁹ (FAQ) For quick sort, merge sort, insertion sort, heap sort, and shell sort, questions generally deal with the concept of each, so you should understand how each sort processes the data.

◆ Brute-Force Search Method

Brute-force search is the method in which the desired string is searched for by comparing the characters one by one from the beginning of the array in order. This is a concept identical to that of linear search. The search ends when the last character of the array is compared with the last character of the string to be found. Below is a specific explanation using an example.⁷⁰

Text	P	Q	A	B	C	Z	X	Y	Z	R	X	Y	Z
Pattern	X	Y	Z										

- (1) The first character of the pattern is compared with the first character of the text.

Text	P	Q	A	B	C	Z	X	Y	Z	R	X	Y	Z
Pattern	X	Y	Z										

- (2) Because of the mismatch, the second character of the text is now compared with the first character of the pattern.

Text	P	Q	A	B	C	Z	X	Y	Z	R	X	Y	Z
Pattern	X	Y	Z										

- (3) Repeat this process, and the first match occurs with the seventh character “X.”

Text	P	Q	A	B	C	Z	X	Y	Z	R	X	Y	Z
Pattern							X	Y	Z				

- (4) Having had a match, now the 8th character of the text is compared with the second character of the pattern.

Text	P	Q	A	B	C	Z	X	Y	Z	R	X	Y	Z
Pattern								X	Y	Z			

- (5) Since the second pair also matched up, the third characters are compared.

Text	P	Q	A	B	C	Z	X	Y	Z	R	X	Y	Z
Pattern									X	Y	Z		

Now we have determined that the string pattern S is present in the text string R.

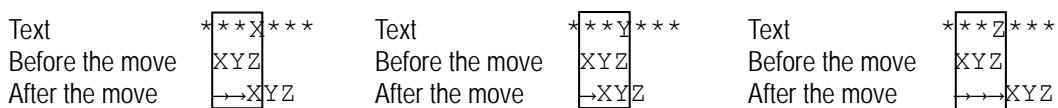
⁷⁰ (Hints and Tips) In string search, there needs to be an index for the string *S* and another index for the string *R*. When answering a question, the crucial point is to grasp how to use the indexes.

◆ Boyer-Moore Method (BM Method)

This is a method that takes the contents of the pattern string into account to eliminate waste. If the pattern and a string of the text do not match up, the number of characters that can be skipped depends on the right-end character of the search range of the text being compared.

Let us explain this specifically, using the same example as in brute-force search.

- (1) If the rightmost character of the portion of the text currently being compared with the string is “X,” the next possible place where the pattern can be matched is two characters ahead, so the next two characters are skipped.
- (2) If the rightmost character of the portion of the text currently being compared with the string is “Y,” the next possible place where the pattern can be matched is one character ahead, so the next character is skipped.
- (3) If the rightmost character of the portion of the text currently being compared with the string is “Z,” the next possible place where the pattern can be matched is three characters ahead, so the next three characters are skipped.



- (4) If the rightmost character of the text is not X, Y, or Z, then the situation is identical to that of (3), so the next three characters are skipped.⁷¹

⁷¹ (Note) In the BM method, it is necessary in advance to calculate the number of characters to be skipped. The example discussed here has a 3-character pattern, so the number is 2 for X, 1 for Y, and 3 for Z or any other characters. These need to be calculated before the string search begins.

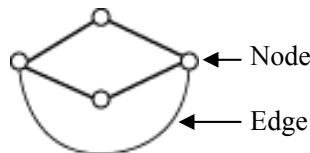
1.4.4 Graph Algorithms

Points

- A tree is a type of graph.
- The order in which tree search is performed can be breadth-first or depth-first.

A graph algorithm is an algorithm where the search is performed on a tree, one of the question-oriented data structures.⁷² Depending on the order of search, a graph algorithm can be breadth-first or depth-first. The depth-first order frequently appears on the exams, so be sure that you understand how to pick out the nodes using this method.

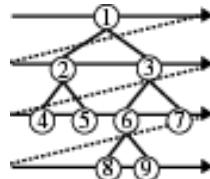
A graph consists of nodes and edges.⁷³ A node is a vertex that forms the graph whereas an edge is a segment connecting a point to a point. Below is an example of a graph.



A tree can be considered a graph in which not all nodes are connected to all others.

◆ Breadth-First Order

The search begins at the root and traverses from lower levels and from left to right. Below, the number at each node indicates the order in which the nodes are traversed.



⁷² **Question-oriented data structure:** A question-oriented data structure is a data structure often used to create a program. Since the algorithms using data is well-established, such a structure enables the programmer to write a program with few errors. Examples of question-oriented data structures include trees, stacks, queues, and lists.

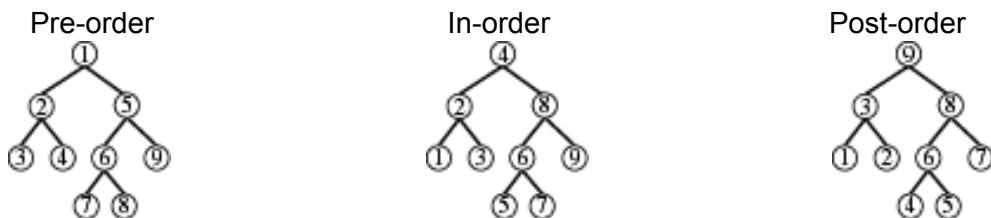
⁷³ (Hints and Tips) When you hear the term graph, you may think of a pie chart, a bar graph, etc., but in the world of mathematics, it refers to a set of points and edges.

◆ Depth-First Order

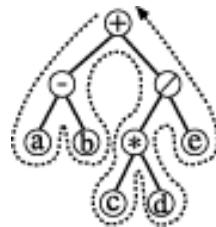
In depth-first search,⁷⁴ we start with the root and traverse from the left child and from leaves. Depending on the timing when the nodes are traversed, it can be classified as shown in the following table.

Search method	Order in which nodes are traversed
Pre-order	Parent, left child, right child, in this order
In-order	Left subtree, parent, right subtree, in this order
Post-order	Left subtree, right subtree, parent, in this order

Below, the number at each node indicates the order in which the node is traversed.



It is probably not very clear yet what the rules are for each of the search types, so let me add some explanation. In depth-first order, the search follows the order as shown below:



In pre-order, the node values are taken out whenever you traverse the left side of the nodes. Hence, the order is “+ – a b / * c d e.” In in-order, the node values are taken out whenever you traverse under the nodes. Hence, the order is “a – b + c * d / e.” In post-order, the node values are accessed whenever you traverse the right side of the nodes. Hence, the order is “a b – c d * e / +.”⁷⁵

⁷⁴ (FAQ) Depth-first order frequently appears on the exams. Understand well how the nodes are taken out in pre-order, in-order, and post-order.

⁷⁵ (Note) Note the result of traversing the tree to obtain symbols and variables. In pre-order, the result is “+ – a b / * c d e,” which is in Polish Notation. In in-order, the result is “a – b + c * d / e,” which is in standard mathematical notation. In post-order, the result is “a b – c d * e / +,” which is in Reverse Polish Notation.

Quiz

- Q1** In binary search, when the number of sorted data values is quadrupled, how much does the maximum number of comparisons increase by?
- Q2** Explain the characteristics of each of the sorting methods: “shell sort,” “bubble sort,” “quick sort,” and “heap sort.”

Question 1

Difficulty: **

Frequency: ***

- Q1.** There is a register which stores values in binary. After entering a positive integer x into this register, the operation “to shift the register value 2 bits to the left and to add x to the value” will be performed. How many times as large as x is the resulting register value? Here, assume that overflow due to shifting will not occur.

a) 3

b) 4

c) 5

d) 6

Answer 1**Correct Answer:** c

In general, if there is no overflow, shifting n bits to the left multiplies the value by 2^n while shifting n bits to the right multiplies the value by $1/2^n$. Shifting 2 bits to the left is to multiply by 2^2 , so if we let y be the calculation result, y is related to x by the following equation:

$$\begin{aligned}y &= x \times 2^2 + x \\&= x \times (2^2 + 1) \\&= 5 \times x \quad (\text{y is 5 times } x)\end{aligned}$$

- a) To make it 3 times as large, we would shift the register value 1 bit to the left and add x to it. Shifting 1 bit to the left multiplies the value by 2^1 , so the result would be as follows:

$$y = x \times 2^1 + x = 2x + x = 3x$$

- b) To make it 4 times as large, we would shift the register value 2 bits to the left, which multiplies the value by 2^2 , and the result would be as follows:

$$y = x \times 2^2 = 4x$$

- d) To make it 6 times as large, we would shift the register value 2 bits to the left and add this result to the result obtained by shifting the original register value 1 bit to the left. Shifting 2 bits to the left multiplies the value by 2^2 , and shifting 1 bit to the left multiplies the value by 2^1 , so the following would result:

$$y = x \times 2^2 + x \times 2^1 = 4x + 2x = 6x$$

Question 2

Difficulty: *

Frequency: ***

Q2. Which of the following is an appropriate description concerning the cancellation of significant digits?

- a) It means that the number of the significant digits is extremely reduced when a floating point number is subtracted by another whose value is almost equal.
- b) It refers to an error which occurs because the calculation result exceeds the maximum numeric value that can be processed.
- c) It refers to an error which occurs when rounding off (up or down) the numbers smaller than the lowest digit when the total number of digits in numerical representation is limited.
- d) It refers to the omission of the low-order digit of an operand when adding floating point numbers.

Answer 2

Correct Answer: a

Cancellation of significant digits is a phenomenon in which higher-order significant digits are lost in subtraction involving two values of the same sign which are close and in addition involving two values of the opposite signs whose absolute values are close. It occurs because computers process all numbers with only a finite number of digits. For instance, it occurs in the following calculation:

$$\begin{array}{r}
 123.4567 \\
 - 123.4556 \\
 \hline
 0.0011
 \end{array}$$

Here, the higher-order digits become 0, reducing the number of significant digits drastically.

- b) This describes an overflow.
- c) This describes a rounding error.
- d) This describes a loss of trailing digits.

Question 3

Difficulty: ***

Frequency: ***

- Q3.** The truth table below shows the results of logical operation “ $x @ y$.” Which of the following expressions is equivalent to this operation?

x	y	$x @ y$
True	True	False
True	False	False
False	True	True
False	False	False

- a) $x \text{ OR } (\text{NOT } y)$
 c) $(\text{NOT } x) \text{ AND } (\text{NOT } y)$
 b) $(\text{NOT } x) \text{ AND } y$
 d) $(\text{NOT } x) \text{ OR } (\text{NOT } y)$

Answer 3

Correct Answer: **b**

In logic operations, we assign “1” for “true” and “0” for “false.” It is easier to use familiar notation, so we shall use the following symbols:

$x \text{ AND } y \rightarrow x \cdot y$ (logical product)

$x \text{ OR } y \rightarrow x + y$ (logical sum)

$\text{NOT } x \rightarrow \bar{x}$ (logical negation)

Then, the logical expressions in the answer group can be rewritten as follows:

- | | |
|---|-------------------------|
| a) $x \text{ OR } (\text{NOT } y)$ | $x + \bar{y}$ |
| b) $(\text{NOT } x) \text{ AND } y$ | $\bar{x} \cdot y$ |
| c) $(\text{NOT } x) \text{ AND } (\text{NOT } y)$ | $\bar{x} \cdot \bar{y}$ |
| d) $(\text{NOT } x) \text{ OR } (\text{NOT } y)$ | $\bar{x} + \bar{y}$ |

Then we check to see which of the expressions in the answer group matches (has the identical results with) the given logic operation:

x	y	\bar{x}	\bar{y}	a)	b)	c)	d)	$x @ y$
				$x + \bar{y}$	$\bar{x} \cdot y$	$\bar{x} \cdot \bar{y}$	$\bar{x} + \bar{y}$	
1	1	0	0	1	0	0	0	0
1	0	0	1	1	0	0	1	0
0	1	1	0	0	1	0	1	1
0	0	1	1	1	0	1	1	0

Hence, the operation whose results match those of $x @ y$ is $\bar{x} \cdot y$.

Question 4

Difficulty: **

Frequency: **

- Q4.** When the syntax for numerical values is defined as shown below, which of the following expressions is treated as <numerical value>?

```

<numerical value> ::= <numerical string>|<numerical string>E<numerical string>|
                     <numerical string>E<sign><numerical string>
<numerical string> ::= <numeral>|<numerical string> <numeral>
<numeral> ::= 0|1|2|3|4|5|6|7|8|9
<sign> ::= +|-
```

- a) -12 b) 12E-10 c) +12E-10 d) +12E10

Answer 4

Correct Answer: **b**

This answer conforms to the third form (<numerical string> E <sign> <numerical string>) of the definition of <numerical value>.

This type of definition is called BNF notation (Backus-Naur Form). BNF notation is used as a way to formally denote the syntax of a programming language.

Overview of BNF notation is as follows:

- $\alpha ::= \beta$ → The left-hand side α is defined as the right-hand side β . In other words, $\alpha = \beta$.
- < α > → This denotes the variable α . < > can be omitted.
- | → This means “or.” “ $\alpha ::= \beta | \gamma$ ” means “ $\alpha ::= \beta$ ” or “ γ .”

“ $::=$ ” can simply be written “ $=$ ”

- a) By the definition of <numerical value>, “-” (<sign>) must follow “E.” The underlined part does not satisfy the definition. -12
- c) By the definition of <numerical value>, “+” (<sign>) must follow “E.” The underlined part does not satisfy the definition. $\pm 12E - 10$
- d) By the definition of <numerical value>, “+” (<sign>) must follow “E.” The underlined part does not satisfy the definition. $\pm 12E10$

Question 5

Difficulty: **

Frequency: **

- Q5.** A key is composed of 3 alphabetic characters. When the hash value h is decided with the following expression, which of the following collides with the key “SEP”? Here, “ $a \text{ mod } b$ ” represents the remainder when a is divided by b .

$$h = (\text{Sum of positions for each alphabetic character in the key}) \text{ mod } 27$$

Alphabetic character	Position
A	1
B	2
C	3
D	4
E	5
F	6
G	7
H	8
I	9
J	10
K	11
L	12
M	13

Alphabetic character	Position
N	14
O	15
P	16
Q	17
R	18
S	19
T	20
U	21
V	22
W	23
X	24
Y	25
Z	26

a) APR

b) FEB

c) JAN

d) NOV

Answer 5**Correct Answer:** **b**

A hash value is the result of converting the key by a hash function, which is used for hashing. The term “hashing” refers to a process of performing some sort of calculation on the key to convert it to an address value in order to obtain the storage address of the record in a direct organization file, for example. Here, the function used to obtain the address is called a hash function. If hashing generates the same hash value for two or more different keys, it is called a collision. Records that came in later when a collision occurred are called synonyms. Calculating the hash value for “SEP” by means of the given hash function, we can obtain the following:

$$\begin{aligned} h &= (\text{sum of positions for each alphabetic character used in the key}) \text{ mod } 27 \\ &= (19 + 5 + 16) \text{ mod } 27 \\ &= (40) \text{ mod } 27 \\ &= 13 \quad (40 \div 27 = 1 \text{ remainder } 13) \end{aligned}$$

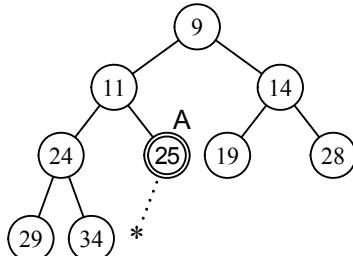
- a) “ARP” $(1 + 18 + 16) \text{ mod } 27 = 8$ ($35 \div 27 = 1$ remainder 8)
- b) “FEB” $(6 + 5 + 2) \text{ mod } 27 = 13$ ($13 \div 27 = 0$ remainder 13) — collision
- c) “JAN” $(10 + 1 + 14) \text{ mod } 27 = 25$ ($25 \div 27 = 0$ remainder 25)
- d) “NOV” $(14 + 15 + 22) \text{ mod } 27 = 24$ ($51 \div 27 = 1$ remainder 24)

Question 6

Difficulty: **

Frequency: **

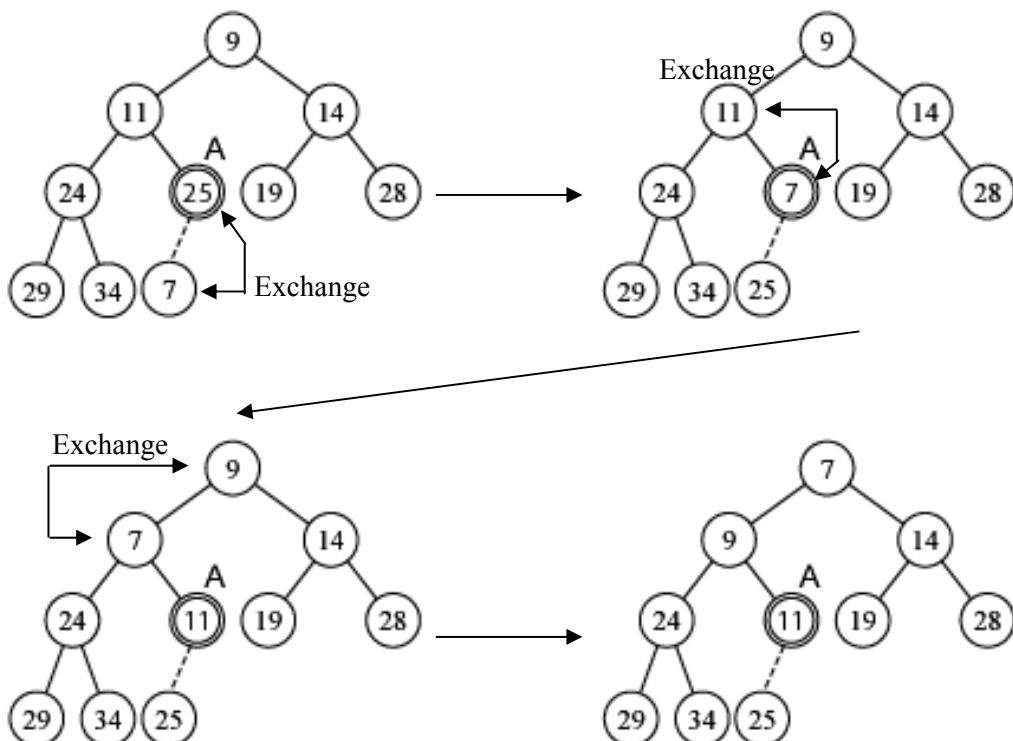
- Q6.** In the heap shown below, the value of a parent node is less than the values of child nodes. When inserting a node into this heap, an element is added at the very end. If that element is less than the parent node, the parent and child are exchanged with each other. If element 7 is added to the heap at the position marked by the asterisk (*), what element will end up at position A?



- a) 7 b) 11 c) 24 d) 25

Answer 6**Correct Answer:** **b**

Add the element to the given position and then repeat the procedure to exchange the child and the parent when the child element has a value smaller than the parent value. “7” is the added element here.



Now, the heap is complete. Hence, the element that ends up at position A (◎) is “11.”

Question 7

Difficulty: *

Frequency: ***

Q7. Which of the following terms expresses a characteristic of stack operations?

- a) FIFO b) LIFO c) LILO d) LRU

Answer 7

Correct Answer: **b**

A stack is a data structure of the type known as Last-In First-Out, where data stored last will be the first data to be taken out. The operation of inserting data into a stack is called a “push,” and the operation of taking data out of a stack is called a “pop.”

- a) FIFO (First-In First-out) is the data structure of a queue, where data stored first will be the first data to be taken out.
- c) LILO (LInux LOader) is a boot loader (program to load the OS into memory) that allows PCs to read Linux.
Translator's note: It seems natural that LILO means “Last-In Last-Out” in this question.
- d) LRU (Least Recently Used) means “least accessed in recent history” and is used as the page-replacing algorithm in a virtual memory system. This is the method of paging-out which discards the least recently accessed page.

Question 8

Difficulty: **

Frequency: **

Q8. The decision table below shows the conditions for creating reports from employee files. Which of the following can be concluded from this decision table?

Under age 30	Y	Y	N	N
Male	Y	N	Y	N
Married	N	Y	Y	N
Output Report 1	—	X	—	—
Output Report 2	—	—	—	X
Output Report 3	X	—	—	—
Output Report 4	—	—	X	—

- a) Report 1 contains the contents of Report 4 except for data on men age 30 and over.
- b) Report 2 contains all unmarried men.
- c) Men in Report 3 are also included in Report 2.
- d) Persons included in Report 4 are not included in any of the other reports.

Answer 8**Correct Answer:** d

Let the negation of “married” be “unmarried” and the negation of “male” be “female.” Now, read the answer group descriptions carefully. In the following explanation, the underlined parts indicate negation (N).

- a) The output conditions for Report 1 are “under 30, not male, married.”
→This is “under 30, female, married.”
The output conditions for Report 4 are “not under 30, male, married.”
→This is “at least 30, male, married.”
So, Report 1 contains females only. Report 4 contains males only, and removing those “males, at least 30” from Report 4 causes it to be the empty set. Hence, this description is wrong.
- b) The output conditions for Report 2 are “not under 30, not male, not married.”
→This is “at least 30, female, unmarried.”
So, report 2 contains females only, so it is not true that “all unmarried men” are included. Hence, this description is wrong.
- c) The output conditions for Report 3 are “under 30, male, not married.”
→This is “under 30, male, unmarried.”
The output conditions for Report 2 are “not under 30, not male, not married.”
→This is “at least 30, female, unmarried.”
So, Report 3 contains males only while Report 2 contains females only. There is no intersection. Hence, this description is wrong.
- d) By elimination this must be the correct answer, but let us check it. Organizing all of the output criteria for all of the reports from a), b), and c) in the answer group, we get the following:
Report 1: “under 30, female, married” (from “a”)
Report 2: “at least 30, female, unmarried” (from “b”)
Report 3: “under 30, male, unmarried” (from “c”)
Report 4: “at least 30, male, married” (from “a”)

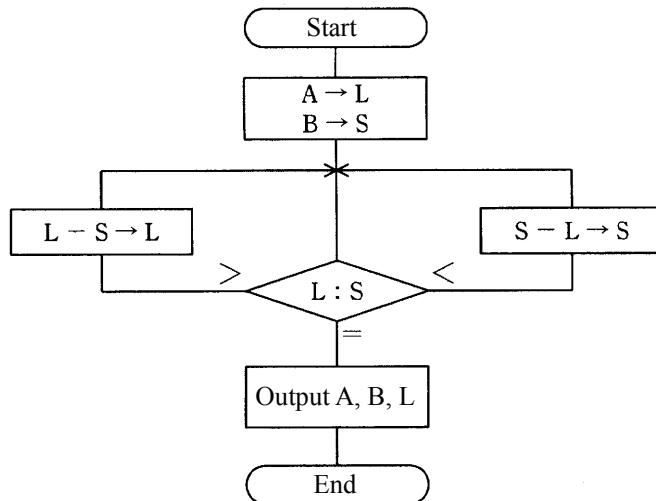
The condition “at least 30” for Report 4 is also for Report 2, but the other conditions are in negation of each other, so no person is included in both. Further, the condition “male” is also for Report 3, but the other conditions are in negation of each other also. Similarly, the condition “married” is also for Report 1, but again the other conditions are in negation of each other. Therefore, no other reports contain those persons contained in Report 4. This is the correct description.

Question 9

Difficulty: **

Frequency: ***

- Q9.** The flowchart below illustrates the Euclidean algorithm for obtaining the greatest common divisor of values “A” and “B,” by repeated subtraction. When “A” is 876 and “B” is 204, how many comparisons are required for completion of this process?



- a) 4 b) 9 c) 10 d) 11

Answer 9**Correct Answer:** **d**

The Euclidean algorithm is an algorithm to obtain the greatest common divisor of two integers A and B. However, you need not know this algorithm; all you have to do is to track how the data changes. First, by “A→L” and “B→S,” the values for which the greatest common divisor is to be obtained are rewritten as L and S. The algorithm then determines which is greater and subtracts the smaller from the larger. Then, in case of “L=S,” the algorithm stops.

Since initially A=876 and B=204, we subtract B (=S) from A (=L) as many times as possible. Note that the values must be compared first before the subtraction takes place.

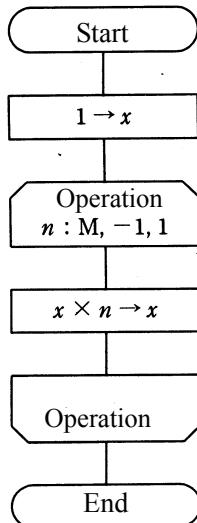
- (1) Under the condition of L=876 and S=204, repeat subtraction until L<S. Since $876 \div 204 = 4$ with remainder 60, the subtraction and replacement “L – S → L” can be executed 4 times before “L < S” is satisfied. Hence, the comparison (L:S) occurs 4 times.
- (2) Under the condition of L=60 and S=204, repeat subtraction until L>S. Since $204 \div 60 = 3$ with remainder 24, the subtraction and replacement “S – L → S” can be executed 3 times before “L > S” is satisfied. Hence, the comparison (L:S) occurs 3 times here.
- (3) Under the condition of L=60 and S=24, repeat subtraction until L<S. Since $60 \div 24 = 2$ with remainder 12, the subtraction and replacement “L – S → L” can be executed 2 times before “L < S” is satisfied. Hence, the comparison (L:S) occurs 2 times here.
- (4) Under the condition of L=12 and S=24, repeat subtraction until L=S. Since $24 \div 12 = 2$ with remainder 0, the subtraction and replacement “S – L → S” can be executed 2 times before “L = S” is satisfied. Hence, the comparison (L:S) occurs 2 times here.
- (5) The number of times of comparison for “L:S” is calculated as follows:
We can now add the numbers from (1) through (4).
Total number of times of comparison = $4 + 3 + 2 + 2 = 11$ (times)

Question 10

Difficulty: ***

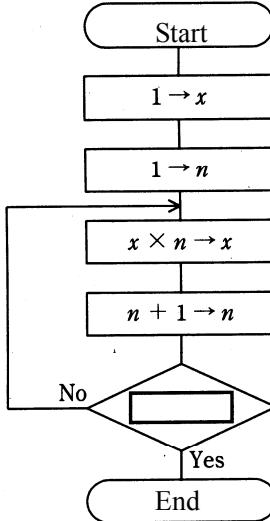
Frequency: ***

- Q10.** When the algorithms described by the two flowcharts below are performed on a positive integer M , which of the following conditions needs to be inserted in the box below so that the same value x can be obtained?



Note: The repetition specification in the loop limit denotes the following.

Variable name: initial value, increment, final value



- a) $n > M$ b) $n > M + 1$ c) $n > M - 1$ d) $n < M$

Answer 10

Correct Answer: a

The notation “ $n: M, -1, 1$ ” at the loop limit means, as explained in the question, the following: let the initial value of n be M , add “ -1 ” (subtract 1) each time the loop is executed, and stop when the final value “ 1 ” is reached. This means that the value of n changes from $M, M - 1, M - 2, \dots, 2$, and 1 .

Let us follow the flowchart on the left first. Starting with $n = M$ and decreasing the value by 1 each time the loop is executed until the value gets to 1 ($n = M, M - 1, M - 2, \dots, 2, 1$), the following operation is going on since “ $x \times n \rightarrow x$ ” is executed within the loop. As “ $1 \rightarrow x$ ” suggests, the initial value for x is 1. Let us track how the value of x changes as n changes:

$$n=M: \quad x \times n = 1 \times M = M \rightarrow x = M \quad (\text{The value of } x \text{ changes to } M.)$$

$$n=M-1: \quad x \times n = M \times (M-1) = M(M-1) \rightarrow x = M(M-1)$$

(The value of x changes to $M(M-1)$.)

$$n=M-2: \quad x \times n = M(M-1) \times (M-2) = M(M-1)(M-2) \rightarrow x = M(M-1)(M-2)$$

(The value of x changes to $M(M-1)(M-2)$.)

and so on...

$$n=2: \quad x \times n = M(M-1)(M-2)\dots2 \rightarrow x = M(M-1)(M-2)\dots2$$

$$n=1: \quad x \times n = M(M-1)(M-2)\dots2 \cdot 1 \rightarrow x = M(M-1)(M-2)\dots2 \cdot 1 = M! \quad (M \text{ factorial})$$

This is calculating $M \cdot (M - 1) \cdot (M - 2) \cdot \dots \cdot 2 \cdot 1$. For an integer value M , the product $M \cdot (M - 1) \cdot (M - 2) \cdot \dots \cdot 2 \cdot 1$ is called $M!$ (M factorial).

On the other hand, consider the flowchart on the right. n changes from 1, 2, ..., M , and the process " $x \times n \rightarrow x$ " is repeated in the loop. This is also factorial calculation, beginning with 1.

Let us now specifically track how x changes with respect to the value of n . As in the flowchart on the left, the initial value of x is 1. Further, each time the loop is executed, the value of n increases by 1. The underlined part in each line below is the previous value of x :

$$\begin{aligned}n=1: \quad & x \times n = \underline{1} \times 1 = 1 \rightarrow x \quad (x = 1) \\n=2: \quad & x \times n = \underline{1} \times 2 \rightarrow x \quad (x = 1 \times 2) \\n=3: \quad & x \times n = \underline{1 \times 2} \times 3 \rightarrow x \quad (x = 1 \times 2 \times 3) \\n=4: \quad & x \times n = \underline{1 \times 2 \times 3} \times 4 \rightarrow x \quad (x = 1 \times 2 \times 3 \times 4)\end{aligned}$$

Let us consider how large n should be in order to make the result identical to the result of the flowchart on the left. The flowchart on the left repeats " $x \times n \rightarrow x$ " to execute the following calculation:

$$\text{Flowchart on the left} = M \cdot (M - 1) \cdot (M - 2) \cdot \dots \cdot 2 \cdot 1$$

The multiplication begins with M here; on the right, the multiplication begins with 1. Hence, as shown below, if the multiplication continues until M , the results of the two flowcharts will be identical:

$$\text{Flowchart on the right} = 1 \times 2 \times 3 \times \dots \times M$$

Therefore, we are to repeat " $x \times n \rightarrow x$ " until $n = M$. Following the flowchart, after the command " $x \times n \rightarrow x$," the program executes " $n + 1 \rightarrow n$," so after $n=M$ is multiplied, we will have $n = (M+1)$. This means that the program should flow to the "end" branch when " $n=M+1$ " is satisfied. Among the options in the answer group, this condition is " $n>M$."

2 Computer Systems

Chapter Objectives

A computer system is composed of hardware and software. There are many types of computer, but the principles of their operation are fundamentally the same. We will learn the mechanism of computers (hardware) in Section 1 and software (operating system) for efficient computer use in Section 2. We will further learn some configurations of computer systems for achieving improved reliability in Section 3 and ways to evaluate the performance of computers in Section 4. Finally, in Section 5, we will learn various systems that use computers.

- 2.1 Hardware**
- 2.2 Operating System**
- 2.3 System Configuration Technology**
- 2.4 Performance and Reliability of Systems**
- 2.5 Systems Application**

[Terms and Concepts to Understand]

Central Processing Unit (CPU), cache memory, input/output interface, auxiliary memory, task management, job management, multiprogramming, virtual memory, dual system, duplex system, client/server system, availability, MTBF, MTTR, Internet

2.1 Hardware

Introduction

Hardware is defined as devices with which a computer is configured. Computers consist of processing units, memory, input/output units, etc. In this section, we will explain these units from the standpoint of hardware.

2.1.1 Information Elements (Memory)

Points

- Information elements include ROM and RAM.
- SRAM and DRAM are typical types of RAM.

Semiconductor memory is memory made of integrated circuits (ICs) using semiconductors. Semiconductor memory includes **ROM**, which is non-rewritable, and **RAM**, which is rewritable.

◆ ROM (Read-Only Memory)

ROM is semiconductor memory that is not erased when the power is turned off.¹ On mask ROM and PROM, data can be written only once; on EPROM, however, data can be written and re-written repeatedly using a special method. Types and characteristics of ROM are shown in the following table.

Type (Name)	Characteristics, etc.
Mask ROM	Data is written at the time of manufacturing. It cannot be re-written later.
PROM (Programmable ROM)	PROM data is written by the user when it is first used. It cannot be re-written later.
EPROM (Erasable PROM)	EPROM data is written by the user electrically. All the data can be erased using ultraviolet rays.
EEPROM (Electrically EPROM)	All the data can be erased and re-written. Data is erased electrically.
Flash memory ²	Erasure and re-writing can be done collectively or on a block basis. Data is erased electrically.

¹ **Volatility and non-volatility:** It is the property that the contents of memory are lost when the power is turned off is called volatility. RAM is a type of volatile memory. On the other hand, the property that the contents of memory are not lost when the power is turned off is non-volatility. ROM is a type of non-volatile memory.

² (Hints & Tips) Flash memory is classified as EEPROM.

◆ RAM (Random Access Memory)

RAM is semiconductor memory that loses its memory contents when the power is turned off.³ Unlike ROM, its contents can be changed, so it is used for main memory, graphics memory,³ and cache memory.

There are two typical types of RAM: **SRAM** and **DRAM**. The characteristics of SRAM and DRAM can be summarized as shown in the following table.

Comparison item	SRAM	DRAM
Level of integration	Low (small capacity)	High (large capacity)
Access speed	Fast	Slow
Price	Expensive	Inexpensive
Usage	Cache memory ⁴ Battery-operated devices	Main memory
Operation	No refresh is required.	Refresh is required.
Structure	Flip-flop Complicated structure	Condensers and transistors Simple structure

SRAM (Static RAM)

SRAM is composed of a flip-flop,⁵ so it does not require any refresh operations and is able to speed up information reading and writing. However, the cost is higher for the same capacity than DRAM, because the SRAM structure is more complicated than that of DRAM. For this reason, it is used mainly in areas where the speed, not the cost, is important, such as in cache memory. It is also used in battery-operated devices.

DRAM (Dynamic RAM)

DRAM consists of condensers and transistors, representing whether or not there is electrical charge in the condensers, by using 1 or 0. As time elapses, the electrical charge in the condensers gets discharged, resulting in memory loss; therefore, it needs to be re-written (refreshed) at certain time intervals (every few milliseconds). Since the structure is rather simple, the manufacturing cost is low, and it is mainly used in the main memory of PCs.⁶

Types of DRAM equipped with high-speed data transfer functions include SDRAM, DDR SDRAM, etc.

³ **Graphics memory:** It is memory used when images and characters are displayed on the display screen using a computer. It is also referred to as video memory (VRAM).

⁴ **Cache memory:** It is high-speed memory placed between main memory and the CPU to speed up data reading from main memory to the CPU.

⁵ **Flip-flop** (also known as bistable circuit): It is an electrical circuit with two stable states, which maintains its state until an input that changes one of the states is entered.

⁶ (FAQ) Frequently there are questions that compare SRAM and DRAM. You should have good knowledge of the level of integration, usage, structure, etc.

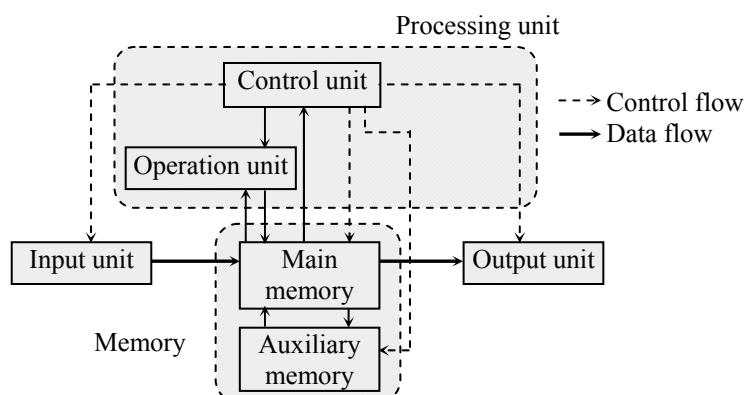
2.1.2 Processor Architecture

Points	<ul style="list-style-type: none">➤ A computer consists of five major units (functions).➤ There are several addressing methods; direct addressing, indirect addressing, etc.
---------------	---

The term architecture refers to “structures or organizations.” The processor architecture refers to the configuration and operating principles of the computer.

◆ Configuration of Computer

Below is a figure showing the basic configuration of a computer. This configuration is called the “big five units” or “big five functions,” because there are five major components.



The control unit and the operation unit are together called the processing unit or the central processing unit (CPU).

◆ Address Modification and Addressing Methods

A **program** is stored in the main memory and is retrieved, one instruction at a time, by the control unit to be executed. Address modification occurs in order to locate the location of data that is subject to processing.

Address modification is a function that obtains the value of the address actually accessed based on the address specified by the instruction. The method for address modification is called an **addressing method**. The address actually accessed as the result of the address modification is called the **effective address**. Addressing methods is described below in detail.

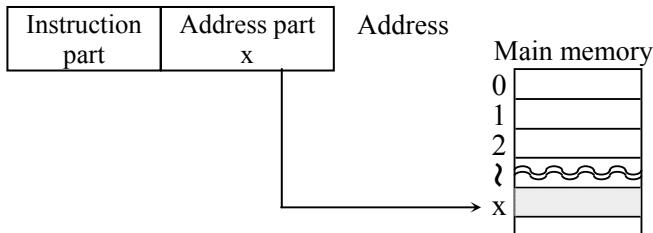
⁷ “The big five units”:

- **Control unit:** It is the unit that controls the entire computer. It extracts and reads instructions of the program stored in the main memory and sends to various units the directions necessary to execute the instruction.
 - **Operation unit:** It is the unit that performs the arithmetic operations, logic operations, and other operations. It consists of adders, registers, complementers (units that convert values to their complements), etc.
 - **Memory:** It is a generic term of the unit that stores data, programs, etc. It can be classified into main memory and auxiliary memory.
 - **Input unit:** It is a generic term of the unit that enters programs and data into the computer.
 - **Output unit:** It is a generic term of the unit that outputs results of computer processing in characters and numbers that we can recognize.

Direct addressing method

In this method, the content stored in the address part of the instruction becomes the data subject to operation.

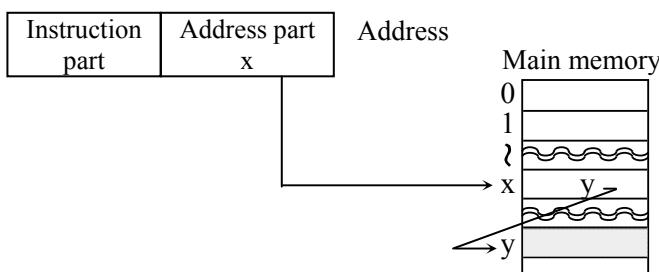
(Direct addressing method)



Indirect addressing method

In this method, the data stored in the address designated by the address part of the instruction are not the data subject to operation; rather, the data stored at the address designated by that content are the data subject to operation.

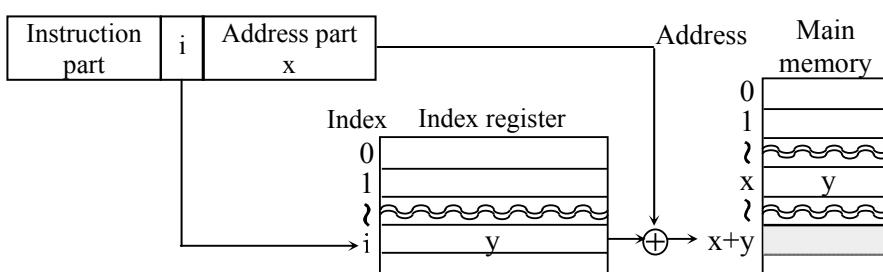
(Indirect addressing method)



Indexed addressing method (index modification)

In this method, the effective address is the sum of the value of the address of the instruction and the value of the index register.⁸ For example, when processing an array, we can look up the content of another address simply by changing the content of an index register.⁹

(Indexed addressing method)



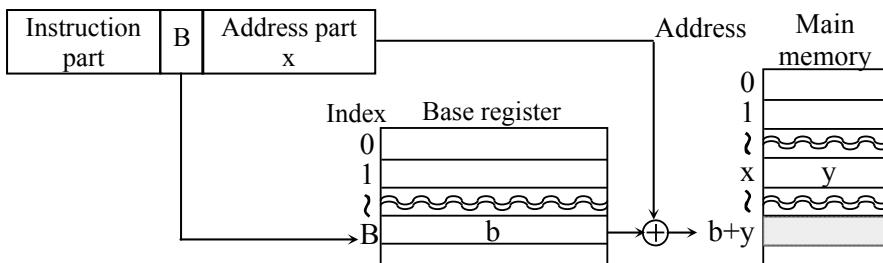
⁸ **Register:** It is low-capacity, high-speed memory where data is temporarily stored. It is located in the CPU. There are various registers, including the following: general registers, for storing intermediate and final results of operations; status registers for indicating the CPU state after an instruction is executed; index registers for address calculations; and base registers.

⁹ (FAQ) There are questions on the concept of addressing methods. An example is “Which of the following is an appropriate description of the direct addressing method?” Be sure to have these methods organized in your mind: the direct addressing method, index addressing method, immediate value addressing method, etc.

Base addressing method

In this method, the effective address is the sum of the address designated by the address part of the instruction and the content of the base address register.¹⁰

(Base addressing method)



Immediate value addressing method

This is the method where the address of the instruction stores the data subject to processing, not an address.¹¹

◆ RISC and CISC

Various types of technology have been used to speed up computer processing time; RISC and CISC are examples of this technology. With the advancement of semiconductor technology, the integration density of integrated circuits has risen continually. At present, computers are made with integrated circuits, and RISC and CISC are the two approaches for developing computers. Computers configured for high speed with simple instructions and simplified hardware are called RISCs. In contrast, those where complicated instructions are configured on one circuit are called CISCs.

RISC (Reduced Instruction Set Computer)

These computers have only a set of simple, frequently used instructions integrated onto a single VLSI (very large scale integration) chip in order to achieve high performance through improved machine cycles (operation speed) and a reduction in instruction processing time. The emphasis is placed on keeping the length of each instruction to a fixed length and limiting the time required to execute each instruction to a fixed amount. By doing so, the technology of pipeline control has been easily implemented. However, the number of instructions to be executed becomes large unless efficient object programs are created, so it is essential that the compiler have an optimization function.¹² Most computers called workstations are of this type.

CISC (Complex Instruction Set Computer)

These computers have complex instructions integrated onto a single VLSI chip in order to achieve high overall performance. Most general-purpose computers are CISCs.

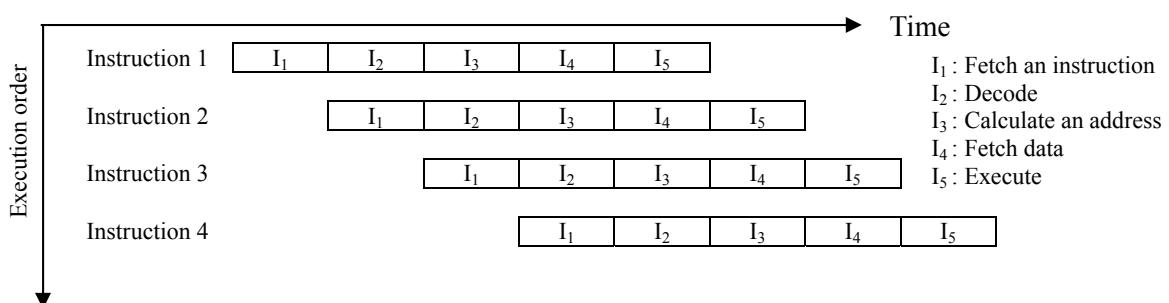
¹⁰ (Note) The base addressing method can be used regardless of where in the main memory the program is stored, simply by changing the value of the base register. Such a structure is called a re-locatable structure.

¹¹ (Hints & Tips) Note that in the immediate value addressing method, the address of the main memory is not designated.

¹² **Optimization:** It is a function of a compiler to eliminate redundancy of a program in order to reduce the execution time of the object program and the size of the program. This is done in a variety of ways, such as calculating constants in advance, simplifying formulas, and eliminating double loops.

◆ Pipeline Control

We have mentioned RISC and CISC as technologies to improve computer processing speed. To further improve the speed, the RISC system uses pipeline control. Pipeline control is a technology to reduce the instruction execution time of the CPU. This is an attempt to do the following: when execution steps of an instruction are divided into 5 or 6 steps, and if each step is completed within a certain fixed amount of time and the instruction steps stay independent of one another, then we can improve the overall processing speed by delaying the execution of each instruction 1 step behind the previous instruction. In reality, however, due to branching instructions, there are times when the next instruction address is not completely determined, and some steps are not completed within the fixed processing time. Everything is not always functioning in an ideal way, but pipeline control does process instructions concurrently, providing one way of speeding up the computer.¹³



2.1.3 Memory Architecture

Points

- A hierarchical memory structure is introduced in order to achieve high-speed, large-capacity memory.
- Cache and interleaving technologies are used for speeding up memory.

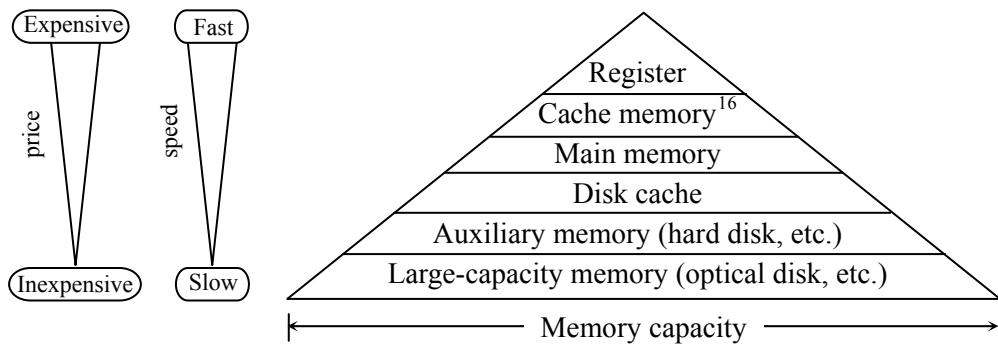
There are many requirements for memory, but requirements for high speed and large capacity¹⁴ are of particular importance. In general, however, high-speed memory is expensive and has small capacity whereas low-speed memory is inexpensive and has large capacity. So, efforts are being made to combine high-speed but small-capacity memory and low-speed but large-capacity memory to develop high-speed and large-capacity memory.

¹³ (FAQ) Questions regarding pipeline control often appear on exams. Most of them are in the form of choosing an appropriate description of pipeline control, so you only have to know that pipeline control executes instructions concurrently.

¹⁴ (Note) Besides these, requirements for memory include reliability, ability of random access, non-volatility, re-writable function, portability, low cost, etc.

◆ Memory Hierarchy

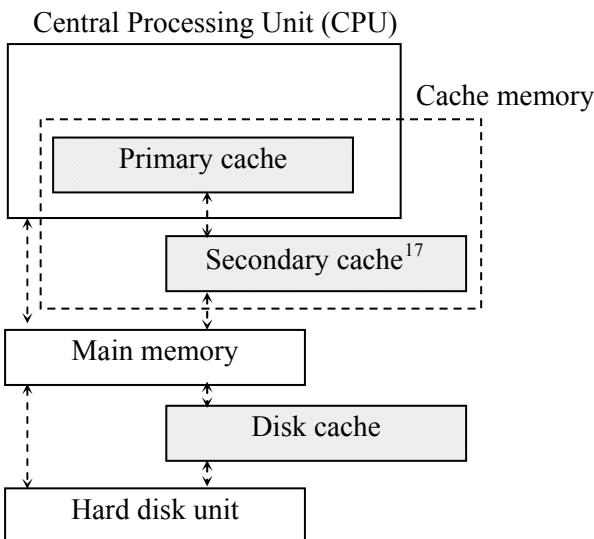
Memory hierarchy is a hierarchical representation of the relationship between the access speed and capacity of various types of memory.¹⁵



◆ Cache Memory (High-Speed Buffer Memory)

Cache memory is high-speed, small-capacity memory that is placed between the CPU or registers and the main memory. The main memory is slower than the CPU or registers, so the CPU process can be made more efficient by storing frequently accessed data and programs of the main memory into the cache memory. Recently, a secondary cache has been installed for further improvements in speed.

The cache memory placed between a hard disk and the main memory is called the **disk cache**. The figure below shows the relationship between the cache memory and the disk cache.



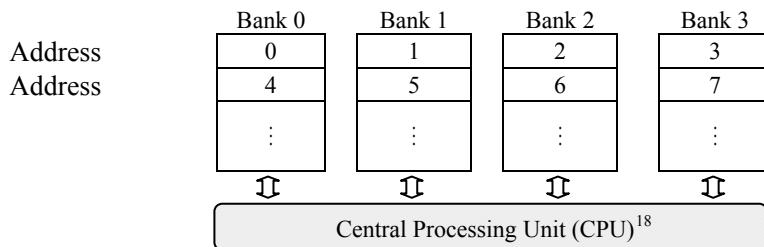
¹⁵ (Note) If t is the average memory access time, t_m is the access time of the main memory, t_c is the access time of the cache memory, and h is the hit ratio, then the following equation holds: $t = t_c h + t_m(1 - h)$.

¹⁶ **Hit ratio:** It is the probability that the portion of a program necessary to execute that program is in the cache memory

¹⁷ **Secondary cache:** The primary cache is the cache memory which is built in the CPU; the secondary cache is the cache memory placed between the primary cache and the main memory.

◆ Interleaving (Memory Interleaving)

The main memory is divided into multiple units called banks, and addresses are assigned across the banks. Often, the main memory is accessed over a sequential range of addresses at a time, so the speed can be enhanced by accessing a sequential range of addresses concurrently. For example, even when the operation of Bank 0 is not yet completed, Bank 1 can be accessed. Below is an example of 4-way interleaving (with 4 banks).



Data and programs are stored over a sequence of addresses (horizontally), but the memory is accessed in bank units (vertically). This allows concurrent access to a sequence of addresses.

2.1.4 Magnetic Tape Units

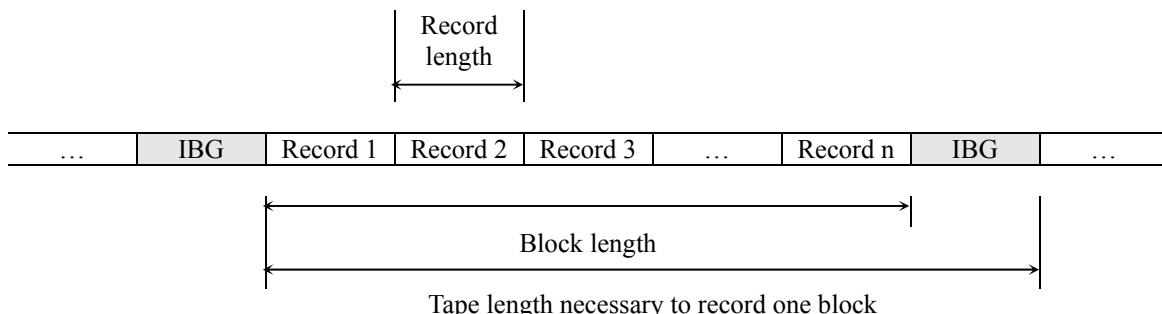
Points

- Capacity calculations require blocking factors and record density.
- Performance calculations do not include the stop time.

Magnetic tape is a medium that records data onto a tape that has been magnetically coated. The unit price of this memory medium is cheap and has a large capacity, so it is used in cases such as backing up entire hard disks.

◆ Capacity Calculation

The record format of a magnetic tape is shown below. As we can see in this figure, in order to record one block, we need to include an IBG (Inter-block gap) which is an area to identify the block and contains a special code. A magnetic tape reader reads data in block units, so this area is crucial even in identifying the end of each block.



¹⁸ (FAQ) Interleaving is a way to speed up memory. Questions on the concept of interleaving have appeared often, so be sure you understand this.

Assuming that the specifications of a magnetic tape are given below, let us calculate precisely the number of records that can be stored on this single magnetic tape.¹⁹

[File specifications]	
Record length	80 bytes
Blocking factor	100

[Magnetic tape specifications]	
Record density	64 bytes/mm
Inter-block gap (IBG)	15 mm
Tape length	730 m

Calculation of block length

The blocking factor is 100. → 100 records can be stored in one block.

Record length → 80 bytes

So the number of bytes L_1 for each block, excluding the IBG, is as follows:

$$L_1 = 80 \text{ (bytes/record)} * 100 \text{ (records/block)} = 8,000 \text{ (bytes/block)}.$$

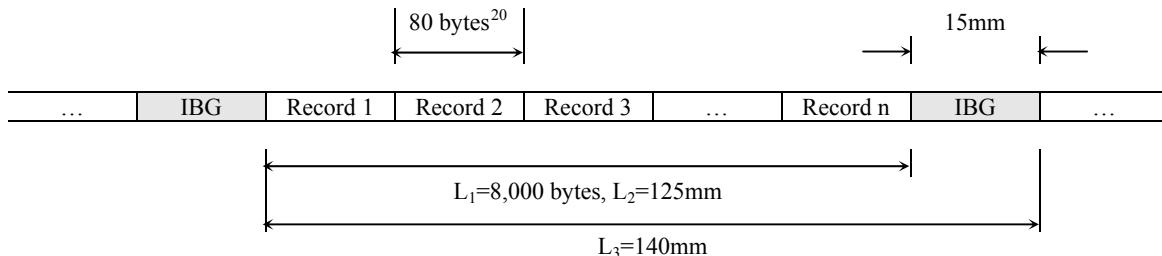
The record density is 64 bytes/mm. → 64 bytes can be stored on 1 mm of tape.

Hence, the length L_2 of a block, excluding the IBG, is as follows:

$$L_2 = 8,000 \text{ (bytes/block)} / 64 \text{ (bytes/mm)} = 125 \text{ (mm/block)}.$$

Therefore, the block length L_3 , including the IBG, is as follows:

$$L_3 = 125 \text{ (mm/block)} + 15 \text{ (mm/block)} = 140 \text{ (mm/block)}.$$



Number of records that can be stored on one magnetic tape

Let us now calculate precisely how many records can be stored on one magnetic tape.

(1) Calculating the number of blocks that can be stored on one tape

Since the length of the tape is 730 m ($730 * 10^3 \text{ mm}$), the number B_1 of blocks that can be recorded on one tape is as follows:

$$\begin{aligned} B_1 &= (\text{Length of the tape}) / (\text{Length of a block}) \\ &= 730 * 10^3 / 140 \end{aligned}$$

¹⁹ (FAQ) On each exam, there is at least one question dealing with the calculation of the capacity or performance of a magnetic tape or a hard disk. If you keep these ideas organized in your mind, you can answer these questions because the difference is only in the numerical values.

²⁰ (Hints & Tips) Besides “bytes/mm,” the record density can be represented in “columns per mm” or “bpi.” A column is the same as a byte. “bpi” stands for “bytes per inch,” which is the number of columns per inch. Converting from inches to mm is necessary here, but you need not remember the formula since the relationship between inches and mm will be given in the question.

$= 5,214.285\dots \doteq 5,214$ (blocks).²¹

The fractional part 0.285 is less than one block, so we truncate it.

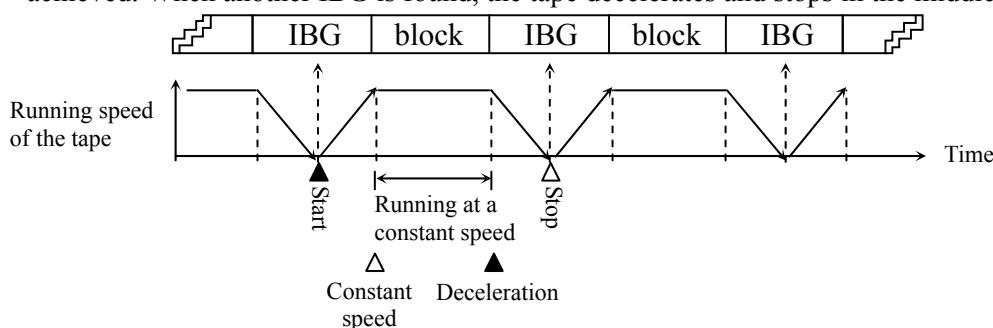
(2) Calculating the number of records that can be stored on one tape

Since one tape can record 5,214 blocks, and each block has 100 records, the number B_2 of records that can be stored on one tape is as follows:

$$\begin{aligned} B_2 &= (\text{number of blocks that can be stored on one tape}) * (\text{blocking factor}) \\ &= 5,214 * 100 \\ &= 521,400 \text{ (records)} \end{aligned}$$

◆ Performance Calculation

The running speed of the tape is constant when reading or writing data. In theory, the tape begins to accelerate in the middle of IBG and starts to read and write when a constant speed is achieved. When another IBG is found, the tape decelerates and stops in the middle of IBG.



Assuming that the specifications of a magnetic tape are given below, let us calculate the time it takes the magnetic tape to read one block.²²

[Specifications of a magnetic tape]

Data transfer speed	320 Kbytes/sec
Record length	80 bytes
Blocking factor	100
Start or stop time	6 milliseconds

Transfer time for one block

Transfer time for one block can be obtained by dividing the length of a block by the data transfer speed:

$$\text{Data transfer time for one block} = (\text{block length}) / (\text{data transfer speed})$$

$$\begin{aligned} \text{Block length} &= (\text{record length}) * (\text{blocking factor}) \\ &= 80 * 100 \\ &= 8,000 \text{ (bytes)} \end{aligned}$$

$$\text{Data transfer time for one block} = 8,000 \text{ (bytes)} / 320 \text{ (Kbytes/sec)}$$

²¹ (Hints & Tips) The fractional part of the number of blocks is discarded here, but it actually becomes a short block, which is a block with fewer records than the other blocks.

²² (Hints & Tips) In performance calculations, data is transferred in blocks, so we do not need to consider the length of IBG.

$$\begin{aligned}
 &= 8,000 / (320 * 10^3) \\
 &= 25 * 10^{-3} \text{ (seconds)} = 25 \text{ (milliseconds)}
 \end{aligned}$$

Time required to read one block

We add the start-up time to the data transfer time for one block.

$$\begin{aligned}
 &\text{Time required to read one block} \\
 &= (\text{start-up time}) + (\text{data transfer time for one block}) \\
 &= 6 \text{ (milliseconds)} + 25 \text{ (milliseconds)} = 31 \text{ (milliseconds)}
 \end{aligned}$$

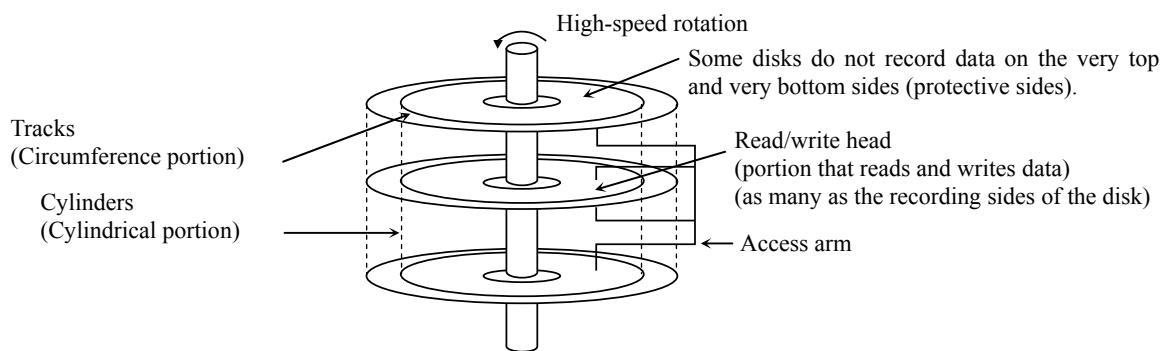
Note that the start-up time is added but not the stop time. When being read, no data is transferred until the beginning of a block is reached. Hence, the time until this is achieved is part of the waiting time. After that, data is transferred, but when the data transfer is completed, the stop operation and the program processing are performed concurrently. Thus there is no need to add the stop time.²³

2.1.5 Hard Disks

Points

- A hard disk is configured with cylinders and tracks.
- Hard disks of the sector type do not have IBGs.

A **hard disk** is a medium that achieves random and high-speed reading and writing of data, consisting of 1 to 10 round disks coated with a magnetic substance on the front and the back sides and rotated at a high speed. If there is only one disk, it is called a **floppy disk**.²⁴

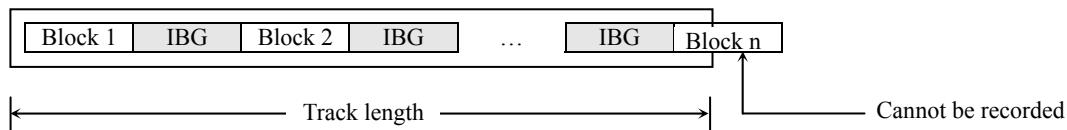


²³ (Hints & Tips) That stop time is not included is a standard assumption in exam questions.

²⁴ (Note) Another medium that, like a floppy disk, can read and write and can easily be carried around is MO (magneto optical disk), which is very popular because its capacity is about 600 times that of a floppy disk.

◆ Capacity Calculation

Just as on a magnetic tape, data is recorded in blocks on a hard disk. However, if a block cannot fit into a track, it cannot be recorded.



First, data is recorded on a track, and when that track is filled, the recording proceeds to the next track which is a track on the corresponding circumference of the next surface. In other words, data is recorded in cylinder units.

Assuming that the specifications of a hard disk are given below, let us calculate actually how many cylinders are necessary to write 100,000 records.

[File specifications]	
Record length	250 bytes
Blocking factor	8

Specifications of a hard disk	
Number of cylinders per disk	400
Number of tracks per cylinder	19
Number of bytes per track	13,000
Block gap	135 bytes

Block length (B)

$$\begin{aligned} B &= (\text{record length}) * (\text{blocking factor}) + (\text{block gap}) \\ &= 250 * 8 + 135 = 2,135 \text{ (bytes)} \end{aligned}$$

Number of blocks that can be recorded on one track (N)

$$\begin{aligned} N &= \text{track length} / 2,135 = 6.008\dots \doteq 6 \text{ (blocks) (truncated)} \\ \text{We truncate the answer because a fractional part does not constitute a block.} \end{aligned}$$

Number of records that can be recorded on one track (Rt)

$$\begin{aligned} R_t &= N * (\text{blocking factor}) \\ &= 6 * 8 = 48 \text{ (records)} \end{aligned}$$

Number of records that can be recorded on one cylinder (Rs)

$$\begin{aligned} R_s &= R_t * (\text{number of tracks per cylinder}) \\ &= 48 * 19 = 912 \text{ (records)} \end{aligned}$$

Number of cylinders required to write 100,000 records (S)

$$\begin{aligned} S &= (\text{number of records}) / R_s \\ &= 100,000 / 912 = 109.649\dots \doteq 110 \text{ (cylinders) (rounded up).} \end{aligned}$$

In general, files are secured in cylinder units, so if there are unused tracks on a cylinder, the

number of cylinders is rounded up to the next integer.²⁵

◆ Performance Calculation

The access time of a hard disk is calculated as follows:

$$\begin{aligned}\text{Access time} &= \text{waiting time} + \text{data transfer time} \\ &= (\text{seek time} + \text{latency time}) + \text{data transfer time}\end{aligned}$$

Translator's note: The waiting time (seek time + latency time) is often called the access time.

The **seek time** is the time during which the read/write head moves to the track where the data is recorded. The **latency time** is the time until the desired data come under the read/write head. The seek time and the latency time are determined by where the head is located, so we use the average values. In actual exam questions, the seek time is always given, and the latency delay can be calculated by the duration of one rotation, which is obtained by the inverse of the number of rotations per time unit. Since the minimum latency time is 0 and the maximum is 1 rotation time, the average latency time is the duration of a half rotation.

Assuming that the specifications of a hard disk are given below, let us calculate actually the access time for reading data contained in one block (5,000 bytes).

[Specifications of a hard disk]²⁶

Number of rotations of the hard disk:	2,500 rotations per minute
Memory capacity per track:	20,000 bytes
Average seek time:	25 milliseconds

Calculation of the average latency time

The fact that the number of rotations of this hard disk is 2,500 rotations per minute means that the disk makes 2,500 rotations every minute. Hence, the rotation time is as follows:

$$\text{Rotation time} = (1 * 60,000 \text{ msec/min}) / 2,500 \text{ revolutions/min} = 24 \text{ msec/revolution}$$

Note carefully these units. The rotation speed is given in rotations per minute, but the rotation time is in milliseconds. Hence, we need to convert minutes to milliseconds (1 minute = 60 seconds = 60,000 milliseconds).

Since the average latency time is $\frac{1}{2}$ of the rotation time, it is 12 milliseconds.

Calculation of data transfer speed

Since one rotation allows the transfer of data contained on one track, 20,000 bytes are transferred in 24 milliseconds. Hence, $20,000 / 24$ (bytes/msec) is the data transfer speed. We can calculate this quotient. But, since it is indivisible, we shall leave it as it is here.

²⁵ (FAQ) On each exam, there is at least one question dealing with the calculation of the capacity or performance of a magnetic tape or a hard disk. If you keep these ideas organized in your mind, you can answer these questions because the difference is only in the numerical values.

²⁶ **Seek time/Latency time:** The seek time is sometimes called the positioning time. The latency time is sometimes called the search time.

Calculation of data transfer time

The time required to transfer 5,000 bytes is then calculated as follows:

$$\begin{aligned}
 \text{Data transfer time} &= (\text{amount of data transfer}) / (\text{data transfer speed}) \\
 &= 5,000 / (20,000 / 24) \\
 &= 5,000 / 20,000 * 24 \\
 &= 6 \text{ (msec)}
 \end{aligned}$$

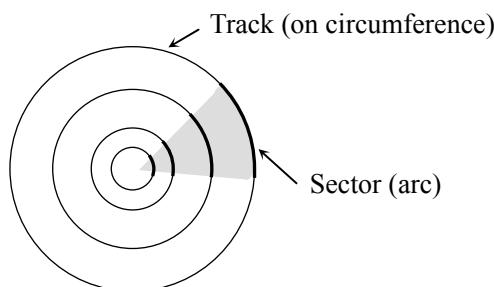
Calculation of access time

$$\text{Access time} = 25 \text{ msec} + 12 \text{ msec} + 6 \text{ msec} = 43 \text{ msec}$$

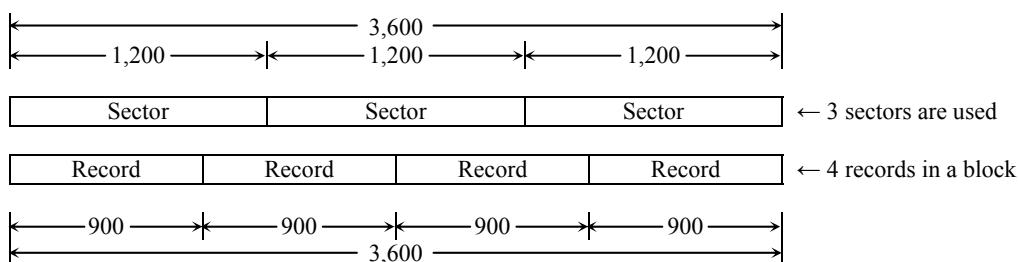
The calculations above are based on the rotation speed of the hard disk. However, auxiliary memory, such as a hard disk and a magnetic tape, exchanges data with the computer through input/output channels.²⁷ It is therefore necessary to install input/output channels with appropriate transfer speeds.²⁸

◆ Capacity Calculation of Sector-based Hard Disk

The term sector refers to the way a magnetic medium is partitioned on floppy disks or hard disks. A sector is an arc of a fan-shaped portion of the disk formed by radial lines drawn from the center of a track in equal intervals. Input/output of a sector-based recording medium is done in sector units, without using IBGs. Each sector is filled with as many records as possible, and the remaining portion of the sector is not used.



For example, suppose that each track consists of 12 sectors, each of which consists of 1,200 bytes on a hard disk. To store files whose record length is 900 bytes, there is no sector remainder, as shown below, if the flocking factor is 4.



²⁷ **Input/output channel:** Data-transfer path for exchanging data between auxiliary memory and the computer

²⁸ (Hints & Tips) The data transfer speed of a hard disk is determined by the rotational speed of the disk, so it is meaningless to have a high-speed channel. Channels are to be selected according to the transfer speed of the disk.

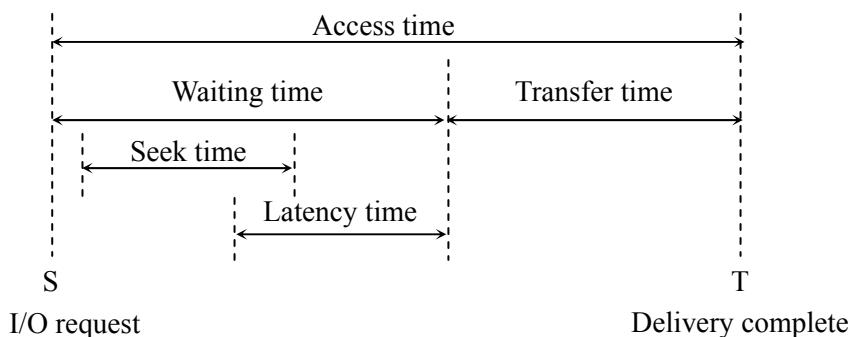
2.1.6 Terms Related to Performance/ RAID

Points	<ul style="list-style-type: none"> ➤ Regarding storage media, there are some terms such as access time, waiting time, transfer time, seek time, and latency time. ➤ RAID is a disk array system for achieving enhanced reliability and/or increased processing speed.
---------------	---

There are some terms related to the performance of storage media, including access time, waiting time, transfer time, seek time, and latency time. **RAID**, sometimes called a **disk array**, is a way to control multiple disks placed in parallel as if they were one unit.

◆ Terms Related to Performance

The figure below shows how the various processing times are related to one another, beginning at time S, when a processing unit requests input/output to a hard disk unit, to time T, when the data delivery is completed.



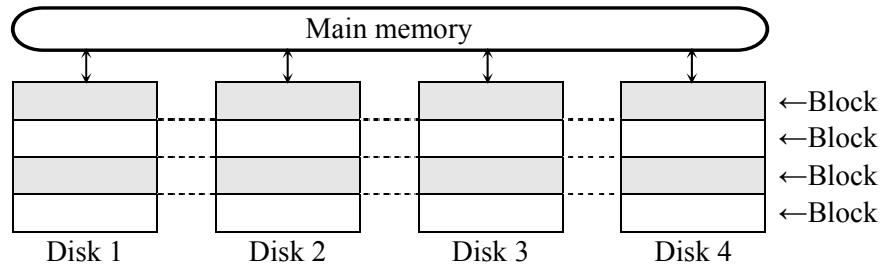
Since the head moves to the desired track while the disk is rotating, the rotation time and moving time overlap.²⁹ However, on IT exams, this overlap is almost always ignored. Hence, there is no problem defining waiting time as follows:

$$\text{Waiting time} = \text{Seek time} + \text{Latency time} \text{ (or Search time)}$$

◆ RAID (Redundant Array of Independent Disks)

RAID describes auxiliary storage in which multiple hard disks are placed in parallel and are controlled as if they were one disk unit so that the input/output speed can be improved and/or reliability can be enhanced. Sometimes the term RAID refers to such an auxiliary storage or a method. It is an attempt to speed up the process by spreading the blocks over multiple disks and reading the blocks simultaneously.

²⁹ (Hints & Tips) A hard disk is rotating as the head approaches the track, so the seek time and the latency time overlap partially.



RAID has 6 levels as described below, from RAID0 to RAID5.

RAID0

This is the method of writing blocks of a fixed size on multiple disks. Access is not centralized on one single unit, so the input/output time can be reduced.³⁰

RAID1

By recording the same data on two disks, this configuration enhances the safety of the data.³¹

RAID2, RAID3, and RAID4

These are configurations where, in addition to data recorded on the hard disk, there is a disk designated as the error-checking disk to prevent failures. RAID 2 can correct errors. RAID3 and RAID4 can detect errors while they cannot correct errors. In RAID3, data is partitioned in bits or bytes whereas RAID4 partitions data in block units.³²

RAID5

This is where each data block is assigned a parity value. Data and parity are written on separate disks, and a failure on a single disk can be recoverable.

Below is an example of RAID5. Here, 4 disks are handled as one group.

Disk 1	Disk 2	Disk 3	Disk 4
Block 1	Block 2	Block 3	Parity ³³ 1~3
Block 4	Block 5	Parity 4~6	Block 6
Block 7	Parity 7~9	Block 8	Block 9
Parity 10~12	Block 10	Block 11	Block 12

Here, the data is divided up into blocks of a certain length, and three blocks are considered to form a unit. For example, Blocks 1 through 3 are a unit, and for each bit, the exclusive logical sum of blocks 1 through 3 is written on a separate disk as parity value 1 through 3. Similarly, the exclusive logical sum of blocks 4 through 6 is written on a separate disk as parity value 4 through 6.

There is also RAID6, in which the parity values are separated as in RAID5 and the data is recoverable even when two disks fail.³⁴

³⁰ (Hints & Tips) The only feature about RAID0 is that I/O is dispersed, so this is not a measure to improve reliability.

³¹ (Note) RAID1 is called mirroring since the same data is recorded on separate hard disks.

³² (Note) There is also RAID0+1, a combination of RAID0 and RAID1, already in use.

³³ (Note) The parity of RAID5 uses the exclusive logical sum of multiple blocks. Hence, even if one of the disks should fail, the damaged data can be recovered by taking the exclusive logical sum of the other blocks.

³⁴ (FAQ) Often on the exam, there are questions of the form: “Which of the following statements is appropriate concerning RAID...?” Remember the difference between RAID0 and RAID1.

2.1.7 Auxiliary Storage/Input and Output Units

Points	<ul style="list-style-type: none"> ➤ Auxiliary storage includes hard disks, magnetic tapes, magneto optical disks, CDs, DVDs, etc. ➤ Input/output units include keyboards, image scanners, tablets, displays, printers, etc.
---------------	--

Any storage excluding the main memory is called **auxiliary storage**. Auxiliary storage can be used to compensate for the insufficient capacity of the main memory. In general, auxiliary storage has larger capacities in comparison with the main memory.

Input and output units include both input units, where data is entered into the computer, and output units, where data is taken out of the computer. A unit equipped with both the input and output functions is called an input/output unit.

◆ Auxiliary Storage

Typical auxiliary storage includes the following media. In the past there was a time when magnetic tapes and floppy disks were the mainstream media; however, recently the main types have been hard disks, magnetic optical disks, CDs, and DVDs.³⁵

Medium	Capacity	Re-writing	Properties, etc.
Hard disk	small to 300GB	Yes	Mostly built-in
DVD	DVD-ROM	4.7 to 9.4BG	Replacement for CD-ROM
	DVD-R DVD+R	3.95 to 7.9GB	Playable on DVD-ROM units
	DVD-RAM DVD-RW DVD+RW	3.95 to 7.9GB	Multiple specifications
CD	CD-ROM	700MB	For software distribution, etc.
	CD-R	700MB	For backup, etc.
	CD-RW	700MB	Requires a dedicated drive for re-writing
Magneto optical disk ³⁶ (MO)	128, 230, 640MB, 1.3GB	Yes	Written with magnetism and light and read with light
Floppy disk	1.4MB	Yes	Good portability
Hard disk	a few GB	Yes	For backup
DAT ³⁷	Max. about 24GB	Yes	For backup

³⁵ (Note) DVDs are optical disks just as CDs are, but with reduced laser-light wavelength, the DVDs have larger capacities. The record density on DVD is also larger.

³⁶ (Hints & Tips) A magneto optical disk uses light and magnetism for writing data but uses only light for reading data.

³⁷ (Note) DAT is a unit that records audio onto a magnetic tape using digital signals. It was originally designed for music, but it is now used as a backup system because of its low cost.

◆ Input and Output Units

Input units include **keyboards, image scanners, tablets, pointing devices**, etc. Output units include **displays, printers, etc.**

Input units

The most common input devices are keyboards and pointing devices. A keyboard is used to enter numerals and characters while pointing devices³⁸ are used to enter coordinate values. Other input units are shown in the following table.

Unit	Functions, etc.
OCR	This reads handwritten characters and printed characters optically.
OMR	This reads handwritten marks optically.
Mouse	This is used to enter coordinate positions of the mouse pointer.
Tablet	Dedicated devices such as a light pen are used to enter coordinate positions.
Barcode reader	This reads barcodes (not the numbers printed).
Image scanner	This reads image data such as pictures and photos, converting them to digital data.

Output units

The most common output devices are displays and printers.³⁹

Unit	Descriptions, etc.
Display	CRT
	This uses a cathode-ray tube; it is inexpensive and comes with a large screen.
Printer	Liquid crystal ⁴⁰
	This uses liquid crystal; it is expensive but thin and saves footprint.
	Laser printer
	The principle is the same as for a copy machine; the print quality is good, but it is expensive.
	Inkjet printer
	Printing is by injection of ink; it is small and inexpensive.
	Thermal printer
	The printing quality is good but requires specific paper.
	Thermal-transfer printer
	Heat melts the ink, resulting in high-quality, but the operating cost is high.
	Dot impact printer
	This is noisy but inexpensive; this has duplicating capability.
	Plotter
	Printer for design drawings

³⁸ (Note) **Pointing device:** A pointing device is any unit designed to enter coordinate positions such as a mouse, a tablet, etc. Other examples include trackballs, digitizers, touch screens, etc.

³⁹ (Note) **OLED:** It is a display using the organic light emitting diode technology. It uses organic materials that emit light when voltage is applied. Compared with LCDs, the viewing angle is larger, the contrast is better, and the response speed is higher, in addition to being thinner and lighter.

⁴⁰ (Note) **Liquid crystal display (LCD):** A liquid crystal display can be of various types such as TFT and STN (currently the mainstream is DSTN). STN has a simple structure with low manufacturing costs, but its resolution and contrast are also low. DSTN is an improved version of STN, where contrast is enhanced. TFT has contrast and resolution equivalent to those of CRT but is expensive.

2.1.8 Input and Output Interfaces

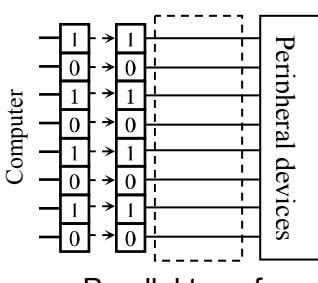
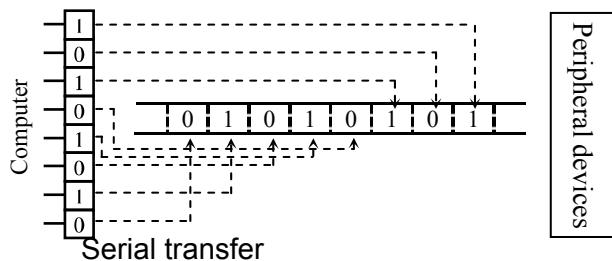
Points

- I/O interfaces include SCSI, USB, etc.
- USB is equipped with a hot plug function and a plug-and-play function.

Input and output interfaces are interfaces for connecting peripheral devices such as printers and hard disk units to the PC and for transferring data. Depending on their types, the transfer may be either serial data transfer or parallel data transfer.⁴¹

◆ Data Transfer Methods

There are two data transfer methods between the computer and its peripheral devices: serial transfer and parallel transfer. **Serial transfer** is the type of transfer in which data output from the computer are serially transferred, one bit at a time. **Parallel transfer** is the type of transfer in which data bits are transferred in parallel from the computer. For example, if the transfer is 8-bit parallel, there are 8 signal lines:



⁴¹ (FAQ) Frequently the exams have questions concerning combinations of I/O interfaces and data transfer methods. It is good to know common I/O interfaces and data transfer methods.

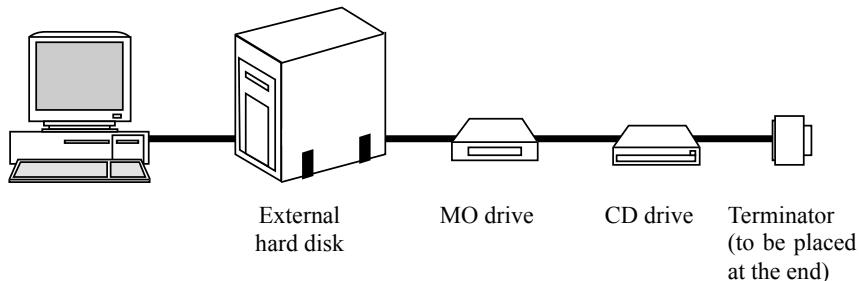
◆ Types of Input and Output Interfaces

Below is a table that summarizes I/O interfaces commonly used today. USB is the most commonly used interface now, equipped with a plug-and-play function⁴² and a hot plug function.⁴³

Type	Transfer method	Properties, etc.
RS-232C	Serial	Originally used to connect a PC to a modem; currently used also for I/O units
SCSI	Parallel	Daisy chain allows connection of up to 7 relatively high speed units
Centronics	Parallel	Connecting a PC to a printer; high speed but not for long distance
GPIB	Parallel	Connecting a PC to a measuring device; also known as IEEE-488
USB	Serial	Connecting up to 127 units in a tree configuration
IEEE1394 ⁴⁴	Serial	Connecting up to 63 units in daisy chain or tree configuration

Daisy chain

This is the connection method used in SCSI and GPIB, where peripheral units are connected along a line. The last unit in the line requires a termination resistor called a terminator.



USB

USB has two modes: the full speed mode of 12Mbps and the low speed mode of 1.5Mbps. In the full speed mode, relatively high speed units such as printers and scanners are connected. In the low speed mode, relatively slow units such as keyboards and mice are connected. Currently, USB 2.0 has increased its high speed mode up to 480Mbps, so most peripheral units can be connected.

Other interfaces

In addition to the above, there are other interfaces such as IDE (connecting a hard disk), ATA (IDE standardized by ANSI), and ATAPI (connecting ATA with units other than a hard disk, such as CD-ROM drive and tape streamer). However, currently CD-R and CD-RW units are commonly connected via SCSI and USB.

⁴² **Plug-and-play (Plug and play):** This refers to the function of automatically installing and setting the device driver when the peripheral unit or extension card is connected to the computer. The OS checks all the units connected to the computer when it is started up, installing the required device drivers. If the OS does not have the device driver of the unit in its own library, it requests installation of the device driver and, if necessary, even re-starts the computer automatically.

⁴³ **Hot plug:** This is the function that enables the plug-and-play function while the computer and peripheral unit power is on.

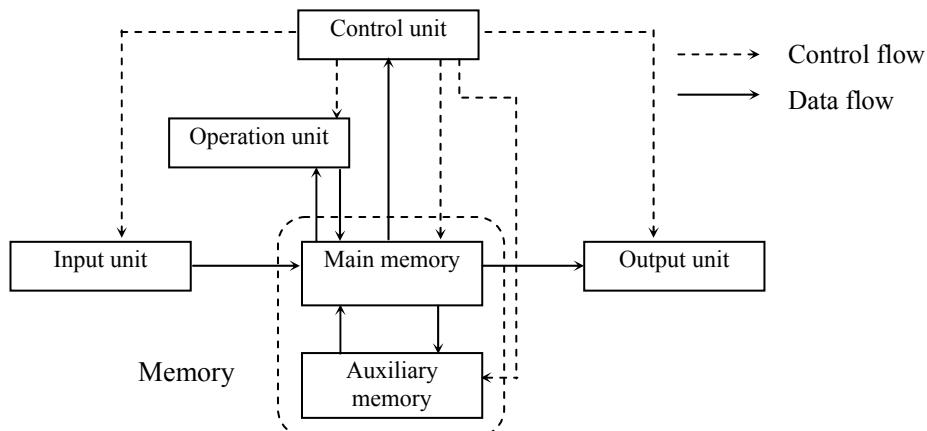
⁴⁴ **IEEE 1394:** The standard where the transfer speed is 100Mbps, 200Mbps, or 400Mbps. This is equipped with a hot swap function (peripheral units can be connected or disconnected without having to turn the power off).

Quiz

Q1 Compare DRAM and SRAM:

Item for comparison	DRAM	SRAM
Refresh		
Level of integration		
Access speed		
Unit price per bit		
Usage		

Q2 Among the components that compose a computer, which ones compose the central processing unit (CPU)?



Q3 Explain the role of cache memory.

Q4 Describe the characteristics of RAID.

Q5 Is USB a serial interface or a parallel interface?

2.2 Operating Systems

Introduction

Computers do not operate only with hardware. They function only with the use of software called an operating system (OS).

2.2.1 Configuration and Objectives of OS

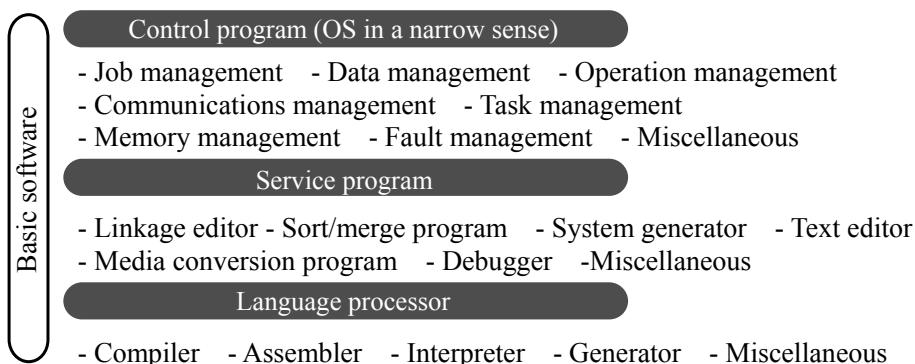
Points

- There is a broadly defined OS and more narrowly defined OS.
- The objective of an OS is the effective use of computers.

The definition of an operating system (OS) is not clear. The basic software is called an **OS in a broad sense** while the control program is called an **OS in a narrow sense**.

◆ Configuration of OS

An operating system is the basic software that comprehensively controls and manages the entire operation of hardware and software of a computer system. A program referred to as the basic software and its role are shown below.⁴⁵



Service programs and language processors are sometimes called processing programs, which run on the control program. For this reason, the control program is called an OS in a narrow sense.

⁴⁵ (Note) Operating systems for personal computers include Windows XP, Mac OS X, and OS/2. Those for workstations include Windows Server 2003, UNIX, and Mac OS X Server. For general-purpose machines, there is also MVS developed by IBM. In addition, there is a free OS program called Linux, which is compatible with UNIX.

◆ Objectives of OS

An OS attempts to improve the productivity of the entire system by eliminating unnecessary operations and waste of various resources surrounding the computer and by operating the computer system efficiently. The objectives of an OS are organized in the following figure.

Objectives of OS	Effective use of hardware resources:	Multiprogramming, ⁴⁶ spooling function, ⁴⁷ etc.
	Response to various processing modes:	Batch processing, online real-time processing, etc.
	Securing reliability and safety:	Improvement in RASIS, etc.
	Load reduction of application software:	Virtual memory, ⁴⁸ library management, ⁴⁹ etc.
	Support of computer control and operation:	Continuous processing, recording the operation data, etc.

Let us take a look at these individual objectives in detail:

Effective use of hardware resources

Hardware resources include the central processing unit (CPU), memory, I/O units (including channels). etc. The OS controls these resources so that they can be used efficiently.

Response to various processing modes

One computer can handle various processing modes such as batch processing, remote batch processing, online processing, real-time processing, and interactive mode processing. In particular, since online processing has become widespread, the scope of computer applications has been dramatically enlarged.

Securing reliability and safety

Indexes for reliability and safety include RASIS. This is a term coined by taking the initial letters of the words Reliability, Availability, Serviceability, Integrity, and Security.

Load reduction of application software

Application software refers to a program which runs under the control of the OS. The OS provides an environment in which application programs can be efficiently executed.

⁴⁶ **Multiprogramming:** It is a mechanism in which programs are processed alternately on one CPU so that it can appear as though multiple programs were operating at the same time.

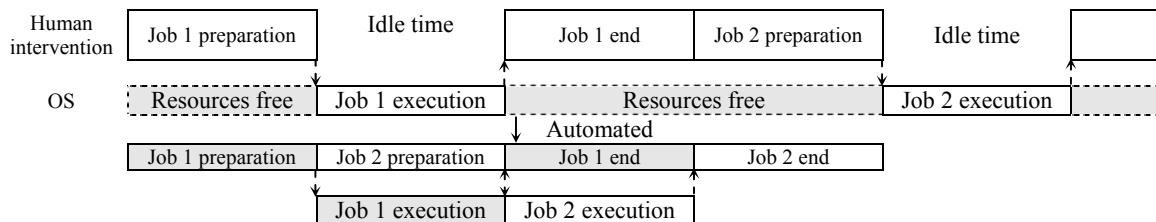
⁴⁷ **Spooling:** It is accomplished by using high-speed hard disks as a virtual I/O unit. For example, directly printing on a low-speed printer slows down the processing speed. Instead, the output results can be recorded on a high-speed hard disk first, and then a service program, dedicated only to output, can do the printing when the CPU is not busy.

⁴⁸ **Virtual memory:** It is a technique to enlarge the apparent capacity of the main memory so that large-scale programs can be loaded in the memory at a time. Often, auxiliary storage such as a hard disk is used as virtual memory.

⁴⁹ **Library management:** It is the function that systematically accumulates primitive programs, object programs, load modules, and other programs developed. This enables integrated management of software assets that are managed individually (by individuals).

Support of computer control and operation

An OS eliminates human intervention as much as possible, as it processes programs (jobs) continuously and records the operation status (log). The record of the operation status is used to check the circumstances under which a fault occurred.⁵⁰



2.2.2 Job Management

Points

- Jobs are units of tasks given to the computer, consisting of multiple programs (job steps).
- Job management has functions such as scheduling and spooling.

One of the functions of the control program, the "OS" in a narrow sense, is "job management." In **job management**, the priorities of jobs are determined, and the jobs are synchronized. In batch processing, the OS analyzes the contents of JCL (job control language) to assign resources⁵¹ and schedule jobs. In interactive mode processing, the OS analyzes instructions entered at the terminal, assigns resources, and performs scheduling. In addition, job management has other functions such as spooling and cataloged procedures.



- Scheduler:** Managing the order of job execution
- Master scheduler:** Interface with the operator
- Job scheduler:** Managing reception, selection, start, and finish of jobs
- Reader:** Reading jobs
- Initiator:** Preparing for the beginning of jobs and programs
- Terminator:** Clean-up after jobs and programs
- Spooling:** Input management for jobs, output management for process results
- Cataloged procedure:** Support for execution of typical jobs

⁵⁰ (FAQ) Concerning operating systems, many exam questions involve knowledge of terms. Be sure to know terms such as multiprogramming, virtual memory, and spooling function.

⁵¹ **Resource:** A resource is a device/unit of various kinds necessary for the computer to operate. It refers to any device related to memory, input, output, control, and other functions; specifically, these include the CPU, main memory, and files.

◆ Scheduler

In job management, jobs are continuously executed under a master scheduler and job scheduler. The master scheduler plays the role of an interface with the operator via the console panel.⁵² The job scheduler manages the reception, selection, start, and finish of the jobs.

◆ Reader

This reads the contents of JCL, analyzes them, schedules jobs, and places them in a queue.

◆ Initiator

This selects the programs with high execution priorities among those in the queue and assigns the resources that those programs need.

◆ Terminator

This releases resources that were used by programs just completed. If there is another program following, the terminator starts up the initiator.

◆ Spooling (Spool)

Spooling is the function of the I/O of jobs independent of the programs. Any output results to low-speed units such as a printer are first stored in a spool file. Then, after the program is finished, the output results are printed on the printer from the spool file by the service program of the OS.⁵³

The reason this is done is that, when the I/O unit is slow, directly performing the I/O process would reduce the processing speed of the computer.⁵⁴

◆ Cataloged Procedures

In job execution directions, typical processing (routine work) such as translation of languages is done in the following way. A set of JCLs is registered together at a separate location, and this registered set of JCLs is called for executing programs. By doing this, the computer prevents JCL errors. This set of registered JCLs is called catalogued procedures.

⁵² **Console panel:** It is a unit where the operator interacts with the computer system via key control and monitors its operation and where the system communicates failures, etc. It consists of a keyboard and a display.

⁵³ (Note) For spool files usually a hard disk is used. The service program that performs the sending of spool files to a printer is often called a writer (output writer).

⁵⁴ (FAQ) Most exam questions on job management are about spooling. They are always in the form of selecting the correct term, so be sure to have accurate understanding of spooling.

2.2.3 Task Management

Points

- A task is the smallest execution unit for using a resource.
- An interruption takes place in order to switch tasks.

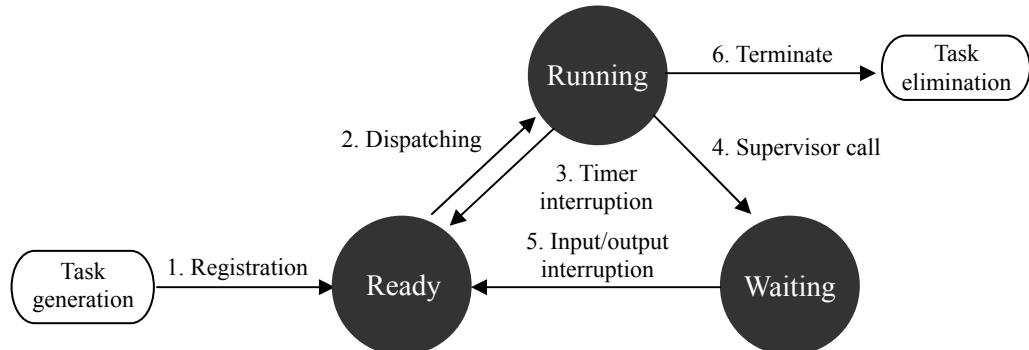
One of the functions of the control program, which is the “OS” in a narrow sense, is “task management.” **Task management** is the function of controlling the execution of programs and consists of various procedures such as synchronization control of programs, dynamic assignment of resources for program execution, and management of execution priorities of the programs. It also conducts various types of interruption control.

◆ Tasks and Jobs

A unit of processing from the perspective of the user (a single program or a set of programs executed consecutively) is called a **job**. In contrast, a unit of internal processing executed under the OS is called a **task**. “Task” refers to a processing unit that subdivides a program process.⁵⁵

◆ Control of Task Execution

Task management generates tasks required in response to a command and monitors the execution process. When a task generated becomes no longer necessary, the task is eliminated. The state transitions for tasks are shown below:



1. A task has been generated, so it is entered into the queue. → Move to the ready state.
2. For execution, move to the running state via the task dispatcher. → Move to the running state.
3. Time has expired; withdrawn to make way for a task with high priority. → Move to the ready state.
4. Withdraw for an I/O instruction. → Move to the waiting state.
5. Again waiting to be executed after completion of I/O. → Move to the ready state.
6. All processes are now completed. → The task is terminated.

Dispatching refers to the step of selecting a task with high priority from among those tasks in the ready state and advancing it to the running state.⁵⁶ **Supervisor call** refers to the step of invoking a function of the OS; in state transfer of tasks, this refers to an I/O instruction. I/O interruption is a

⁵⁵ **Process:** This term refers to a program being executed and is used interchangeably with the term “task.” “Process” is a word used by some operating systems such as UNIX. In recent years the expression “process” is used frequently.

⁵⁶ **Dispatcher:** It is the program of the OS that carries out dispatching; also known as the dispatching routine.

notice that I/O is completed.⁵⁷

◆ Interruption

Interruption refers to temporary suspension of a program currently being executed for any reason and transferring control to the OS to execute some necessary processing program. There are **external interruptions** caused by certain specific states of the hardware and **internal interruptions** caused intentionally when the control program is called from within a program.

The hardware detects interruptions. When the CPU detects an interruption, the OS receives it, changes the program being executed to the necessary state prior to the interruption, examines the cause of the interruption, and transfers control to the corresponding processing routine (program). The area in which this state is stored is called the PSW (program status word). There is also a possibility that while an interruption process is taking place, another interruption occurs. Priorities are given to these interruptions depending on their types so that multiple interruptions can also be controlled. The table below describes main types and examples of causes for interruptions.

Type of interruption cause	Name	Possible causes
External interruptions	Machine check interruption	Malfunction of units, fault, power/voltage trouble
	Clock function interruption (timer interruption)	Elapsed time of a fixed length (interval timer), reaching a designated time
	Input/output interruption ⁵⁸	Input/output completion, input/output unit status change (out of paper, etc.)
	External signal interruption	Instruction from console panel, external signals
Internal interruptions ⁵⁹ (traps)	Program interruption	Overflow, underflow, undefined instruction code execution, division by 0, memory protection violation
	Supervisor calls (instruction interruption)	Input/output operation command, task switching, page fault, control program invoked

2.2.4 Data Management and File Organization

Points

- Data management provides integrated methods for accessing files.
- File organization includes sequential organization, direct organization, indexed organization, partitioned organization, etc.

Another function of the control program, which is the “OS” in a narrow sense, is “data management.” **Data management** is the control program that manages data input and output. It provides various file organization methods such as sequential organization, direct organization, and indexed organization. It works as a bridge between logical files processed within a program and physical files whose structures are different.

Data management allows programmers not to worry about the physical structure of the files.

⁵⁷ (FAQ) State transition of tasks (processes) is almost certain to appear on every exam. Commit the entire figure of state transition to your memory.

⁵⁸ (Hints & Tips) Issuing of an I/O instruction is notified by supervisor call; I/O completion is notified by I/O interruption.

⁵⁹ Internal interruptions are intentionally caused by programs, so they are sometimes referred to as traps.

◆ Access Methods

Access methods include sequential access, direct access, and dynamic access, which is a combination of the first two.

Access methods

Sequential access: Processing files sequentially from the beginning⁶⁰

Direct access: Processing a specific record directly

Dynamic access: Using direct access to find and position a record, followed by sequential access

Sequential access

This is the method of handling records in a file in sequential order from the beginning. This can be performed with almost all recording media. This is suitable for collective processing in which all records in a file are subject to processing.

Direct access

This is the method where a necessary record is directly (randomly) accessed regardless of the order in which the records are stored. This method is used when the file medium is a directly accessible storage medium such as a hard disk. It is used in online real-time systems where only a part of a large number of records stored in a file needs to be accessed for quick update.

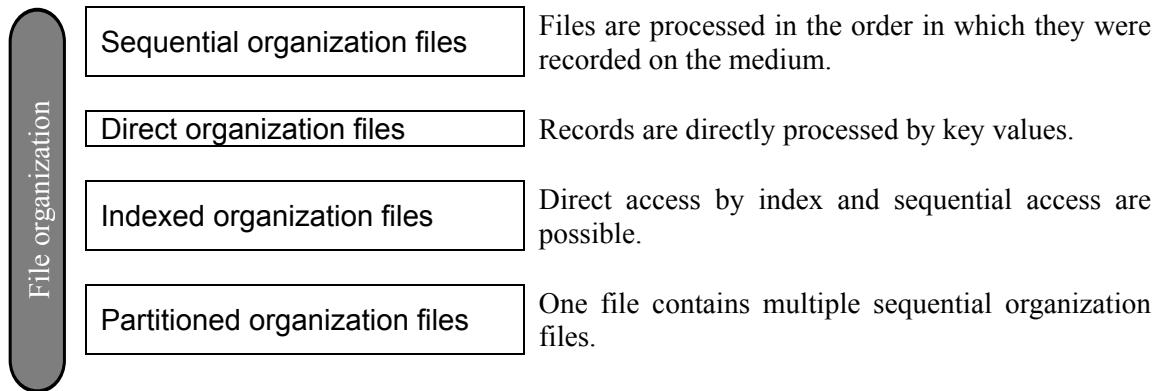
Dynamic access

This is the method where direct access is used to find a specific record and then sequential access follows. Similar to direct access, this is used when the file medium is a directly accessible storage medium.

⁶⁰ (Hints & Tips) Sequential access can be used with most media; however, direct access and dynamic access are limited to directly accessible media such as a hard disk.

◆ File Organization

File organization methods include sequential organization, direct organization, indexed organization, and partitioned organization,⁶¹ etc.

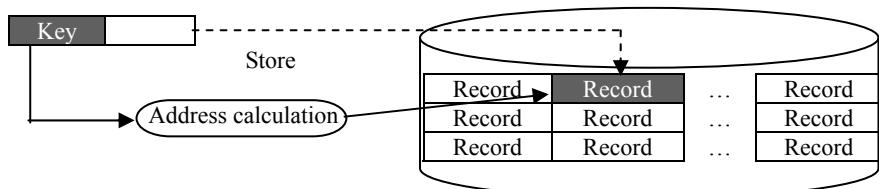


Sequential organization files

The records in the files are stored in consecutive positions following a certain order. Files of this type can be created on almost all media such as magnetic tape, hard disk, and floppy disk. In general, only sequential access is possible with these files.

Direct organization files

A storage address on the medium is calculated based on the key value found in each record, and the record is stored in that position.⁶² To access a record, we first calculate the storage address using the same formula, and the record is read from that location. There is a method where the key value of each record is directly used as the storage address for the record, but this is not very practical, creating a lot of wasted memory if the key values are not consecutive. A more general way is to use a certain type of conversion formula to calculate a storage address from the key value of each record. This is called **address conversion (randomization)**.



Address conversion sometimes produces the same address for different records. In such a case, the record stored first is called the home record while the record assigned to the same address later is called a synonym record.⁶³

⁶¹ (FAQ) Questions concerning the characteristics of these organization methods are frequently asked. Know the characteristics of each type of file organization.

⁶² (Hints & Tips) A special case of direct organization is relative organization, in which the key values are consecutive like 1, 2, 3, ... and the key value is itself the storage address of the record.

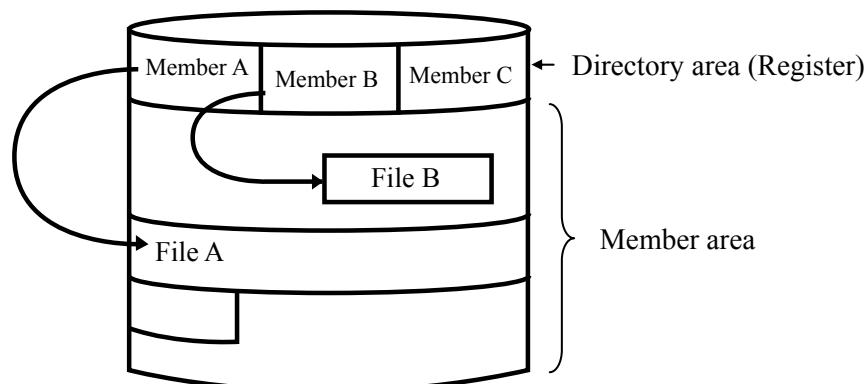
⁶³ **Synonym/Home:** Address conversion can take different key values but produce the same storage value. This is called a "synonym" (word meaning the same thing). If the result of key conversion stores a record, this record is called the home record, and another record that could not be stored there is called a synonym record. A synonym record needs to be stored elsewhere by some other method (e.g. by list).

Indexed organization files

These are files with an index, and they are organized such that the user can access the records by looking up their addresses by index. Sequential access, direct access, and dynamic access are all possible. In each case, the actual records are accessed only after their addresses on the medium are looked up using the index, so not only does the medium contain a basic data area (prime domain) where the data is stored but also an index area. Further, in order to prevent a situation where records cannot be added to the basic data area, an overflow area⁶⁴ is also reserved.

Partitioned organization files

In these files, sequential organization files are grouped into units called members, each of which is given a name. Then a directory containing these names and their leading addresses is created. Access is allowed to these members. Think of a member as a set of multiple files organized sequentially. Direct access can find the beginning of a member, and sequential access can be used to find a record in that member. Partitioned organization files are used as storage locations for program files and libraries but not very much as data files.



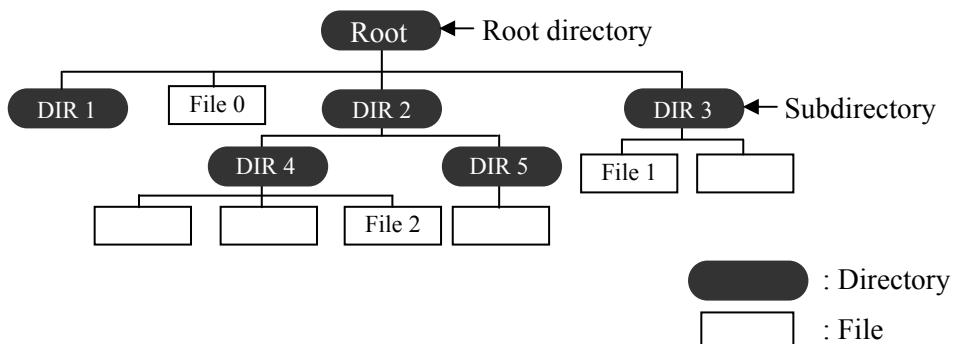
⁶⁴ (Note) There are two types of overflow areas: cylinder overflow area and independent overflow area. Overflow records from various tracks are stored in a cylinder overflow area; if a cylinder overflow area becomes full, additional records are stored in an independent overflow area shared by all of the files.

◆ Hierarchical File Systems

UNIX and Windows⁶⁵ make use of hierarchical file systems as mechanisms to manage files efficiently.

Directories and files

A file system has a hierarchical structure consisting of files and directories (directories are file registers). At the top of the hierarchical structure is the **root directory**, and directories under it are called **subdirectories**.⁶⁶



File manipulations

When searching for a file, we designate the path showing in which directory the file is located. There are two methods for doing this. For instance, if the hierarchical structure is as shown in the figure above, we can designate the path in the following ways:⁶⁷

- Absolute path

Designating a path from the root directory⁶⁸

<Example> Here is a way to designate “file1.”

\DIR3\file1 (The leading symbol “\” indicates the root directory.)

- Relative path

Designating a path from the current directory⁶⁹

<Example 1> Here is the designation of file2 when the current directory is “DIR2.”

DIR4\file2

<Example 2> Here is the designation of file2 when the current directory is “DIR4.”

file2

⁶⁵ (Hints & Tips) What UNIX and MS-DOS call “directory” is called “folder” in Windows and MacOS.

⁶⁶ (Hints & Tips) Whereas directories and files can be made under a directory, files and directories cannot be made under a file.

⁶⁷ (FAQ) Concerning hierarchical file systems, there are exam questions like “Choose an appropriate designation as an absolute path or as a relative path.” In those questions, the symbol for separating directories and files, as well as its use, will be explained in the question text.

⁶⁸ (Hints & Tips) Here we are using the symbol “\” to separate directories and files, but some operating systems use the symbol “/” instead.

⁶⁹ **Current directory:** It is a directory in which the user is working at the moment.

2.2.5 Memory Management

Points

- Memory management uses two types of memory: real memory and virtual memory.
- The basic format of the virtual memory system is the paging method.

Another role of the control program, which is the “OS” in a narrow sense, is “memory management.” **Memory management** makes the most effective use of the memory as well as compensating for any lack of the main memory capacity. To this end, it effectively uses auxiliary storage as part of the memory.⁷⁰

◆ Real Memory System⁷¹

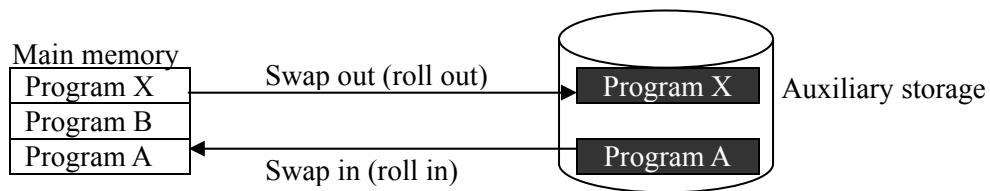
This is the system that manages the physical space of the main memory. Real memory can be controlled in various ways: partition, swapping, relocation, and overlay methods.

Partitioned method

When a program is placed in the main memory, the main memory is partitioned into several partitions, into which the program is loaded.⁷² Without memory management, fragmentation occurs, causing a situation which prevents programs from being stored even though empty space exists. Hence, to combine all empty areas together, compaction⁷³ is necessary.

Swapping (roll-in / roll-out)

Swapping refers to execution as the program keeps switching back and forth between the main memory and auxiliary storage. If a program is entered with higher priority than the priority level of the currently executed program, the new program is immediately loaded into the main memory and is executed. However, if there is no space in the main memory, any program in the main memory can be moved to the auxiliary storage. Hence, this system compensates for a lack of main memory capacity by utilizing auxiliary storage. However, if swapping occurs frequently, it means that programs are switched back and forth many times, thus reducing the processing efficiency of the computer system.



⁷⁰ **Memory leak:** Sometimes, for some reason, memory in the main memory, secured dynamically by an application, may not get released but remains in the main memory. This is called memory leak. To eliminate memory leak, compaction must be performed.

⁷¹ **Real memory system or “Real Storage (RS)”:** This refers to the actually existing memory; it is the main memory.

⁷² (Note) In partitioned method, multiple programs can be stored simultaneously, so multitasking is possible.

⁷³ **Compaction:** It means collecting empty memory areas to form a continuous area; also known as garbage collection.

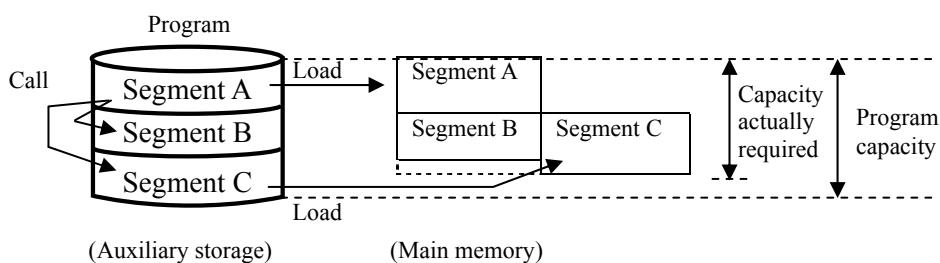
Relocation (Relocatable)

Relocation refers to the function wherein a program already assigned to a certain area is re-stored in another location. A program whose structure allows it to be relocated is called a relocatable program.⁷⁴

Overlay method

The physical limitations of the main memory can be eliminated; that is, programs are divided into segment units, and only the necessary segments⁷⁵ are loaded into the main memory to be executed. The entire program is stored in auxiliary storage, and the main memory contains only frequently used segments. Exclusive segments, which are never used simultaneously, are loaded from auxiliary storage to the main memory on an as-required basis.

For example, suppose that Segment A is a main routine used with high frequency while Segments B and C are subroutines called exclusively by Segment A. While Segment B is being executed, Segment C is in auxiliary storage. When Segment C is called, it is loaded in the area of Segment B. Consequently, the entire memory capacity of the program is “A + B + C,” but the capacity of the main memory is sufficient if it is at least the greater of “A + B” or “A + C.”



◆ Virtual Memory

Virtual memory provides a large capacity of storage space regardless of the size of the main memory.⁷⁶ Programs are stored in virtual memory (normally in auxiliary storage), and only the parts necessary for execution are loaded into the main memory.

Since the program is loaded into virtual memory, the instructions and data is given virtual addresses, which need to be converted to actual addresses (main memory addresses) for the execution of the program. This conversion is implemented by hardware called DAT (Dynamic Address Translator).

If virtual memory is used, it is necessary to convert virtual memory addresses to physical addresses in the main memory. This address conversion is performed at a high speed by DAT. Relocation is also performed if the main memory has fragmentation. Since this relocation is done during the execution of a program, it is called **dynamic relocation**.

We discuss the virtual memory strategies, which include three methods: page, segment, and segment-page.

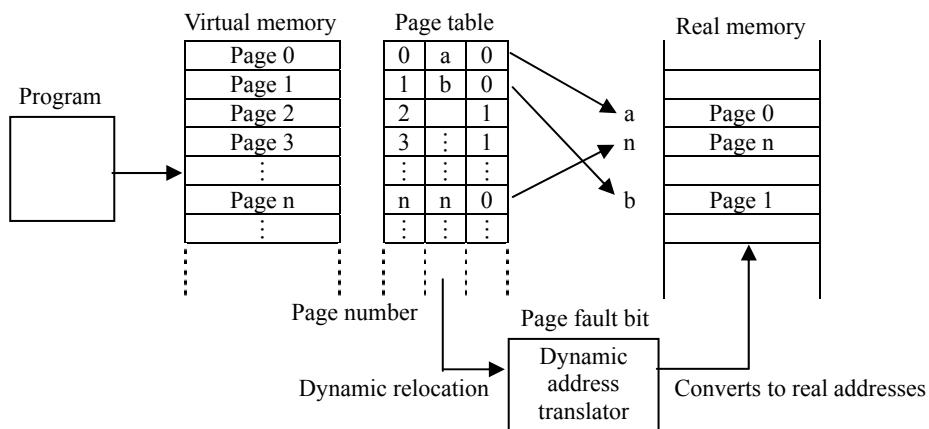
⁷⁴ (Note) “Relocatable” means that compaction is possible.

⁷⁵ **Segment:** It is a logical processing unit of a program. Here, we can regard a segment as a subroutine.

⁷⁶ **Virtual memory system or “Virtual Storage (VS)”:** It is a conceptual storage that does not actually exist. A program to be executed appears to be loaded into virtual memory, which is a large memory space, while only the portions (pages or segments) of the program with high frequency of use, data, and other parts necessary for the execution get loaded into the main memory.

Page method (Paging method)

In this method, the program is partitioned into units of a fixed size, called pages. A page then becomes the unit for loading into the real memory. Pages are managed by a page table, which has one entry for each page of virtual memory. If the corresponding page is in the real memory, the page fault bit becomes 0. This page fault bit then indicates whether or not the corresponding page is in the real memory.



Segment method

In this method, programs and logical sets of data is considered segments. Virtual addresses consist of segment numbers and addresses within the segments. The paging method is only for memory management and as such, the programs do not need to be written with pages in mind. In contrast, in the segment method, in which segments have different capacities, the programs must be written in consideration of the segment sizes.

Segments are logical processing units, so they can be treated as subroutines. However, the flexible lengths are sometimes inconvenient to manage, and the usage efficiency of the main memory may be reduced.

Segment-page method

This is an improved version of the segment method, in which segments are further partitioned into pages. Real addresses are accessed in the order of “segment → page → relative displacement within the page.”⁷⁷

⁷⁷ **Relative displacement within the page:** It is an address assigned such that the beginning of the page has displacement 0.

◆ Paging Algorithms

If a page necessary for processing is not found in the real memory, an interruption called a **page fault** occurs, and the page is read into the real memory from virtual memory. This is called **page-in**. On the other hand, **page-out** is to move an unnecessary page out to virtual memory. Page-in and page-out are together called **paging**.⁷⁸

If paging occurs frequently, the time for executing the control program increases, reducing the performance. This is called **slashing**. To minimize the occurrence of slashing as much as possible, various algorithms are proposed to select pages that are subjects of page-outs.

Common page-out methods and their properties are shown below.⁷⁹

Method	Properties, etc.
LRU (Least Recently Used)	Pages are compared on the time elapsed since the last referencing. The page with the longest elapsed time is paged out.
FIFO (First In First Out)	The page with the longest elapsed time up to the present is paged out.

⁷⁸ (Hints & Tips) Swapping and paging are similar, but note that swapping takes place in program and segment units whereas paging takes place in units called pages, which are parts of a program.

⁷⁹ (FAQ) There are exam questions that require specific tracking of page-ins and page-outs. For example, if the order in which pages are used is “1, 3, 2, 3, 5, 2,” and if the main memory has page 3, which page will be the first to be paged out? For these questions, have clear understanding of ideas like LRU and FIFO.

Quiz

- Q1** Explain the roles of task management.
- Q2** What type of file organization is this? Sequential organization files are grouped into units called members, each of which is given a name. A directory is created, including these names and leading addresses, and access is allowed to these members.
- Q3** Explain swapping.
- Q4** What is the unit of loading into the main memory in the page method?

2.3 System Configuration Technology

Introduction

Various system configurations are being used to reduce the cost and increase the efficiency of computer systems. These include client/server systems to distribute the load, dual systems to improve reliability, and duplex systems.

2.3.1 Client Server Systems

Points	<ul style="list-style-type: none"> ➤ A client server system is a typical example of distributed processing. ➤ Clients request processing, and servers provide services (processing).
---------------	--

A client/server system (CSS) is a form of computer system where a network is used to distribute processing; it is a type of system configuration.⁸⁰

◆ Overview of Client Server System

CSS is configured with computers with roles called clients and servers. The servers provide services (processing) such as file management, database management, modification and supply of data, printing control, and communication functions, as requested by clients.

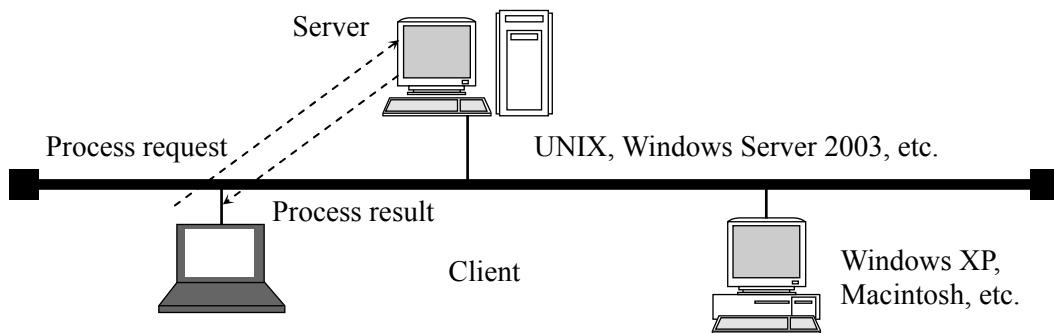
A client sends service requests to a server, receives the results of data processed by the server, and displays the results.

The clients and the servers distribute their processes in an attempt to spread the load of computer processing. In addition, by sharing resources, the user can reduce waste. For instance, by connecting a high-speed printer to a server, the clients can share the high-speed printer. Having one high-speed printer may be less expensive than preparing a low-speed printer for each of the clients, although this depends on the number of clients.

Computers used as a server are generally more high-performance than the client computers. To clarify the functions of various servers, they can be named by their functions, such as file servers, database servers, print servers, and communication servers.⁸¹

⁸⁰ **Client/server system:** In multiprogramming, if there is one operational computer, the user can run both the client and the server within the one computer. In other words, a client/server process does not mean that everything is distributed. It is one method to achieve distributed processing.

⁸¹ (Hints & Tips) If a service request made by a client cannot be provided by the server, that server can become a client and request another server to perform the requested process.



◆ Types of Server

The table below shows the types of server, depending on the provided functions.⁸²

Type of server	Services provided
File server	Managing shared files, controlling file access and writing
Database server	Managing databases, operating databases with DBMS (Database Management System)
Print server	Managing shared printer, printing when a printing request is received
Communication server	Providing communication functions with the outside using a network

◆ Characteristics of Client/Server System

Since CSS provides a typical type of distributed processing, the characteristics of distributed processing apply in a straightforward manner.⁸³

Advantages of client/server system

- When the processing is local at clients, the response is faster.
- Costs for the entire information system can be reduced, achieving a good cost/performance ratio.
- Specialized server functions give the system greater economical efficiency and performance.
- It is easily extended; clients and servers can be flexibly added as needed.
- Even when clients access a server, they are unaware that they are in a distributed environment.

Disadvantages of client/server system

- The system could be confusing unless the server administrator is clearly identified.
- The performance deteriorates if the use gets concentrated on specific servers.
- The performance of the entire system depends on the network performance.⁸⁴

⁸² (Hints & Tips) Clients and servers do not necessarily have to have the same OS. Where there are multiple servers, they do not have to have the same OS either. In addition, if there are multiple clients, they need not have the same OS either.

⁸³ (Note) If client/server type programs are logically divided into three layers (presentation layer, function (application) layer, and data layer), such a system is called a 3-layer client server system. By distinguishing 3 layers by function, such a system strives to enhance system performance and efficiency for development and maintenance.

⁸⁴ (FAQ) There have been many exam questions regarding the knowledge of client/server systems. Most of them are about the role of a client or that of a server, so be sure to know these things well.

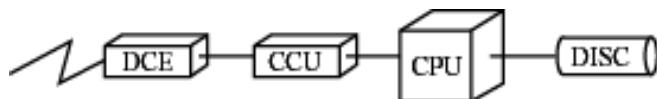
2.3.2 System Configurations

Points	<ul style="list-style-type: none"> ➤ A dual system is a configuration for high reliability; a duplex system is a configuration for high availability. ➤ There are two types of multiprocessor systems: loosely coupled and tightly coupled systems.
---------------	---

A variety of system configuration patterns are available according to the objectives of information processing. For instance, the reliability of a computer system improves by having multiple units installed. We will look at main system configurations and their characteristics from the viewpoints of reliability, efficiency, costs, and other factors.

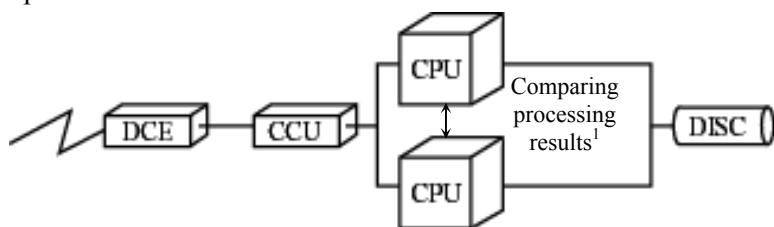
◆ Simplex System

This system consists of one CPU only. Reliability and the processing capabilities are inferior in comparison with other configurations, but it is economical. This configuration is commonly used.⁸⁵



◆ Dual System

This is a system configuration in which two CPUs perform the same processing and compare the processing results to each other. This configuration is applied when the process is not allowed to stop, even for a moment. If one CPU fails, the system cuts off the failed CPU and continues processing on the other CPU. Reliability is extremely high, but this system is expensive.^{86 87}



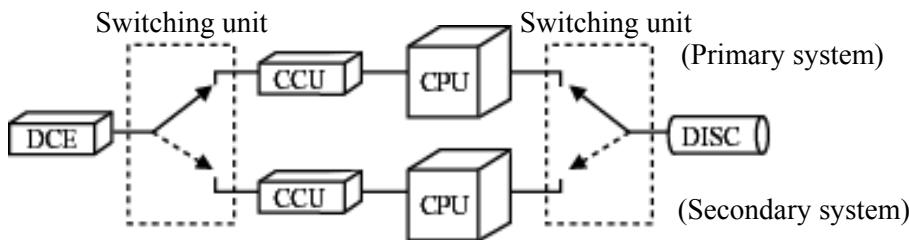
⁸⁵ **DCE (Data Circuit-terminating Equipment):** This unit converts signals received from communication line, sends them to data terminals, and also executes exactly the opposite operation. Normally, this unit is connected at the end of a communication line and functions as an interface with a computer. A modem (modulator-demodulator) is used on an analog line, and DSU (Digital Service Unit) is used on a digital line.

⁸⁶ **CCU (Communication control unit):** This unit controls the reception and transmission of data, performs error control, and assembles and decomposes characters.

⁸⁷ **DISC:** Auxiliary storage

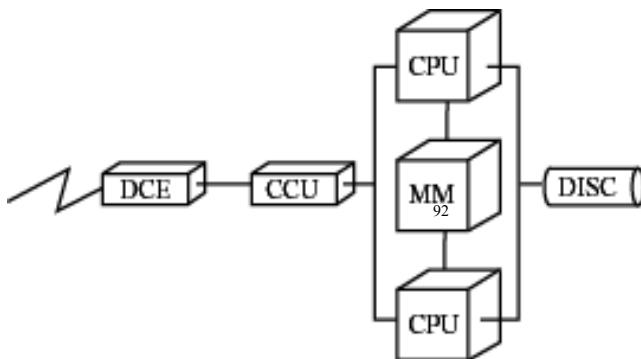
◆ Duplex System (Standby System)

In this system, two CPUs are prepared, where the primary system is used for online processing while the secondary system is used for low-priority processing such as batch processing.⁸⁸ If the primary system fails, the online processing gets switched to the secondary system.⁸⁹ Switching time is said to be anywhere from several tens of minutes to several hours. Reliability is lower than that of a dual system, but this system is better from the standpoint of costs, so it is more commonly used than the dual system.



◆ Multiprocessor System (Concurrent Processors)

In this system, multiple CPUs and CCUs are sharing and processing tasks, so it is a system configuration with high processing efficiency. There are two types of multiprocessor system. One is LCMP⁹⁰ (loosely coupled multiprocessors), in which multiple computer systems are controlled by separate operating systems. The other is TCMP⁹¹ (tightly coupled multiprocessors), in which multiple computer systems share the main memory and are controlled by the same operating system. The figure below shows an example of TCMP.



⁸⁸ (FAQ) There are many exam questions on characteristics of dual, duplex, and multiprocessor systems. The key term for each system is as follows: “comparing process results” for dual, “switching the units” for duplex, and “sharing the main memory” for multiprocessor systems.

⁸⁹ (Note) If a failure occurs in the primary system, it takes time to switch to the secondary system. This is because the batch processing or whatever else is being executed in the secondary system must be suspended, and the OS must be booted for the online system. A hot standby system configuration can solve this by standing by, ready to switch at any time. In this case, the OS for the online system stays on, so switching can occur immediately.

⁹⁰ **LCMP (Loosely Coupled Multiprocessor):** Each CPU has its own main memory and independent OS. CPUs are joined by a high-speed network or shared path. This is a configuration where independent computer systems are connected via a network.

⁹¹ **TCMP (Tightly Coupled Multiprocessor):** One main memory and one OS are shared in this configuration. Each CPU can perform identical processes, so even if one CPU fails, the processing can continue, albeit with lower performance. This configuration is highly reliable and thus is used in systems where a high level of processing capability is required.

⁹² **MM:** Main memory

2.3.3 Centralized Processing and Distributed Processing

Points	<ul style="list-style-type: none"> ➤ Centralized processing has a high level of safety but is inflexible. ➤ Distributed processing is economical, but its real substance is hard to understand.
---------------	---

Depending on how the computers are placed physically, there are two types of processing: centralized processing and distributed processing.

◆ Centralized Processing

Centralized processing is a system configuration in which one computer is connected with many terminals, and the one computer alone does all of the processing. It is easy to maintain the consistency of data, and it is easy to manage the resources. These merits contributed to the popularity of this configuration in which a general-purpose computer is used as the host in centralized processing. Below is a summary of relative comparison with distributed processing.

Advantages of centralized processing	Disadvantages of centralized processing
It is easy to improve the cost/performance ratio. (Grosch's Law ⁹³) Operation and maintenance require a smaller staff. Safety level of the system is high.	Extendability is poor to keep up with new technologies. Backlog can be easily accumulated. ⁹⁴ Overhead of the OS is significant. Recovery of a host failure is time-consuming. A failure has a far-reaching effect.

◆ Distributed Processing

Distributed processing is a system configuration in which multiple computers connected via a network perform the processing. Since the processing is done through a network, the processing time is longer than that of centralized processing. But, the merit is that a failure of one computer does not affect the entire system. Below is a summary of relative comparison with centralized processing.⁹⁵

Advantages of distributed processing	Disadvantages of distributed processing
Management responsibilities are clear. (Management responsibilities can be delegated to each organization.) Effects of system failure are local. Maintenance is easy (locally closed). It is economical as only necessary units are installed	Its real status can be hard to understand. It is difficult to identify trouble spots. Network performance has great impact. Data inconsistency can occur easily. Individual units are managed carelessly.

⁹³ **Grosch's Law:** It states that “performance is directly proportional to the square of the price.” If the price of a computer doubles, the performance quadruples. However, technological advancement has reduced the prices of devices significantly, so this law is no longer applicable.

⁹⁴ **Backlog:** It means systems, software, programs, etc. that are necessary to develop but the development of which has not even begun. The term often refers to those that are held back in the IT department within a company.

⁹⁵ (FAQ) There are exam questions where you are required to identify the characteristics of centralized processing and distributed processing. For example, questions may be of the form “Which of the following is an appropriate characteristic of a centralized processing system?” Know the advantages and disadvantages of each processing type.

As shown in the following table, distributed processing can be classified according to the distribution status of functions and loads. It is said that vertical load distribution does not exist in reality.

Configuration Function	Function distribution	Load distribution
Horizontal distribution	Horizontal function distribution	Horizontal load distribution
Vertical distribution	Vertical function distribution	

Horizontal function distribution

A horizontal function distribution system is a system in which computers are classified according to type of application and type of data; examples include processing function distribution and database distribution. For instance, in financial institutions, host computers may be classified into those in an information system and those in an accounting system; this classification is based on the type of processing, so it is an example of processing function distribution. Database distribution means that computers are located in appropriate locations based on the contents of data.

Horizontal load distribution

This is a system in which multiple computers perform processes jointly when an application is executed. When a process is requested, an idle computer responds. In this mode, if one computer fails, the process switches to another computer and is continued. Hence, this system is quite effective in time of failure. A tightly coupled multiprocessor system is an example of this type.

Vertical function distribution

This is a system where the processing function is shared among workstations belonging to individual users as well as computers shared by multiple users. Here, there is a vertical relationship in regard to the processing function. A client/server system is a typical example of a vertical function distribution system.⁹⁶

⁹⁶ (Hints & Tips) A client/server system appears as if it were horizontal distribution, but it is properly classified under vertical function distribution. Since one server performs processes of multiple clients, there is a vertical relationship in functions.

2.3.4 Classification by Processing Mode

Points	<ul style="list-style-type: none"> ➤ In batch processing, data is stored and processed all at once. ➤ In real-time processing, data is processed at the moment they come into existence.
---------------	--

From the standpoint of processing modes, system configurations can be classified into two categories: batch processing and real-time processing. They can also be classified by whether or not they are connected to a network.

Processing mode	Operation mode	Connection method
Batch processing ⁹⁷	Center batch processing	Offline
	Remote batch processing	
Real-time processing	Interactive mode processing	Online
	Online transaction processing	
	Real-time control	

◆ Batch Processing Systems (One-Time Processing Systems)

The word “batch” means a “bundle.” Batch processing is any method in which data to be processed by a computer are stored for a certain period of time or are saved on original sheets or storage media and are later processed all at once. Such systems have the following characteristics:

- The computer can be used efficiently because the processing is done all at once.
- It is suitable for routine and repetitive processing (standard tasks).
- The results are not immediately obtained because the processing is collectively done.

In **center batch processing** systems, the processing takes place at a central computer center; in **remote batch processing**,⁹⁸ the batch processing is performed from a terminal at hand via a communication line.

◆ Interactive Processing Systems

This is a method in which the processing is performed in the mode of a dialogue with the computer. We can proceed with the processing while communicating with the computer and can correct errors immediately once they are noticed; therefore, it is a processing mode suitable for program development, etc. An interactive processing system requires prompt responses.

Also, an interactive processing system requires an OS in which the multiprogramming function is

⁹⁷ **Batch processing and real-time processing:** Batch processing is where data is stored and processed all at once. Examples include the calculations of electricity payment, water payment, and gas payment. Payroll calculation is also an example of batch processing. Real-time processing is where data is processed immediately when they are generated. Seat-reservation systems of airlines and trains are examples of this type.

⁹⁸ **Center batch process and remote batch process:** Center batch processing does not use communication lines. Remote batch processing is where a terminal unit at hand is used to perform batch processing at a host computer at a remote location. It is sometimes called RJE, which stands for “Remote Job Entry” as the user enters a job at a remote site.

supported, and the processing is performed using TSS.⁹⁹

◆ Online Transaction Processing Systems

A transaction is a series of data or instructions, sometimes called a transaction file or transaction data. Online transaction processing refers to the processing of data updates on an online file at the computer center via a connected online terminal. Examples of online transaction processing include savings account systems in banks and train and airline seat-reservation systems.

Online transaction processing requires specialized (dedicated) terminals with a high level of usability aimed at enhancing the processing efficiency, such as bank ATMs and terminals for issuing reserved-seat tickets at reservation windows. In addition, since data is shared by many terminals, there is a risk that simultaneous access to the same data could cause problems such as deadlock¹⁰⁰ or data destruction. Hence, attempts are made to enhance the data maintainability.

◆ Real-Time Control

In general, real-time control refers to the method by which data is processed in real time once a processing request is made and the result is immediately reported to the requester. This concept also includes online transaction processing. However, in a narrow sense, this refers more specifically to the processing mode at places like manufacturing plants, where the system is interlinked with a sensor system tracking the physical motions of objects to be controlled, processing corresponding to the external signals are immediately executed, and the results are immediately sent back to units on production lines (e.g. robots) as control signals. Production control systems at steel plants and automobile factories are examples of this type. Another example is a 24-hour monitoring process of electrical system managed by computers; if there is something wrong with the system, the system reports it to the maintenance company in real time, or the unit that has detected it gets cut off in real time.¹⁰¹

⁹⁹ **TSS (Time Sharing System):** The system authorizes multiple programs to be executed in a specific order for an extremely short duration at a time (several milliseconds at a time). This procedure is repeated so that the execution of each program can get completed within a certain period of time. From the user's viewpoint, there is no mutual interference, so each user feels as if he or she were the only one exclusively using the computer system.

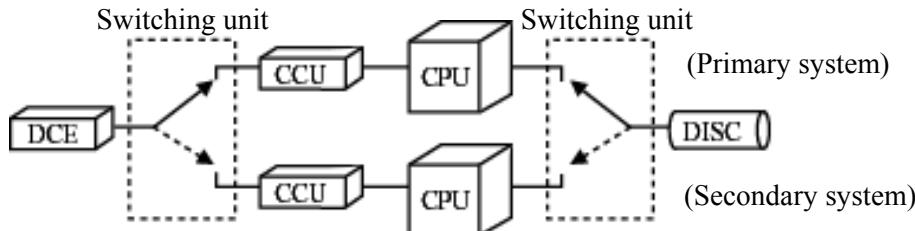
¹⁰⁰ **Deadlock:** It is a situation where the system gets stuck because multiple tasks (programs) try to access the same resource (file, database, etc.) and go into the waiting mode.

¹⁰¹ (FAQ) Many exam questions involve characteristics of remote batch processing and interactive processing. Both are online, but note that remote batch is a type of batch processing while interactive processing is a type of real-time processing.

Quiz

Q1 Explain the roles of clients and servers in a client/server system.

Q2 What is the type of system configuration shown below?



Q3 As shown below, distributed processing is classified by the distribution status of functions and loads. To which category does a client server system belong?

Function	Configuration	Function distribution	Load distribution
Horizontal distribution	Horizontal distribution	Horizontal function distribution	Horizontal load distribution
Vertical distribution	Vertical distribution	Vertical function distribution	

Q4 What is the difference between batch processing and real-time processing?

2.4 Performance and Reliability of Systems

Introduction

To evaluate computer systems, various methods are available. While good performance (high processing speed) is important, fault-tolerance (high reliability) is also significant.

2.4.1 Performance Indexes

Points

- Response time, throughput, and turn-around time are used for performance evaluation.
- Instruction mix and benchmark are used for performance indexes.

To evaluate the comprehensive performance of computer systems, including their software and hardware, we can use various criteria such as response time, throughput, and turn-around time. Indexes to evaluate performance, especially the hardware, include instruction mix and benchmark.

◆ Terms Related to Performance Evaluation of Computer Systems

In evaluating the performance of computer systems, it is important to calculate the processing time of jobs and programs.¹⁰²

Response time

This is the amount of time between the completion of input at an input unit and the beginning of output at an output unit. For example, when a processing request is made at the keyboard of the computer, this time refers to the amount of time it takes until the result is shown on the display unit or until the printing begins. This is mainly used to evaluate the performance of an online system.

Throughput (Processing capability)

This refers to the amount/number of jobs that can be processed by the computer system within a certain unit of time, or the amount of time required to process a certain job. This processing time includes the exclusive CPU time and process-waiting time such as preparation for I/O operation and clean-up time.

¹⁰² (Hints & Tips) There are instruction mixes and benchmarks for evaluating the performance of computers, and instruction mixes are for hardware evaluation. However, even if hardware is very fast, the entire system performance becomes poor if the performance of the OS is poor. Hence, the performance of hardware is often used only for reference.

Turn-around time (TAT)

Technically, this refers to the amount of time it takes information to make the rounds of the system. In batch processing, this is the duration between submission of a program at the window and the time when the results are obtained. In business operations, this is the duration from the time when a client places an order to the time when the ordered product is shipped and reaches the client.

◆ Instruction Mix¹⁰³

An instruction mix is used to compare the performance of hardware in computer systems. Even if the hardware is fast, if the performance of the OS is poor, the performance of the entire system becomes inferior. An instruction mix is to use an average program and calculate the average instruction execution time per instruction and MIPS value,¹⁰⁴ based on the execution frequency of each instruction.

Under these conditions, let us do some specific calculations of the MIPS value.

Instruction group	Execution speed (microsecond)	Frequency of appearance
A	0.1	40%
B	0.2	30%
C	0.5	30%

First, let us calculate the average instruction execution time. The execution speed of each instruction is expressed in microseconds (10^{-6}). The **average instruction execution time** is the sum (over all instructions) of the products of the execution time of instructions and their respective frequencies.

$$\begin{aligned}\text{Average instruction execution time} &= 0.1 * 10^{-6} * 0.4 + 0.2 * 10^{-6} * 0.3 + 0.5 * 10^{-6} * 0.3 \\ &= (0.04 + 0.06 + 0.15) * 10^{-6} \\ &= 0.25 * 10^{-6} \text{ (seconds/instruction)}\end{aligned}$$

The average number of instructions executed per second is the inverse of the average instruction execution time, so it is obtained as follows:

$$\begin{aligned}\text{Average number of instructions executed per second} &= 1 / (0.25 * 10^{-6}) \\ &= 4 * 10^6 \text{ (instructions/second)} \\ &= 4 \text{ (MIPS).}^{105}{}^{106}\end{aligned}$$

FLOPS¹⁰⁷ is used as an index to evaluate the performance of floating-point operations.

¹⁰³ (Note) An instruction mix for scientific calculations is called “Gibson mix,” and one for business calculations is called “commercial mix.”

¹⁰⁴ **MIPS (Million Instructions Per Second):** This is the performance index expressing the number of machine instructions, in millions (10^6), that can be executed per second. This is just for the performance of hardware, so again it is used only for reference.

¹⁰⁵ (FAQ) Exams do have questions where you are asked to calculate MIPS values given an instruction mix or to calculate the average clock count per instruction. You would want to be familiar with these calculation questions.

¹⁰⁶ **Clock:** This refers to the frequency of a clock signal generated by a circuit called a clock generator. Since instructions inside the CPU are synchronized to this clock signal as they are executed, the higher the clock frequency is, the more instructions can be executed in a given period of time. For example, if the clock frequency is 200MHz, there are $200 * 10^6$ clock signals per second. In general, one instruction takes several clocks.

¹⁰⁷ **FLOPS:** It stands for floating-point operations per second. This is an index expressing the number of floating-point operation instructions executed per second. If it is expressed in millions (10^6), it is called MFLOPS.

◆ Benchmark

A benchmark is used to compare and evaluate the comprehensive performance of computers, including the hardware and the OS, by measuring the standard program execution time.¹⁰⁸

2.4.2 Reliability

Points	<ul style="list-style-type: none"> ➤ Reliability indexes include RAS, RASIS, and the bathtub curve. ➤ Points to be aware of in reliability design are fail-safe and fail-soft.
---------------	--

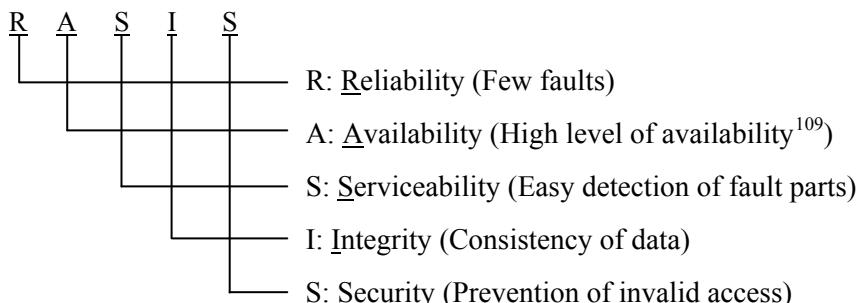
The level of reliability required for information systems varies depending on the purpose for which the systems are used. Sometimes the economical factor must be sacrificed to achieve a high level of reliability. In some other situations, not only the subject of reliability is focused on the operation of the system but also the information handled by the system needs to be reliable as well.

◆ Reliability Indexes

Reliability is the degree to which system operation is stable. The ideal case is that the system does not fail, but there is no system that does not ever fail.

RAS/RASIS

Both of the terms RAS and RASIS are acronyms of elements that help computer systems to operate in a stable manner. RAS stands for the first three elements of RASIS:

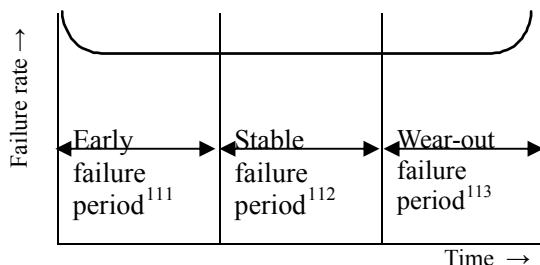


¹⁰⁸ (Note) An example of a technical calculation benchmark is SPECmark, and an example of a transaction benchmark is TPC. TPC-C is the frequently used benchmark under TPC which directly responds to actual business applications.

¹⁰⁹ **Availability:** Availability refers to the probability that the system is maintaining its functions (operating) at any given time or the percentage of the duration when the functions are maintained during a certain period of time.

Bathtub curve

The bathtub curve is used to illustrate the concept of hardware lifecycle. Hardware may fail during the initial period of its operation due to defective parts, etc., but the probability at which these failures occur decreases gradually as repairs and replacements are made. After that, because of wear and tear of various parts, the probability of failures increases, and eventually its life is determined to be over. This curve is shown below.¹¹⁰



◆ Reliability Design Points

A highly reliable system that can continue to operate even when some part of the system fails is called a **fault-tolerant system**. Common technologies for configurations of highly reliable systems include **fail-soft**, the function enabling the system to continue its operation, perhaps with lower performance or fewer functions, when a failure occurs, and **fail-safe**, the function enabling the system to operate safely by avoiding risky conditions when a failure occurs.

Fail-soft

This refers to the function in which, when a failure occurs, the failed part gets cut off and the system continues to operate, perhaps with a lower performance level (fall back¹¹⁴). In a duplex system, normally the two systems are independently processing data, but if one system should fail, the configuration would switch the processing to the other system and would carry on the processing. In addition, when a failure occurs in multiprocessors, the system continues its services by cutting off the failed processor. This too is a system configuration with fail-soft in mind.

Fail-safe

This refers to the function in which, when a failure occurs, the system locks its functions in a safe mode established in advance to control the extent of the impact of the failure.¹¹⁵ This is just like the measure where all railroad lights turn red when an accident has occurred. In system configurations where two systems compare the processing results of each other, such as in a dual system, when the compared results are different, the system in which a failure is determined to have occurred is cut off while the operation continues on.

¹¹⁰ (Note) The bathtub curve is so named because the graph showing the relationship between the failure probability and time resembles the shape of a bathtub.

¹¹¹ **Early failure period:** It is a period of failures at the beginning of unit use. These failures become less frequent as time passes.

¹¹² **Stable failure period:** The unit is stable during this period, with less frequent failures.

¹¹³ **Wear-out failure period:** A certain period of time has passed, and failures become more frequent during this period.

¹¹⁴ **Fall-back:** In a fail-soft computer system, processing continues at a lower level of functionality; this is called a fall-back or a fall-back operation.

¹¹⁵ (Hints & Tips) Fail-soft and fail-safe are similar words, so do not confuse them.

Fool-proof

This term refers to a measure that prevents an unintentional use of a program from causing a failure, especially when indefinitely many users use the same program. If one individual is using a particular program, the way the program is written does not create a major problem, but when there are indefinitely many users, how the program gets used is hard to predict.¹¹⁶

2.4.3 Availability

Points

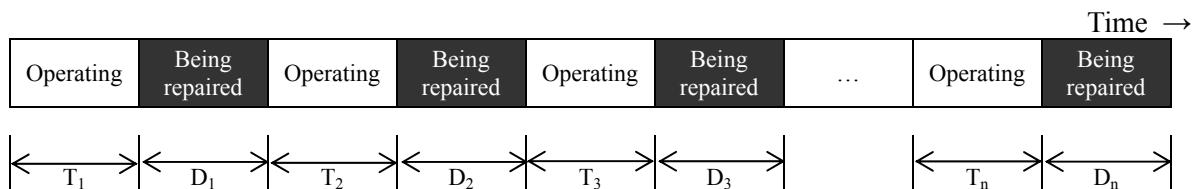
- MTBF is the time when the system is operating properly, and MTTR is the time when it is being repaired.
- Availability is the ratio of the time when the system is operating properly.

One of the indexes in RASIS is “A” for availability, which means the operation rate. The availability is calculated using MTBF and MTTR as follows:

$$\text{Availability: } A = (\text{MTBF}) / (\text{MTBF} + \text{MTTR})$$

◆ MTBF and MTTR

Suppose that the operation status of a computer system is as shown below:



MTBF (Mean Time Between Failures)

This is the average length of time that the system continues to operate without a failure. The larger MTBF is, the more reliable the system is. Therefore, this is used as an index of reliability (“R” in RASIS).¹¹⁷

$$\text{MTBF} = (T_1 + T_2 + T_3 + \dots + T_n) / n$$

(Here, “n” is the number of intervals the system was operating without failure.)

¹¹⁶ (FAQ) There are exam questions concerning what each of the letters RASIS means as an index for computer system reliability. At least know what RAS stands for.

¹¹⁷ (Note) Functions that improve MTBF include error detection, automatic 1-bit error correction, instruction re-try, etc. These are functions that prevent the computer system from coming to a stop. Functions that improve MTTR include log output. By looking up logs, the cause of failure can sometimes be identified. Remote maintenance also helps detect a failure promptly, enhancing MTTR.

MTTR (Mean Time To Repair)

This is the average length of time required for repair when a failure occurs. The shorter the repair time is, the better the system is. Therefore, it is used as an index of serviceability ("S" in RASIS).

$$\text{MTTR} = (D_1 + D_2 + D_3 + \dots + D_n) / n$$

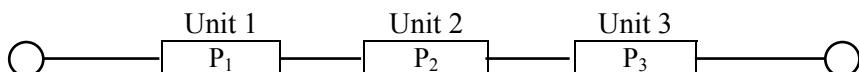
(Here, "n" is the number of intervals the system was operating without failure.)

◆ Calculation of Availability

To calculate the availability, the serial connection and parallel connection sections must be calculated differently. The basic ideas are described below.¹¹⁸

Availability in a serial connection system

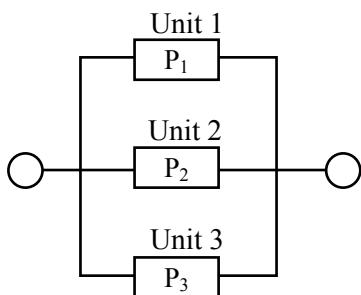
The availability of an entire serial connection system as shown here is the product of the availabilities of each unit. Here, P_1 , P_2 , and P_3 are the availabilities of the respective units shown in the figure.



$$\text{Availability of the entire system} = P_1 * P_2 * P_3$$

Availability in a parallel connection system

Suppose that we have, as shown below, a system in parallel connection where the system operates as long as at least one of Units 1, 2, and 3 is operating. Here, the availability is calculated using the fact that the probability of the entire sample space is 1.



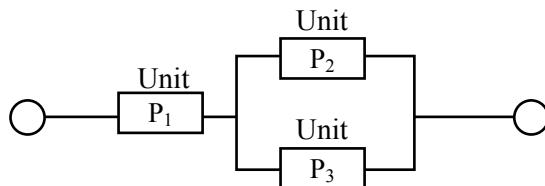
Availability of the entire system

$$\begin{aligned}
 &= 1 - (\text{Probability that all units fail simultaneously}) \\
 &= 1 - (\text{Prob. that Unit 1 fails}) * (\text{Prob. that Unit 2 fails}) * (\text{Prob. that Unit 3 fails}) \\
 &= 1 - (1 - P_1) * (1 - P_2) * (1 - P_3)
 \end{aligned}$$

¹¹⁸ (FAQ) There is always a question involving a calculation of availability. Make sure you understand correctly how to calculate it.

Availability in a system where serial and parallel connections are combined

Suppose that there is a system which operates if Unit 1 is operating AND at least one of Units 2 and 3 is operating. In this case, we consider that Unit 1 and the parallel section (Units 2 and 3) are serially connected.

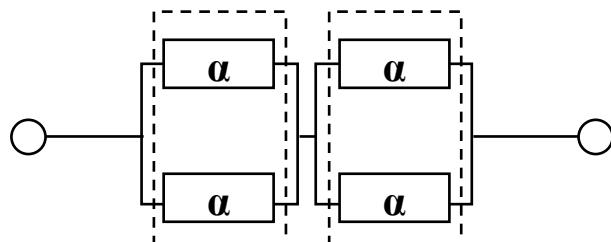


Availability

$$\begin{aligned}
 &= (\text{Availability of Unit 1}) * \{1 - (\text{Prob. that Units 2 and 3 fail simultaneously})\} \\
 &= (\text{Availability of Unit 1}) * \{1 - (\text{Prob. that Unit 2 fails}) * (\text{Prob. that Unit 3 fails})\} \\
 &= P_1 * \{1 - (1 - P_2) * (1 - P_3)\}
 \end{aligned}$$

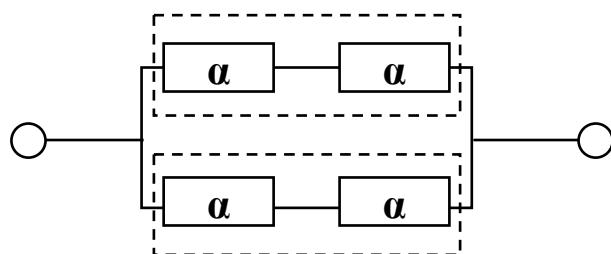
Let us consider more complicated configurations.¹¹⁹ Even though the two systems below may appear similar, the availabilities are different. Here, the letter α in the figure indicates the availability.

[Configuration 1] The two parallel sections (inside the dotted lines) are serially connected.



$$\text{Availability} = \{1 - (1 - \alpha)^2\} * \{1 - (1 - \alpha)^2\} = \alpha^2(2 - \alpha)^2$$

[Configuration 2] The two serially connected units (inside the dotted lines) are connected in parallel.

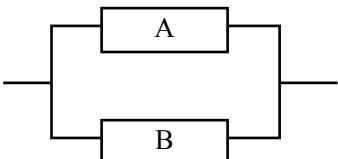


$$\text{Availability} = 1 - (1 - \alpha^2) * (1 - \alpha^2) = \alpha^2(2 - \alpha^2)$$

¹¹⁹ (Hints & Tips) Note that similar configurations have different availabilities.

Quiz

- Q1** Explain the meanings of the following terms: “MIPS,” “response time,” “throughput,” and “turn-around time.”
- Q2** Explain the meaning of RASIS.
- Q3** Explain the meanings of MTBF and MTTR.
- Q4** Express the availability using MTBF and MTTR.
- Q5** Calculate the availability for the entire system configuration shown below. A and B are units, each of which has an availability of 0.97. The entire system is assumed to be in operation if at least one of the units is operating.



System Applications

Introduction

Various systems have been developed using networks and databases. Close to our daily life are the Internet and database services (generally called commercial databases). Examples of applications of multimedia systems include 3D graphics.

2.5.1 Network Applications

Points

- Uses of the infrastructure include the Web, the Internet, intranets, and extranets.
- Application systems include Internet shopping, groupware, and debit cards.

Today, our information society has networks spanning all over the world like a gigantic web. Systems using networks themselves and application systems with add-on values are available.

This is a rather new area, so there are not many exam questions on this topic, and they are relatively easy. Most of the questions simply require knowledge of the terms, so be sure to memorize them to improve your exam scores.

◆ Uses of Infrastructure

“Infrastructure” means “foundation” or “basis.” In computer systems, this word refers to the foundation of software and hardware to form the systems. For instance, in network construction, various components such as communication lines, communication units, and the charge system of the communication lines are parts of what is known as the **communications infrastructure**.

Web (WWW: World Wide Web)

This is the information search system in the hypertext format, developed by researchers at CERN.¹²⁰ Since information distributed all over the world is mutually linked by this network using hypertext,¹²¹ a name meaning “global spider web” was given to it.

WWW is a mechanism on the server that records information in the form of an Internet homepage. Software that accesses WWW and displays it on a screen is called a Web browser or simply a “browser.”

The Internet

It is a collection of networks all over the world connected together by TCP/IP. There is no government organization or designated organization managing it in an integrated manner. Instead, the technical support and resource management are done by volunteer organizations.

¹²⁰ CERN (Conseil European pour la Recherche Nucleaire) (European Council for Nuclear Research): It is a quantum physics research institute jointly funded and operated by 12 European countries, but generally it is known as the institute which developed WWW on the Internet. Its name has now been changed to Laboratoire European pour la Physique des Particules, but the abbreviation remains the same.

¹²¹ **Hypertext:** It is a structure in which pointers are placed within texts so that links can be made to jump from those pointers to other texts and pictures. To create a document in the hypertext format, one uses HTML (HyperText Markup Language). To identify a WWW server address, we can use URL (Uniform Resource Locator).

ARPANET,¹²² created by the United States Defense Department, set the foundations for the Internet.

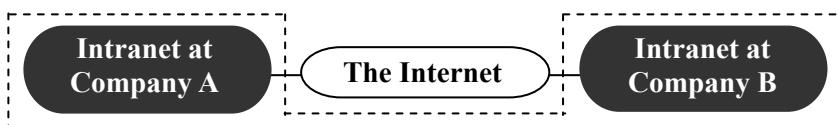
Intranets

An intranet is a company-wide network applying the technology of the Internet. Normally, a firewall¹²³ is set up between the Internet and a company-wide network in order to prevent leakage of confidential information of the company.



Extranets

An extranet is an intranet extended over numerous companies. In general, an extranet is built by connecting intranets to the Internet.



Mobile communication

Mobile communication¹²⁴ is an environment in which the network can be accessed from any location. Today, communication is the mainstream, so we can send and receive e-mails on the Internet and obtain a variety of information all with one telephone.

Satellite communication

Satellite communication is a wireless communications system using a communications satellite. A broadcasting station transmits (uplink) a huge amount of information to a stationary communications satellite located 36,000 km above the equator (in a stationary orbit), and the information is distributed all at once (downlink) to various receiving stations on the earth. A large amount of information can be transmitted to many points.

CATV

New types of services using CATV (Community Antenna TeleVision, or Cable TV) are being considered and discussed commercially, such as Internet connection, telephone services, experiments involving PHS (Personal Handy-phone System), and VOD (Video On Demand). CATV is expected to be a major part of the infrastructure in the multimedia era.¹²⁵

¹²² **ARPANET (Advanced Research Project Agency Network):** This is a nationwide computer network developed under the sponsorship of the Advanced Research Project Agency (ARPA) of the United States Department of Defense. It is the predecessor of the Internet.

¹²³ **Firewall:** It is the mechanism which is located between the Internet and a company-wide intranet to manage data communication and to protect the internal network from external attacks and invalid access. The word could also refer to this functional role.

¹²⁴ **Mobile/mobile computing:** "Mobile" refers to any information device that can be carried around, including cell phones, PHSs (personal handy-phone system), and notebook PCs. "Mobile computing" refers to the mode of using any of these information devices to have access to the company network from the outside.

¹²⁵ (Note) CATV began as a reparation facility in remote areas and a community facility in rural regions. Today, urban CATV, which can provide broadcast services on many channels, is getting attention as a new-generation component of the infrastructure. Coaxial cables are used for distribution so that high-quality images can be received.

◆ Application Systems

A network application system is a social system using a network. Specifically it includes the following:

Internet shopping

This is a system in which the user can shop at a virtual store set up on web pages. To make a payment, the shopper can use his or her credit card or go to a nearby convenience store to pay.

Groupware

This is software for communication within an organization or for information sharing. It has functions such as electronic mail, schedule sharing, document sharing, and workflow.

Debit cards

This is a service whereby a cash card issued by a bank can be used to make payments. The money for the payment is directly withdrawn from the bank account in real time.¹²⁶

¹²⁶ Debit cards have been traditionally called bank POS; cash cards are used instead of cash.

2.5.2 Database Applications

Points

- An example of a database application is a data warehouse.
- Applications in business include corporate accounting, inventory management, document management, and sales support.

One type of database application system is a data warehouse. Application systems in which databases are applied in business include corporate accounting systems,¹²⁷ inventory management systems,¹²⁸ document management systems,¹²⁹ and sales support systems.¹³⁰

◆ Data Warehouse

A data warehouse is a company-wide database to support decision-making. The idea is to have a large amount of data stored, organized, and used to help make business decisions. Sometimes it is called an informational database.

◆ Data Mining

This refers to a technology or method of drawing out tendencies, trends, correlations, and patterns necessary for management and marketing, through dialogues with a large amount of raw data.

Whereas a data warehouse normally analyzes various data based on some hypothesis, data mining discovers trends and patterns in order to establish the hypothesis.

◆ Data Mart

A data mart is a database which stores data obtained from a data warehouse. The data stored in a data mart, is selected and summarized according to the purposes of a specific user group. Whereas a data warehouse contains information for the entire company, a data mart has a relatively small amount of data tailored for the target users.

¹²⁷ **Corporate accounting system:** It is a system in which the accounting procedures of a corporation are computerized in an attempt to make the accounting tasks more efficient and quicker and to obtain timely understanding of the business and managerial records.

¹²⁸ **Inventory management system:** It is a system to keep the production (purchase) and demand in balance, managing the inventory such as products and raw materials kept by the company at an optimum amount. In a retail store such as a supermarket, the sales information entered at POS terminals is collected and analyzed so that the demands can be predicted and more products are automatically ordered, taking into account safe inventory volumes and optimum amounts to purchase.

¹²⁹ **Document management system:** It is a system in which a corporation manages various types of documents and sources; document search is possible from a variety of fields such as the storage location or contents of the document. It is an attempt to make document management and document preparation more efficient, e.g. to avoid duplicate preparation of the same document.

¹³⁰ **Sales support system:** It is a system that supports making sales plans and business plans, based on accumulated sales information.

◆ OLAP

OLAP (OnLine Analytical Processing) is the concept of analytical application in which the end user discovers problems and solutions by directly searching and organizing a database; the goal is to achieve quick data access and to provide a function for easy analysis.¹³¹

◆ OLTP

OLTP (OnLine Transaction Process) is the processing mode in which messages are sent to the host computer from multiple terminals connected online to the host computer, which, according to the message received, in turn performs the process including access to a series of databases and returns the process results immediately to the terminals.

Databases used by OLTP are called business databases or, sometimes, operational databases. These are terms in contrast with informational databases.

Below, we compare informational databases with operational databases:

Item for comparison	Informational databases	Operational databases
Target task	Supporting decision-making	General business tasks
Data addition	Yes	Yes
Data updating	Generally no	Generally yes
Processing mode	OLAP	OLTP
Main users	Management staff	General workers
Business type	Non-standard tasks ¹³²	Standard tasks
Period of data retention	Long term	Short term

◆ Application Systems

Various systems that use databases are developed. Today, it is not an overstatement to say that most of the systems in operation use databases. Some of the great advantages for using databases are as follows:

- Data can be easily accumulated.
- Data can be managed in an integrated manner.
- Data can be easily processed.
- Data can be easily searched.

¹³¹ (FAQ) There have been exam questions concerning data warehouses. Know accurately the meanings of data warehouses, OLAP, and OLTP.

¹³² **Standard/non-standard tasks:** Standard tasks are those for which processing procedures are fixed, such as daily business procedures and daily input of sales data. Non-standard tasks are those for which processing procedures vary case by case. Creation of analysis documents, for instance, requires different processes depending on the purpose of use, so it is considered non-standard.

2.5.3 Multimedia Systems

Points	<ul style="list-style-type: none"> ➤ Examples of multimedia usage include AI, 3DCG, and pattern recognition. ➤ Multimedia application systems include Internet broadcasting and VOD (Video On Demand).
---------------	--

Multimedia refers to handling not just characters and text but also mixtures of still images, moving pictures, audio, and other communication media. The term also refers to devices and software used in multimedia communication.

◆ Artificial Intelligence (AI)

Artificial intelligence refers to a system that performs inference processes such as an expert system¹³³ and a machine translation system. A typical computer does no more than manipulating numerical values and performing logical operations. AI, on the other hand, performs inference processes centered on manipulating character strings. It is said that Lisp and Prolog are languages suitable for developing software using AI technologies.

◆ 3-Dimensional Graphics (3D Computer Graphics, or 3DCG)

3D computer graphics¹³⁴ involves creating virtual 3D space inside the computer, placing solid models in this space, and moving them around. It is embedded in movies, games, and animations. VRML¹³⁵ is used to develop 3DCG.

The following table compares artificial reality and virtual reality.

Type	Explanations
Artificial reality (AR)	<p>It is the technology of creating a virtual world inside the computer, with a sense of reality. <u>Special equipment is not necessary to experience the artificial reality.</u></p>
Virtual reality (VR)	<p>It is the technology of creating a fictitious world and having people experience and feel that world as though it were real. 3D vision using dedicated display units and special input equipment are used. Examples: pre-experience of virtual surgery, flight simulator, etc.</p>

¹³³ **Expert system:** It is a system created with the knowledge base of various specialists (experts) in a variety of fields; given certain conditions, the system applies the knowledge based on certain rules so that problems can be solved as if they were solved by the experts.

¹³⁴ **Computer graphics (CG):** It is the technology of creating images via computers, or images made by such technology. There are methods where the computer processes already existing images, and there are other methods where the computer creates images themselves. The latter method is called CGI (computer-generated images).

¹³⁵ **VRML (virtual reality modeling language):** language specifications to describe 3DCG used on the Internet.

◆ Multimedia Application Systems

There are many systems that apply multimedia. Now that WWW has brought in a graphical environment through the Internet, a variety of multimedia contents are available for use.

Internet broadcasting

This is broadcasting using multimedia on the Internet. With the use of streaming distribution technology,¹³⁶ programs are broadcasted in real-time on the Internet. Compared with conventional broadcasting business, equipment costs much less, and global information transmission is possible. In addition, on-demand service¹³⁷ can also be provided, so we can tune in whenever we wish to watch a particular program.

Internet broadcasting comes in various formats. On-demand broadcasting stores the contents on a server and distributes them per request from a user. Live broadcasting (Internet live) distributes live programs, such as concerts, simultaneously to multiple users.

Non-linear editing

This is a method of video editing where images are digitized and video is produced in free order using a computer. It is easy to correct images, switch the order in which they appear, and create a different version. Incidentally, the conventional method in which video images are dubbed in the order of their completion is called linear editing.

Video on demand

This refers to the service of instantly sending a video program requested by the viewer via, for example, bidirectional CATV. The service provider stores many video programs on its video server and distributes the one requested by the viewer.

A video server can respond to simultaneous access by many viewers, and programs are requested to be sent from the beginning. Hence, it is necessary to construct an image database and connect it to individual mobile terminals and household receivers via broadband communication lines such as cable, wireless, etc.

¹³⁶ **Streaming:** It is the technology of reading data and playing the data back immediately. It enables Internet broadcasting and playback of contents without waiting time. For streaming distribution, the line speed must exceed the amount of data; however, Internet lines are generally slow, so normally the data is compressed to enable real-time transmission. Conventionally, playback used to be time-consuming since the data had to be downloaded first and then played back. With streaming, however, playback is done while data is being received.

¹³⁷ **On-demand:** It is a function to provide what is requested whenever requested.

Quiz

Q1 What is an intranet?

Q2 What is a data warehouse?

Question 1

Q1. There is a system which manages the file area in units of blocks. Each block contains eight sectors, and one sector is 500 bytes. How many sectors in total would be assigned to store two files, one consisting of 2,000 bytes and the other of 9,000 bytes? Here, the sectors occupied by management information, such as directories, can be ignored.

- a) 22 b) 26 c) 28 d) 32

Answer 1

Correct Answer: **d**

Files are saved in units of 8 sectors. Eight sectors, as shown below, are 4,000 bytes.

$$8 * 500 = 4,000 \text{ (bytes)}$$

Hence, if one block is less than 4,000 bytes, all 4,000 bytes are used.

Next, we find the number of sectors necessary for each of the 2,000-byte file and 9,000-byte files.

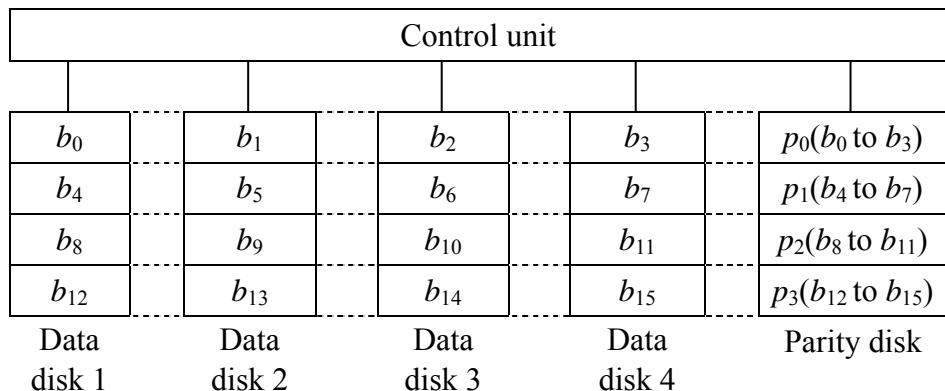
Capacity required for the 2,000-byte file	= $2,000 / 4,000 = 0.5 \text{ (block)}$
	→ 1 block (= 8 sectors) required
Capacity required for the 9,000-byte file	= $9,000 / 4,000 = 2.25 \text{ (blocks)}$
	→ 3 blocks (= 24 sectors) required

Hence, to save files of 2,000 bytes and 9,000 bytes, the total number of sectors allocated to the two files is 32 as shown below.

$$\text{Number of sectors required} = 8 + 24 = 32 \text{ (sectors)}$$

Question 2

- Q2.** Which of the following is the appropriate term for the process of breaking down data and storing it on multiple hard disks, as shown in the figure below? Here, b_0 to b_{15} represent the sequence in which data is stored on the data disk in units of bits, and p_0 to p_3 represent the parity used to identify disk failure.



- a) Striping
- b) Disk caching
- c) Blocking
- d) Mirroring

Answer 2

Correct Answer: a

Striping is to distribute one block of data onto two or more disks and write simultaneously. By striping, each block can be read and written in parallel, so the input/output speed increases.

Striping is defined as a technology for RAID. As shown in the figure above, this configuration contains a disk dedicated to the parity; such a configuration is called RAID2, RAID3, or RAID4.

- b) Disk cache is placed between a hard disk and the main memory; it is a buffer (buffer memory) to improve the apparent speed of the hard disk.
- c) Blocking is to handle each logical set of multiple records as one physical record (block).
- d) Mirroring is to prepare multiple disks and write the same data onto separate disks simultaneously, i.e., a multi-disk configuration. If one of the disks fails, the operation continues with the remaining disks only. This configuration is called RAID1.

Question 3

- Q3.** Which of the following is arranged in the order of the effective memory access speeds from fastest to slowest?

Cache memory			Main memory
With cache or w/o cache?	Access time (ns)	Hit rate (%)	Access time (ns)
A w/o cache	–	–	15
B w/o cache	–	–	30
C with cache	20	60	70
D with cache	10	90	80

Answer 3

Correct Answer: b

Cache memory (buffer memory) is memory which is placed between the CPU and the main memory to adjust speed differences between the two. The effective access speed can be increased by adding high-speed buffer memory, and by reading and writing on this cache memory as much as possible.

Let t_c be the access speed of the cache memory, t_m be the access speed of the main memory, and h be the hit ratio. The effective memory access speed is then calculated as follows:

Effective memory access speed = $t_c * h + t_m * (1 - h)$

The hit ratio is the probability that the data to be read is in the cache memory. The higher the hit ratio is, the faster the effective memory access speed becomes.

For A through D, we need to calculate the effective memory access time. For A and B, there is no cache memory, so the access time of the main memory is the effective memory access time. Here, we can consider $h = 0$. The numbers below indicate the order of each, from fastest to slowest (from the shortest effective memory access time to the longest).

$$\begin{aligned} A: 15 \text{ (ns)} & \quad (1) \\ B: 30 \text{ (ns)} & \quad (3) \\ C: 0.6 * 20 + (1 - 0.6) * 70 = 40 \text{ (ns)} & \quad (4) \\ D: 0.9 * 10 + (1 - 0.9) * 80 = 17 \text{ (ns)} & \quad (2) \end{aligned}$$

Hence, if we arrange Memory A through D from fastest to slowest in terms of the effective memory access time, the order is “A, D, B, C.”

Question 4

- Q4.** When a certain file was copied from one directory to another on a hard disk in a PC, file fragmentation occurred. Which of the following is an appropriate description concerning this situation?
- a) The fragmentation can be eliminated by performing a physical dump of the entire disk and then restoring the disk.
 - b) Access time will be longer even for some files other than the file in which the fragmentation occurred.
 - c) If the file in which the fragmentation occurred is copied again, the fragmentation in the copy destination may get worse, but it will never be eliminated.
 - d) Even if fragmentation has occurred, the size of the file is still the same as that of the original one.

Answer 4

Correct Answer: d

Fragmentation means that this file saved on the hard disk could not secure one continuous area and thus is saved across multiple blocks. When fragmentation occurs, various parts of the hard disk must be accessed, reducing the processing efficiency. However, the file size does not change, as the file is simply saved in a divided manner.

- a) If the disk is physically copied, the situation does not change. It must be copied logically.
- b) One file was copied on the hard disk on which files had already been saved, so no other file except the one that was copied was affected. Physically nothing was changed.
- c) The fragmentation will be solved if the copy destination has a continuous empty area whose size is larger than the file size.

Question 5

- Q5.** The table shown below gives processing times for a CPU and I/O devices to execute 5 stand-alone tasks. Which task can be executed simultaneously with the “High” priority task so that the CPU idle time from task execution start to end can be zero? Here, each task uses a different I/O device and is performed concurrently. The overhead of the OS can be ignored.

Unit: ms		
Priority	Stand-alone task processing time	
High	CPU (3) → I/O (3) → CPU (3) → I/O (3) → CPU (2)	
a)	Low	CPU (2) → I/O (5) → CPU (2) → I/O (2) → CPU (3)
b)	Low	CPU (3) → I/O (2) → CPU (2) → I/O (3) → CPU (2)
c)	Low	CPU (3) → I/O (2) → CPU (3) → I/O (1) → CPU (4)
d)	Low	CPU (3) → I/O (4) → CPU (2) → I/O (5) → CPU (2)

Answer 5

Correct Answer: c

We check the operation of each “low-priority” task to see which one uses the CPU while the “high-priority” task is not using the CPU. The input/output units are different, so there is no waiting for the input/output (I/O) units.

The “high-priority” task uses I/O units twice, each for 3 ms. If the use of the CPU by the “low-priority” task takes exactly 3 ms, the CPU has no idle time. With this said, let us now consider each task listed in the answer group.

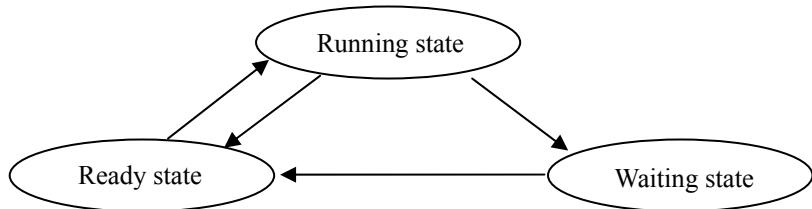
For instance, consider Task (a). During the 3-ms period when the “high-priority” task uses an I/O unit for the first time, the “low-priority” task uses the CPU for 2 ms. Hence, 1 ms of CPU idle time will result.

Similarly, any “low-priority” tasks like (b) and (d), with 2 ms of CPU use, will cause CPU idle time. In contrast, Task (c) has 4 ms of CPU use, but this comes in the end, so no idle time will occur if this is after the completion of the “high-priority” task.

Hence, the “low-priority” task that can completely eliminate the CPU idle time until the execution of both tasks is completed is “CPU (3), I/O (2), CPU (3), I/O (1), CPU (4).”

Question 6

- Q6.** The state transition diagram below shows a task (process) state transition on a multitasking computer. When does the task state change from the running state to the ready state?

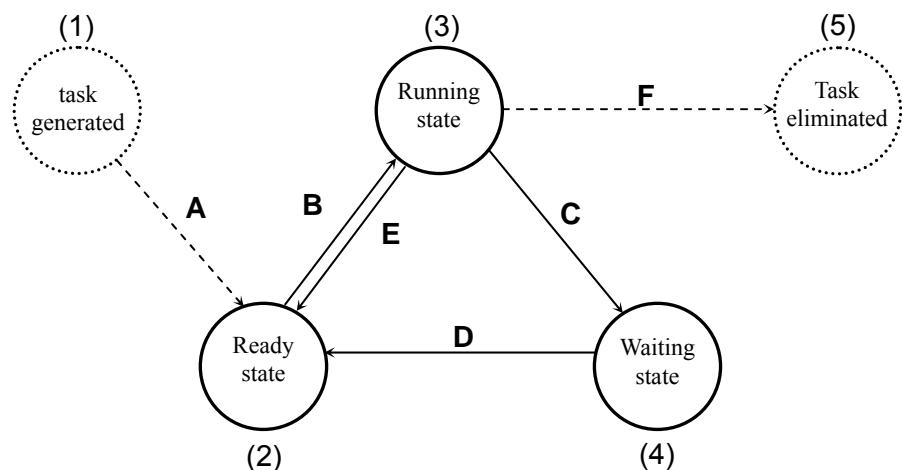


- a) A task with higher priority compared to its own has moved to the ready state.
- b) A task has been generated by the job scheduler.
- c) An I/O operation has completed.
- d) An I/O operation has been requested.

Answer 6

Correct Answer: a

When a task (process) is generated, it proceeds to the ready state and is entered into the queue. After that, it goes into the running state depending on the task priority and goes into the waiting state when an I/O operation occurs. Such changes in the task state are called the state transitions of the task.



We now explain the state transitions of tasks. The numbers in parentheses () are for state explanations, and A through D are for transition explanations.

No.	State	Explanation
(1)	Task generated	A task generated is registered into the task queue.
(2)	Ready	Waiting to be authorized to use the CPU
(3)	Running	Authorized to use the CPU and is currently being executed
(4)	Waiting	Waiting for completion of an event such as an I/O operation
(5)	Task eliminated	A task is completed and no longer necessary; it is removed from the task queue.

Transition	Explanation
A	A task is generated; moved to the ready state
B	By priority, moved to the running state
C	Moved to the waiting state (waiting for I/O, etc.) due to an event
D	Moved to the ready state after completion of an event
E	Moved to the ready state due to another task with high priority
F	The task in the running state is completed.

As you can see from these tables, transition from the running state to the ready state is transition E. This is when a task whose priority is higher than the task being executed goes into the ready state.

- b) This describes transition A in the state transition figure for the tasks.
- c) An I/O operation is an operation of input or output. When this I/O operation is completed, an I/O interruption occurs and the task moves into the ready state. This describes transition D in the state transition figure for the tasks.
- d) An I/O operation is requested when the task issues an instruction for input or output. If this happens, a supervisor call occurs, and the task moves from the running state to the waiting state to wait for the completion of the I/O. This is transition C in the state transition figure for the tasks.

Question 7

- Q7.** Which of the following is an appropriate statement concerning a client/server system?
- a) The client and the server must use the same kind of OS.
 - b) The server sends data processing requests and the client processes those requests.
 - c) A server can support a client function that enables it to request processing of another server if necessary.
 - d) The server functions must be allocated to different computers, such as a file server and print server.

Answer 7

Correct Answer: c

A client server system is a system made up of processing units called clients and servers. A client performs data input/output and other processes through a server while a server controls all input and output that depend on the hardware according to its type. Normally, a client unit is a unit equipped with data processing functions such as a personal computer or a workstation, so it can perform applications on its own. In fact, it also performs processes only a client can perform, such as displaying text and drawing figures. A server, on the other hand, accesses databases and performs printing processes in response to requests by clients. Further, if a server is not able to perform a process, it can request another server to do that. Here, the first server becomes a client because it is requesting another to perform a process.

- a) Different operating systems do not cause any problems as long as the protocol is established. We can have a combination of servers with UNIX and clients with Windows.
- b) It is a client that sends requests for processing. It is a server that performs the requested processes.
- d) In a small-scale system, a server and a client can even be the same. X Windows of UNIX is an example of this type. If the clients and servers are built on the same platform (OS) and can be connected via a network, there is no inhibitory effect.

Question 8

- Q8.** When comparing a distributed processing system, which consists of multiple computer systems located in a wide area, with centralized processing systems that operate in a single center, which of the following is the most appropriate feature of centralized processing systems?
- In the event of a disaster or a failure, recovery work can be conducted in a centralized fashion in the center, avoiding the risk of a long shutdown of the entire system.
 - Since the system is collectively managed, it is easy to satisfy requests for additions or changes to system functions, and the accumulation of backlog seldom occurs.
 - Data consistency is easily maintained and managed through the centralized implementation of measures in the center.
 - Although the operation and management of hardware and software resources become complex, expansion taking advantage of new technologies is easy.

Answer 8

Correct Answer: c

A centralized processing system is the idea of centralizing processes to a general-purpose large computer (host computer). Since the equipment is centralized at one location, the management is easy. However, if the host computer fails, it is possible that the system gets shut down for a long period of time.

- Since the host computer performs all processing, the system gets shut down until the host computer is recovered. If this is a serious failure, it is possible that the downtime of the system becomes long.
- In a centralized processing system, all requests must be answered by the host computer alone. If the contents of the requested items vary significantly in level, the host computer cannot meet all those requests, causing backlog accumulation. Incidentally, backlog can also refer to a system waiting to be developed.
- A centralized processing system processes all tasks at the host computer, so it is cumbersome to respond to each type of business task separately. Even if we want to extend the system, often there are tasks that cannot be suspended. With the introduction of a new technology, some tasks may not be able to be processed any longer.

Question 9

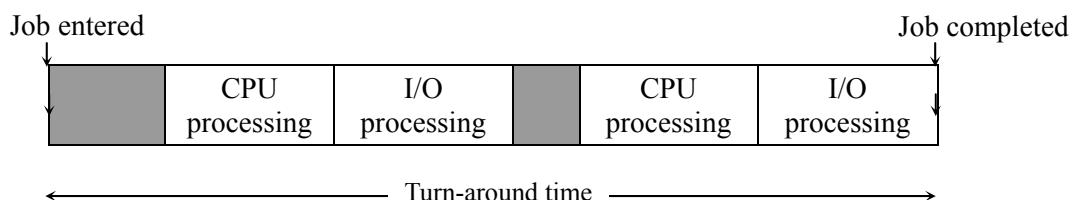
- Q9.** For one job, which of the following formulas appropriately expresses the relationship between turnaround time, CPU time, I/O time, and process waiting time? Here, other types of overhead time are ignored.
- Process waiting time = CPU time + Turnaround time + I/O time
 - Process waiting time = CPU time - Turnaround time + I/O time
 - Process waiting time = Turnaround time - CPU time - I/O time
 - Process waiting time = I/O time - CPU time - Turnaround time

Answer 9

Correct Answer: c

Process waiting time is the time until the start of a CPU process or an I/O process of the job entered into the computer system. Turn-around time (TAT) is the time interval from submitting a job to receipt of the results. This concept is mainly used in batch processing. TAT includes both the CPU processing time and I/O time. Hence, process waiting time is obtained by subtracting CPU processing time and I/O processing time from TAT.

In the figure below, the shaded areas represent process waiting time.

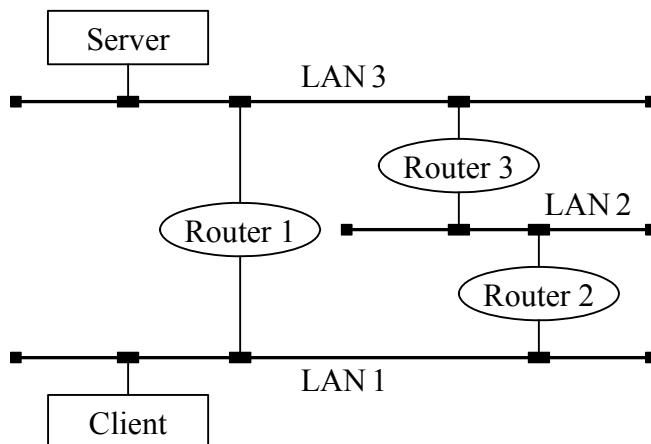


Hence, for one job, the formula expressing the relationship among turn-around time, CPU time, I/O time, and process waiting time is as follows:

Process waiting time = Turn-around time - CPU time - I/O time.

Question 10

Q10. LAN facilities are installed as shown in the figure below. Using the server connected to LAN3, the client on LAN1 is performing a business application. Data transmission is normally performed via Router 1. If a failure occurs in Router 1, Routers 2 and 3 are used for transmission between LAN 1 and LAN 3. What is the availability of the LAN equipment connecting LAN1 and LAN3? Here, the failure rate of each router is 0.1, no switch-over time is required in case of a failure, and failures in LAN facilities other than the routers are not taken into account.

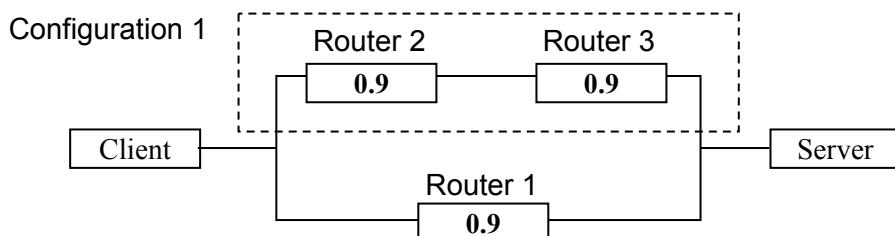


- a) 0.729 b) 0.981 c) 0.990 d) 1.000

Answer 10

Correct Answer: b

Note that “Router 1” and “the serial connection of Routers 2 and 3” are in parallel connection. Since the failure rate of each router is 0.1, the availability is 0.9. Hence, the configuration is as shown below. The values in the boxes are availability of each unit.



$$\text{Availability of Configuration 1} = 0.9 * 0.9 = 0.81$$

$$\text{Availability of the entire system} = 1 - (1 - 0.81) * (1 - 0.9) = 0.981.$$

Question 11

Q11. Which of the following is an appropriate statement concerning VR (Virtual Reality)?

- a) Using technology such as CG, VR expresses the world created inside a computer as if it were the real world.
- b) For the purpose of improving GUI, it does not display an image incrementally from the top, but first displays a rough mosaic-like image and gradually sharpens it.
- c) VR tests whether or not hypothetical results can be obtained from computer simulations such as those of wind tunnel tests used for automobile or aircraft design.
- d) VR makes abilities such as human recognition and inference possible on a computer.

Answer 11

Correct Answer: a

Virtual reality (VR) is to artificially create a sense of reality by combining CG (Computer Graphics) and sound effects. To appeal to the senses, we may also use dedicated display units such as a head-mounted display (HMD) for 3D vision and special input units such as data gloves. In addition, the response to the images is returned to the user, further enhancing the sense of reality.

- b) This describes interlace.
- c) This describes simulation.
- d) This describes AI (Artificial Intelligence).

3 System Development

Chapter Objectives

System development means the creation of software to operate computers. In general, this is performed in the order of requirement analysis, external design, internal design, programming, and testing, but various methodologies have been proposed, depending on the situation of system development. In Section 1, we will learn the methodologies of system development as well as programming languages, groups of tools, and evaluation of software quality, all of which support system development. In Section 2, we will learn specific procedures of system development and methods of testing.

3.1 Methods of System Development

3.2 Tasks of System Development Processes

[Terms and Concepts to Understand]

Programming language, compiler, subroutine, recursion, reentrant, CASE, ERP, waterfall model, prototyping model, function point method, DFD, E-R diagram, review, white box test, black box test, independence of modules

3.1 Methods of System Development

Introduction

To develop systems, we need to know the methodologies of system development. The methodologies can be classified into process models and cost models. A process model is a method of development procedures while a cost model is a method of estimating the cost. To apply these methods, we must first know the system development environment. A system development environment is a group of tools that support system development and includes programming languages and CASE.

3.1.1 Programming Languages

Points

- Types of programming languages include procedural, functional, logic, and object-oriented types.
- Typical languages include COBOL, C, Java, and SGML.

A programming language is a language that describes the processes (programs) we want computers to perform. We select appropriate programming languages depending on the application.

◆ Classification of Programming Languages

Here is a way to classify programming languages and some typical languages.

Type	Characteristics	Programming languages
Procedural ¹	The procedures are expressed as specific algorithms. Each procedure to be executed by the computer is written, one instruction at a time.	COBOL, C, Fortran, Pascal, etc.
Functional	Process steps are expressed by composition of basic functions (list processing)	Lisp, etc.
Logic	Relations are defined by basic logic formulas (inferential processing)	Prolog, etc.
Object- oriented	Operation is conducted by objects which integrate the data and their processing.	Java, C++, Smalltalk, etc.

¹ **Non-procedural:** Sometimes programming languages that are not procedural are called non-procedural programming languages. They are characterized by the property that the order in which instructions are written in the program does not match the order of execution. Generally, parameters are given, and the processes are executed according to the contents of the parameter definitions.

◆ Programming Languages

The characteristics of commonly used programming languages are organized below.

Procedural/functional/logic/object-oriented

Language	Characteristics
COBOL	A business-processing language The language specifications were established by CODASYL.
C	Developed by AT&T ² to write OS for UNIX ³ Allows easy portability
Fortran	Developed by IBM as a computing language for science and technology
Pascal	A structured programming language developed for the purpose of teaching students
Lisp	A list-processing language developed at MIT ⁴ Used for research in artificial intelligence, etc.
Prolog	A language with an inferential mechanism Developed at the University of Marseille in France
C++	An object-oriented language and an extension of C Completely upward-compatible with C
Java	Developed by Sun Microsystems, based on C++ Runs on any OS
Smalltalk	Developed by Xerox at its Palo Alto laboratory Dialogue-type and programmable

Markup languages (document-formatting languages)

These are languages where layout information, font size, formats, and other specifications are directly embedded for display on the screen or for printing. Inserting symbols (tags) such as <TITLE> and </TITLE> in a paragraph is called marking up (tagging). The following table shows main markup languages.

Language	Characteristics
SGML	Standard Generalized Markup Language Logical structure and semantic structure of documents are described with simple marks.
HTML	HyperText Markup Language This is the language that is used in creating Web pages on the Internet.
XML	eXtensible Markup Language This is an extension of HTML hyperlink function, extended so that SGML can be sent and received via a network. Tags are not fixed but can be defined arbitrarily.

Other programming languages

The following table shows other programming languages.

Language	Characteristics
PostScript	A page description language ⁵ developed by Adobe Systems of the U.S.
Visual Basic	A programming language for Windows, developed by Microsoft of the U.S.
Perl	A script language that describes access counters and CGI ⁶ of Web pages.

² AT&T: American Telephone and Telegraph, a telecommunications company, oldest in the world and largest in the United States.

³ (Note) C is a language developed to write an operating system for UNIX, but since it is so easy to use, today a wide range of programs are written in it, including business applications and operating systems.

⁴ MIT: Massachusetts Institute of Technology.

⁵ Page description language: It is a language used to define printing image for the printer when printing a document using a page printer. Identical images can be printed even if printers have different resolutions.

⁶ CGI (Common Gateway Interface): It is a mechanism that takes requests from a WWW browser, calls an external program requested, and returns the execution results to the WWW browser.

◆ Script Languages

A script language is a language that uses text (characters) to describe procedures to be executed by the computer. The processing procedures described by a script language are called **scripts**. Mainly these are in database software and spreadsheet software used as macros. In the sense that these languages describe procedures, they are similar to procedural programming languages; however, scripts are characterized as being **event-driven**.⁷ Also, often a development environment using GUI is provided so that the end user can easily write programs.⁸

3.1.2 Program Structures and Subroutines

Points

- Program structures include reentrant, reusable, and recursive types.
- Subroutines can be open subroutines or closed subroutines.

Processes frequently used in a program or processes shared by multiple programs are set aside as separate programs and are shared among many programs. Such a shared program is called a subroutine (subprogram), and a variety of structures are used according to the conditions of use.

◆ Program Structures

According to the structure, programs can be classified as shown below.

Program structures	Recursive	Structure that calls itself
	Reusable	Can be used repeatedly without reloading
	Reentrant	Multiple tasks ⁹ can use the program at the same time
	Serially reusable	Multiple tasks can use the program serially
	Non-reusable	Must be reloaded for each use

Recursive

A procedure is said to be **recursive** if the definition of that procedure refers to the procedure itself. A program in which the definition of a subroutine or a function uses the subroutine or the function itself is called a **recursive program**. Such a reference within itself is known as a **recursive call**. It can be used in most programming languages, but COBOL and Fortran are exceptions.

⁷ **Event-driven:** It is a program that is triggered by an event and starts up to respond to and process the event. An event is any conditional change, such as a press on the keyboard. Programs that start up when the user clicks on an icon are event-driven.

⁸ (FAQ) There are exam questions on combinations of common languages and their classification. For instance, know that COBOL is procedural, Lisp is functional, and Java is object-oriented.

⁹ **Task:** It is a processing unit obtained when program processes are minutely divided

Reusable

This term refers to the program structure that allows multiple programs (tasks) to share the use of the program without reloading the program into main memory each time. If the program can be used simultaneously by multiple tasks, it is called **reentrant**,¹⁰ otherwise, it is called **serially reusable**. A program with a structure to allow reentry is called a **reentrant program**.

◆ Subroutines (Subprograms)

A subroutine refers to a part of a program which is repeatedly used within the program to execute common procedures. If multiple programs execute the same procedures, those procedures can be combined as one program so that the multiple programs can share their use. Such a program is also called a subroutine.

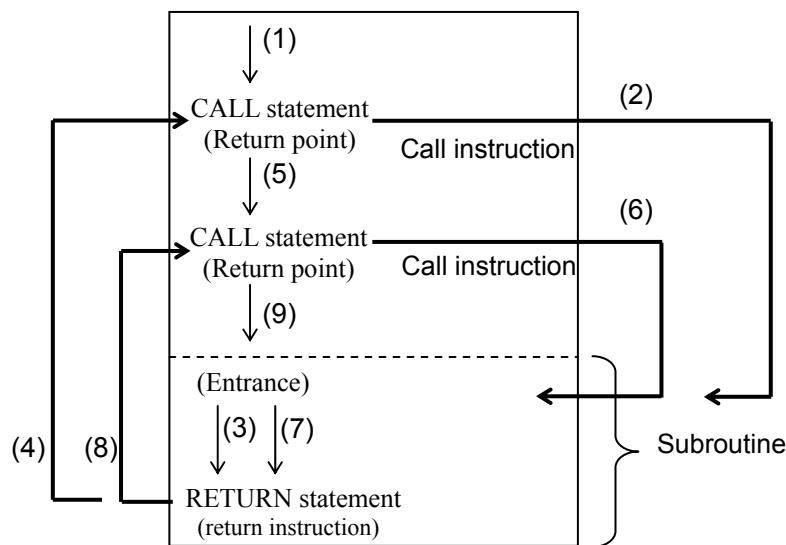
Open subroutine

An open subroutine¹¹ is a subroutine that is embedded wherever a program needs it as many times as the program needs it.

Closed subroutine

A closed subroutine is created independently of programs that need it as a subroutine. If a program needs the subroutine, it executes a subroutine-call instruction (usually a CALL statement) to deliver the control to the subroutine.

The figure below illustrates the concept of a closed subroutine. The processes are executed in the order (1), (2), (3),... By the CALL statement, the program jumps to the entrance of the subroutine, and by the RETURN statement, it returns to the instruction following the CALL statement (return point).¹²



¹⁰ (Note) In a reentrant program, the unchangeable parts (mainly procedural parts) and changeable parts (mainly data) are separated so that multiple programs can use it at the same time by sharing the use of the unchangeable parts while securing only the changeable parts according to the programs that call the reentrant program. In general, most online-processing programs have the reentrant structure.

¹¹ **Open subroutine:** It can be implemented as a macro in assembler language, a copy library in COBOL, and "%include" in C.

¹² (FAQ) With regard to program structure, many recent exam questions have involved recursion and reentry. Recursion is calling itself, and reentry is being simultaneously called up by multiple programs.

3.1.3 Language Processors

Points

- Language processors include compilers, interpreters, etc.
- Load modules are generated by linkage editors.

A programming language uses expressions similar to daily language so that programs can be easily written. However, computers cannot understand the instructions of any programming language as is. Hence, it is necessary to convert those programs written in programming languages into a format that computers can understand. This conversion is performed by what is called language processors.¹³

◆ Language Processors

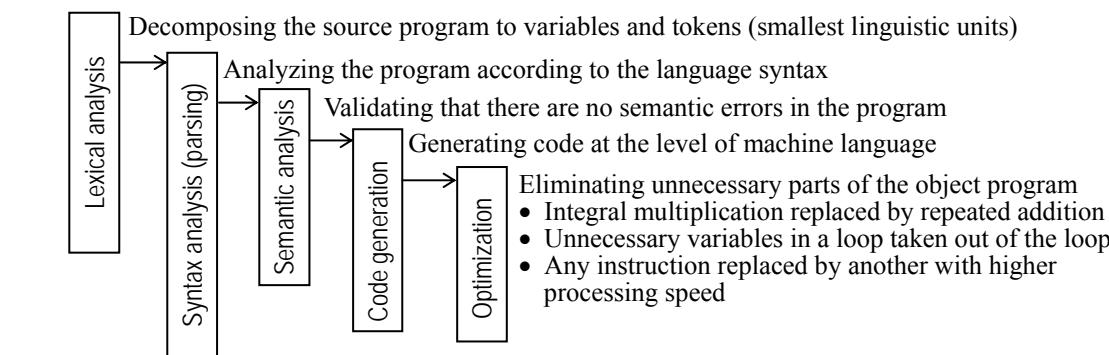
A language processor is a program that translates (converts) source programs to machine language. Language processors are as follows:

Language processors	<table border="0"> <tr> <td style="border: 1px solid black; padding: 2px;">Assembler</td><td>Translates assembler language to machine language</td></tr> <tr> <td style="border: 1px solid black; padding: 2px;">Compiler</td><td>Translates compiler language¹⁴ to machine language. (COBOL, Fortran, C, etc.)</td></tr> <tr> <td style="border: 1px solid black; padding: 2px;">Generator</td><td>Creates programs by giving parameters. (RPG, etc.)</td></tr> <tr> <td style="border: 1px solid black; padding: 2px;">Interpreter</td><td>Executes while translating instructions. (BASIC, APL, LOGO, etc.)</td></tr> </table>	Assembler	Translates assembler language to machine language	Compiler	Translates compiler language ¹⁴ to machine language. (COBOL, Fortran, C, etc.)	Generator	Creates programs by giving parameters. (RPG, etc.)	Interpreter	Executes while translating instructions. (BASIC, APL, LOGO, etc.)
Assembler	Translates assembler language to machine language								
Compiler	Translates compiler language ¹⁴ to machine language. (COBOL, Fortran, C, etc.)								
Generator	Creates programs by giving parameters. (RPG, etc.)								
Interpreter	Executes while translating instructions. (BASIC, APL, LOGO, etc.)								

In addition, there are also preprocessors,¹⁵ which convert source programs to a compiler language, not to machine language.

◆ Procedures of Compiler

A compiler language is translated to machine language in the order below. A program translated into machine language is called an object program (also an object module).



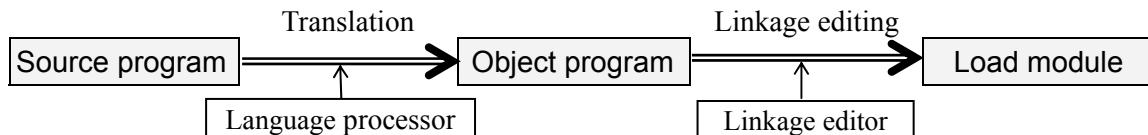
¹³ (FAQ) There have been many exam questions where you are to select characteristics of interpreters and compilers. Be sure to completely understand the characteristics of each. Questions concerning the procedures of a compiler have also been frequently asked on the exams.

¹⁴ **Compiler language:** It is a programming language that generates object programs from source programs using a compiler. It is also called a higher-level language and includes COBOL, Fortran, Pascal, PL/I, and C. A compiler language uses expressions similar to what humans use in daily living, so they are easy to understand and easy to learn.

¹⁵ **Preprocessor:** It is a program that takes source programs before they are translated by a compiler to machine language and makes them execute various processes. For example, a preprocessor for the C language supports functions such as defining numerical values found in the source programs as character strings and obtaining library files referenced by the source programs. They are designated by “include.”

◆ Creation of Load Modules

Computers can read and execute only machine language, so any program written in a language other than machine language must somehow be translated into machine language. One way to do this is to use a compiler.¹⁶



Load modules (executable programs) are the programs that can actually be executed. Object programs, which are simply translated by a language processor, cannot be executed. Through a linkage-editing program (linkage editor), what is required for execution needs to be added to the object program.

A **linkage editor**, in linking two or more object programs, calls function programs and subroutines used by the object programs from the software library and links them to the object programs. This is also called a **linker**.

Interpreters do not have object programs. Rather, they execute each instruction as they translate the instructions one by one. Generators directly create load modules by giving parameters.

◆ Execution of Program

To execute a program, it is necessary to store the program to be executed in the main memory or in a virtual memory. This function is performed by a loader.

A loader stores a load module in the main memory, and then the computer takes out one instruction at a time from the load module, interprets it, and executes it.

3.1.4 Development Environments and Software Packages

Points

- CASE tools and test support tools support system development.
- ERP is a software package designed to make business processes more efficient.

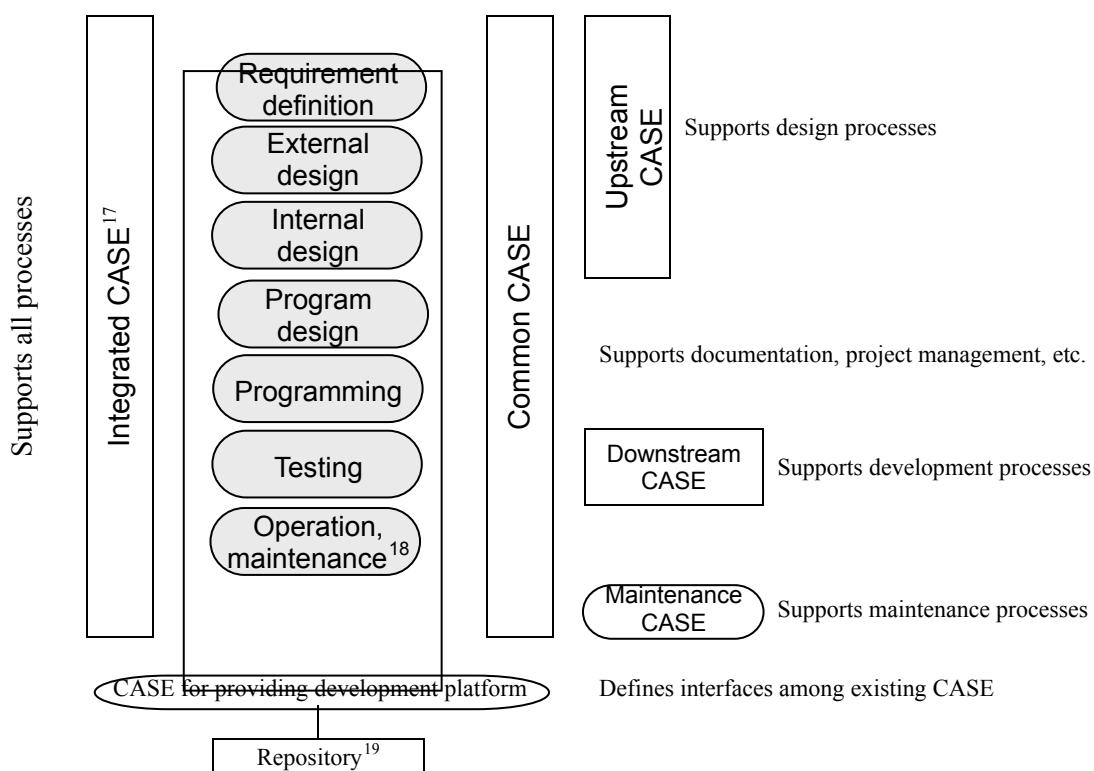
A development environment includes hardware necessary to build a system and software such as system development support tools.

¹⁶ (Note) Linking a subroutine during the creation of a load module is called a static link. In contrast, it is also possible to link a necessary subroutine when the program is executed. This method is called a dynamic link.

◆ CASE (Computer Aided Software Engineering) Tools

CASE is a group of software that supports system development and automation of maintenance work. CASE contains shared databases which store information necessary for development, such as requirements and design information for system development. It also performs consolidated management of the entire process of system development. In addition, the design results can be illustrated by easy-to-understand figures.

Specific CASE (Supports specific processes)



¹⁷ **Integrated CASE**: These are tools that support the entire system development process. Initially, the idea of integrated CASE was to have one CASE that covers all the processes; however, the reality was that partial CASE was in use, and the idea that it is better to use these existing tools became more popular. Therefore, integrated CASE is now developed as a means to provide interfaces between various tools so that design information can be communicated smoothly.

¹⁸ (Hints & Tips) Some common CASE tools have functions to support the entire development process. However, these are to be distinguished from integrated CASE. Common CASE manages areas besides design information, such as documentation (tables, graphs, figures) support, project management, and systems configuration management.

¹⁹ **Repository**: It is a database in CASE tools storing a variety of information, also known as a software engineering database or storage. By consolidated management of the design information using a repository, it is possible to check for consistency and completeness as well as to automate development processes.

◆ Test Support Tools

Test support tools analyze a program and monitor the operation of the program during execution. Main tools and their functions are shown in the following table.

Tools	Functions
Memory dump	Outputs the memory contents immediately after a program is abnormally terminated
Snapshot dump	Outputs the memory contents during program execution
Tracer	Outputs executed instructions and contents of the register at that time
Test data generator	Automatically generates various data for testing
Simulator	Simulates module functions that are not executable independently; online simulator, unit simulator, etc.
Miscellaneous	Driver stub tool, media conversion tool, inspector ²⁰ ²¹

◆ EUC (End User Computing)

EUC (End User Computing) refers to the idea that the end user himself or herself performs processes such as design, development, operation, and maintenance of information systems. Advancement in PC performance, lower costs, development of distributed processing based on networks, and popularity of package software have all contributed to the progress of EUC.

Traditionally, user departments used to request that development and operation management be done by information systems departments; however, the work of information systems departments has increased, making it difficult to provide individual services to end users. Then, within scopes limited to their departments and groups, the end users themselves began to perform the operation and development of systems. As the end users conduct their own operation, design, development, and integration, they can customize systems that fit their specific individual needs.

System integration and development by end users are called EUD (End User Development), but in practice the difference between EUC and EUD is not clearly identified.

◆ Software Packages

A **software package** is “general-purpose software that general users can commonly use.” Amid the various kinds of software packages, business packages have received greater attention recently, as they support efficient business processes.

ERP

ERP (Enterprise Resource Planning) is the concept of planning optimization of management resources through company-wide understanding of business information.²² Integrated (cross-sectoral) software to achieve this goal is called an ERP package. An ERP package is a software package integrating, in one database, all the common tasks regardless of the task type such as production management, accounting, sales management, personnel, and payroll.

²⁰ **Inspector:** It uses dialogues to force-change data and look up contents during the execution of a program.

²¹ (FAQ) Questions about test support tools do appear on exams. Know correctly the functions of snapshot dump and tracer.

²² (FAQ) Recently, questions on software packages have appeared frequently. In particular, questions on ERP stand out. Understand the characteristics of ERP. To implement ERP, a review of business processes is necessary.

CRM

CRM (Customer Relationship Management) is the concept that all departments which have contact opportunities with customers should share and manage customer information and contact history so that any questions from the customers can be answered appropriately. Companies attempt to promote the expansion of their customers by integrating all communication channels including telephone, fax, the Web, and e-mail, reinforcing the relationships with their customers and providing services that meet individual customer needs.

SFA

SFA (Sales Force Automation) is the basic concept of information systems that facilitate work restructuring of the entire sales activities that support corporate profits by using information technology. For example, more sales can be generated by managing previous contact records for each customer on computers. Moreover, transfer of work to new staff can be made more smoothly.

CTI

CTI (Computer Telephony Integration) is technology that provides a high level of telephone services by combining the information processing functions of computers and the communication functions of telephone switches.

3.1.5 Development Methods

Points	<ul style="list-style-type: none"> ➤ Process models include the waterfall, prototyping, and spiral models. ➤ Cost models include the COCOMO model and the FP method.
---------------	--

A **process model** is a model of system development method seen from the perspective of the processes involved; a **cost model** is a model from the perspective of the costs.

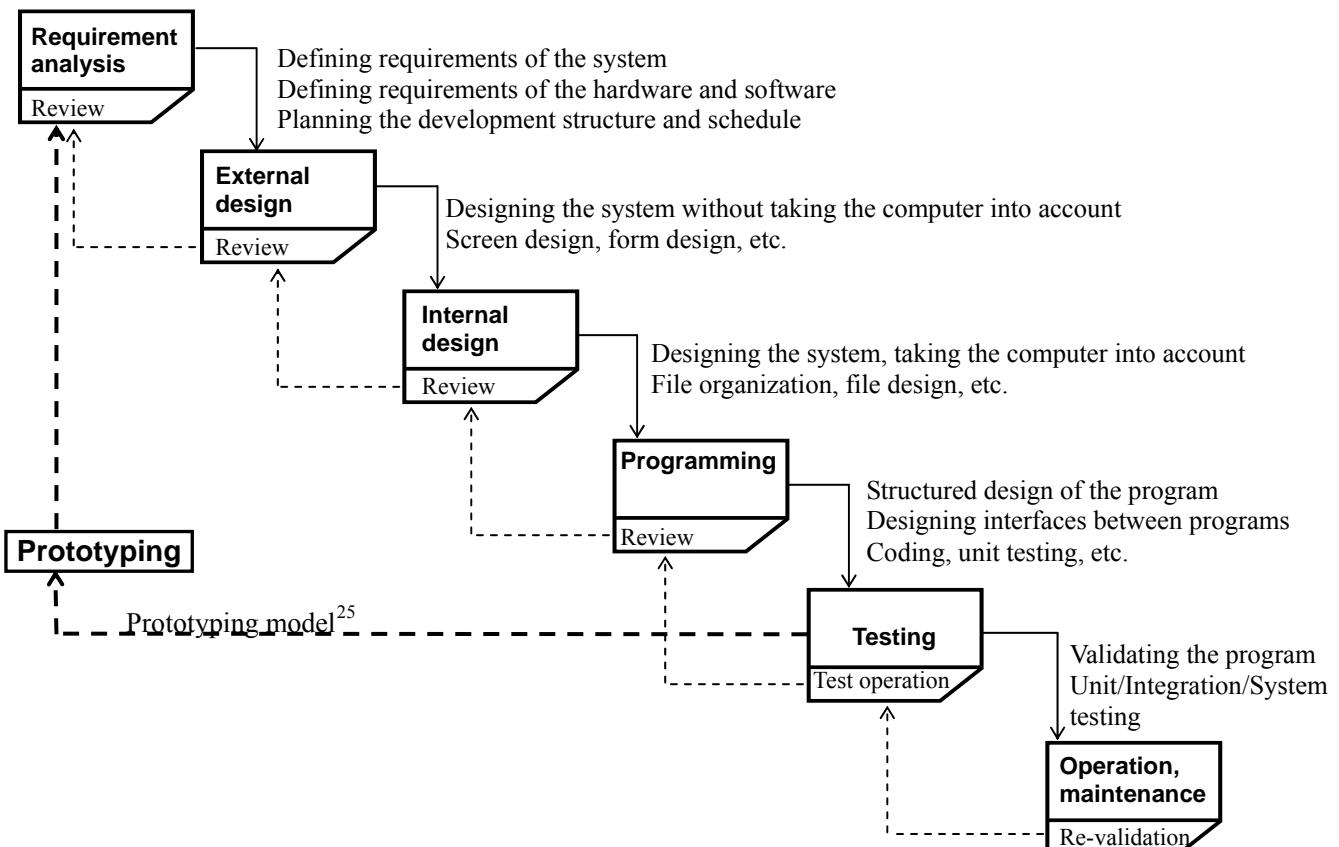
◆ Process Models

A **process model** is a model abstracting the process of system development. By establishing a process model, the procedures of system development are given a direction or a guideline. The table below shows various models and their characteristics.

Name	Characteristics
Waterfall model	<p>Each phase of the development process flows from upstream to downstream, without going back.</p> <ul style="list-style-type: none"> • Each phase is reviewed at the time of its completion for quality management. • It is difficult to clarify all of the requirements in the initial stages of the development. • There are always activities that require iteration.
Prototyping model	<p>A prototype of the user interface is developed to clarify the requirements.</p> <ul style="list-style-type: none"> • The requirements are clarified in early stages. • Final stages will have few corrections and reviews.
Spiral model ²³	<p>Subsystems are developed independently.</p> <ul style="list-style-type: none"> • Scales of simultaneous development can be controlled. • Development staff can be secured in a stable manner.

²³ **Spiral model:** It is a process model in which the methods of both the waterfall model and the prototyping model are incorporated. If a large-volume application can be divided into mutually highly independent components, for each component, either the waterfall model or the prototyping model is applied.

The figure below shows examples of development phases using the waterfall model and the prototyping model. The task contents during the development process in the prototyping model are identical to those in the waterfall model. As for the user interface, requirement analysis and testing are repeated until the specifications are finalized. For other parts, the waterfall model is used.²⁴



◆ Cost Models

Software cost is the cost incurred in each process of the lifecycle of software development (Software Development Life Cycle: SDLC). A **cost model** is a model to quantify the cost (i.e., productivity) such as the productivity and quality measures of the software. Cost models and characteristics are shown in the following table.

Name	Characteristics
COCOMO model	The programmer's work load is calculated in terms of cost based on a mathematical formula, using a statistical model, consisting of basic, intermediate, and advanced (detailed) levels.
FP (Function Point) method	The numbers of the five elements—input, output, inquiry, logical files, and interfaces—are obtained and added up with weights. Based on the assumption that this weighted sum is in correlation with the scale of the software development, the development size is estimated. The view held by this method is that what the users really need is not the programs but the functions.

²⁴ (FAQ) Many questions on the waterfall model and the prototyping model have appeared. Be sure to correctly understand the characteristics of each.

²⁵ (Hints & Tips) Since the idea of the waterfall model is so clear and easy to understand, many projects have applied this method. Since the work proceeds step by step, it is used in relatively large-scale projects. On the other hand, the prototyping model shows its effectiveness in developing relatively small-scale applications.

3.1.6 Requirement Analysis Methods

Points	<ul style="list-style-type: none"> ➤ DFD and E-R diagram are used to represent the results of requirement analysis. ➤ Terms related to object orientation include encapsulation and inheritance.
---------------	--

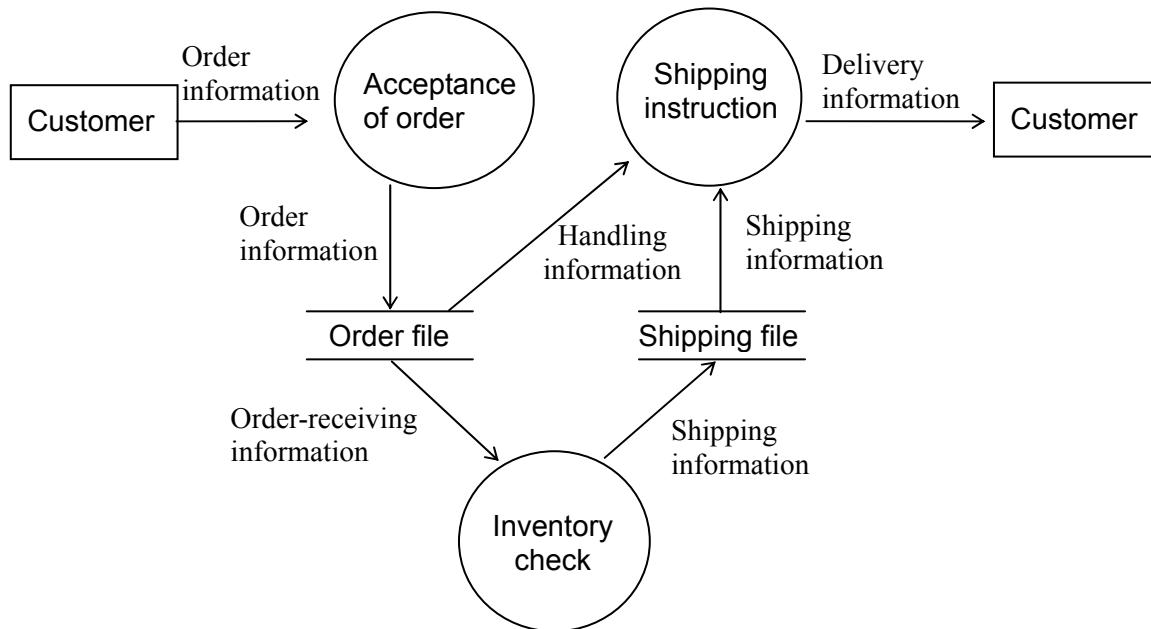
Requirement analysis means carefully identifying and organizing the requirements of a system. The results of requirement analysis are visualized by DFDs and E-R diagrams. Another method of analysis conducted from a completely different perspective is analysis by object orientation.

◆ DFD (Data Flow Diagram)

DFD is a diagram showing the flow of data (data flow). Flow of materials (objects) and money is not included. DFD is data-oriented approach. In DFD, the system is expressed using the symbols shown in the table below.

Symbol	Name	Explanation
□	External entity	Data generation (source), destination (sink, absorption)
○	Process	Data processing such as modification and conversion
→	Data flow	Data flow
=	Data store	Data storage (file)

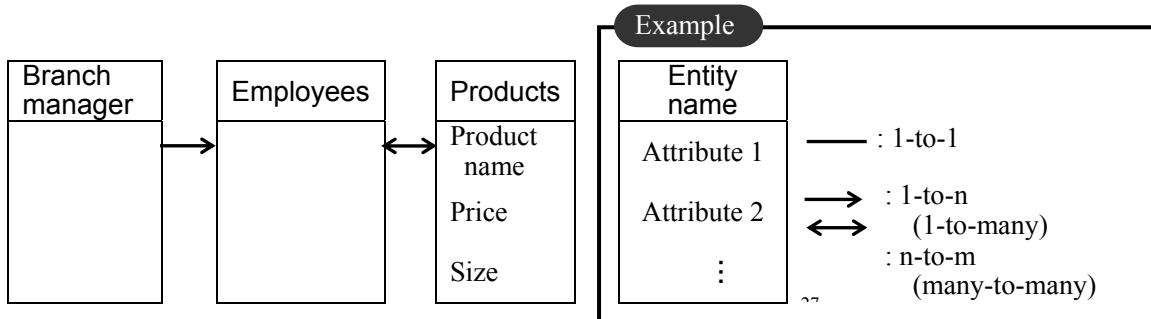
The process expressed by DFD is characterized by having not only input data flow but also output data flow which delivers the results of processing. Neither process appears alone.²⁶



²⁶ (FAQ) Questions involving DFD have often appeared on the exams. Many of them ask about the meanings of the symbols, so at least understand the meaning of each symbol. Others include questions regarding items to note concerning statements in DFD. Each process has input and output, so a diagram always shows an input data flow and an output data flow, such as “→ ○ →.” If either one is missing, it is an error as DFD.

◆ E-R Diagram (Entity-Relationship Diagram)

An **E-R diagram** is a figure that shows **relationships** between **entities**. An entity can be a person, an object, an event, or a concept which needs to be managed in business processes, and each entity has attributes. Below is an example of an E-R diagram. Note that the attributes of the branch manager and the employees are omitted.

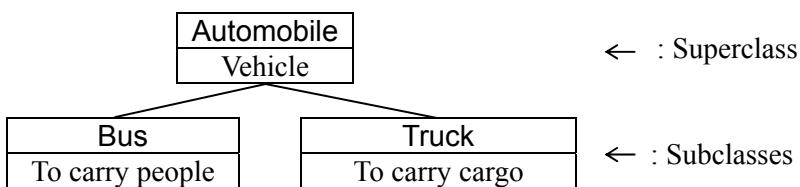


The above diagram indicates that one branch manager (1) manages several employees (n). Also, each employee handles several (m) products and each product is handled by several (n) employees. Each product has the attributes of product name, price, and size.

◆ Object Orientation

Object orientation means to model the data and their operation (process) together. Integrating data and operation is called **encapsulation**. Properties shared by the data are extracted, and the data is organized into classes in a hierarchical structure. The lower-level classes inherit the properties of the upper-level classes in this structure, and the properties thus inherited are referred to as “**inheritances**.”²⁸

For example, consider the relationship between the automobile and the bus. If the automobile is defined as a “vehicle,” and the bus is defined as a “vehicle to carry people,” the property of “being a vehicle” is common to both, so the automobile is the superclass while the bus is its subclass.²⁹ Using the inheritance function, then, the bus can simply be defined in terms of “carrying people.”



²⁷ (Note) Relations between entities are called correspondences (cardinalities). The relation between the branch manager and the employees is 1-to-n or 1-to-many, and the relation between the employees and the products is n-to-m or many-to-many. This relation between the branch manager and the employees indicates that each employee has one branch manager but the branch manager has multiple employees. If an employee is picked, then the branch manager of the employee is uniquely identified, but picking a branch manager does not identify one unique employee associated with him/her. In contrast, the relation between the employees and the products is such that neither does picking an employee identify a unique product, nor does picking a product identify a unique employee.

²⁸ (Hints & Tips) In object orientation, we need to design only the part(s) to be added. For instance, to add the truck, only the part “to carry cargo” needs to be designed. However, in reality, it is difficult to identify the common properties and properties to be added.

²⁹ **Superclass/Subclass:** An upper-level class is called a superclass whereas a lower-level class is called a subclass.

3.1.7 Software Quality Management

Points	<ul style="list-style-type: none"> ➤ Reviews include walk-through and inspection. ➤ The bug-detection rate of software is similar to the growth curve.
---------------	--

Quality management of software refers to evaluation and management of the quality of software in order to satisfy the user's requirements. Methods for this management include reviews, reliability prediction, etc.

◆ Review Methods

A **review** is a discussion meeting held at the end of each process in order to avoid carrying the existing problems over to the next process in system development.

Points to note for a review are as follows:

- There should be 4 to 6 participants (if too many are present, there may be no consensus reached).
- Documents should be distributed in advance. (Problems should be listed in advance.)
- The purpose is to find errors. (Measures to eliminate them are to be discussed later.)
- The meeting should be limited to 1 to 2 hours. (If a longer meeting is necessary, have it on another day.)
- Management should not attend the meeting. (This could lead to personnel evaluation.)

A review has several types, depending on the level at which it is applied:³⁰

Type	Functions
Design review	This is for each design process (external, internal, programming) of system development. This is for evaluation of various design documents and validation of interfaces, etc.
Walk-through	This is for all processes of system development. In early phases, not only the development personnel but also end users participate in it.
Inspection	This is for all processes of system development. It is to be performed systematically under the direction of a moderator. ³¹ Problems pointed out should be made known to the entire project. Inspection conducted in the programming phase is specifically called code inspection. ³²

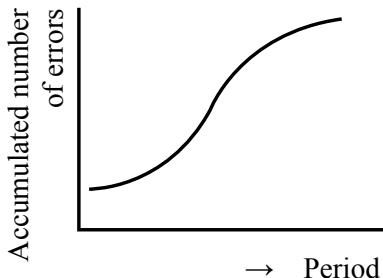
³⁰ (FAQ) The meaning of the term “review” and the difference between walk-through and inspection are often asked in exam questions.

³¹ **Moderator:** A moderator is a manager who is trained to conduct reviews and can handle errors detected. The moderator selects reviewers called inspectors who have the ability and expertise to assess the deliverables of each process.

³² **Code inspection:** Code inspection specifically refers to inspection of source programs. In code inspection, the source programs are checked and validated on a line-by-line basis.

◆ Growth Curve

In the testing phase, the relationship between the accumulated number of detected errors and time (period) is said to be similar to the growth curve. Characteristics of the growth curve are as shown below. The growth curve is sometimes called the **S-shape curve**.



- In the beginning, a gradual increase continues.
- The number increases sharply at a certain time.
- Finally it reaches saturation.

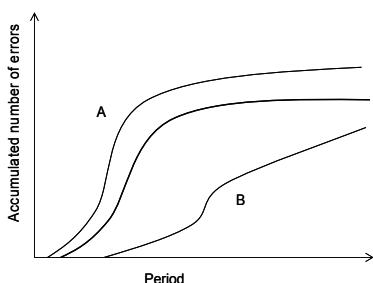
This growth curve indicates that errors are not easily detected in the beginning, that after that the number of the detected errors gradually increases, and that the number of errors decreases ideally in the end.³³ ³⁴

◆ Error-Planting Model

The error-planting model is also called the **error-spreading model** or the **bug-embedding model**. In this model, errors are intentionally placed in the program. Then the ratio of the number of errors planted and the number of errors detected is used in proportional distribution to obtain an estimate for the total number of errors in the program. Today, this model has been improved so as not to spread errors into the program. Rather, two independent testing groups perform testing on the same program, and the number of errors detected by each group is used to estimate the total number of errors.

³³ (FAQ) There are exam questions to select the correct growth curve. For instance, there may be several graphs in the answer group, and the question may say, "Which of the following curves shows that the testing is performed as planned?" Know the characteristics of the S-shape curve.

³⁴ (Hints & Tips)



Graph "A" shows that the number of errors is larger than the standard bug curve. We can infer that the test data is so good that many errors are detected. However, we can also infer that errors are found early because the quality of the software is poor. On Graph "B," the accumulated number of errors does not stabilize, so we can infer that the software quality must be very poor.

◆ Software Quality Characteristics

Quality characteristics of software are the standards by which the quality of the software is evaluated. ISO/IEC9126-1, which includes international standards of quality characteristics of software, includes the six characteristics listed in the following table.

Quality characteristics	Definition
Functionality	Functions and purposes match up.
Reliability	Specified functions work under specified conditions, and recovery from a failure is easy.
Usability	The purpose and function of use are clear, and the operation is easy.
Efficiency	The execution time is fast, using the resources effectively.
Maintainability	Changes and repairs are easy.
Portability	It can easily be moved to another environment.

Quiz

- Q1** List programming languages that are each classified as “procedural,” “functional,” “logic,” and “object-oriented.”
- Q2** Explain “reentrant.”
- Q3** What are the types of language processor?
- Q4** List three typical process models and explain the characteristics of each.
- Q5** List two differences between inspection and walk-through.

3.2 Tasks of System Development Processes

Introduction

The most typical method of system development is the waterfall model. Here, we follow the phases of the waterfall model and organize the contents of activities at each phase.

3.2.1 External Design

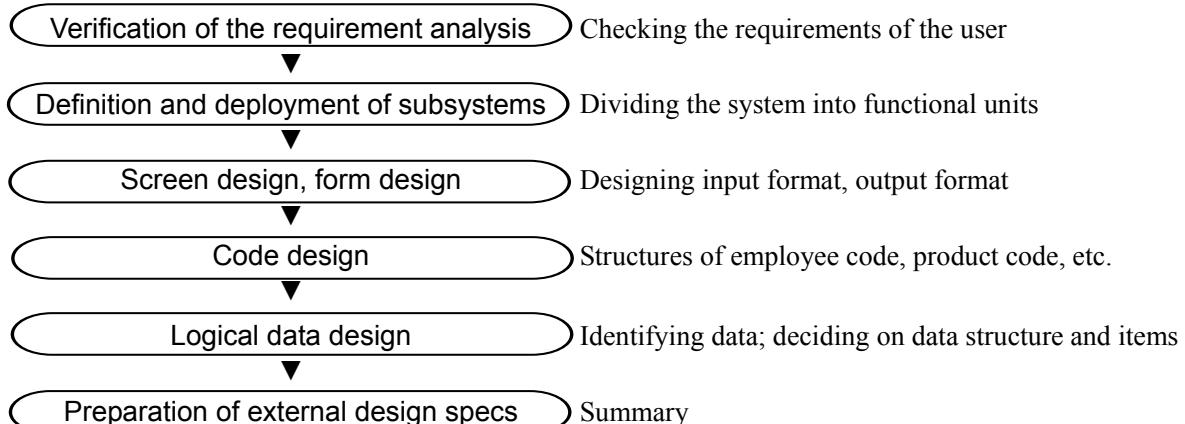
Points

- External design includes screen design, form design, and code design.
- In code design, expandability is essential, and ease of learning is of primary importance.

External design refers to designing the system without taking the computer into account. The results of activities are summarized in the external design specifications for review.³⁵

◆ Procedure of External Design

External design is the activities of collecting and analyzing the system requirements given by the user and establishing the system functions based on the results. Without considering the specifications of the computer, the design is performed around the application functions seen from the user's standpoint. Here, the main activities include screen design, form design, and code design. The flow of external design is shown below.



³⁵ (Hints & Tips) On the IT Engineer Exams, this is called "external design," but some reference books may call it "functional design," "overall design," or "outline design." The contents of activities are essentially the same.

◆ Points to Remember in Screen Design

In screen design, it is important to design with the idea of making the interface compatible with the user (human interface). Thus, it is necessary to consider simplifying input and having a screen that is easy to follow. Specifically, the following points need to be considered:³⁶

- Consider the source of data, amount of data, number of items and digits, attributes, etc.
- Place input items so that they can flow from top to bottom and left to right.
- Try to standardize the screen layout and operability.
- Keep message representations consistent.
- Consider the possibilities of aborting an operation midway or re-starting from the previous screen.

◆ Points to Remember in Output Design

Output design refers to designing the output format of the system, including screen display and printing. Of these, printing (reports) is central. As with screen design, it is necessary to consider the human interface and to make the reports easy to read. More specifically, the following points need to be considered:

- Set up sequence and positions in consideration of the relationship among the items.
- Ensure that the title appropriately expresses the printed contents.
- Clearly distinguish various dates such as the date prepared, date reported, and date approved.
- Position the items with appropriate spacing.
- Plan for the entire report to have sufficient empty space.
- Consider the design so that critical items can be immediately identified.

◆ Code Design

Codes need to have the functions listed in the table below. When designing, we must consider various properties such as commonality, systematization, expandability, and clarity.³⁷

Function	Explanation
Identification	A function to distinguish data. Codes can distinguish between two people with the same first and family names.
Classification	A grouping function; classifying by affiliation code, etc.
Listing	A sorting function: If the digits are aligned, data can be sorted by date of birth.
Checking	A function to check input values; for instance, by adding a check digit. ³⁸

³⁶ (FAQ) The following type of questions has been frequently asked on the exams: "Which of the following is an appropriate description on points to remember in screen design and form design?" Read the descriptions carefully to answer them.

³⁷ (Note) Examples of codes include consecutive codes and digit-specific codes. Consecutive codes are consecutive numbers assigned to data listed in order from the beginning. These are useful when the number of data is fixed. In digit-specific codes, the data is classified into large classes, middle classes, and small classes with certain standards in a hierarchical structure, and each group has consecutive codes. Digits can be lengthy, but they are suitable for computer processing. The zip code system is an example of this type.

³⁸ **Check digit:** It is a 1-digit code obtained by performing a certain calculation, defined in advance, on each digit of a numerical item. When a code is entered, the same calculation is performed to obtain the 1-digit number, which is then compared with the check digit. If they are the same, the code is considered valid; otherwise, it is considered invalid. It is a method for detecting errors.

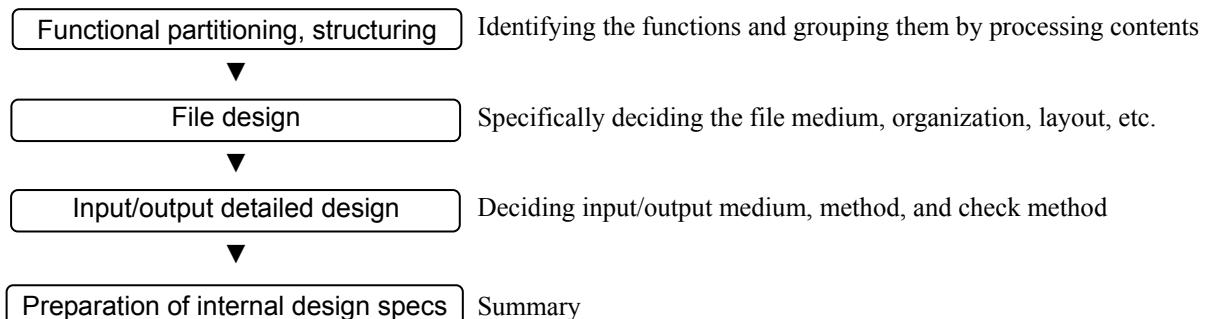
3.2.2 Internal Design

Points	<ul style="list-style-type: none"> ➤ Internal design includes functional partitioning, structuring, file design, and input/output detailed design. ➤ In input/output detailed design, the main focus is on the input data check method.
---------------	---

Internal design refers to designing the system taking into account the computer that is planned to be used. The results of activities are summarized in the internal design report for review.

◆ Procedure of Internal Design

From the standpoint of the systems developer, an optimum system is designed based on the computer specifications. The flow of internal design is shown below.



◆ File Design

The organization and medium of files need to be decided according to the purpose of their use. For backup, various media are used, including magneto optical disks (MO), floppy disks (FD), CD-R, CD-RW, DVD-R, DVD-RAM, DAT, etc.³⁹

According to the purpose of their use, the organization method of files is selected from options including sequential, direct, indexed sequential, and partitioned organizations.⁴⁰ If a large number of records exist and most of them are to be read and updated, sequential organization is suitable. For random processing, direct organization is suitable.

³⁹ (Hints & Tips) An appropriate medium is chosen based on what is being backed up. The approximate capacity of each medium is as follows:

MO: 128, 320, 540, 640 MB/ 1.3 GB

FD: 1.2/ 1.44 MB

CD-R: 640/700 MB

CD-RW: 700 MB

DVD-R: 4.7/ 8.5 GB

DVD+R: 4.7/8.5 GB

DVD-RAM: 4.7 GB

DVD-RW: 4.7 GB

DVD+RW: 4.7 GB

DAT: 24 GB maximum

⁴⁰ (Hints & Tips) Partitioned organization is hardly ever used in a data file. It is almost always used in library files.

◆ Check Methods

Among the methods used for checking input data, some of the typical methods are organized in the table below:

Method	Check contents
Numeric check	Is the numerical item really a number?
Format check	Is it in the predefined format? Are the digits not misaligned?
Limit check	Is the value within the upper and lower bounds?
Range check	Is the value within the correct range? (this can be considered as a type of limit check)
Validity check	Is the value logically valid? (e.g., Feb. 29 on a non-leap year)
Sequence check	Are the key item values listed in sequence?
Balance check	Are the paired items matched up correctly? (e.g., renters and landlords)
Collation check	Is the code value contained in the master file?

◆ Check Digit Method

This is the method which determines whether an input error is made when a code is entered by performing the same calculation when the code was created and by comparing the calculation result with the code. In general, one digit (check digit) is added to the base code at the end to form a code.

XXXXX Y → XXXXXY
Base code Check digit Code

The check digit method often uses *modulus 11*. In *modulus 11*, weights 2, 3, ... are assigned to each digit of the base code from the lowest digit. The product of the weight and the numerical value for each digit is then calculated and the sum of the products is found. Finally, this sum is divided by 11, and the remainder is the check digit. If a resulting product is a 2-digit number, the digits are separated. Below is an example when the base code is “12345.”

[Base code]	<u>1</u>	<u>2</u>	<u>3</u>	<u>4</u>	<u>5</u>	→ 14 / 11 = 1 remainder 3
[Weights]	x	x	x	x	x	↓
	6	5	4	3	2	<u>12345</u> 3
	■	■	■	■	■	Base code
Products (multiplication results) →	6	10	12	12	10	↓
Separated into digits →	6	1+0	1+2	1+2	1+0 = 14	<u>123453</u> The code

Since the division is by 11, this method is called *modulus 11*. Now, dividing a number by 11 may cause a remainder of 10. In this case, the check digit is defined as 0. Besides this, there is also a method called *modulus 10*.⁴¹

⁴¹ **Modulus 10:** The idea of calculating the check digit by *modulus 10* is exactly like the idea of *modulus 11*, except that the weighted sum is divided by 10, not 11. *Modulus 11* and *modulus 10* are both considered able to detect most input errors.

3.2.3 Software Design Methods

Points

- Software design methods include structured design and design by data structure.
- Structured design is process-oriented design.

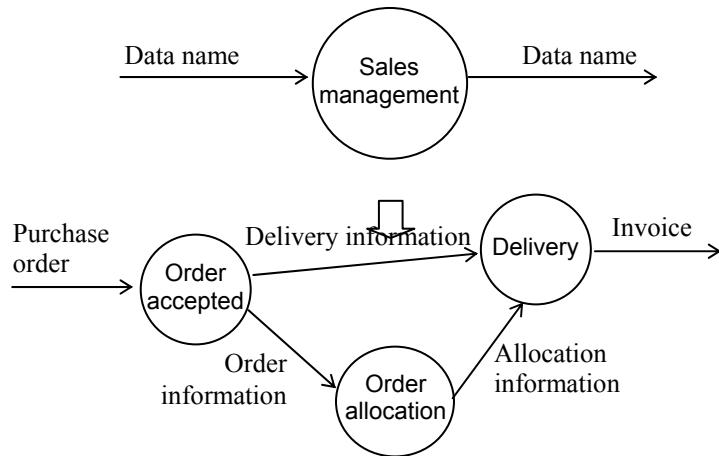
When software is modeled in terms of processes and data separately, structured design is the process-oriented technique. Alternatively, there is another technique in which design activities proceed, focusing attention on data structure. Structured design techniques include bubble charts, STS partitioning method, and TR partitioning method. Techniques focusing attention on data structure include the Jackson method and the Warnier method.⁴²

◆ Structured Design Techniques

The structured design is the technique of designing based on the approach of structuring. The structure chart is used for this method. In this method, the design activities proceed from general overall ideas to specific details, so it is sometimes called “stepwise refinement” (top-down approach or module partitioning⁴³).

Bubble charts

These charts use bubbles (circles) to represent processes that convert input data into output data and are the same as DFDs. For every system, there is only one initial bubble, but as break-down (structuring) continues, bubbles and data flow become more complex. Below is an example of a sales management system broken down.



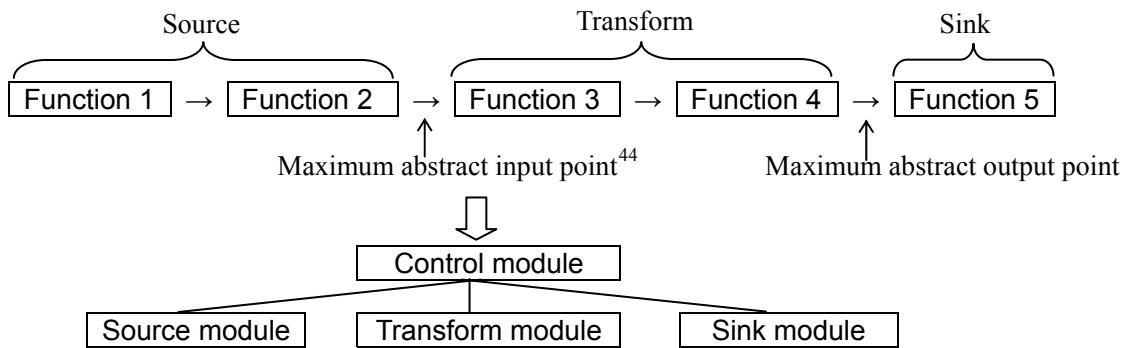
Like DFD, a bubble chart is used to partition the functions of the system. It is also used to partition the functions of a program to construct a hierarchy.

⁴² (FAQ) Along with structured design techniques and the Jackson method, exam questions covering an overview of the software methods have frequently appeared; these include questions like “Which of the following is an example of ...?” and the answer group often lists terms.

⁴³ **Module partitioning:** It is a process whereby the program functions are discussed, divided into functional units, and put into a function hierarchy. Each functional unit to which the functions are partitioned is called a module. A program thus consists of modules.

STS partitioning (source, transform, sink)

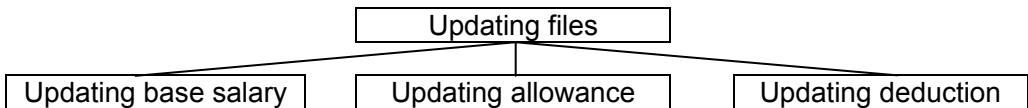
In this technique, the program structure is divided up into a source (input), a transform (processes), and a sink (output), and each of these is defined as one function. Once STS partitioning is done, other techniques such as DFD and bubble chart are used.



In STS partitioning, after dividing into three modules, we add a *control module* to control these modules. This method is used when writing programs for batch processing.

TR Partitioning (Transaction Partitioning)

In this technique, the transactions are partitioned by function with respect to the branching flow of data, and they are formed into modules. The example below is dealing with a part of a payroll program. If the function of “updating files” has three functions—updating the base salary, updating allowances (stipends), and updating deductions—, then each function constitutes a module:



TR partitioning is used frequently in programs in online systems. For example, in a seat reservation system, transactions occur frequently according to each function, such as “inquiry about seat availability” and “issuing tickets.”

⁴⁴ **Maximum abstract point:** It is a concept in STS partitioning where the program is partitioned into three modules. The boundary between the source and the transform is called the maximum abstract input point, and the boundary between the transform and the sink is called the maximum abstract output point. The former is the point where the input data is maximally abstracted, at which the input data is transformed and cease to be input data. The latter is the point where the output data is maximally abstracted, at which the output data (going backward) were first recognized in the form of output data.

◆ Design Techniques based on Data Structure

With the idea that clarifying the input/output data structure naturally determines the processes, these techniques compare the input data structure with the output data structure to induce the structure of the processes. A typical method is the Jackson method (Jackson Structured Programming: JSP). In the Jackson method, the data structure and the program structure are expressed as tree structures.

The table below shows symbols used in the Jackson method and their meanings.

Symbol	Name	Meaning
A	Basic element	One item in data structure; a processing unit in the program structure
	Sequence	A consists of B and C; B and C are executed in that order. Each element appears only once in that order.
	Iteration	Within A, B is repeated at least 0 times (perhaps none). The "*" symbol means repetition.
	Selection	Either B or C is selected by A (one at a time).

The Jackson method is used frequently in programs of business-processing systems. Another method based on data structure is the Warnier method.⁴⁵

⁴⁵ **The Warnier method:** It is a design technique that is based on data structure, like the Jackson method. It is characterized by drawing the so-called Warnier diagram, similar to a flowchart.

3.2.4 Module Partitioning Criteria

Points	<ul style="list-style-type: none"> ➤ The module independence is used for criteria for module partitioning. ➤ Module strength and module coupling (cohesion) are used for criteria for module independence.
---------------	--

In structured design, the validity of the finally partitioned modules is evaluated based on various evaluation criteria such as structure and independence.

◆ Structural Evaluation

For module partitioning, the following characteristics need to be considered:

- **Size:** Are they too small or too big? (Criteria should be set.) The proper size differs depending on the language used (about 300 steps for COBOL).
- **Function:** Are there unnecessary functions? Are there multiple functions? (If there are multiple functions, partition them again.)
- **Interface:** Are there too many parameters? (In such a case, review the partitioning.)

◆ Evaluation of Independence

To evaluate the independence of modules, there are two measures—module strength and module coupling. Partitioning is considered good if its modules have a high level of independence. The weaker the module coupling⁴⁶ is and the stronger the module strength⁴⁷ is, the more independent the modules are.

Types of strength	Strength	Independence	Coupling	Types of coupling
Coincidental strength	Weak	Low	Strong	Content coupling
Logical strength				Common coupling
Classical strength				External coupling
Procedural strength				Control coupling
Communicational strength				Stamp coupling
Informational strength				
Functional strength	Strong	High	Weak	Data coupling

⁴⁶ **Module coupling:** It is a measure of how closely modules are related to one another; the weaker the module coupling is, the more independent they are.

⁴⁷ **Module strength:** It is a measure of how closely the component elements within a module are related to one another; the stronger the module strength is, the more independent the modules are.

Module strength has the seven levels as shown in the following table.

Module strength	Contents
Coincidental strength	The program is simply divided or duplicate functions are eliminated. There are no special relationships among the functions within the module.
Logical strength	The module has multiple related functions and chooses the processing based on parameter (argument) conditions.
Classical strength	The module unifies various modules executed at a designated time and executes multiple functions sequentially. An initial-setup module is an example.
Procedural strength	The module executes multiple serial functions; the relationship within the module is close, and the various functions cannot be executed independently.
Communicational strength	The module executes multiple serial functions just as in procedural strength, but data is transferred between functions.
Informational strength	The module unifies multiple functions that handle the same data structure, has an input point and an output point for each function, and can call each function separately.
Functional strength	The module consists of one function only, and all the instructions are to execute the one function and therefore are closely related.

Module coupling has the following six levels.⁴⁸

Module coupling	Contents
Data coupling	Only the data necessary for processing are passed. Neither the calling module nor the module being called has functional relations.
Stamp coupling	Data structure itself is passed as arguments. The module being called uses part of the structure.
Control coupling	The function code is passed to a subprogram as an argument, influencing the execution of the subprogram.
External coupling	Data externally declared are shared. The difference between common coupling and this is that only the necessary data is externally declared in external coupling.
Common coupling	Data defined in the common area are shared.
Content coupling	A module directly references another module and changes it.

3.2.5 Programming

Points

- Structuring of logic improves productivity and maintainability.
- Programs can be written only with basic control structures.

Programming requires structured logic. By structuring the logic, the level of complexity can be reduced and programs can be written so that they will be easy to understand. This leads to higher productivity and improved maintainability in development.

◆ Structure Theorem

“In a valid program, in which there is a pair of an entrance and an exit, no infinite loop, and no statement that is not executed, we can write the logic only using three basic control structures: sequence, selection (decision), and repetition (iteration).” This is called the *Structure Theorem*. It is very important to keep “structuring” in mind when writing a program. The basic principle is to write a “goto-less program” (program without using a “goto” statement).

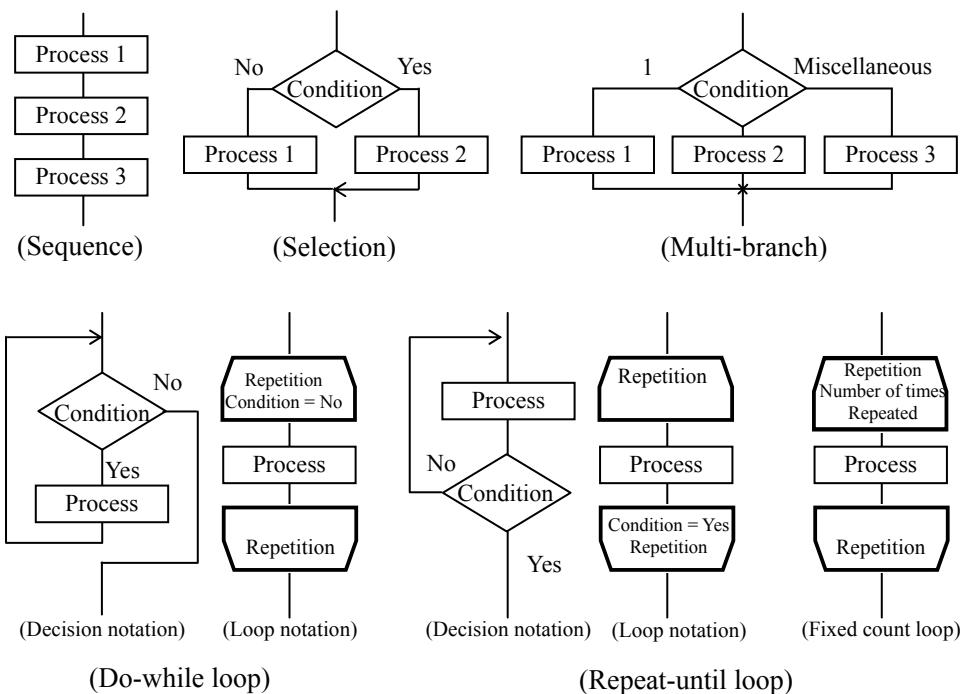
⁴⁸ (FAQ) There are many exam questions on the independence of modules. Be sure to correctly understand the contents of module strength and module coupling.

◆ Basic Control Structures

The Structuring Theorem proposes three basic control structures, but in practice, the following six basic control structures are used.⁴⁹

Name	Explanation
Sequence	The program consists of sentences executed sequentially without “goto” statements and logical decisions.
Selection (“if- then-else” type)	The function to be executed depends on whether or not a certain condition holds.
Multi-branch (“case” type)	Multiple branches are designated depending on the value of a variable or a processing result.
Do-while loop (Pre-test loop)	The condition is determined at the beginning of a repeated process (repeated while the condition holds); depending on the condition, the repeated process may not occur at all.
Repeat-until loop (Post-test loop)	The condition is determined at the end of a repeated process (the loop is terminated if the condition holds); regardless of the condition, the repeated process occurs at least once.
Fixed count loop	The process is repeated a certain fixed number of times on entry into the loop.

The following is a set of flowcharts showing the basic control structures detailed above.⁵⁰



The results of module design are expressed using flowcharts. However, various problems have been pointed out concerning flowcharts, and some companies use different methods of expressions. Yet, on the IT Engineer Exams, mostly flowcharts are used.

⁴⁹ (Hints & Tips) The three basic control structures are “sequence,” “selection,” and “repetition.” “Sequence” and “selection” are as shown in the table, but “repetition,” when first proposed, meant do-while loop.

⁵⁰ (Hints & Tips) The loop notation shows the condition to end the repetition. In other words, it shows the condition under which the repetition is terminated. However, in a do-while loop, the repeated process is executed as long as the condition holds, so we must be careful in denoting the condition using the loop notation and the decision notation.

◆ Graphic Expressions of Module Design

Besides flowcharts, there are other methods of expressing the results of module design, as shown in the table below.

Method	Explanation
Pseudo-coding (Pseudo language)	Pseudo-code is similar to a program code, but allows the use of natural language (e.g., English) for abstraction of functions.
Decision table	Relations between the conditions for and the contents of the processing are expressed in the form of a table.
NS chart	The logical structure is expressed without using arrows. As a visual aid, this is easy to read.
Structure chart	A tree structure is used to express the logic. ⁵¹

3.2.6 Types and Procedures of Tests

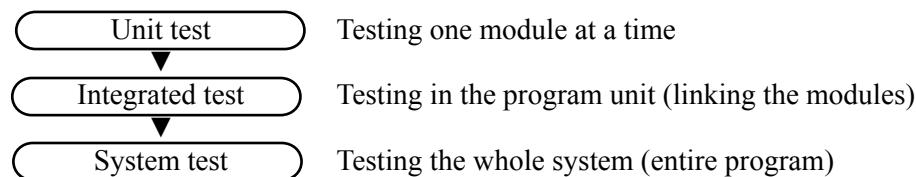
Points

- Types of tests include unit tests, integrated tests, and system tests.
- The order of tests can be top-down or bottom-up.

In system development, the design work is performed top-down, but testing is performed bottom-up. In other words, we take the approach of starting with details and moving toward the whole. This is called stepwise integration. In integrated tests, in order to perform the tests efficiently, we must carefully choose the order in which the modules are tested.

◆ Order of Tests

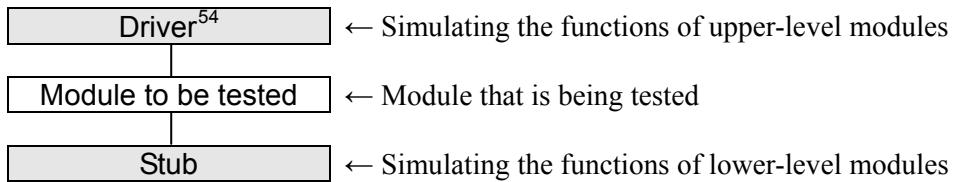
Tests are conducted in the following order:



⁵¹ **Structure chart:** It is also known as a tree-structure chart. To achieve structured programming, various types of structure charts have been proposed by developers and research institutes. Some well-known examples are PAD (Hitachi), SPD (NEC), YAC (Fujitsu), and HCP (NTT).

Unit test

A unit test is a quality test for modules (smallest units within a program). In a test of the entire program it can be difficult to identify the cause of an error, so unit tests are performed for each module as a unit. In unit tests, we perform function tests⁵² and structure tests.⁵³ Since modules do not function by themselves, we prepare drivers and stubs.



Integration test

With multiple modules linked together, we test these linked programs (load modules) in integrated tests. The main goal here is to examine the interface between modules as well as the input and output.

Methods for integration tests include top-down tests, bottom-up tests, big-bang tests,⁵⁵ and sandwich tests.⁵⁶

System test

This is the last test conducted in the development division and is used to examine whether the required specifications are satisfied. For instance, we address questions such as “Are there any performance problems (performance test)?” and “Can it endure heavy loads (load test)?” We also test exceptional items and measures to be taken when a failure occurs.

⁵² **Function test:** It is a validation test, based on module specifications, to verify that all functions that the module is supposed to have are satisfied.

⁵³ **Structure test:** It is a validation test, based on module specifications and source program, to verify that the logic of the module is sound.

⁵⁴ (Hints & Tips) A driver is a program that simulates the functions of an upper-level module, and a stub is a program that simulates the functions of a lower-level module. In general, a stub simply returns a value and therefore is easy whereas a driver controls calls and is therefore often complicated.

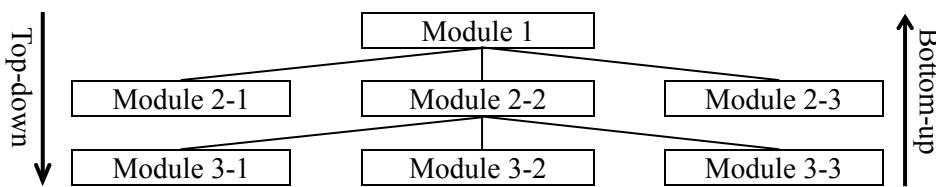
⁵⁵ **Big-bang test:** It is a test wherein all the modules that have completed the unit tests are linked all at once and tested. If the program is small-scale, this could reduce the number of testing procedures; however, if an error occurs, it is difficult to identify where the error has occurred.

⁵⁶ **Sandwich test:** It is a test where lower-level modules are tested bottom-up and higher-level modules are tested top-down. This is the most realistic type of testing.

◆ Order of Integrated Tests

Integrated tests can be top-down or bottom-up. The characteristics of each are shown in the following table and figure.⁵⁷

Type	Characteristics
Top-down test	Testing from upper-level modules to lower-level modules Requires stubs to simulate lower-level modules not yet tested. Interfaces between modules can be sufficiently tested. Initially, parallel work is difficult. Effective in testing newly developed systems
Bottom-up test	Testing from lower-level modules to upper-level modules Requires drivers to simulate upper-level modules not yet tested. Functions of the program can be sufficiently tested. Parallel work is possible from the initial stages of the test. Effective in developing new systems by modifying existing systems



3.2.7 Test Techniques

Points

- Test techniques include black box tests and white box tests.
- Black-box tests are used except for unit tests.

The purpose of testing a program is to verify that the program runs according to the specifications and to eliminate errors embedded in the program. To this end, sometimes error data is intentionally entered. There are two test techniques that are proposed: black box tests and white box tests.

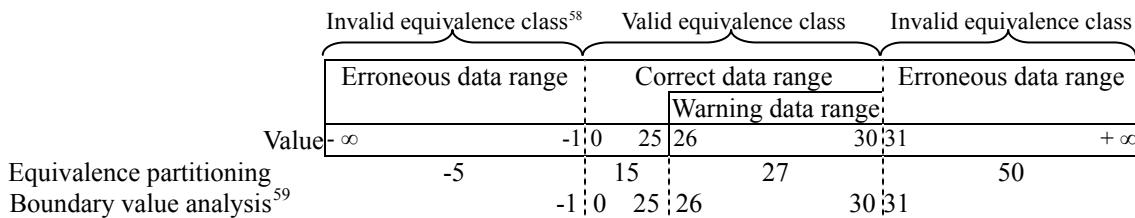
◆ Black Box Tests

A **black box test** is the method whereby test cases are designed based on the external specifications of the program. Regardless of the program logic, test data is prepared based on the external specifications. Test criteria are shown in the following table.

Name	Explanation
Equivalence partitioning	The range of input values is partitioned into several classes, and a test value is picked from each class as a representative value (e.g., the median value of the class).
Boundary value analysis	The range of input value is partitioned into several classes, and the boundary values (limit values) for the classes are picked as test values.

⁵⁷ (FAQ) Frequently we see exam questions such as “Which of the following is an appropriate description of a top-down test?” and “Which of the following is an appropriate description of a bottom-up test?” Be sure you understand the difference between the methods of top-down and bottom-up tests as well as the roles of stubs and drivers.

For example, suppose that in a numerical (integer) item, 0 through 30 are valid data values and other integers are erroneous, prompting an error message to be displayed. Further, suppose that values 26 through 30 prompt a warning message to be displayed as warning data values. Here, under equivalence partitioning, for instance, the set of test data values (-5, 15, 27, 50) may be selected. Under boundary value analysis, the set of the boundary values of the classes (-1, 0, 25, 26, 30, 31) is selected.



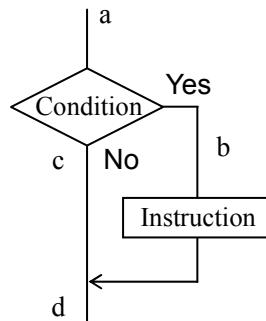
◆ White Box Tests

A **white box test** is the method whereby test cases are designed based on the internal specifications of the program. Below are some of the test criteria: instruction coverage, decision condition coverage (branch coverage), condition coverage (branch condition coverage), decision /condition coverage, and multiple condition coverage. These are listed in ascending order of rigidity (strictness).⁶⁰

Name	Explanation
Instruction coverage	Test cases are designed such that every instruction is executed at least once.
Decision condition coverage (branch coverage)	Test cases are designed such that true and false cases in the decision are executed at least once.
Condition coverage (branch condition coverage)	Test cases are designed such that in multiple conditions every combination having true and false cases is satisfied.
Decision/condition coverage	This is the combination of branch coverage and condition coverage.
Multiple condition coverage	Test cases are designed such that every combination of true/false cases in every condition is tested.

For instance, suppose there is a program with a structure shown in the figure below. Data prepared for each of the test criteria is as follows:

In instruction coverage, data going through the path “a, b, d” are prepared since this path goes through every instruction. In other words, only the data following the “Yes” case in the “condition” is prepared. In decision condition coverage, data for the “Yes” and “No” cases, i.e., data going through both “a, b, d” and “a, c, d” is prepared.



⁵⁸ **Valid equivalence class/ Invalid equivalence class:** In a black box test, a range of correct data values is called a valid equivalence class, and a range of erroneous data values is called an invalid equivalence class.

⁵⁹ (FAQ) There are exam questions in which you are to prepare test data for equivalence partition and boundary value analysis. Understand fully what these terms mean, and make sure that you are able to prepare test data.

⁶⁰ (FAQ) Every exam has questions on the meanings of black box tests and white box tests. Be sure to know these.

The others, i.e., condition coverage, decision/condition coverage, and multiple condition coverage are techniques used for multiple conditions. For instance, suppose that we consider the multiple conditions “a and b” here.

Number	a	b	a and b	Condition coverage	Decision/condition coverage	Multiple condition coverage
(1)	True	True	True		X	X
(2)	True	False	False	X	X	X
(3)	False	True	False	X	X	X
(4)	False	False	False			X

In condition coverage,⁶¹ as combinations having true and false cases in the multiple conditions, numbers (2) and (3) are tested. However, the multiple conditions “a and b” are false in both of these numbers, so the case in which both are true is not tested. In decision/condition coverage, the case in which both are true is included, so numbers (1), (2), and (3) are all tested. In multiple condition coverage, every combination, i.e., (1), (2), (3), and (4) are tested.

⁶¹ (FAQ) Concerning instruction coverage and decision condition coverage, exam questions ask for specific test data. It is best to actually prepare test data and check.

Quiz

- Q1** List the types of tasks done in external design.
- Q2** List the types of tasks done in internal design.
- Q3** To increase the level of independence of modules, what should one do with module strength and module coupling? Also, name the type of strength and coupling referred to here.
- Q4** Describe briefly each of the following checking methods: “numeric check,” “format check,” “limit check,” “range check,” and “sequence check.”
- Q5** Describe the characteristics of a black box test.
- Q6** Describe the characteristics of a white box test.

Question 1

Difficulty: **

Frequency: **

Q1. Which of the following is an appropriate statement concerning the optimization of a compiler?

- a) It generates intermediate code for the interpreter instead of generating object code.
- b) It generates object code that runs on a machine different from the computer on which the compiler runs.
- c) It generates object code that displays the name of the routine to which the control is passed or the content of a variable at a certain point in time when the program is executed.
- d) It analyzes program code and generates object code so that the processing can become more efficient during execution.

Answer 1

Correct Answer: **d**

The optimization of a compiler means eliminating the redundancy of the object program.

- a) Optimization means the elimination of redundancy; it is the compiler that generates intermediate code. Hence, this statement is not an explanation of optimization.
- b) This is an explanation of cross compilers.
- c) This is an explanation of tracers.
- d) Optimization increases the processing efficiency during execution through various means including removing unnecessary parentheses and pre-calculating operations involving only constants.

Question 2

Difficulty: **

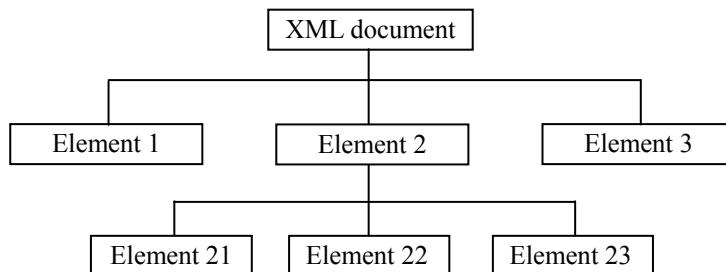
Frequency: ***

- Q2.** Which of the following is an appropriate statement with regard to the method for defining an element, which is a minimum unit for constructing an XML document?
- A start tag and an end tag are paired up for the construction, and neither tag can be omitted.
 - A data element is constructed so that it can be placed between a start tag and an end tag. In some cases, however, no data exists.
 - In an XML document, multiple root elements can be defined to represent a hierarchical structure.
 - Comment information is added to represent the type of element. This is identified as the element name.

Answer 2**Correct Answer:** b

XML (eXtensible Markup Language) is an extension of the functions of HTML, eliminating unnecessary functions of SGML and optimized to be used on the Web as well. XML, like HTML, is used on the Web; the difference is that whereas the tags of HTML are fixed, the tags of XML can be defined uniquely by the user by means of the definition called DTD (Document Type Definition).

- It is correct that the structure is such that the data is surrounded by a start tag and an end tag, but when there is no data, to indicate the empty element, a special tag such as <element name/> can be designated, distinguished from the start tag. For instance, this may be a description like . Hence, the start tag and the end tag may not form a pair.
- In principle, the data is surrounded by a start tag and an end tag. If there is no data, that is acceptable.
- In XML, all elements are in nested structure. An element may directly contain other multiple elements, but there is no element that is directly contained in multiple elements. Here, the one that includes others is called a “parent,” and that which is included is called a “child.”



- Comment information is not contained in the data and is practically ignored. Comments are surrounded by “<!--” and “-->.”

Question 3

Difficulty: *

Frequency: ***

Q3. Which of the following statements describes the characteristic of the waterfall model, which guarantees the consistency of system development?

- a) In principle, it is not allowed to go backwards across development phases.
- b) System development is divided into multiple phases to be managed.
- c) It is absolutely necessary to create a project organization.
- d) The development activities in the next phase are based on the results passed down from the preceding phase.

Answer 3

Correct Answer: **d**

The waterfall model is a process model in which system development proceeds from upstream phase to downstream phase in sequence: “basic planning → external design → internal design → program design → programming → testing → installation, operation, maintenance.” Since the flow of the development process is divided for each phase, it is easy to grasp an overview of the entire project. Project management is also considered easier because the work flows from upstream to downstream sequentially. However, since there is basically no going back, it has the disadvantage that the development efficiency drops if the process requires regression.

In the waterfall model, a review is conducted at the end of each phase so that a bug is not carried on to the next phase. If there is a bug discovered in a downstream phase (phase after programming), the cost required for system modification (cost for regression) is extremely high. Therefore, bugs must be discovered in an upstream phase (a design phase between basic planning and program design).

- a) This is one of the characteristics of the waterfall model, but this simply describes how the process proceeds; it does not guarantee the consistency of system development.
- b) This is one of the characteristics of the waterfall model, but it describes the classification of the contents of activities; this does not guarantee the consistency of system development.
- c) A project team is organized for system development in general, not just in the waterfall model.
- d) The design phase of the waterfall model is stepwise refinement. Contents of the previous phase are carried over to the next phase; this guarantees the consistency of system development.

Question 4

Difficulty: * Frequency: ***

- Q4.** Which of the following is an appropriate statement concerning “prototyping,” a method of software development?
- a) Since activities proceed sequentially through basic planning, external design, internal design, program design, programming, and testing, it is possible to get a good overview of the entire project and it is easy to determine the schedule and allocate resources.
 - b) Since a trial model is created at an early stage of system development, it is possible to eliminate vagueness and differences of perception between the user department and the developing department.
 - c) The characteristics of the software are divided into those for which the specifications are fixed and do not require changing and those for which the specifications require changing. Then, the process of creating, reviewing, and changing the code according to those specifications is repeated.
 - d) A large application is divided into highly independent components; then processes of design, coding, and testing are repeated for them, gradually expanding the scope of development of the program.

Answer 4

Correct Answer: **b**

Prototyping is a method in which a prototype (trial model) is made for the parts directly visible to the system user (screen, form, etc.) and systems are developed based on the feedback obtained from users who have tested the prototype. Hence, the statement b) is appropriate.

a) explains the waterfall model, and d) explains the spiral model.

Question 5

Difficulty: * Frequency: ***

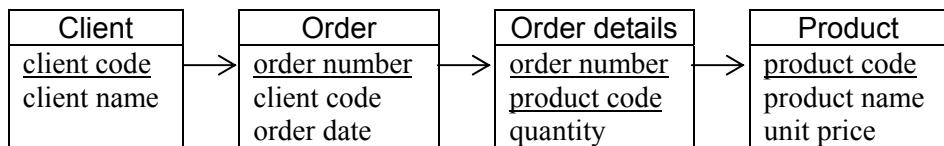
Q5. Which of the following is shown by an E-R diagram?

- a) Relationships between entities
- b) Relationships between entity types and their instances
- c) Relationships between data and processes
- d) Relationships between processes

Answer 5

Correct Answer: a

An E-R diagram shows the relationships between entities (actual objects), indicating an entity with a () and corresponding relations between entities with arrows (→ , ←→ , —). The following is an example of an E-R diagram:



Example

Entity name	Relationship	Description
attribute 1	—	: 1-to-1
attribute 2	→	: 1-to-many
:	←→	: many-to-many

An underlined attribute is a primary key attribute. An entity type is an entity having data subject to management. Normally, the entities are expressed with nouns such as “client” and “product.” “Instances” are entities with values.

Client

Client code	Client name	Client address
1011	George Bush	Crawford, Texas
1021	William Clinton	Hope, Arkansas

← Entity Type

← Instance

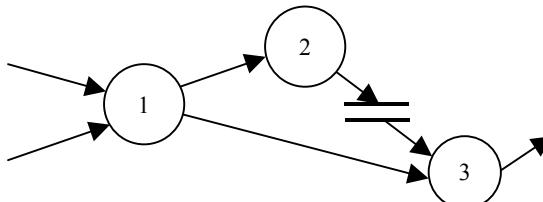
← Instance

Question 6

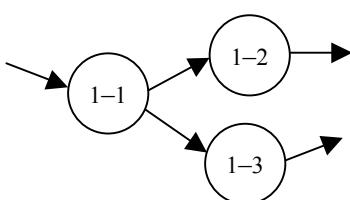
Difficulty: **

Frequency: ***

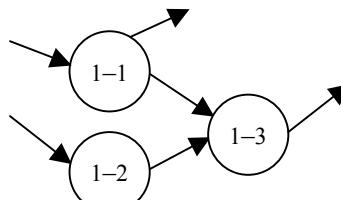
- Q6.** The figure below shows a certain level in a hierarchical DFD. Which is the most appropriate method of describing DFD of the level immediately below? Assume that the processes in the level immediately below *Process n* are numbered processes of the form *n-1*, *n-2*, etc.



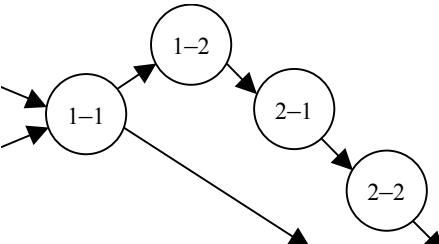
a)



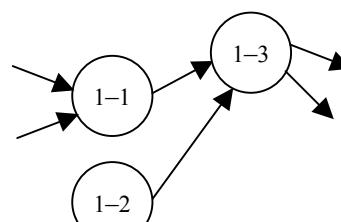
b)



c)



d)

**Answer 6****Correct Answer:** b

In DFD, the processes (circles) get broken down in order. Hence, when they are broken down, multiple processes on a higher level cannot be merged together on a lower level.

In DFD given in this question, note that *Process 1* has two input dataflow arrows as well as two output dataflow arrows.

- While *Process 1* has two input dataflow arrows, each of its child processes has only one input dataflow arrow.
- Child Process 1-1* has one input dataflow arrow, and so does *Child Process 1-2*. The total is 2. As for output dataflow, there is one arrow from *Child Process 1-1* and another from *Child Process 1-3*, a total of 2. Hence, this may be a break-down of DFD in the question.
- Since DFD breaks down one process, combinations like {(1-1), (1-2)} and {(2-1), (2-2)} are acceptable whereas a combination like {(1-1), (1-2), (2-1), (2-2)}, in which multiple processes on an upper level are combined, is not.
- The numbers of input dataflow arrows and output dataflow arrows are correct, but every process must have at least one input dataflow arrow and at least one output dataflow arrow. There is no input dataflow arrow for *Process 1-2*.

Question 7

Difficulty: **

Frequency: ***

- Q7.** The following table gives the number of items by category and the weighting factor for user functions of an application program. Information is based on the function point method. How many function points does this application program have? Here, the correction coefficient of complexity is 0.75.

User function type	Number of items	Weighting factor
External input	1	4
External output	2	5
Internal logical file	1	10
External interface file	0	7
External inquiry	0	4

a) 18

b) 24

c) 30

d) 32

Answer 7**Correct Answer:** a

In the function point method, the number of function points is obtained as follows:

- Multiply the number of functions (number of items) by the corresponding weighting factor
- Find the sum of these products
- Multiply the sum by the complexity (correction coefficient of complexity) to obtain the answer

The number of function points is then as follows:

$$\begin{aligned} \text{Number of function points} &= (1 * 4 + 2 * 5 + 1 * 10 + 0 * 7 + 0 * 4) * 0.75 \\ &= 18 \end{aligned}$$

Question 8

Difficulty: *** Frequency: ***

- Q8.** Which of the following is the preferred procedure for improving reliability and maintainability in software module design?
- a) Increasing both module strength and coupling
 - b) Increasing module strength while decreasing coupling
 - c) Decreasing module strength while increasing coupling
 - d) Decreasing both module strength and coupling

Answer 8

Correct Answer: b

In module design, increasing the independence of modules should be considered in order to improve their reliability and maintainability. If modules are highly independent, they are unaffected by other modules, thereby enhancing their reliability. Furthermore, their maintainability can be improved because a modification made on one module does not affect the others.

Evaluation criteria to measure the independence of modules include module strength and module coupling.

Module strength measures the level (strength, height) of relations within each module. The stronger the relations within modules are, the more independent the modules are. Module coupling measures the level (strength, height) of relations between modules. The smaller (weaker) the relations between modules are, the more independent the modules are.

Question 9

Difficulty: * Frequency: ***

Q9. Which of the following is an appropriate statement concerning the white box test?

- a) Tests are performed sequentially combining modules from the lower level to the higher level.
- b) Tests are performed sequentially combining modules from the higher level to the lower level.
- c) Tests are performed while paying attention to the internal structure of the module.
- d) Tests are performed to check whether or not functions work according to the specifications, regardless of the internal structure of the modules.

Answer 9

Correct Answer: c

The white box test is a method which focuses on the control flow of the program, prepares the test data going through critical paths of the program, and performs the test. Since the internal structure and the logic of the program are carefully examined, we can test detailed functions from the standpoint of the programmer, but the functions that are in the specifications but are not yet implemented in the program are not selected as test data.

- a) This is an explanation of the bottom-up test.
- b) This is an explanation of the top-down test.
- c) This is an explanation of the black box test.

4 Network Technology

Chapter Objectives

Today, many types of network, such as LANs, WANs, and the Internet, are appearing. In this chapter, we will learn the basic technology concerning information communication networks. In Section 1, we will learn by focusing on protocols. By setting protocols, different types of computer can communicate with one another. In Section 2, we will study specific communication technologies, including how data is sent and received, etc. In Section 3, we will learn the structures and usage of a variety of networks including LANs and the Internet.

- 4.1 Protocols and Transmission Control**
- 4.2 Transmission Technology**
- 4.3 Networks**

[Terms and Concepts to Understand]

TCP/IP, OSI basic reference model, IP address, basic procedures, HDLC, parity check, CRC, bit synchronization (start-stop synchronization), character synchronization, LAN, block synchronization, Internet, CSMA/CD, token passing, inter-LAN connection equipment

4.1 Protocols and Transmission Control

Introduction

In order to enable communication between a sender and a receiver, it is necessary to establish a common set of rules. These rules include communication conventions and transmission control, called protocols.

4.1.1 Network Architectures

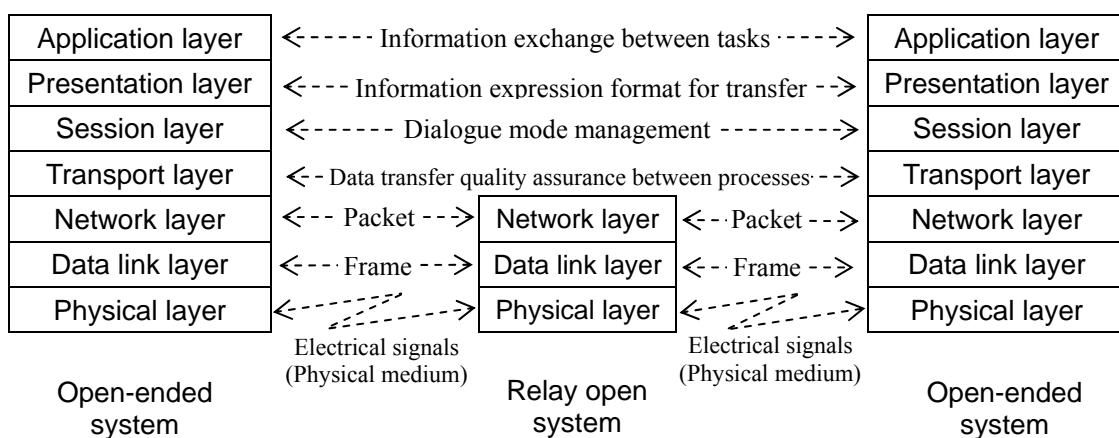
Point

- The OSI basic reference model and TCP/IP are the typical protocols.
- TCP/IP is used on the Internet.

A **network architecture** is a systematically organized form of logical structures and communications protocols¹ to be observed as a standard in a network system.

◆ OSI Basic Reference Model

The **OSI (Open Systems Interconnection) basic reference** model is a model of complex protocols, in which a network is partitioned into seven independent layers from a functional standpoint.²



¹ **Protocol:** It is a set of rules (conventions) for communication. A protocol stipulates the types, semantics, expression formats, and exchange procedures of control messages for communication. Typical protocols include TCP/IP and OSI. Observing a common protocol makes it possible to communicate between different types of computer.

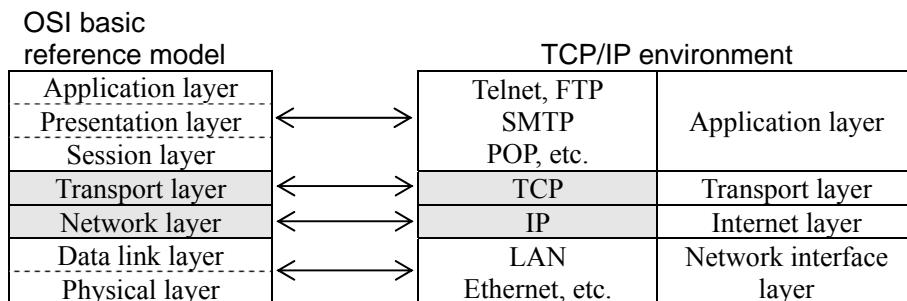
² (FAQ) The roles of each layer of the OSI basic reference model are almost always on the exams. In particular, the functions of the network layer, transport layer, and session layer often appear on the exams.

Layer 7	Application Layer	Deciding data format and contents between users
Layer 6	Presentation Layer	Deciding character set, data format, and data expression format for encryption and compression
Layer 5	Session Layer	Deciding the control methods such as connection and disconnection of lines for proper conversation between users, including the starting and ending of the communication
Layer 4	Transport Layer	Absorbing the difference between communication networks and achieving a communication function that is highly reliable and economical Stipulating the control of detection of transfer errors and their correction on a transmission route
Layer 3	Network Layer	Selecting relays and routes in communication networks to provide network service between terminals
Layer 2	Data link Layer	Stipulating the detection of transmission errors, way of synchronizing, and control of re-sending data so that data can be correctly transmitted
Layer 1	Physical layer	Stipulating the connector shape/type so that terminals can be connected to a communication line, as well as electrical conditions and physical properties for bit transmission

◆ TCP/IP (Transmission Control Protocol/ Internet Protocol)

TCP/IP is the protocol widely used on the Internet and other networks. UNIX workstations are equipped with this protocol as a standard feature. Programs used on the Internet, such as FTP, use services provided by TCP/IP.

The following figure shows the correspondence between the OSI basic reference model and TCP/IP.³



³ (FAQ) The correspondence between TCP/IP and the OSI basic reference model has frequently appeared on past exams. Know that TCP corresponds to the transport layer while IP corresponds to the network layer.

◆ IP Addresses

An IP address is a 32-bit network address used on the Internet and can be classified into several classes according to the network size. Each class is identified by the leading bit pattern of 1 to 3 bits. The network part is unique in the world, and the host part can be systematically defined by each network separately. Below is a schematic figure of how an IP address is structured. Class A has a leading bit of “0,” Class B has two leading bits of “10,” and Class C has three leading bits of “110.”⁴ ⁵ ⁶ ⁷

Class A	0	Network part, 7 bits	Host part, 24 bits	Applied to large networks
Class B	10	Network part, 14 bits	Host part, 16 bits	Applied to medium-size networks
Class C	100	Network part, 21 bits	Host part, 8 bits	Applied to small networks

Since IP addresses identify all computers on the Internet by using 32 bits, it is pointed out that the number of usable IP addresses is insufficient. Hence, 128-bit IP addresses called IPv6 are now in use to a certain extent.

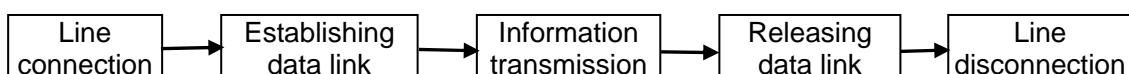
4.1.2 Transmission Control

Point

- The basic procedure is for character transmission.
- HDLC can transmit any bit pattern. (transparent transmission)

Transmission control refers to control used to transmit data between communication devices via a transmission line. Specifically, it includes line control, synchronization control, error control, and data link control.

Transmission control is performed in the following steps:



Establishing a data link means to establish a communication line and to identify the other party (transmission destination). Mutual communication becomes possible only after establishing a data link.

Typical procedures include the basic procedure (BSC) and the HDLC procedure.

⁴ **FTP:** File Transfer Protocol

⁵ **SMTP:** Simple Mail Transfer Protocol

⁶ **POP:** Post Office Protocol

⁷ **Telnet:** It is a virtual terminal protocol for a computer at a remote location.

◆ Basic Procedure (Basic mode data transmission control procedure)

The **basic procedure**⁸ is a procedure of control using 10 transmission control characters. Basically, it transmits characters, and the information being transmitted is called messages. A message contains a special bit pattern called transmission control characters before, in the middle of, or after transmitted data. Below is an example of transmitted data. The text is the transmitted data, consisting of a set of 8-bit character codes. Each transmission control character is also 8 bits long.

S	S	S	S		E	S	S	S	
Y	Y	Y	T	Text	T	Y	Y	Y	STX: Start of TeXt
N	N	N	X		X	N	N	N	ETX: End of TeXt

SYN: SYNchronous idle (idle time for synchronization)⁹
 STX: Start of TeXt
 ETX: End of TeXt

In the basic procedure, synchronization with the transmission destination is performed by attaching several transmission control characters (8 bits long) called “SYN” at the beginning of the text. Later, the receiving party accepts them in 8-bit increments.

To control transmission rights, various methods are used, including the contention method and the polling/selecting method. Data is transmitted in block units while transmission and reception are being verified.

Contention method

The contention method works as follows: between two computers connected point-to-point,¹⁰ one wishing to transmit data sends a transmission request. When a positive response is received from the other party, transmission privilege is given, and data transfer begins.

Polling/selecting method

The polling/selecting method is used in a multi-drop system.¹¹ A host surveys (polls) each terminal in sequence to see if the terminal requests transmission. If so, the terminal is given transmission privilege, and data is received by the host. The host then asks the terminal if reception is possible. If the terminal gives a positive answer (or the terminal is selected), the data is sent.

⁸ (FAQ) The basic procedure is also known as BSC (Binary Synchronous Communication). Many exam questions involve the meaning of polling and selecting in the basic procedure. Know the meanings of these terms well.

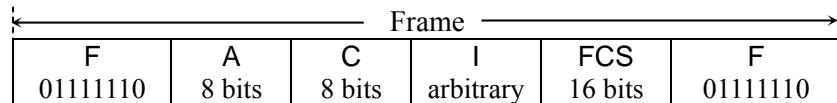
⁹ **Synchronization:** It is required to match the timing of sending and receiving of signals when data is transmitted and received between communication units

¹⁰ **Point-to-point:** It is a two-point system, or direct connection system. Two or more terminals are connected to a computer, and each terminal has a dedicated line.

¹¹ **Multi-drop system:** Multiple terminals are connected to a single line. A “control station” manages data communication with all terminals, and this station controls all sub-stations (terminals) centrally.

◆ HDLC Protocol (High-level Data Link Control)

HDLC is a transmission control procedure aiming to achieve highly efficient and reliable data transmission between computers. Transmission is performed by blocks of data called frames. The mechanism is as shown below.



F	Flag sequence: A bit string showing the beginning and the end of a frame
A	Address field: Address of the transmission destination
C	Control field: Various control information
I	Information field: Data transmitted
FCS	Frame check sequence: Check bit by the CRC method ¹² using A through I

HDLC has the following characteristics:¹³

- Bit-oriented (possible to transmit an arbitrary bit pattern)¹⁴
- Continuous transfer (possible to transmit without getting a response within the limits of the certain number of frames)
- Strict error check (using CRC)
- Full duplex communication (cf. Sec. 4.2.3) is possible even in multi-drop lines.

¹² **CRC (Cyclic Redundancy Check):** It is a code used to detect an error in one block of data

¹³ (Note) In HDLC, the bit “0” is inserted whenever there are at least five consecutive 1's. In so doing, it ensures that no bit pattern is identical to the flag sequence. For instance, if a data sequence is “0111110,” the bit “0” is inserted so the sequence becomes “01111010.”

¹⁴ (FAQ) There are exam questions concerning the roles of each field of HDLC and characteristics of HDLC. Be sure to know that HDLC is bit-oriented (anything can be sent).

Quiz

Q1 Show the correspondence between the OSI basic reference model and TCP/IP.

Q2 List the characteristics of HDLC.

4.2

Transmission Technology

Introduction

Transmission technology is used to transmit data at high speed, efficiency, and quality. More specifically, it includes technology in error control, synchronization control, and duplexing.

4.2.1 Error Control

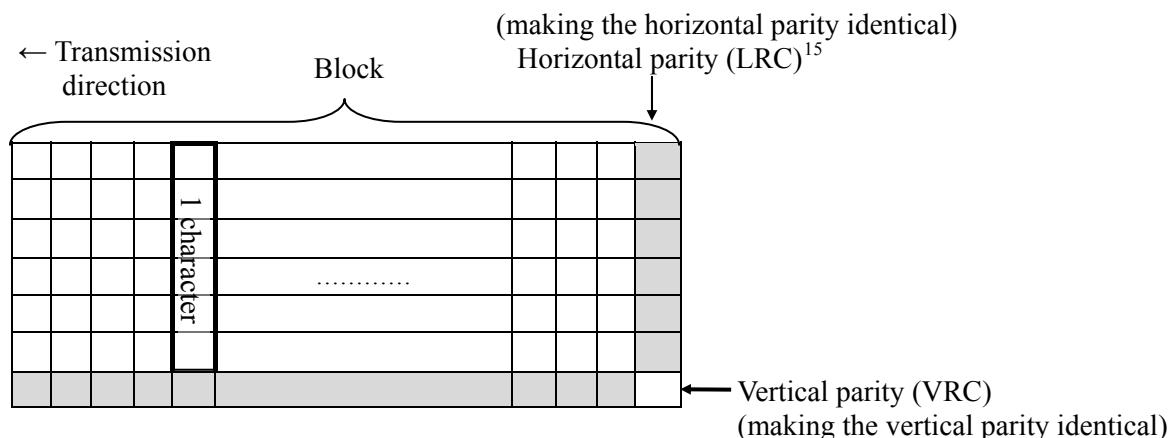
Point

- Methods of error control include parity check and CRC.
- CRC is a high-performance error detection method used in HDLC and other protocols.

Error control refers to improving the quality of data transmission through detecting errors in data transmission and, in some cases, correcting errors. Typical checking methods include **parity check** and **CRC**.

◆ Parity Check Method

Parity check is the error detection method in which the number of 1's is set to be even or odd by adding one bit, horizontally or vertically, to characters transmitted in binary code. Making the number of 1's even is called **even parity check** while making it odd is called **odd parity check**.

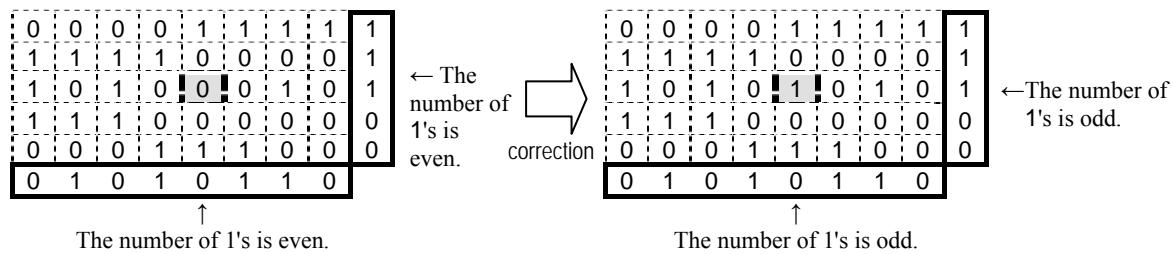


Characteristics of parity check combining LRC and VRC are as follows:

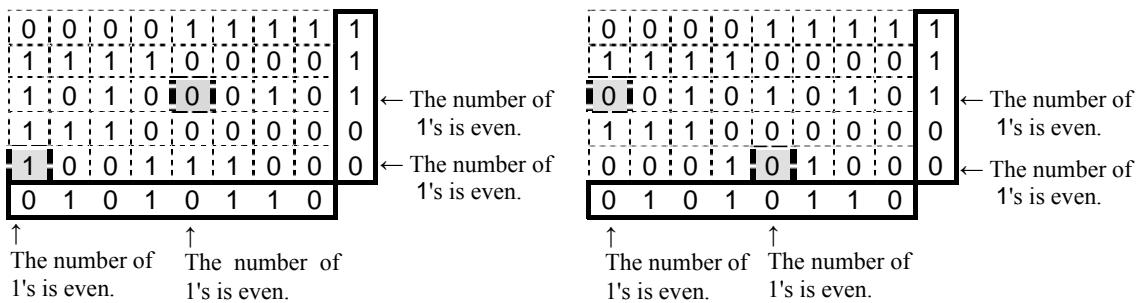
- 1-bit errors can be detected and corrected.
- 2-bit errors can be detected but cannot be corrected.

¹⁵ **LRC/VRC:** Parity check applied to each string of bits in the same horizontal position of each character (horizontal parity check) is called LRC (Longitudinal Redundancy Check); parity check applied to each character in the vertical direction (vertical parity check) is called VRC (Vertical Redundancy Check).

Below is a figure where the bits in the shaded area are erroneous in odd parity. Normally, the number of 1's should be odd, but here it is even, indicating that there is an error.¹⁶



If 2 bits are erroneous, as shown below, the number of 1's is even while it should be odd, both in horizontal and vertical parities. However, there are two possible combinations of errors, making it impossible to correct them.



A code in which a bit is added for error detection is called a Humming code.¹⁷

◆ CRC (Cyclic Redundancy Check)

CRC is a method of using the remainder resulting from division by a certain polynomial as the check bit. For each transmission unit, the bit string is considered a binary number. Take a polynomial, established in advance ($X^{16} + X^{12} + X^5 + 1$ is recommended by ITU-T),¹⁸ divide the binary number by this polynomial, and get the remainder, which is used as the check bit and added to the end of the transmission unit. The receiving party divides the transmitted information by the same polynomial and, if the remainder is 0, determines that there is no error. This method is effective in detecting errors in a block (one piece), burst errors (consecutive bit errors), and random errors (errors without patterns).

¹⁶ (FAQ) Questions concerning parity check do appear on the exams, like “Which bit column has erroneous data if odd parity is used?” These are very easy questions since all we have to do is to count the number of 1's.

¹⁷ Humming code: It is a code in which a check bit is added to the information bits is generally called a Hamming code. Not only can it detect errors, but it can also correct them. Parity check is a specific example of Hamming code check.

¹⁸ **ITU-T** (International Telecommunications Union-Telecommunications Standardization Sector): As one sector of the ITU, this organization considers technology, operation, and fees concerning telecommunications, prepares standards, and issues the standards as recommendations.

4.2.2 Synchronization Control

Point

- There are two types of synchronization: asynchronous method and synchronous method.
- In the asynchronous method, there are two more bits for each character.

To send and receive data correctly, the sender and the receiver adjust the timing of transmission; this is referred to as **synchronization**. The computers or terminals of the sender and the receiver must perform synchronization according to the data contents.

There are a few types of synchronization method, depending on how it is performed.

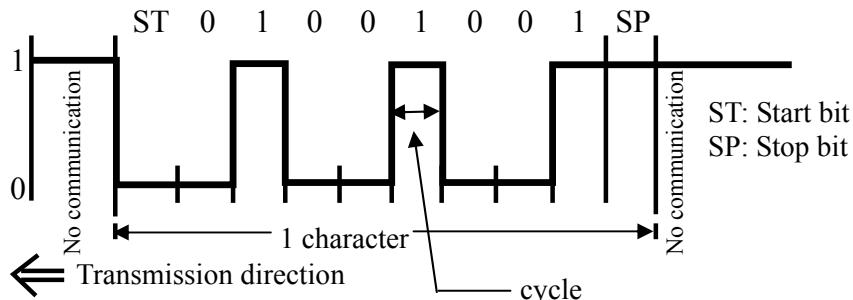
Synchronization methods	Bit synchronization	Synchronizing by bit units
	Character synchronization	Synchronizing by character units (synchronizing by SYN code)
	Block synchronization	Synchronizing by block units using flag sequences

◆ Bit Synchronization (Asynchronous)

Bit synchronization is a synchronization method which designates a start bit indicating the beginning of data (one character) and a stop bit indicating the end of the data.¹⁹ It is also called start/stop synchronization method. Because of the two extra bits, each character will require 10 bits, 2 more than the conventional expression. The start bit is expressed by “0,” and the stop bit by “1.”

The line is always in the condition of “1,” identical to the stop bit. When the start bit “0” is received, reception takes place with a certain cycle. For this reason, the cycle must be determined between the sender and the receiver in advance.

Here is an example when the 8-bit character “01001001” is received.

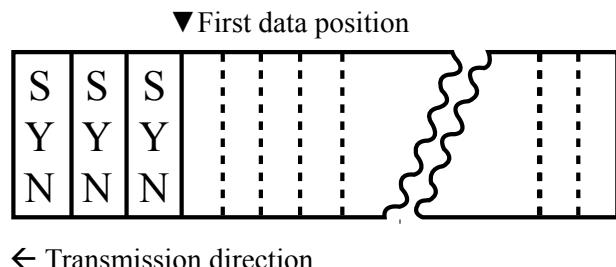


Bit synchronization adds a start bit and a stop bit for each character, so the overall transmission efficiency is quite low, but it is used in low-speed terminals because the mechanism is simple.

¹⁹ (Hints & Tips) Bit synchronization is sometimes called the asynchronous method or start-stop synchronization. As a means of synchronization, this method uses the so-called “asynchronous” method, which does NOT mean “not synchronizing.” Be careful not to misinterpret this term.

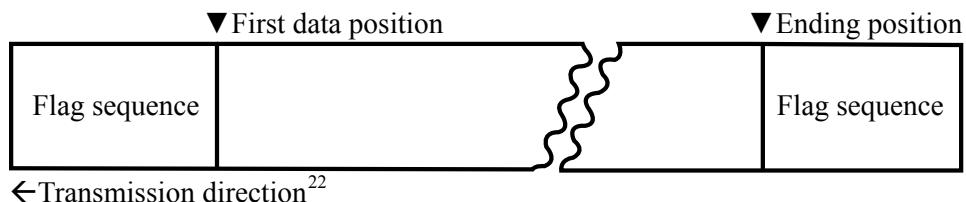
◆ Character Synchronization (Synchronous)

Character synchronization is the method in which an SYN code (10010110_2) is placed before a data block as a synchronization code.²⁰ The SYN code is sent consecutively multiple times just to be sure. At the receiving end, when the SYN code is received, the following bits are divided into 8-bit units, each of which is recognized as a character.



◆ Block Synchronization (Synchronous)

In block synchronization, a special bit string for synchronization is sent attached to the front and the end of a series of transmitted blocks.²¹ This bit string is called a flag sequence, indicating the first and the last positions of the transmission block. Hence, regardless of the character boundaries, data with a flexible number of bits can be sent. Block synchronization is more efficient than character synchronization, so it is used in terminals that perform high-speed transmission. It is used in the HDLC procedure.



²⁰ (Note) Character synchronization is also called the continuous synchronization method or the SYN synchronization method. Since the SYN code is formed with 8 bits, the same number as for a character, the data following the SYN is received in units of 8 bits. This system is used in mid- to high-speed terminals. This method is the synchronization method used in the basic procedure.

²¹ (Note) Block synchronization is also called flag synchronization or frame synchronization. In HDLC, the bit pattern "01111110" is used as the flag sequence.

²² (FAQ) Questions concerning bit synchronization are frequently seen on the exams. Remember that the first bit is "0" and the last bit is "1" for each character. Further, there have been exam questions that give the number of bytes (number of characters) of data as well as the line speed and ask you how many seconds it takes for the data to be transmitted. In bit synchronization, a start bit and a stop bit are added to each character, so remember that each character takes 10 bits.

4.2.3 Multiplexing and Communications

Point

- FDM and TDM are the basic types of multiplexing.
- There are three transmission methods: simplex, half-duplex, and full-duplex.

Multiplexing refers to the communications among multiple computers through one transmission line simultaneously. We can reduce the communication costs by using one high-speed line by multiplexing it into multiple low-speed lines. There are three transmission methods: simplex, half-duplex, and full-duplex, depending on the types of data flow.

◆ Multiplexing Methods

There are two types of multiplexing: **FDM** and **TDM**.

FDM (Frequency Division Multiplexing)

FDM is the method of multiplexing with a frequency division multiplexer²³ to divide the transmission frequency bandwidth of an analog line into multiple small bands and to use each channel as an independent communication channel. For example, a line with a bandwidth of 48 kHz may be partitioned into 12 channels, each of which has a bandwidth of 4 kHz, so that they can be used as 12 telephone lines. Each of the divided channels can then be used for either analog transmission or digital transmission. In digital mobile phones and digital television broadcasting, digital transmission is performed in communication channels resulting from frequency partition.

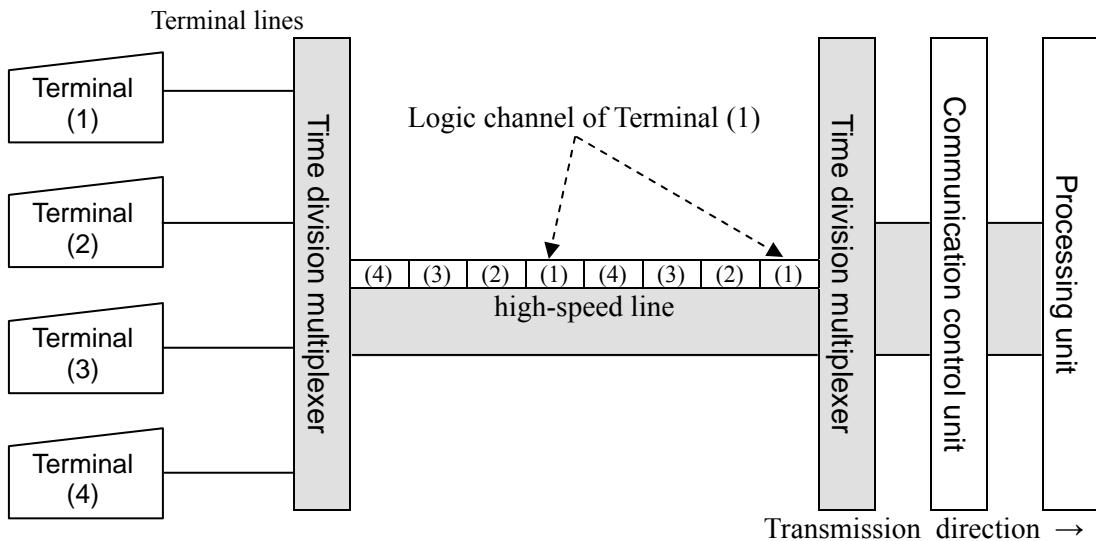
TDM (Time Division Multiplexing)

TDM is the scheme of dividing one digital line into multiple low-speed channels. For instance, if a line whose speed is 64 Kbps is connected to 16 terminals, then each terminal has a speed of up to 4 Kbps.

In TDM, one digital line is partitioned by time, and transmission and reception alternate (get switched) in time intervals of certain length. This switching unit is called a TDM (Time Division Multiplexer).²⁴

²³ **Frequency Division Multiplexer (FDM):** It is a multiplexing unit used for frequency division multiplexing.

²⁴ **Time Division Multiplexer (TDM):** It is a unit used to partition one digital transmission line by time so that the line can be used as multiple communication channels.



◆ WDM (Wavelength Division Multiplexing)

Whereas optical fibers can provide high-speed transmission (several Gbps), optical signals of one wavelength have the disadvantage of not being capable of bidirectional transmission. **WDM** eliminates this disadvantage; it is the method of transmitting multiple optical signals with different wavelengths on one optical fiber.²⁵

For instance, if a channel with a transmission speed of 2.5Gbps per wavelength is multiplexed into 4 channels, transmission at a total speed of 10Gbps can be achieved.²⁶

◆ Transmission Methods

Transmission can be classified into three methods by the way data flow; they are **simplex**, **half-duplex**, and **full-duplex**. One transmission line consists of a pair of two communication media; it is called the two-wire system. There is another system, called the four-wire system, in which there are two pairs of communication lines (4 media): one pair for sending, and the other for receiving. In general, the four-wire system is used for full-duplex while the two-wire system is used for half-duplex.²⁷

Transmission methods	<table border="1" style="width: 100%; border-collapse: collapse;"> <tr> <td style="padding: 5px;">Simple</td><td style="padding: 5px;">Communication where the data flow in one direction only</td></tr> <tr> <td style="padding: 5px;">Half-duplex</td><td style="padding: 5px;">Communication where sending and receiving occur alternately and repeatedly</td></tr> <tr> <td style="padding: 5px;">Full-duplex</td><td style="padding: 5px;">Communication where sending and receiving can occur simultaneously</td></tr> </table>	Simple	Communication where the data flow in one direction only	Half-duplex	Communication where sending and receiving occur alternately and repeatedly	Full-duplex	Communication where sending and receiving can occur simultaneously
Simple	Communication where the data flow in one direction only						
Half-duplex	Communication where sending and receiving occur alternately and repeatedly						
Full-duplex	Communication where sending and receiving can occur simultaneously						

²⁵ **DWDM:** The DWDM (Dense WDM) technology is an area of current research; it is a way to achieve even higher-capacity data transmission by increasing the number of wavelengths of WDM or narrowing the gaps between channels. It is said that using DWDM, super-high capacity data transmission, replacing Gbps with Tbps (terabytes per second, where one tera is 10^{12}) is possible.

²⁶ (FAQ) There seem to be no new exam questions on FDM and TDM available, as they have been used up in past exams. Any question on TDM can be answered as long as you know that multiple logic channels can be used because of time-partition of one line. Future exam questions will more than likely involve WDM.

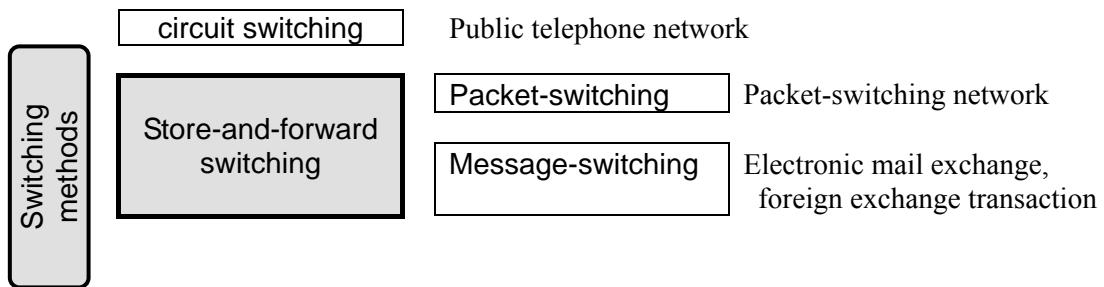
²⁷ (Note) Multiplexing enables a two-wire system to be used for full-duplex communication.

4.2.4 Switching

Point

- There are two types of switching: circuit switching and store-and-forward switching.
- There are two types of store-and-forward switching: packet switching and message switching.

The line to be used in communication differs depending on whether or not the party with whom we are communicating is fixed. If the party is fixed, a dedicated circuit²⁸ is used. If the party changes, a switching circuit is used, as represented by the public telephone network.



◆ Circuit Switching

Under circuit switching, the transmitter calls up the other party by dialing to set up a physical circuit, as represented by the telephone service. This enables high-speed and high-quality data transfer, but both parties are required to use the same speed and same transmission control system.

◆ Store-and-Forward Switching

Under store-and-forward switching, the transmitted data is first stored in a switching unit, the receiver is selected, and then the stored data is transferred to the next switching unit or to DTE.²⁹ Although the transmission speed and quality are poorer than those of circuit switching systems, it is not necessary that the transmitter and the receiver have the same speed nor that they use the same transmission control system. It is suitable when the amount of data transmitted at one time is small and when the communication traffic is light.

Among store-and-forward switching systems, there are message exchange systems, where storing and switching occur in message units, and packet exchange systems, where messages are partitioned into packets of a fixed size and transferred in packet units.

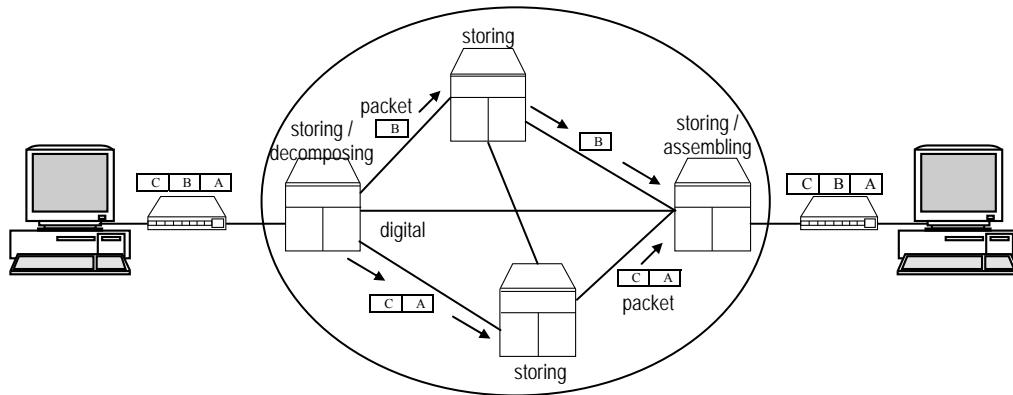
In **message exchange**, generally the message contents are transmitted without modification. For instance, this is used in electronic mails on the Internet and foreign exchange dealing systems between banks.

²⁸ **Dedicated line:** It is a communication line which is set up between communication points desired by the users and can be used exclusively by these users. Generally the fees for dedicated lines are charged on a monthly basis, determined by the communication distance and transmission speed. There are analog dedicated lines stipulated by frequency bands and digital dedicated lines stipulated data transmission speed.

²⁹ **DTE (Data Terminal Equipment):** It is a unit that has the functions of a data transmitter or a data receiver or both and is equipped with the data communication function. In general, these include computers and terminals that can be connected to modems (modulator-demodulators).

In **packet exchange**, data is divided into packets³⁰ of certain size (a block of data); then to each packet, the forwarding address, data attributes, and error check codes are added before the packet is transmitted onto the communication medium. Since the lines are not exclusive to any user except when the data is actually being transmitted or received, the channels can be multiplexed, and the lines can then be used efficiently.³¹

Packet switching network



Quiz

- Q1** List the methods of synchronization control.
- Q2** Describe the characteristics of packet exchange.

³⁰ **Packet:** In data communication, it is a block of data along with added control information such as the forwarding address. By transmitting and receiving data by partitioning them into multiple packets, one prevents intermediate communication lines between the two locations from being exclusively used, resulting in more efficient use of the communication circuits. Further, since the route can be selected flexibly, when a part of one line fails, another route can be used as a replacement.

³¹ (FAQ) Questions on packet exchange will appear on the exams. Know that communication is possible between different computers and terminals with different speeds.

4.3 Networks

Introduction

A network is a collective term referring to a connecting organization. An information communication network consists of communication lines for data transmission and nodes that connect these communication lines. A LAN is a small-scale network whereas the Internet is a large-scale network.

4.3.1 LANs

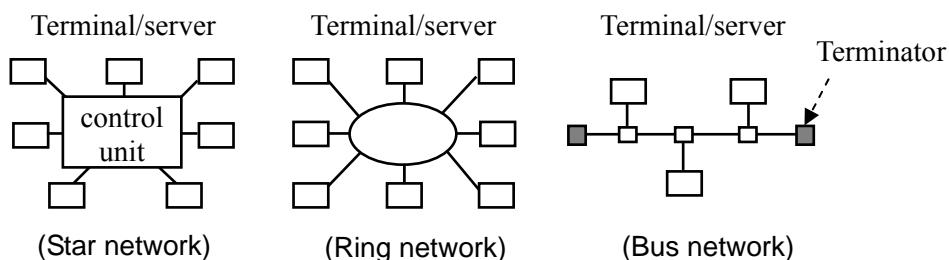
Point

- The connection topologies of LAN include star, bus, and ring.
- The access control methods of LAN include CSMA/CD and token-passing.

LAN stands for “Local Area Network.” It is a network connecting various units that are spread out over a relatively small area, such as within one building or site.

◆ Topology of LAN

The word “topology” here refers to a connection configuration of a network. Typical topologies of LAN include the star, ring, and bus networks.³²



◆ Access Control of LAN

Access control methods of LAN can be classified into the following. The bus and ring networks have only one transmission channel, so it is necessary to control communication to prevent collision between the transmitted signals.

³² (Note) **Star network:** Terminals are connected to the unit that controls communication.

Ring network: Terminals are connected to form a ring (circle).

Bus network: Terminals are connected to transmission routes called buses.

CSMA/CD (Carrier Sense Multiple Access with Collision Detection)

The computer which is about to transmit data checks whether or not data is being transferred on the transmission channel and then sends the data. If data is being transferred, the computer waits for a certain amount of time and then re-sends the data. This method is used for a bus-type or a star-type network. If the network is already busy (in use) when another set of data is to be transmitted, we say that a collision has occurred.

Token passing

In this method, control information called a token is circulated in a certain direction on LAN. The computer that receives the token gets the transmission privilege, adds the destination address and the data to the token, and sends them out. This is used for a ring-type or a bus-type LAN.³³

◆ Specifications and Transmission Media for LAN

Concerning the transmission media (cables) for LAN, there are several specifications including the 10BASE established by the IEEE802 Committee and the FDDI (Fiber Distributed Data Interface) established by the ANSI.

LAN standards	Medium	Transmission speed	Topology	Maximum length	Control method	Remarks	
10BASE2	Thin coaxial	10Mbps	Bus	185m	CSMA/CD	Small-scale LAN	
10BASE5	Standard coaxial			500m		Backbone LAN	
10BASE-T	Twisted pair cable		Star ³⁴	100m		Maximum 4 stages	
10BASE-F	Optical fiber			2km		Maximum 22 stages	
100BASE-T	Twisted pair cable			100m		T2, T4, TX	
100BASE-FX	Optical fiber			Maximum 20km		Good quality	
1000BASE-X	Coaxial cable	1000Mbps (1Gbps)		25m		1000BASE-CX	
	Optical fiber			Maximum 5km		LX, SX	
1000BASE-T	Twisted pair cable			100m		Maximum 2 stages	
FDDI	Optic fiber	100Mbps	Ring	200km	Token passing	Backbone LAN	

The maximum length is the length of the cable between the two terminators in bus-type LAN; the length of the ring in ring-type LAN; and the maximum transmission distance in star-type LAN. The maximum length of FDDI is stated as 200km, but in ring-type LAN, sometimes the cables are doubled up as a precaution against failures. In such a case, the maximum length will be 100km.

◆ Wireless LAN

Wireless LAN uses transmission channels other than cables, such as radio waves and infrared rays. Most of the cables can be eliminated, so it hardly takes any labor to install or move terminals. However, there are limitations in speed and distance, and it may be affected by interference from electro-magnetic noise generated by other devices. Other disadvantages include the high cost per terminal.³⁵

³³ (Note) When the token-passing method is applied to a ring-type LAN, it is called the token ring method; if it is applied to a bus-type LAN, it is called the token bus method. In the token passing method, it is necessary to decide the order in which the token is circulated.

³⁴ (Hints & Tips) Note that 10BASE-T, etc. is a star-type LAN. The device that plays the role of the control unit is called a hub.

³⁵ (Note) Specifications for wireless LAN, established by the IEEE802 Committee, include IEEE802.11a, IEEE802.11b, etc.

4.3.2 The Internet

Point

- The Internet is a network composed of existing networks connected mutually.
- TCP/IP is a protocol predominately used on the Internet.

The term **Internet** means “a network of networks” and is a global scale network of organizations. For the protocol, TCP/IP is used, and communication is based on IP addresses.^{36 37} Intranets and extranets using Internet technologies have also been widely used.

◆ WWW (World Wide Web)

The **WWW** (Web) is the concept of creating a gigantic information space by mutually connecting information spread apart on the Internet like a spider “web.” The links of information on the WWW are accomplished by hypertext. If there is a link within a text, further information can be reached, and, consequently, all computers in the world should be accessible.

The WWW provides mechanisms such as hypertext mentioned above. To view the contents, we need browsing software called a WWW browser, such as Internet Explorer and Firefox.

◆ Internet Services

The Internet uses TCP/IP, so most of the services that can be used with TCP/IP are available on the Internet. Main services are shown the following table.

Name	Explanation
Telnet	Standard protocol for virtual terminals Used to interact with remote computers
FTP	File Transfer Protocol Standard protocol for transferring files Both text files and binary files can be transmitted.
Electronic mail	Function by which the user can send/receive messages to/from one or more people Transmission is possible even when the other party is not connected to a computer. However, for sending and receiving messages, a mail address is required. Protocols used for electronic mails include SMTP and POP3. ^{38 39}

³⁶ (Note) On the Internet, to identify a network or a terminal, a 32-bit IP address (IPv4) is commonly used, but each of these must be unique in the whole world. As the Internet gains popularity, running out of IP addresses is a real issue. Currently, 128-bit addresses called IPv6 are also in use.

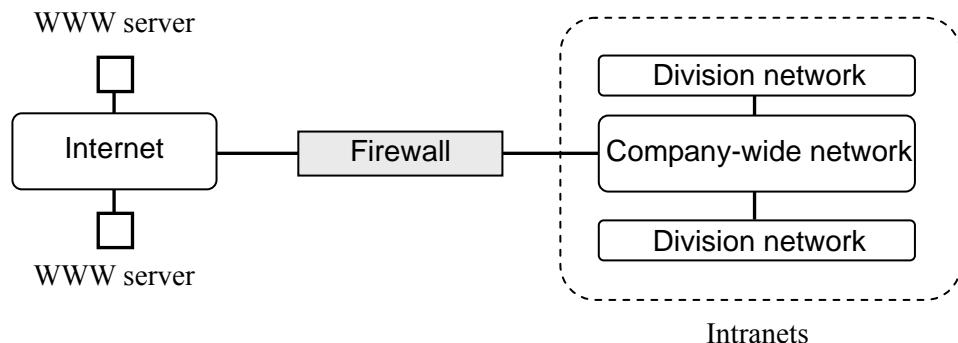
³⁷ One of the ways to address the insufficiency of IP addresses is DHCP (Dynamic Host Configuration Protocol). DHCP is a protocol that automatically assigns IP addresses and necessary information to computers that are temporarily connected to the Internet. When the communication is over, the IP address is automatically collected, and the same IP address is assigned to another computer.

³⁸ **SMTP/POP3:** SMTP is the protocol for sending e-mails, and POP3 is for receiving e-mails. POP3 is the latest version of POP.

³⁹ (FAQ) Frequently there are exam questions on SMTP and POP. Remember that SMTP is a protocol for transmitting e-mails while POP is for receiving them.

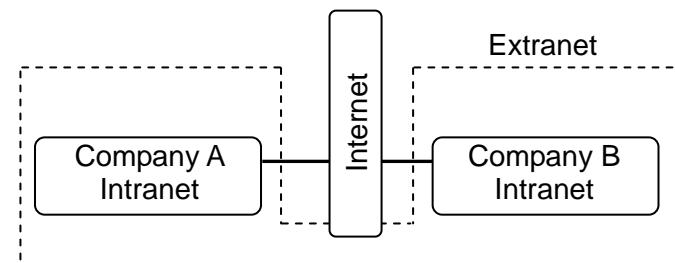
◆ Intranets

An **intranet** is an in-house (company-wide) network using Internet technologies. Normally, between an in-house network and the Internet, a defense system called a **firewall** is installed to prevent critical in-house information from leakage. With the popularity of the Internet and introduction of user-friendly WWW browsers, it is now possible to construct systems such as document sharing, electronic bulletin boards, and electronic mail systems at low costs.



◆ Extranets

An **extranet** is a network in which intranets are extended between companies. In general, intranets are connected to the Internet to construct an extranet.



◆ HTTP (HyperText Transfer Protocol)

HTTP is the communication protocol for sending and receiving HTML documents between a WWW server and a WWW client on the Internet. The WWW client sends URL⁴⁰ of the HTML document⁴¹ it wishes to post. In response, the HTML document possessed by the WWW server is sent to the WWW client.

⁴⁰ **URL** (Uniform Resource Locator): This is the information that identifies the location of a homepage on the Web, consisting of a protocol name, host name, file name, etc.

⁴¹ **HTML (HyperText Markup Language)**: It is the means to write a document in the hypertext format. It uses reserved words contained between “<” and “>” called tags to specify the text formatting, image file display position, link designation, and script declaration. If this is opened using a WWW browser, the browser interprets and displays its contents. To specify the address of a WWW server, URL (Uniform Resource Locator) is used.

4.3.3 Various Communication Units

Point

- Connectors between LANs include routers, bridges, repeaters, and gateways.
- Modems are used for analog communication while DSU and TA are used for digital communication.

A variety of communication units are necessary to carry out communication. To connect multiple LANs, appropriate units are installed according to their purpose. In a data communication system, units to be used differ depending on whether the system uses an analog line or a digital line.

◆ Connection Units between LANs

A **connection unit between LANs** is a device mutually connecting multiple LANs or networks of different protocols. The following figure shows the correspondence between each unit and the OSI basic reference model⁴²:

OSI reference model	Connection units between LANs
Application layer	
Presentation layer	
Session layer	Gateway
Transport layer	
Network layer	Router
Data link layer	Bridge
Physical layer	Repeater
	Hub

```

graph LR
    subgraph OSI_Layers [OSI reference model]
        direction TB
        L1[Application layer] --- L2[Presentation layer]
        L2 --- L3[Session layer]
        L3 --- L4[Transport layer]
        L4 --- L5[Network layer]
        L5 --- L6[Data link layer]
        L6 --- L7[Physical layer]
    end

    subgraph Connection_Units [Connection units between LANs]
        direction TB
        R1[Repeater] --- LAN1[LAN]
        R1 --- LAN2[LAN]
        BR1[Bridge/router] --- LAN3[LAN]
        BR1 --- LAN4[LAN]
        R2[Router] --- LAN5[LAN]
        R2 --- LAN6[LAN]
        H1[Hub] --- LAN7[LAN]
        H1 --- LAN8[LAN]
        H1 --- LAN9[LAN]
        H1 --- LAN10[LAN]
    end

    LAN1 --- R1
    LAN2 --- R1
    LAN3 --- BR1
    LAN4 --- BR1
    LAN5 --- R2
    LAN6 --- R2
    LAN7 --- H1
    LAN8 --- H1
    LAN9 --- H1
    LAN10 --- H1

    BackboneLAN[Backbone LAN] --- H1

```

⁴² (FAQ) The correspondence between the OSI basic reference model and connection units between LANs appears often on the exams. Be sure to know that the router corresponds to the network layer, the bridge to the data link layer, and the repeater to the physical layer.

⁴³ **Filtering:** It is the function whereby the system, based on the transmitter's address, decides whether or not to accept the packet (allow it through) and discards unnecessary packets. By the filtering function, extraneous packets are prevented from entering LAN.

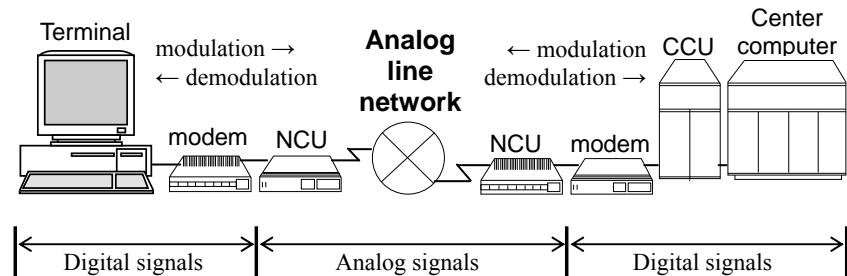
⁴⁴ **MAC address:** It is a 48-bit (6-byte) device number assigned to a LAN card used when a terminal is connected to a network. In principle, there are no two cards in the world with the same MAC address.

⁴⁵ (Hints & Tips) A hub that relays packets with the protocol of the data link layer is called a switching hub. Meanwhile, a hub that relays packets with the protocol of the physical layer is called a repeater hub.

⁴⁶ **Backbone LAN / Branch LAN:** The backbone LAN refers to the transmission routes constituting the main section of the network with in an organization. For the transmission medium, optical fiber cables are used, so the communication is high-speed and high-capacity. This plays the role of connecting two or more branch LANs. A branch LAN is LAN set up for a division or a department of the organization. It is mid- to small-scale, and it is LAN for communication between workstations, PC communications, and file/printer sharing in a system spread out on the premises.

◆ Communication Units for Analog Lines

A **communication unit for analog lines** is used in a data communication system that utilizes the public telephone network (analog) as its transmission route. Since computers are digital and the public telephone network is analog, mutual conversion between digital and analog is needed. Since it is an exchange circuit, the dialing function is also necessary. Communication units for analog lines are shown in the figure below.

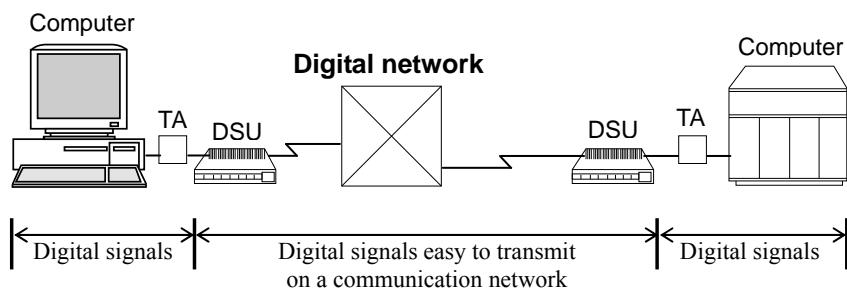


Unit name	Explanation
Modem	MODEM (MOdulation DEModulation): A device that converts digital to analog (modulation) and analog to digital (demodulation)
NCU	Network Control Unit: A device with the function of making phone calls to other parties
CCU	Communication Control Unit: A device for transmission control, error control, decomposing and assembling of transmitted/received characters

◆ Communication Units for Digital Lines

A **communication unit for a digital line** is a transmission system using a digital line as its transmission route. Unlike analog lines, a modem is unnecessary. Instead, what is needed is **DSU** (Digital Service Unit), which converts the digital signals inside the computer into a form easy to transmit on a digital line.

The following figure shows communication units for digital lines.



A unit called TA (Terminal Adapter) may be required between DSU and the terminal.⁴⁷ TA allows telephones, fax machines, and PCs, which have traditionally been used on analog lines, to operate on ISDN lines. Most of the time, TAs are necessary.

⁴⁷ (Hints & Tips) DSU is often installed inside TA and thus is not directly visible.

4.3.4 Telecommunications Services

Point	<ul style="list-style-type: none"> ➤ The ATM performs transmission and reception in units of 53-byte cells.
--------------	--

◆ ISDN (Integrated Services Digital Network)

ISDN is a communications network integrating a variety of services including telephone, data, and fax services. It provides the basic rate interface and the primary rate interface. The basic rate interface can be used on existing telephone lines, but the primary rate interface uses optical fibers. The characteristics of each are shown below.⁴⁸

Name	Explanation
Basic rate interface	consisting of two B channels and one D channel (2B+D); maximum 144Kbps
Primary rate interface	consisting of multiple B channels and one D channel (23B+D, 24B, 4H0, etc.); maximum 1,536Kbps ⁴⁹

Having multiple channels means the option of having multiple lines. For instance, in the basic rate interface, there are two B channels, so they can be used as two lines of 64kbps each or one line of 128Kbps. The details of each channel are shown below.

Name	Explanation	
D channel	Signal channel for control information Can be used as a B channel in packet switching	Basic rate interface: 16Kbps Primary rate interface: 64Kbps ⁵⁰
B channel	User information channel	64Kbps
H channel	User information channel exceeding 64Kbps	H0 (384Kbps) H11 (1,536Kbps) H12 (1,920Kbps) ^{51 52}

◆ ATM (Asynchronous Transfer Mode)

ATM partitions all information into cells of fixed length (53 bytes) for transmission and reception. With the assumption that a high-quality line is used, this mode has achieved high speeds by simplifying protocols such as error control and performing the partitioning process on the hardware.

⁴⁸ (FAQ) There will be exam questions on characteristics of the basic rate interface of an ISDN. Remember that there are two B channels and one D channel in addition to the fact that the D channel is 16kbps.

⁴⁹ (Hints & Tips) The maximum transmission speed is the total speed of all channels. In the basic rate interface, the total is 144kbps because there are two B channels, each with 64kbps, and one D channel with 16kbps.

⁵⁰ (Hints & Tips) In the primary rate interface, D channels are not required, such as in 24B. However, as a whole, at least one D channel is necessary. For instance, if two primary rate interfaces are leased (contracted), the structure needs to be 24B and 23B+D.

⁵¹ (Hints & Tips) In the primary rate interface, one line will be 1,536kbps using the H11 channel. H12 is used in Europe while H11 is used in Japan and the U.S.

⁵² **bps (bits per second):** It means the number of bits that can be transmitted in one second.

◆ **ADSL**

ADSL (Asymmetric Digital Subscriber Line) is the technology for high-speed data transfer using existing telephone lines. This can be used simply by connecting an ADSL modem to the conventional equipment. The speeds are 0.5M to 1Mbps upstream and 1.5M to 40Mbps downstream. The transmission speeds upstream and downstream differ in this “asymmetric” digital subscriber line. It shows its power in downloading massive data such as video-on-demand and Web pages containing video data.

Quiz

- Q1** Explain the CSMA/CD method.
- Q2** List three basic services of the Internet and explain each.
- Q3** Describe the channel structure of the basic interface of ISDN.

Question 1

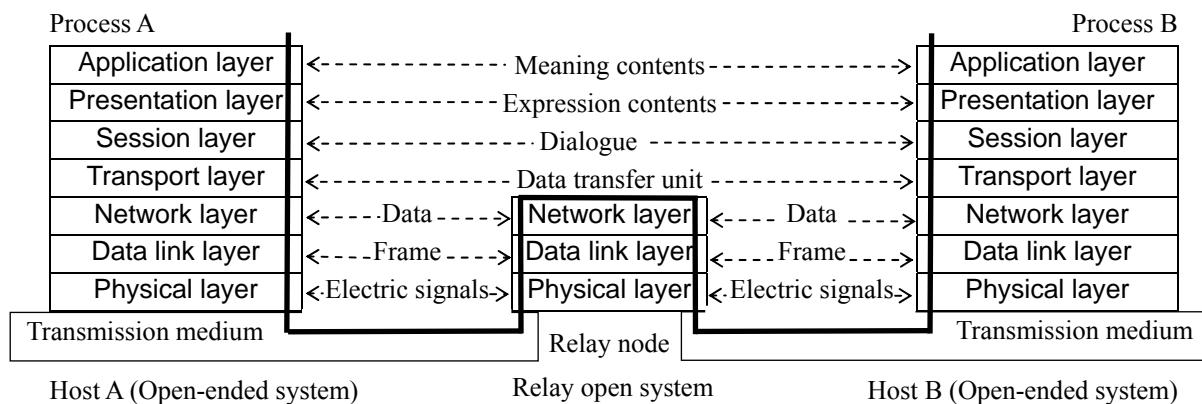
Difficulty: **

Frequency: ***

- Q1.** Which of the following is an appropriate description concerning the network layer of the OSI basic reference model?
- The network layer performs the routing and relaying so that data can be transferred between end-systems.
 - Among the various layers, the network layer is closest to the user and provides functions such as file transfer and e-mail.
 - The network layer absorbs the differences in characteristics of physical communication media and provides a transparent transmission route to upper-level layers.
 - The network layer provides a transmission control protocol (error check, re-transmission control, etc.) between adjacent nodes.

Answer 1**Correct Answer:** a

OSI (Open Systems Interconnection) is structured by partitioning a data communication system by function into seven independent layers in order to simplify connection between different models of computers and between networks. OSI merely provides a basic framework; the 7-layer OSI basic reference model is given as a guideline.



The network layer stipulates the method of selecting the communication route and the relay method. It models a communication network; it selects the communication route between end nodes, relays data, and sends along them. It is processed by protocols such as the X.25 protocol in switching functions such as packet switching and circuit switching. The function of the network layer is, therefore, to select the route to the other computer.

- This is an explanation of the application layer.
- This is an explanation of the physical layer.
- This is an explanation of the data link layer.

Question 2

Difficulty: *

Frequency: ***

Q2. Which of the following protocols is used to automatically set up the IP address that a PC uses to connect to LAN at startup time?

- a) DHCP
- b) FTP
- c) PPP
- d) SMTP

Answer 2

Correct Answer: a

DHCP (Dynamic Host Configuration Protocol) is a protocol that automatically sets up network parameters. When terminals (clients) start up, an IP address is dynamically assigned to each client, and when the session ends, the assigned IP addresses are collected.

- b) FTP (File Transfer Protocol) is a protocol for transferring files on a TCP/IP network.
- c) PPP (Point-to-Point Protocol) is a protocol for WAN used for network connection, not necessarily on TCP/IP. Generally, PPP is used for dial-up connection to the Internet; the user does not need to obtain an IP address.
- d) SMTP (Simple Mail Transfer Protocol) is a protocol for sending and receiving electronic mails between mail servers on a TCP/IP network. This is also used when an electronic mail is sent from a mail client (terminal) to the mail server.

Question 3

Difficulty: **

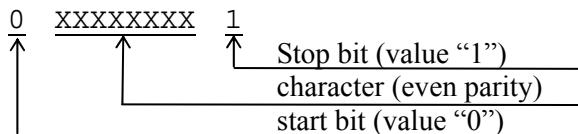
Frequency: *

- Q3.** The character “T” (ASCII code 1010100) was sent via a data transfer using start/stop synchronization with even-parity error detection. If the character is received correctly, what is the bit string that is received? Here, the bits are sent in the following order: start bit (0); the character code, from the least significant bit to the most significant bit; parity bit; and stop bit (1). The bits are written in the sequence in which they are received, starting from the left.

- a) 0001010101 b) 0001010111 c) 1001010110 d) 1001010111

Answer 3**Correct Answer:** b

Since the character length is 7 bits and one parity bit is added, a character is 8 bits long. The start bit “0” is also added before the bit string for the character, and the stop bit “1” is added at the end. Hence, altogether, the character will be 10 bits long. Since even parity is used, the number of 1s in the 8 bits (for the character itself) will be even (possibly 0).



- a) 0001010101
The number of 1s in this part is 3—odd parity.
- b) 0001010111
The number of 1s in this part is 4—even parity.
- c) 1001010110
The stop bit is “0.”
The start bit is “1.”
- d) 1001010111
The start bit is a “1.”

Hence, the bit string, when correctly received, is 0001010101, which is (b).

Question 4

Difficulty: **

Frequency: **

Q4. Audio is sampled 11,000 times per second, and sampled values are each recorded as 8-bit data. In this system, how many seconds of audio can be recorded on a floppy disk whose capacity is 1.4×10^6 bytes?

a) 15

b) 127

c) 159

d) 1,272

Answer 4**Correct Answer:** b

Since the audio is sampled 11,000 times per second, and each sampling produces 8 bits of data, the amount of data transferred per second is as follows:

$$\begin{aligned}\text{Number of bits transferred per second} &= 11,000 \text{ (times/sec)} \times 8 \text{ (bits/time)} \\ &= 88,000 \text{ (bits/sec)}\end{aligned}$$

The capacity of a floppy disk is stated to be 1.4×10^6 bytes, so to use consistent units, we convert the number of bits of data transferred per second into bytes as follows:

$$\begin{aligned}\text{Number of bytes transferred per second} &= \frac{88,000 \text{ (bits/sec)}}{8 \text{ (bits/byte)}} \\ &= 11,000 \text{ (bytes/sec)}\end{aligned}$$

Since 11,000 bytes are transferred every second onto a floppy disk whose capacity is 1.4×10^6 bytes, the number of seconds of the audio data that can be recorded on this floppy disk is as follows:

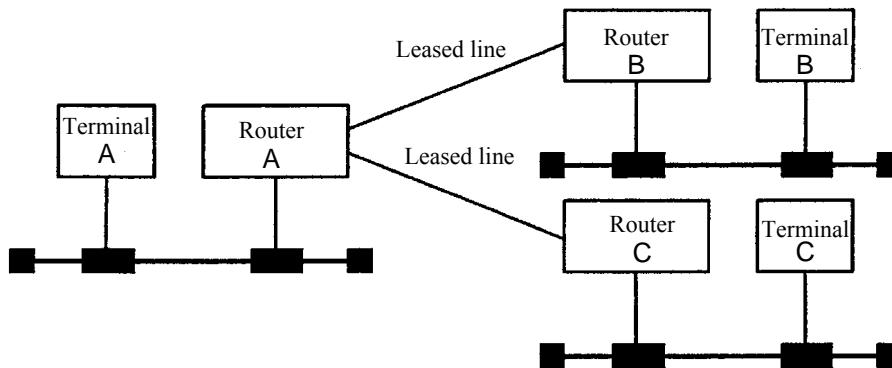
$$\begin{aligned}\text{Amount that can be recorded on a floppy disk} &= \frac{1.4 \times 10^6}{11,000} \\ &= \frac{1.4}{1.1} \times 10^2 \\ &= 1.272727... \times 10^2 \\ &= 127 \text{ (rounded to the nearest integer).}\end{aligned}$$

Question 5

Difficulty: **

Frequency: **

- Q5.** Three IP routers are connected by leased lines as shown in the figure below. Which of the following statements appropriately describes the operation of router A in relaying a TCP/IP packet from terminal A to terminal B?



- a) Router A relays all packets to both router B and router C.
- b) Router A relays packets to router B only according to the relay route specified in the packet.
- c) Router A relays packets to router B only based on the destination IP address in the packet.
- d) Router A learns the location of terminal B from the MAC address of the destination in the packet and relays the packets to router B only.

Answer 5

Correct Answer: c

A router verifies the IP address of the addressee to which the received text (packet) is sent, determines an appropriate route, and delivers it to the destination. In the data link layer of the OSI basic reference model, data can be transferred only between adjacent nodes or on the same segment, but a router sends packets to a designated router by relaying them through the network layer.

- a) If the access control methods of LANs are all identical, a bridge performs this function. A router can connect a network with LAN whose access control method may be different, and it only relays to a designated route.
- b) The relay route of a packet is not fixed. Routes are determined based on those which are set up in the routers or based on the information exchanged between routers, and an best route is selected as the relay route.
- c) It is a bridge that performs relays using MAC addresses.

Question 6

Difficulty: *

Frequency: ***

- Q6.** Which of the following medium access control methods in LAN provides the function of detecting a data frame collision on transmission media?
- a) CSMA/CA
 - b) CSMA/CD
 - c) Token-passing bus
 - d) Token-passing ring

Answer 6**Correct Answer:** b

A medium access control method is a method for sending frames (transmission units) on LAN. As a rule, when a terminal transmits a frame, other terminals need to hold all transmission until the frame reaches the destination. CSMA/CD (Carrier Sense Multiple Access with Collision Detection) is a medium access control method for bus-type LAN and star-type LAN. The terminal that wishes to transmit data checks to see if any communication established by other terminals is being done on the transmission medium; the terminal then sends the data if there is no communication taking place. If there is communication taking place, the terminal waits for a certain period of time and then attempts to re-send the data.

- a) CSMA/CA (Carrier Sense Multiple Access with Collision Avoidance) is a medium access control method for mid-speed LAN whose transmission speed is 1Mbps to 2Mbps.
- c) Token passing bus (token bus) is an application of token passing on a bus-type LAN. Token passing is a medium access control method for ring-type LAN and bus-type LAN. Transmission authorization data, called a token, is constantly going around LAN, and the terminal that has obtained the token gets the authorization for data transmission. A terminal wishing to transmit data gets the token and, in its place, releases the data it wishes to send. Once the transmission is finished, the token is released to the network again.
- d) Token passing ring (token ring) is an application of token passing on a ring-type LAN.

Question 7

Difficulty: *

Frequency: **

Q7. Which of the following is an appropriate description concerning the function of a proxy server used on the Web?

- a) A proxy server converts private IP addresses used on an intranet into global IP addresses, and vice-versa.
- b) A proxy server dynamically assigns an IP address to a client when the client connects to the network.
- c) When a client connected to an internal network communicates with an external server, a proxy server acts as a relay and establishes connection to the server on behalf of the client.
- d) A proxy server has a correspondence table of host names and IP addresses, and it notifies a client of the IP address of a host when the client sends a query.

Answer 7

Correct Answer: c

A proxy server is a server set up to maintain security and achieve high-speed access when making connection to the Internet from an internal network. It prevents unauthorized access into the internal network, and it also relays and manages access from the internal network to the outside Internet.

- a) The function that converts private IP addresses to global IP addresses and vice-versa is a function of IP masquerade or NAT (Network Address Translation). These functions are normally supported on routers (gateways).
- b) This is an explanation of DHCP (Dynamic Host Configuration Protocol).
- c) This is an explanation of DNS (Domain Name System).

5 Database Technology

Chapter Objectives

A database is an organized set of data which is accumulated collectively for purposes of data sharing, integrated management, and a high level of independence. Databases have several categories, but presently, hierarchical database, network database, and relational database serve as the major databases. Among these, relational database is the mainstream database today. In Section 1, we will learn about databases from a theoretical viewpoint, discussing their structures and development methods. In Section 2, we will learn how to make use of SQL, the programming language used to manipulate relational databases. In Section 3, we will mainly learn about DBMS, the software for efficient use of the databases.

- 5.1 Data Models
- 5.2 Database Languages
- 5.3 Database Manipulation

[Terms and Concepts to Understand]

3-layer schema, hierarchical database, network database, relational database, E-R diagram, normalization, selection, projection, join, cursor, DDL, DML, SQL, 2-phase commitment, replication

5.1 Data Models

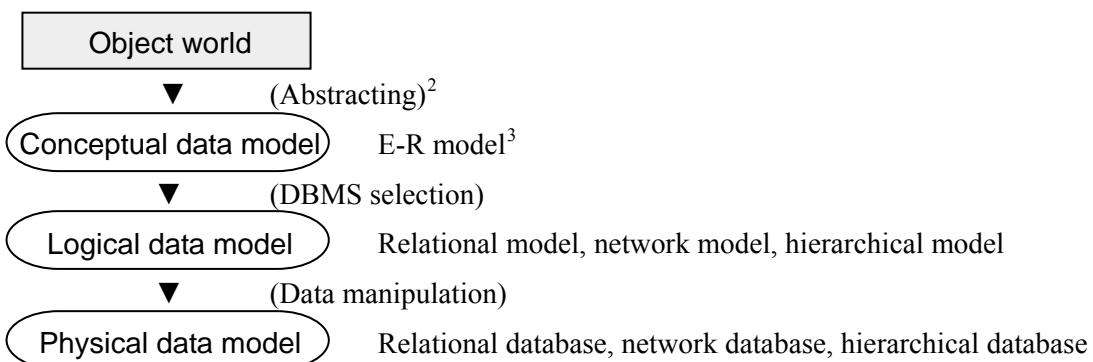
Introduction

A data model is an expression summarizing data items related to one another under certain rules. To create a database, one first produces a data model and normalizes it by eliminating unnecessary information and duplicate items. Then, the database is specifically designed so that search, update, and deletion of data can be performed efficiently. Types of data model include conceptual data models, logical data models, and physical data models. The rules that implement each of these are called conceptual schema, external schema, and internal schema.

5.1.1 3-layer Schema

Points	<ul style="list-style-type: none"> ➤ The framework (definition) of a database is called 3-layer schema. ➤ 3-layer schema consists of conceptual schema, external schema, and internal schema.
---------------	---

Abstracting and organizing the structure of real-world information, which is the object to be made into a database, and then expressing it, is called **data modeling**. A data model can be a conceptual, logical, or physical data model. These are related as shown in the figure below.¹



¹ (Note) A data model is a conceptual expression (model) of data; it could also refer to the rules of expression.

² **Abstracting:** It means extracting the most characteristic elements of the object and removing everything else. We can create a database that can be shared if we, in creating the database model, first extract those elements common to all tasks from among all the data subject to the tasks that need to be systematized.

³ (Hints & Tips) Conceptual models include E-R models discussed in Section 5.1.3. Logical data models have relational models, network models, and hierarchical models. Further, if we take a logical data model and make a database specifically from it, we can get a physical data model.

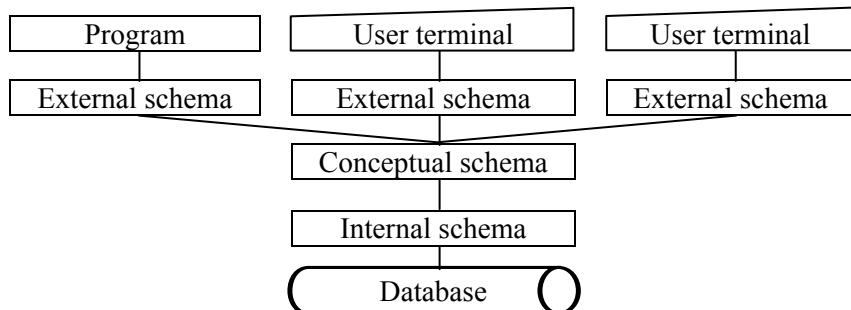
◆ Independence of Data

The independence of data means that the “program does not get changed when data changes.” Since multiple programs share the same set of data, it is not necessary to create as many data sets as the number of individual programs. Hence, the data must be organized systematically.

One type of software to achieve the independence of data is a database management system (DBMS), which keeps the data independent by using 3-layer schema.

◆ 3-layer Schema

A schema is a description of the framework of a database. In ANSI/X3/SPARC,⁴ schemata are classified into conceptual schema, external schema, and internal schema. These are called **3-layer schema**. The figure below shows their relations.



In general, the user of a database utilizes the database through an external schema.⁵

Name	Explanation
External schema	Definition of the database seen from the program or the user. This uses a part of the conceptual schema. In relational databases, this is called a view; in network databases, this is called a subschema. This exists for each program and user.
Conceptual schema	This is the data to be contained in the database, defined according to the data model; a definition of the real data as a whole. It is called a table in a relational database and a schema in a network database.
Internal schema	This is a definition to specifically achieve the conceptual schema for an external storage. It consists of information such as the medium, organization method, and buffer length.

Concerning relational database and network database, see Section 5.1.2. “Logical data models.”

⁴ **ANSI/X3/SPARC:** The ANSI (American National Standards Institute) is a non-profit organization that establishes the industrial standards of the United States. X3 is the committee within the ANSI which discusses the standards associated with information processing. The SPARC (Standards Planning And Requirements Committee) is the committee that is involved with international issues.

⁵ (FAQ) Many exam questions ask about the schema types and their characteristics. Know clearly the differences among conceptual schema, external schema, and internal schema.

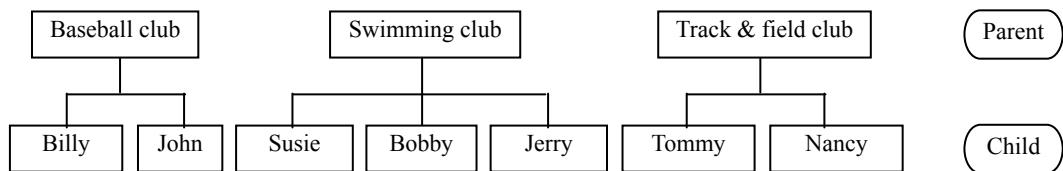
5.1.2 Logical Data Models

Points	<ul style="list-style-type: none">➤ Logical data models contain relational model, network model, and hierarchical model.➤ A database is the result of implementing a logical data model on a storage medium.
---------------	---

Logical data models contain **relational model**, **network model**, and **hierarchical model**. These data models, when they are implemented, become **relational databases**, **network databases**, or **hierarchical databases**.⁶

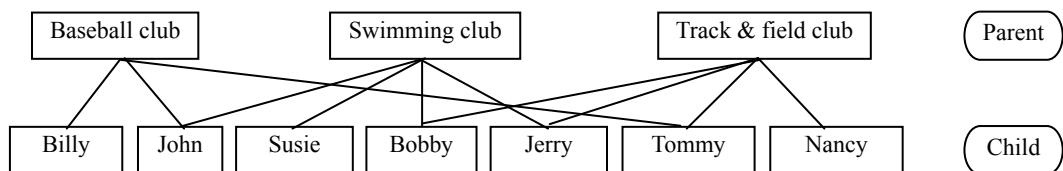
◆ Hierarchical Database (Tree-Structure Database)

A **hierarchical database** divides records into parents and children and shows the relationship with a hierarchical structure. It is characterized by 1-to-many (1:n) correspondences between parent records and child records. In other words, one parent record may have multiple child records, but one child record corresponds only to one parent record. However, hierarchical databases are treated as a special case of network databases, so they are no longer used very much. The structure of a hierarchical database is shown below.



◆ Network Database

A **network database** is different from a hierarchical database in that the parent records and child records do not have 1-to-n (1:n) correspondences; rather, they are in many-to-many (m:n) correspondence. In other words, a parent record may have multiple child records, and conversely, a child record may have multiple parent records.⁷ A network database is sometimes called a CODASYL database.⁸ The structure of a network database is as shown below.



Here, for example, "Susie" belongs to the "swimming club" only, but "Tommy" belongs to both the "track & field club" and the "baseball club."

⁶ (Note) Hierarchical databases and network databases together are sometimes called structure databases.

(Hints & Tips) A structure database is a network database in which each child has only one parent.

⁸ **CODASYL database:** A network database refers to any database based on the language specifications proposed by CODASYL; hence, a network database is also called CODASYL database. CODASYL stands for the Conference On DAta SYstems Languages. This organization consists of the United States government, computer manufacturers, and users. This is the organization that has developed and is maintaining the business-oriented programming language COBOL. It developed COBOL in 1960 and conducted research in database languages later.

◆ Relational Database

A **relational database** is a database in which data is expressed in a two-dimensional table. Each row of the table corresponds to a record, and each column is an item of the records. The underlined columns indicate the primary key.⁹

Name of the table:

Employee_tbl

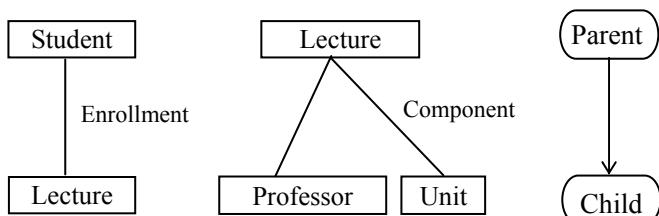
Columns (items, attributes,)		
Employee_number	Name	Tel_number
00100	Paul Smith	03-3456-0001
00200	Rick Martin	03-3456-0011
00300	Billy Graham	03-3456-0010
00400	John Wilson	03-3456-0200

← Row (pair, tuple, record)

Each table is always named. In the above example, the name is “Employee_tbl.” The columns are “Employee_number,” “Name,” and “Tel_number.” A row is a set of data like “00100, Paul Smith, 03-3456-0001.” In other words, we can say that “Employee_tbl consists of 4 rows and 3 columns (4 by 3).”

◆ Bachman Diagram

A Bachman diagram describes the parent-child relation between records in a network database. A parent is called an owner while a child is called a member. Below, terms like “enrollment” and “component” describe the parent-child relations and are called parent-child set types. The actual contents (values) in Bachman diagrams are called **occurrences**.¹⁰



⁹ **Primary key:** It is a column or a set of columns that uniquely identifies a row of the table. In the same table, primary key values cannot be repeated. Here in the “Employee table,” “Employee number” is the primary key. If the values of one column are not unique, a combined key can be defined by combining multiple columns.

¹⁰ **Occurrence:** It is a specific value in a Bachman diagram. For example, if A, B, and C are three of the “students” in the example here, A, B, and C are occurrences.

5.1.3 E-R Model and E-R Diagram

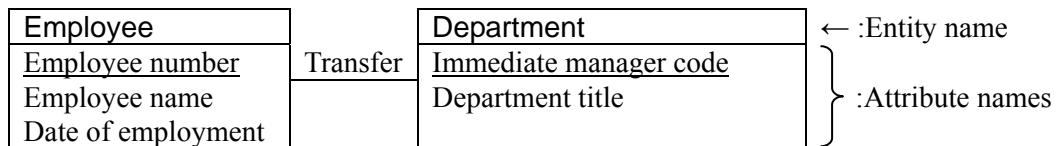
Points	<ul style="list-style-type: none"> ➤ An E-R model is a data model that does not take DBMS into account. ➤ The elements of an E-R diagram are “entities” and “relations.”
---------------	--

Hierarchical models, network models, and relational models are all data models with the assumption that DBMS is used.¹¹ However, data and information used in the real world are not necessarily limited to those compatible with DBMS. One of the methods for expressing real-world data structures as faithfully as possible is the **E-R model**. An E-R model is expressed by using an E-R diagram.

An **E-R diagram** is a technique, used in designing files or databases, for expressing results obtained by grasping the objects to be managed and data items. The objects of management and analysis are referred to as entities, which are associated with one another by relationships. The elements constituting entities and relationships are called attributes.

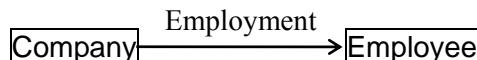
◆ Rules for E-R Diagram

In an E-R diagram, entities are represented in rectangular boxes while relationships are indicated by line segments or arrows (\leftarrow , — , \rightarrow). Attributes are also shown in boxes. In the example below, it is indicated that employees and department are linked by a relationship called transfer. The entity “employee” has attributes called employee number, employee name, and the date of employment. In some cases, the primary keys are underlined.



◆ Correspondence Relations

In an E-R diagram, the 1-to-many relation “one company has multiple employees” is indicated by the following diagram. Note that, as seen here, sometimes the attributes are omitted.¹²



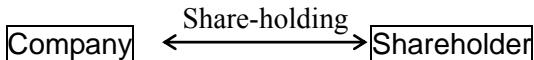
Here, “Company” and “Employee” are linked by the relation called “employment.” The relation between “Company” and “Employee” is 1-to-many (\rightarrow), so for one company, there are multiple (many) employees. If an employee is chosen, there is only one company related to him/her, so one can know which company is associated with the employee. However, choosing one company does not uniquely identify its employee because there are multiple employees. Hence, if unique identification is possible, the identified party is the “1” in the “1-to-many.”¹³

¹¹ **DBMS (DataBase Management System)**: it is a software dedicated to the maintenance and operation of databases.

¹² (FAQ) There are exam questions on how to interpret E-R diagrams. Be sure that you can identify 1-to-many, 1-to-1, and many-to-many relations. The relation between employees and departments in a company with multiple employees, some of whom may belong to multiple departments, is “many-to-many.”

¹³ (Hints & Tips) Be careful as it is easy to reverse “1-to-many” relations. If picking one data value can uniquely identify an associated member, the identified member is the “1.” If unique identification is not possible, the other is “many.”

Below is an E-R diagram showing a many-to-many relation.



Here, the double arrow indicates a many-to-many relation between “Company” and “Shareholder.” This suggests that a shareholder may hold shares of multiple companies and that a company may have multiple shareholders. In other words, we can interpret this figure as follows: “There are multiple companies, each of which has multiple shareholders.”

5.1.4 Normalization and Reference Constraints

Points

- Normalization means eliminating data redundancy.
- There are first, second, and third normal forms.

Normalization (data normalization) means to maintain the consistency and integrity of data by eliminating redundant data. There are first normal forms, second normal forms, and third normal forms. Normalization is a concept that is used only for relational databases. The mainstream databases used today are relational databases, so normalization is an extremely important theme.

◆ Non-normal Form

A non-normal form is a form in which items are simply listed. In general, repeated items are also included. In the figure below, the combination (ProductNumber, Quantity, UnitPrice) is repeated. In the following explanations, the underlined items indicate the primary keys. Here, the fact that the InvoiceNumber is used as the primary key assumes that there is no duplication of the InvoiceNumber.¹⁴

Product information (1)						Product information (2)		
<u>Invoice Number</u>	Customer Number	Customer Name	Product Number	Quantity	Unit Price	Product Number	Quantity	Unit Price

¹⁴ (Hints & Tips) There cannot be two records whose primary key values are the same. In this example of the non-normal form, since the InvoiceNumber is the primary key, there are no duplicate InvoiceNumbers.

◆ First Normal Form

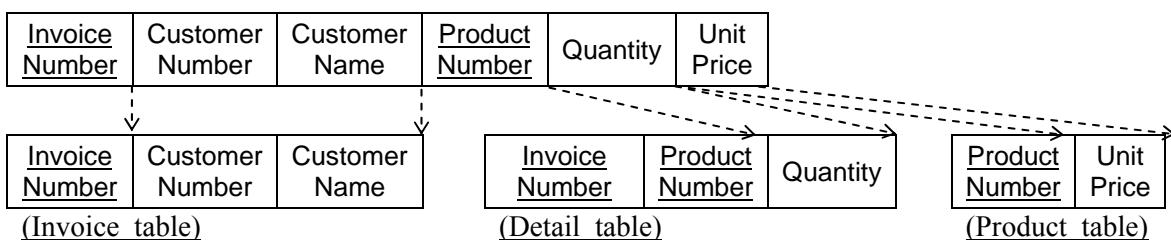
A **first normal form** has no repeated items in the table. To convert a non-normal form into a first normal form, we separate the combination “ProductNumber, Quantity, UnitPrice” in the non-normal form. Then, it is possible that there are multiple records in which “InvoiceNumber, CustomerNumber, CustomerName” are all the same but “product information” is different, so “InvoiceNumber” by itself cannot be the primary key. Hence, we include “ProductNumber” also, and we may use both of these items together as the primary keys.

In the example of the non-normal form above, there is repetition, so the first normal form has only two records as shown below.

If there are multiple pieces of product information, the same contents may be repeated.		There should be only one piece of product information.
<u>Invoice Number</u>	Customer Number	Customer Name

◆ Second Normal Form

In the first normal form, the two items “InvoiceNnumber” and “ProductNumber” are together used as the primary keys. In databases, all non-key attributes must be functionally dependent on the entire primary key.¹⁵ However, the item “UnitPrice” is not related to “InvoiceNumber.” The item “UnitPrice” is determined only by “ProductNumber.” Hence, we now separate “UnitPrice”; in doing so, since “ProductNumber” and “UnitPrice” need to be in correspondence, we use “ProductNumber” as the primary key. Further, “CustomerNumber” can be determined uniquely if “InvoiceNumber” is selected, so “ProductNumber” is unnecessary. Yet, to determine “Quantity,” we must have “InvoiceNumber” and “ProductNumber.” For later explanations, we name these new tables “Invoice_table,” “Detail_table,” and “Product_table.”¹⁶ The result of separating the data into these three tables is called a **second normal form**.



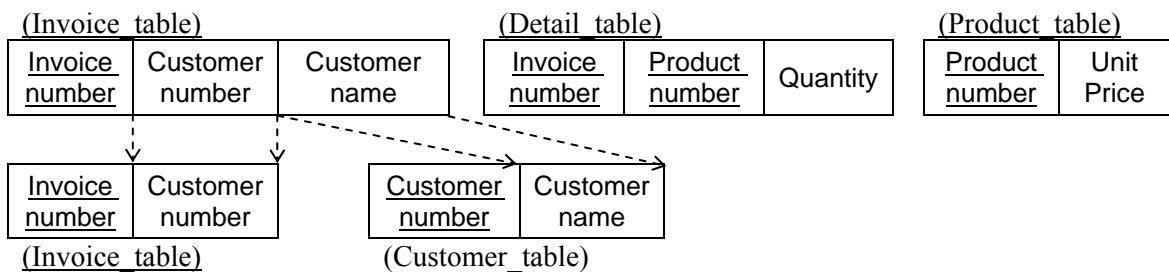
¹⁵ (Note) Dependency on the primary key means that each item can be identified by the values of the primary key.

¹⁶ **Complete functional dependency/ Partial functional dependency:** In the first normal form, “Quantity” is determined for the entire primary key “InvoiceNumber + ProductNumber.” As seen in this example, the dependency on the entire combination of primary-key items is called complete functional dependence. The unit price, on the other hand, depends only on one of the primary keys (in this example, on “ProductNnumber”); When an item is determined by one of the primary keys, we call it partial functional dependency.

Strictly speaking, a second normal form can be defined as “a first form in which all non-key items are in complete functional dependence.”

◆ Third Normal Form

Note that in the second normal form, in the “Invoice_table” in particular, the “InvoiceNumber” uniquely determines the “CustomerNumber.” Furthermore, the item “CustomerNumber” uniquely identifies the “CustomerName.” Hence, as shown below, we now separate “CustomerName” and prepare a new table, “Customer_table,” in which “CustomerNumber” is the primary key. At this time, we do not make any changes in “Detail_table” or “Product_table.”



So, when the records of a non-normal form are modified into a **third normal form**, the data gets separated into four records: “Detail_table,” “Product_table,” “Invoice_table,” and “Customer_table.” A third normal form is characterized by the property that no items are duplicated except for the primary key items.¹⁷

◆ Reference Constraints

If there is no contradiction in the data contained in a database, we say that the database is **consistent**. Various conditions to verify the completeness of data are called **integrity constraints**. Consistency constraints include **reference constraints**, **existence constraints**, **update constraints**, and **format constraints**.¹⁸ ¹⁹

A **reference constraint** is a constraint concerning the consistency between multiple items. If a table contains data that looks up another table, the other table must have the referenced data registered in advance. For example, in the third normal form explained above, to register the **Invoice_table**, it is necessary that the information on customer numbers compatible with the customer numbers in the **Invoice_table** be registered in the **Customer_table**. Here, the customer numbers in the **Invoice_table** are referred to as an **external key** of the **Customer_table**.

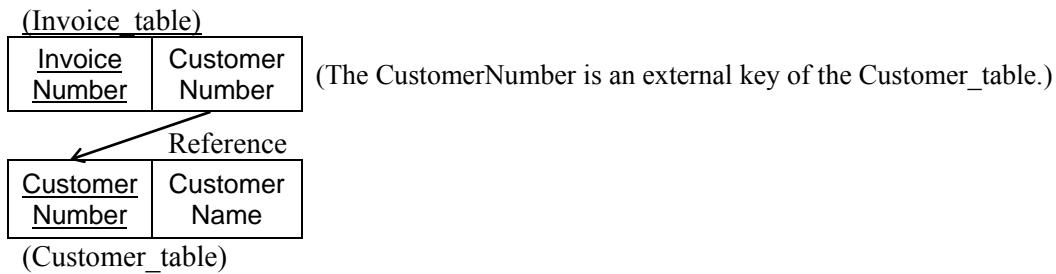
¹⁷ **Transitive functional dependency:** Customer names in the second normal form can be identified because the primary key “InvoiceNumber” identifies the customer number, which identifies the customer name. In other words, the invoice number indirectly identifies the customer name. This type of indirect dependency is called transitive functional dependency. Strictly speaking, a third normal form can be defined as “a second normal form in which no non-key items are in transitive functional dependence.”

¹⁸ (FAQ) There are exam questions that give records in a non-normal form, as well as some assumptions, and then ask you to choose the third normal form from the answer group. If you follow the procedures described in this book to obtain the third normal form, you will certainly get the answer, but you may run out of time. So it is necessary to intuitively find the correct third normal form. You should try many questions for practice, but you can identify the third normal form by the property that “there are no duplicate items except for the primary key items.”

¹⁹ **Existence constraints:** It means constraints that the existence of particular data requires the existence of some other data. For instance, a child record cannot be added unless there is a parent record in existence.

Update constraints: It means constraints that a new item must satisfy certain given conditions in order to be registered. For instance, the value “6” cannot be registered if the value must be between 1 and 5, inclusive.

Format constraints: It means constraints that an item must be in a format that satisfies certain given conditions. For instance, text cannot be registered in an item that requires numerical entry.



5.1.5 Data Manipulation in Relational Database

Points

- Data manipulation includes relational operations and set operations.
- Relational operations include selection, projection, and join.

Among the various kinds of operations on relational databases, relational operations and set operations are the most important. In a relational database, a table, a row, and a column are all treated as a set which extracts values. Extracting processes include manipulation such as **selection**, **projection**, and **join**. These are called **relational operations**. In contrast, there are other operations whereby two tables in a relational database are used to create a new table; these are called **set operations**. Set operations include **union**, **intersection**, and **difference**.

◆ Relational Operations

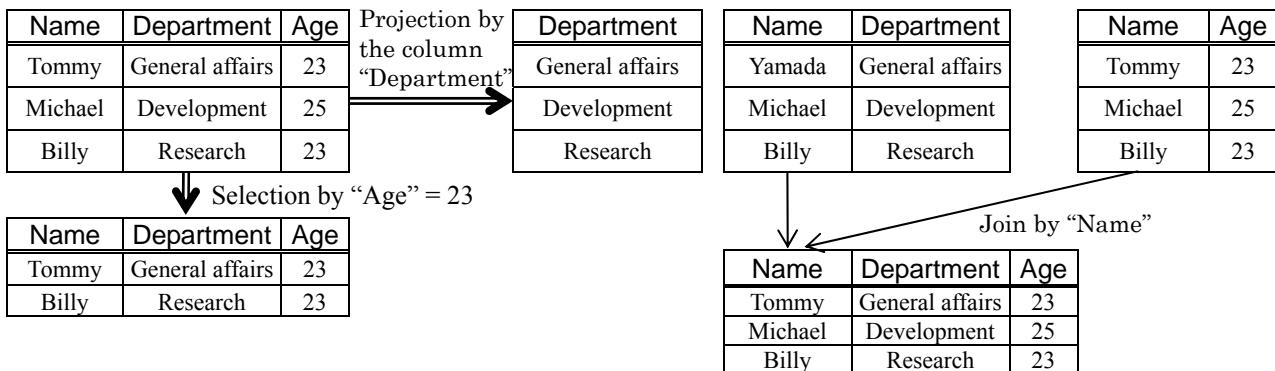
The meanings of relational operations are listed in the following table. Various data is extracted by combining these basic operations.

Operation	Function
Selection	Extracting rows satisfying certain conditions
Projection	Extracting specific columns (attributes)
Join	Connecting multiple tables for equivalent columns

Projection extracts a specific column. In the following figure, for instance, only “Department” is extracted. Selection extracts certain rows, so, for instance, every row whose “Age” is “23” is extracted. Join connects equivalent columns, so, for instance, two tables are joined by “Name.”^{20 21}

²⁰ (Hints & Tips) Results of relational and set operations are displayed as new tables but are NOT actually saved in the database. These are simply stored in the work area as intermediate results.

²¹ (FAQ) Many exam questions involve the meanings of relational operations. Be sure you know that selection extracts “rows” and projection “columns.” Be sure to know also that join is an operation that combines multiple tables.



Projection can also extract multiple columns.

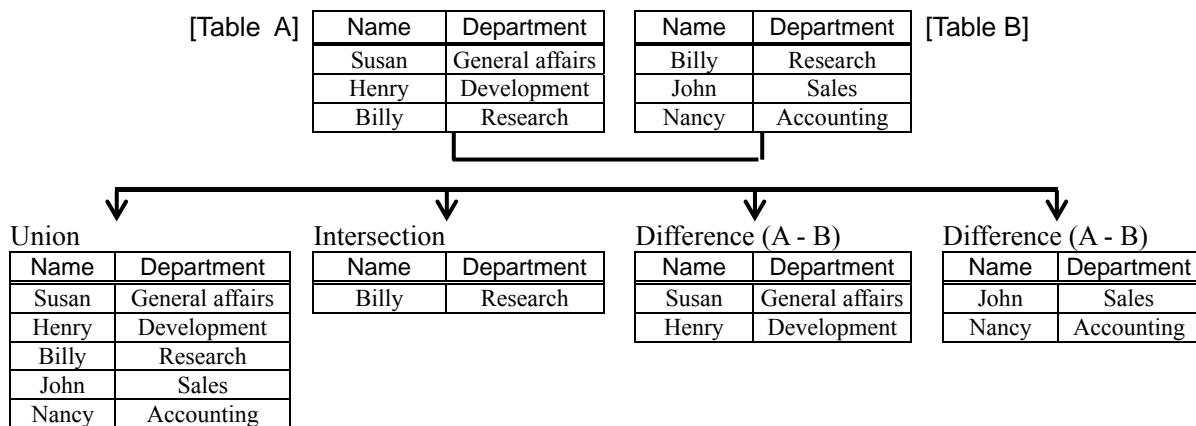
◆ Set Operations

A **set operation**, based on the mathematical theory of sets, includes the following:²²

Operation	Function
Union	Extracting rows which are in at least one of the two tables
Intersection	Extracting rows which contain the same value in both tables
Difference	Extracting rows that are common in both tables

For instance, union extracts rows that appear in Table A or Table B; note that “Billy” appears in both tables, so it is extracted only once. Intersection extracts rows that appear in both Table A and Table B. In this case, only “Billy” is extracted.

The order of operations does not matter in union or intersection, but in difference, the order does matter. Different orders produce different results. “A – B” produces rows that are in Table A but not in Table B. Here, “Billy” is excluded, so “Susan” and “Henry” are extracted. In contrast, “B – A” produces those in Table B and not in Table A, so “Billy” is once again excluded, resulting in “John” and “Nancy” being extracted.



²² **Sorting/ the four basic operations:** A relational database is equipped not only with relational and set operations but also with the sorting functions and the four basic operations. Sorting is the function of ordering data in ascending or descending order of a certain column. The four basic operations apply to numeric attributes and extract the results of doing certain arithmetic (four basic) operations. For instance, it can extract the results of multiplying the values of a certain column by 10.

Quiz

- Q1** List the three-layer schema and explain the roles of each schema.
- Q2** List logical data models and explain briefly the characteristics of each.
- Q3** Describe the characteristics of first, second, and third normal forms.
- Q4** List the types of relational operations and explain the process of each.

5.2 Database Languages

Introduction

A database language is a language used in defining and deleting databases and tables and searching and updating data.

5.2.1 DDL and DML

Points

- Database languages include DDL and DML.
- SQL is a database language for relational databases.

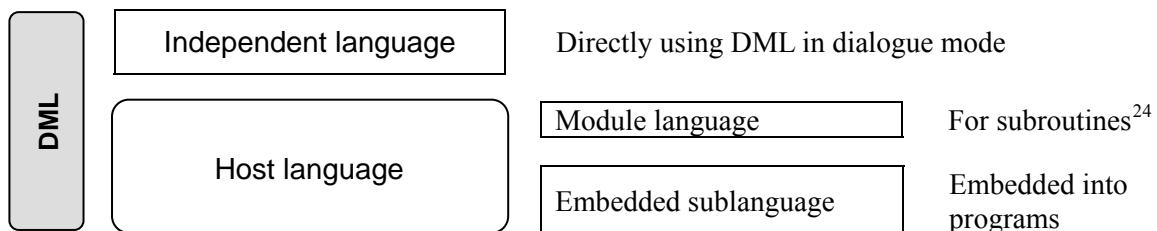
A language used in defining and organizing databases is called a **data definition language (DDL)** while a language used to search, update, add, and delete data is called a **data manipulation language (DML)**. In general, the database administrator uses DDL to edit the database while the system developer uses DML to develop systems using the database. Typical database languages include SQL for relational databases, and NDL²³ for network databases.

◆ DDL (Data Definition Language)

DDL is a language system that defines schema based on the data model. For relational databases, they are stipulated as SQL-DDL.

◆ DML (Data Manipulation Language)

DML is a language system used to manipulate databases by the user. For relational databases, they are stipulated as SQL-DML. According to how they are used, DMLs are classified as shown below.



²³ **NDL (Network Database Language)**: It is a database language for network databases, used to define schema and manipulate databases. NDL consists of the following functions: schema-defining language to define the structure of the database; subschema-defining language to define views; data-manipulating language to manipulate the data in the database; and module language to execute the procedures of a variety of data-manipulating languages.

²⁴ **Module language**: It is a language written in a data manipulation language; it processes databases when it is called from a higher-level language such as COBOL.

Independent language

This refers to a system that provides a programming language different from general-purpose programming languages so that the functions provided by the data-management system can be used within the language functions. It uses SQL and NDL in dialogue style like commands.

Host language

This refers to a system for database manipulation wherein DML is embedded into programs written in higher-level languages such as COBOL, Fortran, and C. Here, the higher-level languages are called the host languages.

Methods for embedding DML into a program include the module language system and the embedded system. In the module language system, we can develop a subroutine which forms the database-manipulation section of the program, and the program calls the subroutine by a “call” statement. The other way, the embedded sublanguage, is where DML is directly written within the program.

◆ Cursor Function

The **cursor function** is used when processing rows (records) of a relational database by using a procedural language. It considers a query result (derived table) by DML as a file so that it can be processed using a programming language.

With the cursor function, files used by existing programs can be switched to databases easily. In SQL, the following manipulation statements are available.²⁵

Manipulation statement	Meaning
DECLARE CURSOR	Declaration of the cursor function
OPEN CURSOR	Start of the cursor operation
FETCH	Read one row
CLOSE CURSOR	End of the cursor operation

²⁵ (Note) In DECLARE CURSOR, the cursor name is defined. Following DECLARE CURSOR, a SELECT statement is written, which is a query written in DML. This procedure produces a resulting table. Then, by the FETCH statement, the table is read beginning at the first row.

5.2.2 SQL

Points

- The statement that extracts data is “SELECT.”
- In subqueries, designate “SELECT” using a “WHERE” phrase.

Here, we explain SQL in detail, which is the database language for relational databases. In SQL, SELECT statements are used to extract data from a database.

◆ Structure of SELECT Statement

A SELECT statement is structured as follows:²⁶

SELECT	item 1, item 2, ...
FROM	table 1, table 2, ...
WHERE	condition

◆ Projection and Selection

Below is an example showing the result of extracting data by projection and selection using a SELECT statement on the table “Employee_tbl.” If “*” is designated after a SELECT statement, all the items in the table will be extracted. We may also combine selection and projection.

[Table] Employee_tbl

Name	Department	Home Country	Age
Jimmy	Sales	Japan	28
Frank	Human Resources	USA	22
Billy	Sales	France	35
Harry	Sales	Italy	43
Josh	General Affairs	Germany	48
Randy	Human Resources	USA	36
Steve	Sales	UK	31

SELECT Name, Department FROM Employee_tbl

Projection (extracting names and departments)

Name	Department
Jimmy	Sales
Frank	Human Resources
Billy	Sales
Harry	Sales
Josh	General Affairs
Randy	Human Resources
Steve	Sales

SELECT * FROM Employee_tbl WHERE Age>=35

Selection (extracting rows where the age is 35 or above)

Name	Department
Billy	Sales
Harry	Sales
Josh	General Affairs
Randy	Human Resources

Name	Department	HomeCountry	Age
Billy	Sales	France	35
Harry	Sales	Italy	43
Josh	General Affairs	Germany	48
Randy	Human Resources	USA	36

Selection and projection

SELECT Name, Department
FROM Employee_tbl
WHERE Age>=35

²⁶ (Hints & Tips) In a SELECT statement, the WHERE phrase can be omitted. If omitted, the conditions for extraction are dropped, so all designated items will be extracted.

◆ Join

Using a SELECT statement, we can join multiple tables through specified columns. Below is an example joining “Employee_tbl” and “Department_tbl.” It is not permitted to have the same column name in the same table, but if there are identical column names in different tables, they are distinguished in the form “**tablename.columnname.**” “Employee_tbl.Department = Department_tbl.Department” is the key joining the two tables.

[Table] Employee_tbl

Name	Department	Home Country	Age
Jimmy	Sales	Japan	28
Frank	Human Resources	USA	22
Billy	Sales	France	35
Harry	Sales	Italy	43
Josh	General Affairs	Germany	48
Randy	Human Resources	USA	36
Steve	Sales	UK	31

[Table] Department_tbl

Department	DeptLeader	Location
Human Resources	Randy	Headquarters
General Affairs	Josh	Headquarters
Sales	Harry	Branch office

```
SELECT Name,
       Employee_tbl.Department,
       HomeCountry, Age, DeptLeader,
       Location
  FROM Employee_tbl,
       Department_tbl
 WHERE Employee_tbl.Department
      =Department_tbl.Department27
```

Name	Department	Home Country	Age	DeptLeader	Location
Jimmy	Sales	Japan	28	Harry	Branch office
Frank	Human Resources	USA	22	Randy	Headquarters
Billy	Sales	France	35	Harry	Branch office
Harry	Sales	Italy	43	Harry	Branch office
Josh	General Affairs	Germany	48	Harry	Headquarters
Randy	Human Resources	USA	36	Randy	Headquarters
Steve	Sales	UK	31	Harry	Branch office

◆ IN and BETWEEN

In **WHERE**, we can specify complex conditions combined with **AND** or **OR**. **IN** designates an **OR** condition while **BETWEEN** designates an **AND** condition.

To extract from “Employee_tbl” those names of people whose ages are 22, 28, and 35, there are two methods, as shown below. Both methods produce the same result.

- SELECT Name FROM Employee_tbl WHERE Age IN (22, 28, 35)
- SELECT Name FROM Employee_tbl WHERE Age = 22 OR age = 28 OR age = 35

To extract from “Employee_tbl” those names of people whose ages are 22 to 28, inclusive, there are two methods, as shown below. Both methods produce the same result.

- SELECT Name FROM Employee_tbl WHERE Age BETWEEN (22, 28)
- SELECT Name FROM Employee_tbl WHERE Age >=22 AND age<=28

²⁷ (Note) If columns have the same name, variables can be used in the way shown below. Here, variable X is assigned to the employee table and variable Y to the Department_tbl.

```
SELECT Name, X.Department, HomeCountry, DeptLeader, Location
  FROM Employee_tbl X, Department_tbl Y
 WHERE X.Department = Y.Department
```

◆ ORDERED BY

ORDERED BY is used to extract data in ascending or descending order by a certain column. Below is an example where the names are to be sorted from “Employee_tbl” in ascending or descending order. **ASC** is used for ascending order and can be omitted. For descending order, **DESC** is used.

- SELECT Name FROM Employee_tbl ORDERED BY Name ASC
- SELECT Name FROM Employee_tbl ORDERED BY Name DESC²⁸

◆ Set Functions and GROUP BY

In the “Sales_tbl” below, we show an example of an SQL statement to extract the total of the sales amount in rows where the “ProductName” is the same. To calculate the total, the set function **SUM** is used. The function **GROUP BY** consolidates all rows that have the same value in a certain column. Here, we consolidate the data by product number.²⁹

[Table] Sales_tbl

```

    graph LR
      A[Table] --> B["SELECT ProductNumber, SUM(SalesAmount) FROM Sales_tbl GROUP BY ProductName"]
      B --> C[Product Number | Sales Amount]
      C -- "Total of G01" --> D[Product Number | Sales Amount]
      C -- "Total of G02" --> E[Product Number | Sales Amount]
  
```

The diagram illustrates the process of executing an SQL query on a table. On the left, a table named "Sales_tbl" is shown with columns "Product Number" and "Sales Amount", containing four rows: G01 (100), G02 (50), G03 (200), and G04 (100). An arrow points from this table to the SQL query: "SELECT ProductNumber, SUM(SalesAmount) FROM Sales_tbl GROUP BY ProductName". Another arrow points from the query to the resulting table on the right, which has two rows: G01 (300) and G02 (150). To the right of the second table, the labels "Total of G01" and "Total of G02" are placed.

Product Number	Sales Amount
G01	100
G02	50
G03	200
G04	100

Product Number	Sales Amount
G01	300
G02	150

²⁸ (Hints & Tips) The column names designated in “ORDERED BY” or “GROUP BY” must be contained in the column names designated by SELECT. This is a syntax requirement of SQL.

²⁹ **Set function:** It is a function prepared by database software. It can find various values such as the total and maximum values for a specific column. Set functions include the following: SUM (total), MAX (maximum value), MIN (minimum value), AVG (average value), and COUNT (number of values).

◆ Subqueries

We can make a query on one table and then use the result of that query to make another query. The first of these queries is called a subquery, which is performed by using IN.

[Table]Order_tbl	
OrderNumber	CustomerNumber
100	A100
101	B200
102	C300

[Table]Order_detail_tbl	
OrderNumber	ProductNumber
100	301
100	302
100	301
100	401
100	402
101	301

[Table]Product_tbl	
ProductNumber	ProductModel
301	television
302	television
401	VCR
402	VCR

First, let us describe the method which does not use subqueries. The SQL statement that extracts the “ProductModel” of the product ordered by “CustomerNumber” A100 from tables “Order_tbl,” “Order_detail_tbl,” and “Product_tbl” is as follows:³⁰

```
SELECT ProductModel
FROM Order_tbl, Order_detail_tbl, Product_tbl
WHERE CustomerNumber = 'A100' AND
      Order_tbl.OrderNumber = Order_detail_tbl.OrderNumber AND
      Order_detail_tbl.ProductNumber = Product_tbl.ProductNumber
```

Here, the “OrderNumber” 100 by “CustomerNumber” A100 in the “Order_tbl” is joined with the “OrderNumber” of the “Order_detail_tbl.” As a result, the first five lines of the “Order_detail_tbl,” i.e. “ProductNumber” 301, 302, 301, 401, and 402 are extracted; in addition, joined with the “ProductNumber” of the “Product_tbl,” the “ProductModel” is extracted.

Let us now write this using the format of a sub-inquiry with IN.

```
SELECT ProductModel FROM Product_tbl
WHERE Product_tbl.ProductNumber
IN (SELECT Order_detail_tbl.ProductNumber
     FROM Order_tbl, Order_detail_tbl
     WHERE CustomerNumber = 'A100' AND
           Order_tbl.OrderNumber = Order_detail_tbl.OrderNumber)
```

As shown above, we can make one query and, using the result of that query, make another query.³¹ The query contained in IN is called a subquery. The SELECT statement of this subquery is executed first, and the extracted information, “Order_detail_tbl.ProductNumber” and “Product_tbl.ProductNumber” are joined. Here, the product numbers extracted by IN are “301, 302, 301, 401, and 402,” but since “301” is duplicated, the duplication is eliminated. Consequently, the four lines “301, 302, 401, and 402” are extracted.³²

³⁰ (Note) Sometimes EXISTS is used for subqueries. IN and EXISTS are different functions, but the execution results are almost identical.

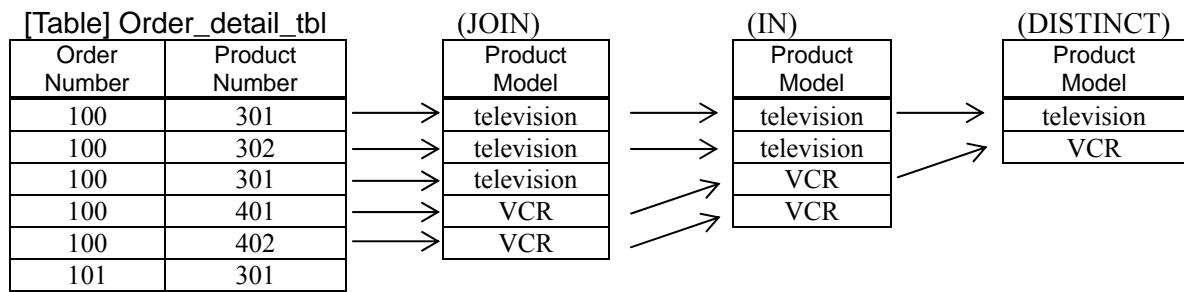
³¹ (Note) We can write NOT before IN. In this case, NOT IN (subquery) gives the negation of the subquery result.

³² (FAQ) Every exam is certain to have questions on the extracted result by a SELECT statement in SQL. Be thoroughly familiar with the use of the SELECT statement.

Further, duplication can be removed. As shown below, one can remove the product model duplication by designating DISTINCT before the product model. Television and VCR are duplicated, so each of these can be consolidated.³³

```
SELECT DISTINCT ProductModel
FROM Product_tbl
WHERE Product_tbl.ProductNumber
IN (SELECT Order_detail_tbl.ProductNumber
     FROM Order_tbl, Order_detail_tbl
     WHERE CustomerNumber = 'A100' AND
          Order_tbl.OrderNumber = Order_detail_tbl.OrderNumber)
```

The table below summarizes the results explained thus far.



³³ (Hints & Tips) Note that the extracted number of records varies with the conditions specified by WHERE. IN is the same as the OR condition, so identical values are discarded. DISTINCT also removes identical values, but it is designated immediately before the column name designated by SELECT.

Quiz

- Q1** Explain the roles of DDL and DML.
- Q2** Explain the cursor function.
- Q3** Give the data extracted from the table “Student_list_tbl” by the following SQL statement:
SELECT Name FROM Student_list_tbl
WHERE Major = 'Physics' AND Age < 20

[Table] Student_list_tbl

Name	Major	Age
Paul Newman	Physics	22
John Wayne	Chemistry	20
Tom Hanks	Biology	18
Robert Redford	Physics	19
Clint Eastwood	Mathematics	19

5.3 Control of Databases

Introduction

In order to ensure the reliability of data, various controls are applied to a database system. In a distributed database, it is important to maintain its consistency

5.3.1 Database Control Functions

Points

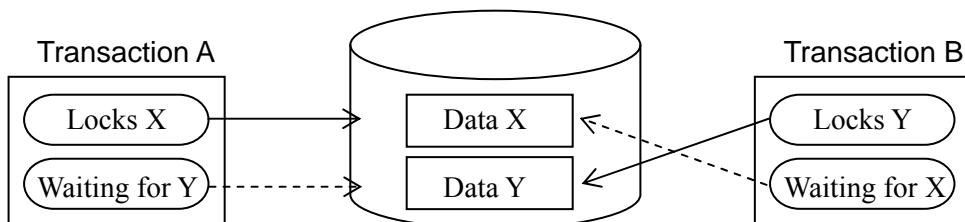
- Exclusive access control is conducted to maintain data consistency.
- Recovery methods from a failure include roll-back and roll-forward.

Database control functions include the access control function and shared resources management. They also support recovery from database failures.

◆ Access Control

In general, for access control to maintain the integrity of a database, exclusive access control³⁴ is conducted. **Exclusive access control** prohibits multiple users from accessing the same data at the same time. Through this control, multiple users can use the same database without causing contradictions. However, if all transactions³⁵ are only referential, exclusive access control is not necessary.

In some instances, exclusive access control may cause a deadlock. A **deadlock** is a situation in which two transactions are waiting for each other to release the lock. An image of a deadlock is shown below.³⁶



³⁴ **Exclusive access control:** It means locking a part of the database while it is updated by one transaction so that other transactions can be prohibited from accessing the same part of the database.

³⁵ **Transaction:** It is a processing unit of the data which is sent from a terminal to the host computer. It is also called a message.

³⁶ (Note) If each transaction locks all necessary resources at the beginning of its processing and does not lock them during the processing, a deadlock can be avoided. Or, if all transactions have an identical order of locking, a deadlock can be avoided.

◆ Shared Resources Management

Multiple users access a database through applications; here, data editing and updating need to be completed by only one application. For this reason, the transaction management function is introduced. The **transaction management function** serves to maintain data consistently without contradiction by dividing up the data into transaction units so that the database can handle updates from multiple users. This type of concept includes the **ACID characteristics**.³⁷

Name		Explanation
A	Atomicity	Property that there is no intermediate stage at the end of processing; either all processes are complete or nothing is being done.
C	Consistency	Property that, regardless of the completion condition of a transaction, the contents of a database cannot have contradictions.
I	Isolation	Property that the processing results cannot be different whether multiple transactions are executed simultaneously or sequentially.
D	Durability	Property that results are not ruined by failures and other factors once the transactions are finished.

◆ Recovery Management

There are various recovery methods, depending on the situation of database failure.

Failure type	Recovery method	Explanation
System failure	System restart	A computer system failure such as physically erroneous operation (1) Back up to the point in time when the data was backed up ³⁸ (2) Rewrite sequentially using post-update information of the log ³⁹
Transaction failure	Roll-back	A logically erroneous operation due to program failure, etc. (1) Roll back the failure data only, using the pre-update information of the log. (2) Re-execute the transaction(s).
Medium failure	Roll-forward	A problem with a medium such as a magnetic disk (1) Replace the medium (2) Back up to the point in time when the data was backed up (3) Rewrite sequentially using post-update information of the log

³⁷ (FAQ) There are exam questions on the meaning and necessity of exclusive access control. Remember that exclusive access control is the function that prohibits access to the same location (record) at the same time. Understand also that exclusive access control is carried out to maintain the integrity of data.

³⁸ **Back up:** it is a duplicate of the contents of the entire database on a medium such as a magnetic tape, copied at regular time intervals. Normally the copied contents are the data immediately before the startup or immediately after the shutdown of an online system. If the system is operating 24 hours a day, backup is often carried out when the transactions are the fewest, such as around midnight.

³⁹ **Log file (journal file):** It means data record in which conditions of the database before and after the updating are recorded whenever the contents of the database are updated. In the operation of some systems, only the contents prior to the updating are recorded.

5.3.2 Distributed Databases

Points	<ul style="list-style-type: none"> ➤ Two-phase commitment is the technology that maintains the integrity of a distributed database. ➤ Replication is the technology that enhances the response performance of a distributed database.
---------------	---

Distributed database is the technology of taking databases kept on multiple computers connected to a network and making them appear as if they were a single database. Therefore, it is not necessary for users to be aware of which computer actually has the necessary data.

◆ Two-Phase Commitment

Two-phase commitment is a mechanism that ensures the integrity of a distributed database and is important in ensuring the integrity when the database is updated.⁴⁰

Two-phase commitment has two phases. In the first phase, the party requesting synchronization makes a request to processing parties for a guarantee of update operation. At this point, all processing parties are secure.⁴¹ Then, each processing party returns either COMMIT or ROLLBACK⁴² to the party requesting synchronization. In the second phase, the synchronization-requesting party decides whether to commit or roll back, considering the response from each processing party. Specifically, even if only one of the processing parties returns ROLLBACK, the requesting party chooses rollback.

The figure in the next page shows the process flow under normal circumstances. We now use this figure to explain two-phase commitment. “ACK” in the figure is a response message indicating normal completion.

Site A and Site B are the locations where the distributed database is located. The host is the computer that controls this distributed database. When the database is updated, the host gives an updating command (1) to each site. Upon receipt, each site temporarily updates the database. This is a condition where the database can be updated any time, but the database has not yet been updated physically. Further, each site prepares itself for the updating and deleting at any time, once it receives a secure command (2) from the host. This is the first phase. If any of the sites have trouble in this phase, the database updating is cancelled at all of the sites.

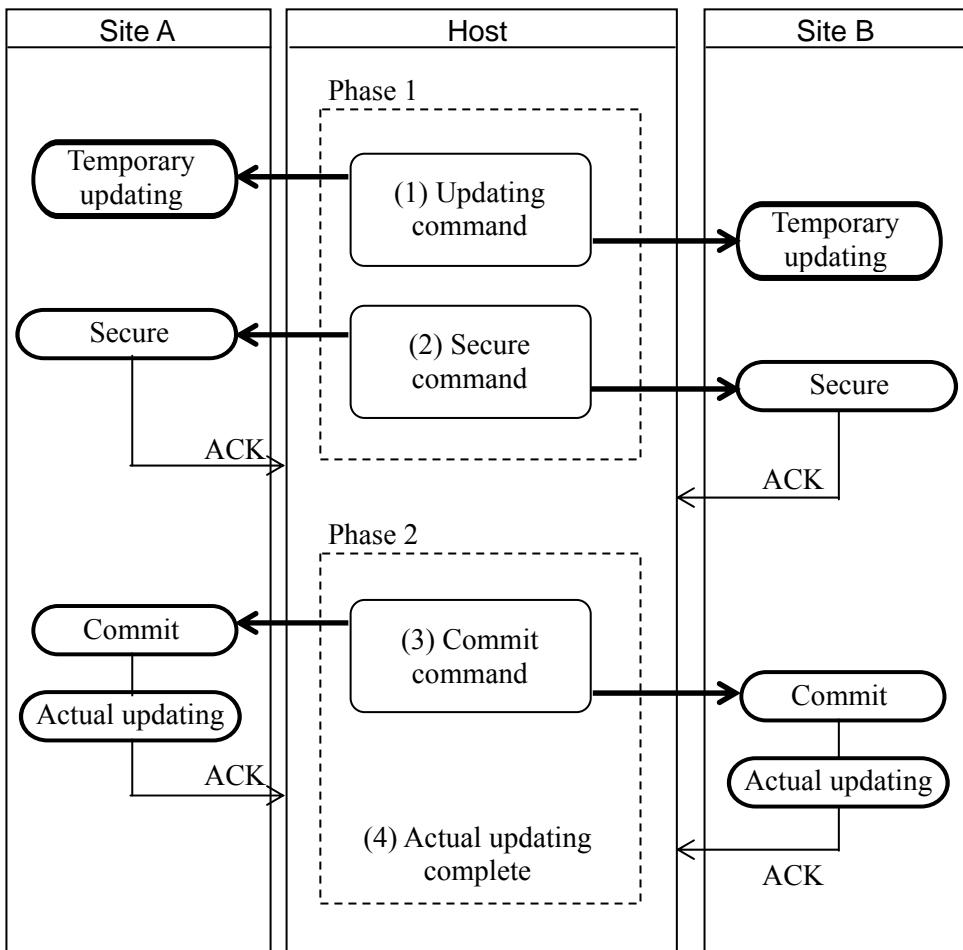
Next, after confirming that each site is ready for the updating (ACK), the host sends a commitment command (3) to all of the sites sequentially. Site A, upon receiving this commitment command, carries out the actual updating and reports the normal completion of the process to the host (ACK). The host then sends the commitment command to Site B, which, likewise, carries out the actual updating and reports the normal completion to the host (ACK). The host then confirms that the entire database is actually updated (4). This is the second phase.⁴³

⁴⁰ **Commitment:** It means finalizing a database updating. Only after this, the process result is maintained. When the application executes a COMMIT command, the update becomes finalized.

⁴¹ **Secure status:** It is a status in which it is possible to complete a processing or to return to the previous status.

⁴² **ROLLBACK:** It means stopping processing and returning related information back to what it was before the processing. This is performed when some trouble occurs during the transaction processing, causing the processing not to be completed normally. Rollback may be done by DBMS but can also be executed by a ROLLBACK command through an application.

⁴³ (FAQ) Many exam questions link database failure conditions with roll-back and roll-forward. These are sure points you can earn if you simply understand the meanings of roll-back and roll-forward.



◆ Replication

Replication is the mechanism of automatically reflecting updated contents in a copy (replica) of the database on the network. The objective of replication is to enhance the responsiveness of the database access in a distributed database environment.

Replicas of the master data are placed on other servers on the network, and when data is updated, the change is automatically reflected onto the replicas. However, the updating of the replicas is carried out asynchronously from the master data.

In general, updating can be only performed on the master, and replicas are only for reference, in order to maintain database integrity.

Quiz

- Q1** Explain the roles of DDL and DML.
- Q2** Explain the ACID characteristics.
- Q3** List what is necessary for recovery when a database fails due to a transaction failure. What is this recovery method called?
- Q4** Explain the two-phase commitment.

Question 1

Difficulty: *

Frequency: ***

Q1. Which of the following operations extracts specific columns from tables in a relational database?

- a) Join b) Projection c) Selection d) Union

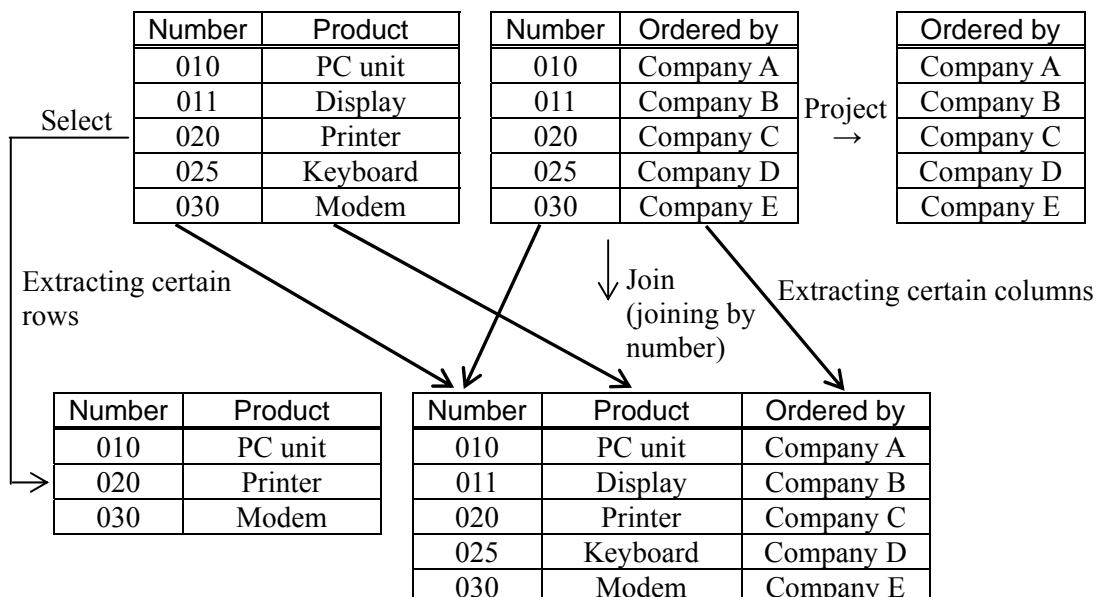
Answer 1

Correct Answer: b

Manipulation of a relational database has the following types.

Types of operation		Contents of operation
Relational operations	Select	Extracts rows satisfying certain conditions
	Project	Extracts columns satisfying certain conditions
	Join	Links tables by column with the same values
Set operations	Union	Extracts rows that are in at least one of the two tables (only one for duplicated rows)
	Intersection	Extracts rows that are in both tables
	Difference	Extracts rows that are in one but not in both tables

The concepts of relational operations are shown below:



Question 2

Difficulty: **

Frequency: ***

- Q2.** Which of the following is the third normal form of the “Skill_record” table? Here, the underlined items represent the primary keys.

Skill_record (Emp_ID, Name, {Skill_code, Skill_name, Experience_years})
 Note: The braces “{” and “}” denote repetition.

- | | | | | | |
|----|---------------|------|-------------------|------------|------------------|
| a) | <u>Emp_ID</u> | Name | <u>Skill_code</u> | Skill_name | Experience_years |
|----|---------------|------|-------------------|------------|------------------|
-
- | | | | | | |
|----|-------------------|------------|-------------------|------------------|--|
| b) | <u>Emp_ID</u> | Name | <u>Skill_code</u> | Experience_years | |
| | <u>Skill_code</u> | Skill_name | | | |
-
- | | | | | | |
|----|-------------------|-------------------|------------------|--|--|
| c) | <u>Emp_ID</u> | <u>Skill_code</u> | Experience_years | | |
| | <u>Emp_ID</u> | Name | | | |
| | <u>Skill_code</u> | Skill_name | | | |
-
- | | | | | | |
|----|-------------------|-------------------|-------------------|------------------|--|
| d) | <u>Emp_ID</u> | <u>Skill_code</u> | | | |
| | <u>Emp_ID</u> | Name | <u>Skill_code</u> | Experience_years | |
| | <u>Skill_code</u> | Skill_name | | | |

Answer 2

Correct Answer: c

Since the skill records come with no explanations, we have no choice but to speculate on the primary key. Here, we may assume it is the employee ID. From the repeated part of the skill records, we can see that one ID number can correspond with multiple skill codes. This implies that one employee can have multiple skills. Hence, the years of experience cannot be determined until two items (the employee ID and skill record) are known. We make the assumption that one employee does not have the same skill code more than once. Otherwise, as explained later, the skill code cannot be the primary key by itself.

First, remove repetition of “Skill_code, Skill_name, Experience_years” from the skill records. The data is then partitioned into multiple records, but if the employee ID is made the primary key by itself, duplication would occur. Hence, it is necessary to define the combination of the employee ID and the skill code as the primary key. This is a first normal form.

In explanations below, underlined items are the primary key.

Skill_records (Emp_ID, Name, {Skill_code, Skill_name, Experience_years})
 → Skill_records (ID, Name, Skill_code, Skill_name, Experience_years)

Next, note that the name is functionally dependent only to the employee ID (partially functionally dependent with respect to the primary key), the skill name is functionally dependent only to the skill code (partially functionally dependent with respect to the primary key), and the “Experience_years” is functionally dependent (completely functionally dependent with respect to the primary key) to “ID, Skill_code” since an employee could have multiple skills.

Now, we need to eliminate partially functional dependency. This is a second normal form. For explanations, we assign a name to each table.

Skill_records (Emp_ID, Name, Skill_code, Skill_name, Experience_years)

- Employee (Emp_ID, Name)
- Skills (Skill_code, Skill_name)
- Skill records (Emp_ID, Skill_code, Experience_years)

After this partitioning, each table has only one non-key item, so there cannot be any transitional functional dependency. Hence, this is a third normal form.

- a) This is a first normal form.
- b) The name can be uniquely determined by the employee ID only, so there is partial functional dependence. Hence, this is a first normal form.
- c) “Experience_years” is functionally dependent on “Emp_ID, Skill_code”. Hence, “Experience_years” is not functionally dependent with respect to the primary key “Emp_ID.”

Question 3

Difficulty: *

Frequency: ***

- Q3.** Which of the following is an appropriate description concerning the primary key of a relational database?
- a) Rows cannot be searched unless a search condition is specified for a column specified in the primary key.
 - b) If a column storing numerical values is specified in the primary key, then that column cannot be used as a subject of arithmetical operations.
 - c) Rows with identical primary-key values are not present in a single table.
 - d) It is not possible to form the primary key comprising multiple columns.

Answer 3**Correct Answer:** c

In a relational database, the primary key is set by combining one or more items in order to identify a row (record). The primary key cannot contain duplicate values within one table.

- a) Rows can be read from the beginning in order.
- b) A primary key is an item or a set of items (columns) that have no duplicate values in the table and does not determine the attributes of the data. Hence, the values can be used for operations as well.
- c) A primary key can be formed by combining multiple columns.

Question 4

Difficulty: **

Frequency: **

Q4. Which of the following is an appropriate description concerning relational database views?

- a) It is not possible to define a view from multiple tables.
- b) When a column is added to an original table, the view must be redefined.
- c) Users must know not only the view structure, but also the structure of the original table itself.
- d) Views are helpful in terms of data protection and data integrity, since the scope in which the data is used can be limited.

Answer 4

Correct Answer: **d**

A view (virtual table) is a description of the database seen from the standpoint of an application. It is a table separately defined by combining necessary items from multiple tables or a single table. Since only the items necessary for the users are defined, the scope in which the data is used can be limited, enabling the protection and integrity of the data. Further, a view derived from a single table can be updated, but a view derived from multiple tables cannot be updated. Even if a view is derived from a single table, it cannot be updated under the following conditions:

- DISTINCT: In DISTINCT, rows with duplicate values are combined into one row, so which row of the original table needs to be updated cannot be determined.
- Set functions, calculations: Values obtained by calculation (for example, the value obtained by the SUM operation, i.e. the total) cannot be updated. In this case, the original data should be updated.
- Subqueries
- GROUP BY, HAVING

A view is called an external schema in a relational database and a sub-schema in a network database.

- a) A view can be defined by combining multiple tables.
- b) A view is not affected when a column is added to the original table.
- c) Since only necessary items are extracted and defined from the original table, the view is not related to the structure of the original table.

Question 5

Difficulty: *

Frequency: ***

Q5. Which of the SQL statements below can acquire Table B from Table A?

[Table] A

emp_ID	name	dept_code	salary
10010	Lucy Brown	101	2,000
10020	Mike Gordon	201	3,000
10030	William Smith	101	2,500
10040	John Benton	102	3,500
10050	Tom Cage	102	3,000
10060	Mary Carpenter	201	2,500

[Table] B

dept_code	emp_ID	name
101	10010	Lucy Brown
101	10030	William Smith
102	10040	John Benton
102	10050	Tom Cage
201	10020	Mike Gordon
201	10060	Mary Carpenter

- a) SELECT dept_code, emp_ID, name FROM A
GROUP BY emp_ID
- b) SELECT dept_code, emp_ID, name FROM A
GROUP BY dept_code
- c) SELECT dept_code, emp_ID, name FROM A
ORDER BY emp_ID, dept_code
- d) SELECT dept_code, emp_ID, name FROM A
ORDER BY dept_code, emp_ID

Answer 5

Correct Answer: d

Table B extracts the department codes, employee IDs, and names from Table A. This is specified as follows:

SELECT dept_code, emp_ID, name FROM A

Note that the records are sorted by department code and, within the same department, by employee ID. This is specified by the following statement. Remember that ASC means “in ascending order”; if this is omitted, the default is ASC.

ORDER BY dept_code, emp_ID

Hence, to get Table B from Table A, we use the following SQL statement:

SELECT dept_code, emp_ID, name ; picking dept. code, emp. ID, and name
FROM ; from Table A
ORDER BY dept_code, emp_ID ; sorting in ascending order of the dept. code and the emp. ID.

Question 6

Difficulty: **

Frequency: *

- Q6.** In a distributed database system, which of the following methods is used to inquire whether multiple sites performing a series of transaction processes can be updated, and can perform a database updating process after confirming that all sites can be updated?
- a) Two-phase commit
 - b) Exclusive access control
 - c) Roll-back
 - d) Roll-forward

Answer 6**Correct Answer:** a

A distributed database is theoretically accessible as one database although it is connected to multiple computers in geographically separate locations such as plants, business offices, and research centers. In a distributed database, if one site finalizes an update process while another site cancels the update process, the data loses integrity. Hence, a special mechanism is necessary in order to ensure the integrity of the databases, which are geographically spread out.

Two-phase commitment is a mechanism by which the integrity of data is maintained when a database is distributed (i.e., distributed database). It has two phases. In the first phase, a party requesting a database update inquires to each of the distributed databases if commitment is possible. Here, each database responds with COMMIT or ROLLBACK, but at this point, the distributed databases hold the update. This situation, in which COMMIT can be performed but the actual update is being held, is called the secure (intermediate) status.

In the second phase, the requesting party examines the response contents of the distributed parties and instructs either COMMIT or ROLLBACK to each database. If any one of the databases has responded with ROLLBACK, the requesting party instructs ROLLBACK to all of the distributed parties. On the other hand, if they all respond with COMMIT, then the requesting party sends a COMMIT command to all of the distributed parties. At this point, the distributed databases are actually updated.

6 Security and Standardization

Chapter Objectives

When information is exchanged through a network, security needs to be ensured, as there is a risk of information leakage and tampering that can occur out of the user's sight. In addition, computer viruses are highly rampant, and, therefore, it is imperative that computer systems and data be protected from these threats. In Section 1, we will learn methods for ensuring security. Meanwhile, the software and data need to be standardized in order to exchange information via a network. Standardization means to set common formats and structures for information. Information can be exchanged without performing any special operations if the information is assembled in accordance with certain standards. In Section 2, we will learn the trends in standardization.

6.1 Security

6.2 Standardization

[Terms and Concepts to Understand]

Encryption, public key cryptography, secret key cryptography, authentication, computer virus, vaccine, firewall, tampering, disguising, ISO9000, MPEG, JPEG, Unicode, EDI, CORBA

6.1 Security

Introduction

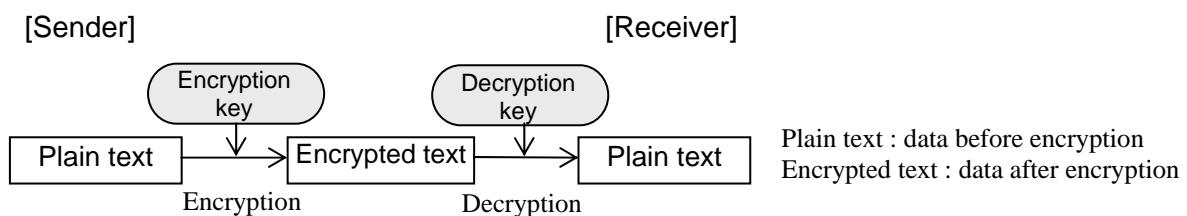
Security means maintaining the safety of computer systems and network systems. One way to prevent unauthorized access is by requiring the user to enter his or her user ID and password. It is also effective to encrypt data to prevent data leakage to a third party.

6.1.1 Security Protection

Points	<ul style="list-style-type: none">➤ Security protection methods include encryption, authentication, and access management.➤ Encryption methods include private key cryptography and a public key cryptography.
---------------	---

◆ Encryption

Encryption means to scramble information by a certain pattern so that a third party cannot understand its contents. It is a highly effective method for protecting information saved in a computer system. Depending on the combination of keys for encryption and for decryption, there are **private key cryptography** (common key cryptography) and **public key cryptography**. The concept of encryption is shown below.¹



Cryptography	Explanations
Private key cryptography	The same key is used as the encryption key and the decryption key (symmetric conversion). The key must be kept secret from others. DES and FEAL are examples of this system.
Public key cryptography	The encryption key is public while the decryption key is kept secret from others. The message is encrypted by the public key of the receiver and decrypted by the private key of the receiver (non-symmetric conversion). RSA is an example. ^{2 3}

¹ (Hints & Tips) Encryption cannot prevent data falsification because it only makes the data unreadable. Also, there is a risk of the encryption pattern breaking if the same key is used for a long time. Be aware that encryption does not offer a perfect solution.

² (Note) Be aware that in the public key cryptography, encryption is performed using the receiver's public key. The public key is available on the network, and anyone can obtain it. However, the decryption key is kept secret. This system is characterized by the fact that symmetric conversion is almost impossible and that the decryption is not possible using the encryption key.

³ **DES/RSA:** An example of a private key cryptography is DES (Data Encryption Standard). An example of a public key cryptography is RSA (Rivest-Shamir-Adleman), named after the initials of the three people who invented it.

◆ Authentication

Authentication means verifying that the user is, without a doubt, a valid user. There are various authentication methods as listed in the table below.

Method	Explanations
Entity authentication	A technology of identifying whether the party with whom we are communicating is valid Often the combination of a user ID and a password is used. Various means are used, including call-back, ⁴ private key cryptography, and public key cryptography.
Message authentication	A technology of detecting any possible falsification in a transferred text or file If falsified, the check bit will be changed.
Digital signature ⁵	A technology of assuring the validity of a document, typically using public key cryptography

Sometimes a one-time (disposable) password is used to enhance security. User ID is also a means for authentication; in general, the user ID is assigned by the system administrator while the password is managed by the user.

◆ Access Management

Access management means preventing unauthorized access to resources (such as data) in a computer system.

To this end, user IDs and passwords are registered in advance in the system, and the user is required to enter his or her user ID and password to gain access to resources and the network. The right of being able to access into resources is called the **access right**.

The following table shows methods of access management to identify individuals.

Type	Explanations
Individual's knowledge	Password, etc.
Individual's possession	ID card, IC card, optical card, etc.
Individual's characteristics	Fingerprint, voice print, hand shape, retina pattern, signature, etc.

Access management by individuals' characteristics is called biometric authentication.

⁴ **Call-back:** It is a method by which the receiver disconnects communication and then reconnects by calling the sender back.

⁵ **Digital signature:** It is a method which applies a public key cryptography. There are various ways to implement this, but the simplest method is to use the public key cryptography in reverse.

The sender encrypts the text with his or her own private (secret) key and adds his or her name in plain text. The receiver, based on the plain name, obtains the public key of the sender (or assumed to be the sender) and uses that public key to decrypt the text. If the decrypted message is readable, the sender is verified; otherwise, the receiver determines that it was sent by someone in disguise.

6.1.2 Computer Viruses

Points

- Computer viruses are programs that execute invalid action.
- A firewall protects unauthorized access from the outside.

With the growing popularity of the Internet, there is a trend that occurrences of various types of damage through networks are on the rise. Hence, it has become increasingly crucial to take measures against computer viruses and to protect against unauthorized access from the outside. Let us explain these issues in detail, including specific measures to be taken.

◆ Computer Viruses

A **computer virus**, or, simply, a **virus**, is a malicious program that enters a system via networks or storage media, and destroys, falsifies, or steals data. A computer virus reproduces itself through networks and storage media. In addition, macro-viruses⁶ can be made easily by almost anyone, so the damage is spreading widely.

In general, a computer virus includes at least one of the following functions.

Function	Characteristics
Self-contagion function	It causes infection by reproducing itself by its own function or by reproducing itself onto another system using a system function.
Incubation function	It contains certain conditions for attack, such as a specific date and time, duration of time, and number of processes; then the virus keeps itself hidden until the attack begins.
Symptom-presentation function	It has functions to destroy files such as programs and data or to execute operations which are not intended by the creator.

◆ Vaccines (Vaccine Software, Computer Vaccines)

A **vaccine** is a program that discovers and eliminates computer viruses. Basically, it contains a prepared database (pattern file) in which patterns of already discovered computer viruses are registered, compares various data on disks and memory with these patterns, and detects any computer viruses. For this reason, it is necessary to have up-to-date vaccines at all times.^{7 8}

⁶ **Macro-virus:** It is a computer virus that abuses macro functions of spreadsheet and word-processing software. Conventional computer viruses required knowledge at the level of machine language, so it was difficult for common end users to create them. However, macro-viruses can be written in programming languages, so they can be created relatively easily. Often they are hidden in files attached to an e-mail.

⁷ (FAQ) In the past exams, many questions involving computer viruses have appeared. Be sure you know well the definition of a computer virus, measures to avoid virus infection, measures to take in case of an infection, and matters related to vaccines.

⁸ (Hints & Tips) A vaccine recognizes patterns of already discovered viruses. Hence, it may not be able to handle new viruses. So the basic idea is to take precaution so that the computer does not get infected with a virus. Once an infection is discovered, it is crucial to take immediate actions so that the infection does not spread any further.

◆ Anti-Virus Measures

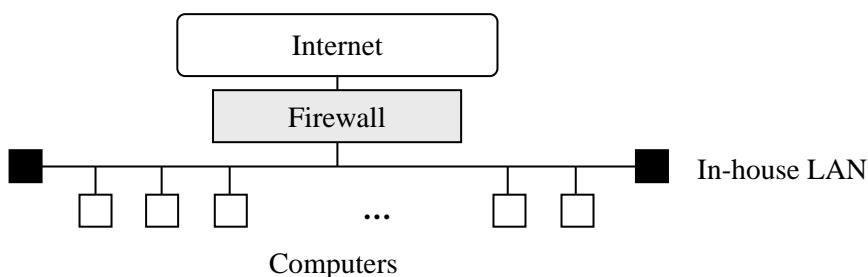
In order to prevent computers from being infected by computer virus, the following points need to be observed:

- Have a vaccine
- Do not copy software illegally
- Do not execute suspicious programs
- Set up passwords and access privileges
- Perform backup periodically
- Do not share disks (clarification of management)
- Do not open suspicious e-mails

If we detect a virus infection, we need to contact the administrator immediately to ask for instructions. Acting on our own judgment can cause further damage by the computer virus.

◆ Firewalls

A **firewall** is a system (mechanism) that protects an internal network such as an in-house LAN from unauthorized access from the outside. More specifically, it is installed between the internal network and the external network, such as the Internet. All communication between the inside and the outside takes place through the firewall. A firewall restricts services available for each user and identifies access from the outside to determine whether or not to allow access into the internal network. Sometimes a computer installed as a firewall is equipped with the functions of a proxy server⁹ as well.



⁹ **Proxy server:** It is a server installed for security protection and high-speed access when an internal network is connected to the Internet. It prevents unauthorized entry into the internal network and relays and manages access from the internal network to the Internet. This function is identical to that of a firewall, so frequently the proxy-server function is carried out by the firewall machine. Additionally, a proxy server has the proxy-response function: data sent from the Internet can be stored here temporarily. Later, when the same Web site is accessed, the access to the Web site can be made faster by turning it at the proxy server.

6.1.3 Computer Crime

Points	<ul style="list-style-type: none"> ➤ Computer crime is an act of entering an information system with a malicious intent. ➤ Computer viruses are one type of computer crime.
---------------	---

Computer crime is the act of entering an information system with a malicious intent and carrying out an action such as destroying data. With the spread of networks, unauthorized users who access networks have begun to appear. Computer viruses are one common means of computer crime.

Computer crimes include manipulating online systems of banks, hacking into remote computers via networks, and placing traps on public domain software.

◆ Falsification

Falsification refers to the act of intentionally changing data or programs in a computer and includes falsifying or modifying a document, replacing storage media with a false version prepared in advance, rewriting data, and erasing data. It is difficult to prevent tampering, but one effective way to detect tampering is message authentication. An example of tampering is the salami method.

The salami method is a way to steal assets little by little from a large quantity of resources. For instance, a user may open a fictional bank account and transfer one or two cents from the other accounts to his or her own account.

◆ Destruction

Destruction is the act of erasing critical data or a program stored in a computer or disabling system devices or storage media by physically destroying them. Examples of destruction include the Trojan horse, logic bombs,¹⁰ and e-mail bombs.¹¹

The **Trojan horse** hides in a program a special instruction not to affect the usual processing and then executes unauthorized functions while letting the program perform its usual objectives. Once a certain condition is satisfied, it may destroy all the files in the computer or steal user IDs and passwords. To prevent a Trojan horse, one can carefully save a backup copy as a real copy and compare a suspicious program with the backup copy to discover the virus. However, it is said that there is no real effective method to prevent the Trojan horse besides checking the source program at the time the program is written or modified.^{12 13}

¹⁰ **Logic bomb:** It is an application of the method used by the Trojan horse. It embeds into the system a process to destroy the system when a certain condition is satisfied (time, situation, frequency, etc.).

¹¹ **E-mail bomb:** It is an act of sending a large number of e-mails or large-size e-mails to a particular person within a short period of time, leading to a failure of the e-mail system.

¹² (FAQ) In the past exams, questions related to computer crime, including items regarding computer viruses, have appeared. Know well the Trojan horse and scavenging.

¹³ (Note) Another type of computer crime is superzapping. This is an abuse of the special function that the system has for emergencies (for instance, a utility program which has access to all the files and through which one can change the contents of those files).

◆ **Leak**

A **leak** is a collective term referring to the robbery of data or copies from an information system. The methods include placing a transmitter on an output unit, mixing confidential data into an output report, and making confidential data appear to be something else by encrypting it. An example of a leak is scavenging (trash hunting).

Scavenging is the act of stealing information from a computer after a job is executed. One may steal information from a document thrown away as trash or information left on the hard disk or in memory. One effective method to prevent scavenging is erasing all information in memory used for temporary storage and on the hard disk upon completion of a job.¹⁴

◆ **Tapping**

Tapping is the act of illegally intercepting data on a network and stealing information or illegally accessing a computer system. Targets of tapping include not only computer data, but also audio data. Encryption is an effective way to prevent tapping.

◆ **Disguise**

Disguise is the act of stealing someone else's user ID and password and acting on a network using the stolen identity. By doing so, the unauthorized user steals confidential information that only the authorized user should access, or commits wrongdoing and blames the authorized user for what he or she has done. Digital signatures are effective in preventing disguise.

¹⁴ (Note) Various methods are in use to maintain security in communication, including “calling number identification” and “closed user group.” “Calling number identification” is a way to notify the call-receiving party of the telephone number of the party making the call. “Closed user group” is to register, in advance, the addresses of the terminals that are allowed to make connection with the electronic exchange unit and make connection only with those terminals.

Quiz

Q1 Enter appropriate words in the blank boxes in the table below.

	Encryption key	Decryption key
Private key cryptography		
Public key cryptography		

Q2 In general, a computer virus is defined as having at least one of three functions. List the three functions.

6.2 Standardization

Introduction

Standardization means to decide on common stipulations, requirements, specifications, structures, and/or formats. Standardized objects can be used without special adjustments. Objects of standardization related to information processing include hardware, software, procedures for systems development, and programming conventions.

6.2.1 Standardization Organizations and Standardization of Development and Environment

Points

- Standardization organizations include ISO, ITU, ANSI, etc.
- Standards include ISO 9000, ISO 14000, etc.

Standardization for information technology is implemented mainly by ISO. Standardization in telecommunications is implemented by the ITU. Other organizations include ANSI, which establishes domestic standards in the United States. Specific standards include the ISO 9000 series for software development and ISO14000 for environmental consideration.¹⁵

◆ Standardization Organizations

The following table includes well-known standardization organizations.¹⁶ ¹⁷

Name	Explanation
ISO	International Organization for Standardization This is an international organization that works to unify and stipulate standards in the industry-related fields. In each field, there is a technical committee (TC), and under TC are subcommittees (SC) and working groups (WG).
ITU	International Telecommunications Union This is an international organization that standardizes telecommunications technologies as well as standardizes and recommends international standards for communications of all kinds. ITU-T ¹⁸ is responsible for telecommunications while ITU-R ¹⁹ is responsible for radio and wireless systems.

¹⁵ (Hints & Tips) Among the standardizing activities of ISO, work in electrical and electronic fields is jointly done with the IEC. Work in telecommunication is jointly done with ITU. Sometimes ISO cooperates with local organizations like ANSI as necessary.

¹⁶ **IEC (International Electrotechnical Commission):** It is a standardization organization set up for the purpose of unifying international standards in electrical and electronic fields. It has now become the telecommunications department of ISO (ISO/IEC), working together as one organization.

¹⁷ **IEEE (Institute of Electrical and Electronics Engineers):** This group has powerful influence in setting standards in areas such as LAN and various interfaces.

¹⁸ **ITU-T (International Telecommunications Union – Telecommunications Standardization Sector):** This is the sector that discusses technologies, operation, and fees related to telephone and telegraph; it also issues its own standards as recommendations. Its major recommendations include the I series for ISDN, the V series for analogue lines, and the X series for digital lines.

¹⁹ **ITU-R (International Telecommunications Union – Radiocommunication Sector):** This is the sector that assigns radio

ANSI	American National Standards Institute This is a non-profit organization involved in establishing industrial standards in the United States and is a member of ISO.
------	---

◆ ISO 9000 Series

The ISO 9000 series is a collective title given to the multiple international standards established by ISO concerning the quality-assuring structure of corporations. This was established in 1987, revised in 1994, and revised once again in 2000 to today's version. The standard for authentication is ISO 9001; other standards indicate items that serve as guidelines to obtain the authentication.

ISO 9001 stipulates the requirements concerning quality management systems for corporations and organizations in the following cases:

- When it is necessary to verify that the company has the ability to provide products that satisfy the client's requirements or applicable required standards
- When the company wishes to achieve improved customer satisfaction

Hence, it is a standard for quality management systems, not a standard for products. Since ISO 9001 is internationally recognized, companies that obtain this can gain international trust.

The structure of the ISO 9000 series is partially shown below:

ISO 9000	Quality management systems - Fundamentals and vocabulary
ISO 9001	Quality management systems — Requirements Management responsibility (customer focus, quality policy, review, etc.) Resource management (provision of resources, human resources, work environment, etc.) Product realization (planning of product realization, customer-related processes, design and development, etc.) Measurement, analysis, and improvement (monitoring and measurement, control of nonconforming products, etc.)
ISO 9004	Guidelines for performance improvements

◆ ISO 14000 Series

The ISO 14000 series is a group of international standards set by ISO concerning environmental protection management. This sets guidelines for measures to be taken by companies to address problems leading to global environmental deterioration such as energy consumption and industrial waste. This is equivalent to the environmental version of the ISO 9000 series and systemizes compliance certification by a third-party organization.²⁰

frequencies and standardizes radio systems, handling satellite communication, fixed wireless communication, mobile communication, television broadcast, etc.

²⁰ (FAQ) Some exam questions ask about the roles of international standardization organizations including ISO and ITU. However, most exam questions assume that you have prior knowledge of these organizations. A good example of such a question is, "Which image compression format is being standardized by a joint organization of ISO and IEC?" Be sure to know at least the names and roles of the standardized organizations mentioned in this book.

6.2.2 Standardization of Data

Points	<ul style="list-style-type: none"> ➤ Character codes include EBCDIC, Unicode, etc. ➤ File formats include JPEG, MPEG, SGML, CSV, etc.
---------------	---

When exchanging data, one of the following processes is necessary: to deliver data after making the data compatible with the format of the receiving party; or to receive data in a different format and then change the format to the receiver's format. Either way, if the formats are inconsistent, it is necessary to change them for the other party. However, this "change of format" can be eliminated if there are standard formats and everyone uses those formats. The data to be standardized here include character codes and file formats.

◆ Character Codes

A character code is a code assigned to each character and symbol for the purpose of processing those characters and symbols on computers. Character codes that can be processed vary depending on the computer.

Code	Explanations
EBCDIC	Extended Binary Coded Decimal Interchange Code Character code established by IBM for general-purpose computers. A set of 8 bits represents one character.
Unicode (UCS-2)	Standard for expressing the characters used all over the world in one integrated character code All characters are expressed using 2 bytes. This is adopted as part of international standards by the ISO.

In addition to the codes listed in the table above, there are ASCII²¹ and EUC²² among others.²³

²¹ **ASCII (American Standard Code for Information Interchange)**: It is a character code set established by ANSI, setting codes for the alphabet letters, numerals, special characters, and control characters such as the new-line (return) code, each using 7 bits. ASCII is adopted as part of international standards by the ISO (ISO 646).

²² **EUC (Extended Unix Code)**: It is any character code, used mainly by UNIX. It can process 2-byte characters as well as 1-byte characters. It is an international standard code established by AT&T and includes Japanese EUC, Korean EUC, and Chinese EUC, etc.

²³ (Note) Unicode was extended after its initial standardization to use 3 or more bytes. Hence, today it is defined such that every character uses 4 bytes in Unicode (UCS-4).

◆ Image Files

An image file is a file in which a still image like a photograph or an illustration is digitized as a file. There are various file formats as listed below.

Format	Explanation
JPEG	Joint Photographic Experts Group: A joint organization of ISO and ITU-T for coding color still images, or the compression/decompression method established by this organization
GIF	Graphic Interchange Format: An image format developed by CompuServe, a large online service company in US It is compatible with color or monochrome image files with 256 or fewer colors.
BMP	Format to save images as bitmap data, standard graphics format used by Windows.
TIFF	Target Image File Format: Expressing data using tags in data blocks within files By using tags, the data format is specified.

◆ Moving Picture Compression

One coding system for moving pictures is MPEG (Moving Picture Experts Group). MPEG is the name of a lower organization of JTC1, which is a joint organization of ISO and IEC. It is a method for compression/decompression of moving pictures, which was established by this group.²⁴

◆ Document File Formats

Document files are standardized so that document data and data prepared using an application can be exchanged easily. Conventionally, document files were exchanged in a format called text file (*.txt). However, text files can only express character strings, not various writing formats such as character thickness, size, color, and document structure. Today, the formats that are standardized include the document styles as well. By standardizing document file formats, documents can be processed in an integrated manner and can also be made into databases.

Currently, the following table shows standardized document file formats.

Format	Explanation
SGML	Standard Generalized Markup Language A language expressing the logical and semantic structures of documents with symbols; a document can be used as a database.
XML	eXtensible Markup Language A language that came after HTML in which the extended functions of SGML can be used on the Web. Users can define their own tags.
HTML	HyperText Markup Language. A language used to make Web pages on the Internet. Tags are surrounded by "< >" to designate character size, color, and hyperlinks.
TeX	Pronounced as "tek" or "tef" A document-formatting and finalizing program to lay out and to print documents with complex mathematical and chemical formulas.
CSV ²⁵	Comma Separated Value Format Each item of the data is followed by a comma and listed. It is mainly used to save data from database software and spreadsheet software.
PDF	Portable Document Format A format developed by Adobe Systems for electronic documents. Documents can be exchanged regardless of the computer model and the platform.

²⁴ (Note) MPEG also codes audio and sound along with the moving images. MPEG has MPEG-1, MPEG-2, MPEG-3, MPEG-4, and more. MPEG-1 is for storage media such as CD-ROM and is a standard for ISO/IEC (ISO 9660). MPEG-2 is an upgrade version of MPEG-1 and is for HDTV (high-definition television) as well as image transmission using broadband ISDN. MPEG-4 is a high-performance code of moving images and audio designed for the Internet and radio communication (mobile communication).

²⁵ (FAQ) Exam questions on the characteristics of SGML, HTML, and CSV have frequently appeared. The key word for each is "markup" for SGML, "hyperlink" for HTML, and "comma" for CSV.

6.2.3 Standardization of Data Exchange and Software

Points	<ul style="list-style-type: none"> ➤ Standardization of data exchange includes EDI, STEP, CALS, etc. ➤ Standardization of software includes CORBA, RFC, OMG, etc.
---------------	---

In order to exchange transaction data between companies, the data being exchanged need to be standardized. One of these standardization concepts is **EDI**.²⁶ Another is **STEP**, which is an exchange standard for product model data. If optimum software products can be freely joined together, including software written by various software manufacturers, a better system can be constructed. For this, software also needs to be standardized.

◆ Standardization of Data Exchange

With the growing popularity of the Internet, a new type of business format has been born. It is to exchange transaction information electronically via a network. For this reason, standardization of data exchange is being considered. The following table shows main standardization methods.

Standard	Explanation
EDI	Electronic Data Interchange (electronic transaction, electronic data exchange)
CALS	Commerce At Light Speed All product-related information is shared from specifications, development, and design to procurement, operation, and maintenance. It is designed to improve productivity, shorten the development period, and reduce costs.
EC	Electronic Commerce Sales are made not at a store or through mail order but on the Internet. Electronic money ²⁷ is being considered as a means of electronic payment.
STEP	Standard for the Exchange of Product Model Data ISO 10303 standard International standard for the exchange of product model data

◆ Open Systems

An open system is a computer system constructed in such a way that, by standardizing the specifications, hardware and software can function without conflict, regardless of the manufacturer. In distributed processing systems, hardware from different manufacturers is often connected together to construct a system; thus, hardware and software need to be standardized.

²⁶ (Hints & Tips) EDIFACT adopted in the U.S. and Europe is used to make data exchange overseas more efficient.

²⁷ **Electronic money:** It is a method of payment using IC or PC communication, characterized by the feature that physical bills and currency are not used. It is used as a means to make payments in e-commerce on a network such as the Internet. In addition, there are IC cards, as small as a business card, equipped with a microprocessor on which an amount is recorded, so that the user can carry it just like cash.

◆ Standardization of Software

For standardization of object-oriented software, there are following software, standards, and standardization organizations.

Name	Explanation
CORBA	Common Object Request Broker Architecture Shared specifications so that objects can exchange messages with each other in a distributed system environment. This is established by OMG (Object Management Group).
EJB	Enterprise JavaBeans Standard specifications to construct Java distributed object-oriented applications. It is possible to combine components using tools from different vendors. It is compatible with CORBA.
RFC	Request for Comments A group of documents on technical proposals and comments which is compiled by IETF. ²⁸ It is available on the Internet and can be obtained by FTP or e-mail. TCP/IP-related protocols, etc., are written in RFC.
OMG	Object Management Group A non-profit organization promoting the popularization and standardization of object-oriented technology. It establishes the industrial standards (OMA) ²⁹ in the field of object-oriented technology.

²⁸ **IETF (Internet Engineering Task Force):** It is an organization for designing and developing Internet protocols and architectures. This group is open to network designers and researchers, and anyone can join.

²⁹ **OMA (Object Management Architecture):** It consists of ORB (object request broker, functions and software used by objects to exchange information with each other), which exchanges messages between objects, a fundamental concept in distributed object orientation; a group of object services, which provide services (CORBA-services) based on ORB; objects that make up application parts; and other components. The common specifications of ORB, which is central, are CORBA.

Quiz

Q1 Describe the contents of the ISO 9000 series.

Q2 (1) What is the compression format for color still images?
(2) What is the compression format for moving pictures?

Question 1

Difficulty: *

Frequency: ***

- Q1.** Which of the following procedures enables a sender to send an encrypted document to a receiver by using a public key cryptography?
- The sender encodes the document by using his own public key, and the receiver decodes the document by using his own private key.
 - The sender encodes the document by using his own private key, and the receiver decodes the document by using a public key.
 - The sender encodes the document by using the receiver's public key, and the receiver decodes the document using his own private key.
 - The sender encodes the document by using the receiver's private key, and the receiver decodes the document by using his own public key.

Answer 1**Correct Answer:** c

A public key cryptography is a system in which the encryption key is publicly released while the decryption key is kept secret. The released key is called the public key, and the one kept secret is called the private key. Unlike in a private key cryptography (common key cryptography), only one encryption key and one decryption key are necessary, so the management of keys is easier. Further, since the decryption key is publicly shared, the key does not have to be sent. But, since the public key cannot be used to decrypt the text, the encryption and decryption can be time-consuming.

In a public key cryptography, the sender encrypts the message by using the receiver's public key while the receiver decrypts the message using his or her own private key.

- In a public key cryptography, what has been encrypted by the public key is decrypted by the private key paired up with the public key. In this option (a), the public key belongs to the sender whereas the private key belongs to the receiver, so they do not make a pair.
- This describes a digital signature.
- A private key is kept secret. In this answer, it is stated that "the sender encrypts the document by using the receiver's private key," but the sender does not have the private key of the receiver.

Question 2

Difficulty: **

Frequency: ***

Q2. Which of the following is an appropriate description concerning computer viruses?

- a) Even if a program file in which a virus lies hidden exists on the computer, as long as the user does not intentionally activate the file, the computer will not be infected.
- b) Viruses destroy the main memory physically and trigger operations not intended by the computer user.
- c) A computer that is updated with the latest engines and signature files for detecting and exterminating viruses will not become infected.
- d) In the virus extermination process, the user can avoid infection from the boot sector by using an OS startup disk that is not infected by the virus.

Answer 2

Correct Answer: d

When a computer is turned on, the first drive where the system goes to read the program is called the startup drive, and the hard disk or the floppy disk used as the startup drive is called the startup disk. A startup disk is prepared in advance to be used in case of emergencies. When the boot sector is infected with a virus, the OS is to be started up from the startup disk which is not infected by the virus.

- a) A boot sector virus (or simply called “boot virus”) enters the boot sector in which the boot program (the program that starts up the OS from the hard disk) is stored and attacks the computer when it is turned on. Thus, virus infection can occur even when the user is unaware.
- b) Since a virus is a program, it does not destroy hardware although it does destroy software.
- c) An engine (software) that detects and eliminates viruses is called a vaccine (vaccine software, computer vaccine). A signature file stores information on previously discovered viruses. Vaccines detect and exterminate viruses by cross-checking the target programs and data with the signature file. Therefore, a vaccine cannot detect or remove a new virus not yet registered in the signature file.

Question 3

Difficulty: *

Frequency: ***

Q3. Which of the following is an appropriate description concerning ISO 9001:2000 certification?

- a) Once certified, the qualification is semi-permanently valid.
- b) There is one certification body per country.
- c) It is a certification for the industrial sector and does not apply to the service sector.
- d) It certifies organizations whose quality management systems meet international standards.

Answer 3

Correct Answer: d

The ISO 9000 series is a collective term referring to the multiple international standards established by ISO concerning the quality management systems of companies. The standard for certification is ISO 9001, and other standards are items that show guidelines for obtaining the ISO 9001 certification. It is not a standard for products; rather, it certifies internationally the quality processes of the companies or organizations based on the following view points:

- They have the ability to provide products that satisfy the customer's requirements or applicable required standards.
- They are working to improve customer satisfaction.

Incidentally, ISO 9001 was revised in December 2000. The required items that used to be distributed are organized into four categories: "management responsibility," "resource management," "Product realization," and "measurement, analysis, and improvement." It is characterized by the concept of a quality management system and continuous improvement of the quality management system.

The "2000" in "ISO 9001: 2000" denotes the fact that it was revised in the year 2000.

- a) A certified company must be examined every year or every six months, and a complete re-assessment is required every three years. Hence, continuous activities are necessary.
- b) There are many certifying organizations. An organization wishing to be certified can choose any certifying organization at its discretion. However, each country has only one accreditation organization that examines and approves certifying bodies.
- c) It applies to all industries.

7 Computerization and Management

Chapter Objectives

In modern society, computers are used in various fields. In our daily lives, we use personal computers at home and work. Computers are also used in corporate accounting systems and production management as well as train seat- and ticket-reservation systems. In this chapter, we will acquire knowledge concerning the development of such systems. In Section 1, we will study information strategies used by companies. Section 2 covers corporate accounting, and Section 3 covers business management. In Section 4, we will study specific examples of information systems using computers.

- 7.1 Information Strategies**
- 7.2 Corporate Accounting**
- 7.3 Management Engineering**
- 7.4 Use of Information Systems**

[Terms and Concepts to Understand]

Chief information officer (CIO), KJ method, brainstorming, decision-support system (DSS), strategic information system (SIS), BPR, balance sheet (B/S), profit and loss report (P/L), depreciation, break-even point analysis, ABC analysis, schedule control, linear programming, inventory control, normal distribution, CAD, FA, POS

7.1 Information Strategies

Introduction

An information strategy is defined as strategic computerization for the purpose of differentiation against competitors in the marketplace. A variety of activities are carried out to achieve this objective.

7.1.1 Management Control

Points

- The chief officer for computerization is a CIO (chief information officer).
- The KJ method identifies essential issues from free discussions.

Management control is an activity that integrates an organization to direct it toward the next action to be taken. It is often said that a company consists of human resources, products, finances, and information. Management control relates the flow of these components with one another, coordinates them, and generates higher value by a guideline called a “management strategy.”

◆ CIO

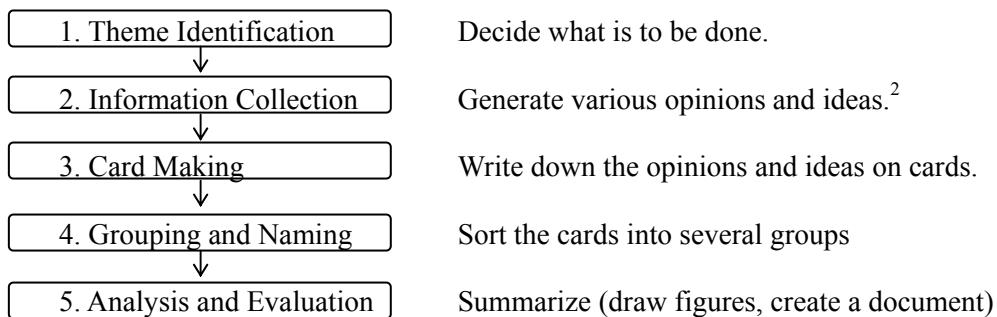
CIO (Chief Information Officer) is the highest-ranking officer in charge of overseeing information systems. Unlike an officer simply in charge of managing the information systems department, CIO is responsible for developing information strategies to effectively utilize the information resources in corporate management. In general, an officer in charge of supervising the information systems department serves as CIO.¹

◆ KJ Method

The **KJ method** is named after the initials of Jiro Kawakida, the inventor of the method. In this method, various ideas are generated to solve one problem, and these ideas are grouped together and related with one another. When an information system is designed, the first step is to listen to the opinions given by the users. The various comments and information collected during this step include many contradictions and conflicts, especially when the number of the users becomes large. The method can be used to effectively identify a true universal need among these contradictions and conflicts.

¹ (FAQ) CIO's meaning and roles have been asked in the past exams. In addition to being knowledgeable about the information systems, CIO is also required to take responsibility for developing information strategies. Therefore, CIO is required to possess a wide range of knowledge including the industry in general, the business of the company, and general administrative functions.

General procedures of the KJ method are illustrated below.



◆ Brainstorming

Brainstorming is a type of meeting which is held under the guideline that absolutely no criticism is allowed on remarks made by the participants. It is characterized by four principles: criticism is forbidden; comments are freely made; quantity is more important than quality; and piggybacking on someone else's idea and position-switching are welcome. These principles facilitate participants to freely express their own ideas and opinions without any restrictions, and, therefore, innovative ideas are expected to be generated during the course of the brainstorming.

◆ OJT and Off-JT

OJT (On the Job Training) is training in an actual work setting. A direct supervisor or superiors provide instructions to subordinates or those with less experience on the skills related to the specific work, including knowledge, techniques, and attitudes, through daily work under clear plans and objectives. Since OJT is not standardized training such as classroom-style training, trainees can have closer interactions with trainers. In addition, it is effective for the trainees in acquiring, improving, and developing the know-how that has accumulated in the company over its history.³

On the other hand, there is **Off-JT (Off the Job Training)**, which is generally considered as typical classroom-style education. This training is targeted for certain employees and is conducted outside the usual workplace separate from their daily work.

² (Note) One method of learning how to conduct information-gathering interviews is role-playing. Four people form one team; one of them acts as the interviewer while another acts as the respondent. Then, the remaining two serve as observers and make comments after completing the role-playing.

³ (Note) A project is an organization that is formed to achieve clearly defined objectives in terms of schedule, cost, and technical performance under predefined time limits; it is dissolved upon achieving its objective. Unlike typical corporate organizations, a project has an objective, a beginning, and an end. In most cases, the job is done by a group of people.

7.1.2 Computerization Strategies

Points	<ul style="list-style-type: none"> ➤ DSS stands for “decision-support system.” SIS is “strategic information system.” ➤ BPR is the restructuring of business processes.
---------------	---

A computerization strategy is to conduct various strategic acts of computerizing to outperform competition, including ERP, CRM, SFA, and CTI. As these matters have already been explained in Chapter 3, we will discuss other topics in this chapter.

◆ Information Systems for Computerization Strategies

There are various information systems for computerization strategies. Appropriate use of these information systems can give a significant competitive advantage.

DSS

DSS (Decision Support System) is a system that supports decision-making by managers and administrators facing non-structured problems (non-routine task). For decision-making in non-routine task, it is difficult to have necessary information defined in advance or to have solution models prepared. Hence, DSS is equipped with a database function,⁴ model base function,⁵ and human interface function.⁶ Using these functions, the user can search for a solution to non-routine task.

SIS

SIS (Strategic Information System) is an information system that actively uses information technology as a part of its corporate strategy to obtain a competitive edge. This includes home-delivery systems for courier services and POS analysis systems at convenience stores.

◆ Work Improvement, Analysis, Design

In order to establish the optimum work flow, it is necessary to review and redesign the work processes.

BPR

BPR (Business Processing Reengineering) is the work of modifying the actual business contents and/or organization, restructuring the business field, based on an analysis of the business contents and business flow, and redesigning for optimization in order to achieve the target level profit or customer satisfaction.

⁴ **Database function:** It is a function that allows free search and analysis of necessary data when a problem occurs.

⁵ **Model base function:** It is a function that chooses appropriate solution models as needed, such as a simulation model or a mathematical model, and performs trial and error.

⁶ **Human interface function:** It is an interface function that allows the database function and model base function to be used easily and interactively.

Business models

A **business model** is a framework for making business concepts concrete. In other words, it is a template for how to carry out business to generate profits. It receives greater public attention as it promotes more distinction (application for patent) by combining business with computers and the Internet through the advancement of IT (Information Technology). A patent of a business model is called a business-model patent.

◆ Business Using the Internet

With the growing popularity of the Internet, new businesses and business structures that had not been commonly seen in the past have been emerging.⁷

e-business/ Dot com business

e-business is a new business structure which takes advantage of the Internet and computers. In an environment of expanding networks, this is an innovative business structure connecting that expansion with the expansion of transactions. It is achieved by defining a business model and making changes in business processes, rules, and organization.

Dot com business (.com business) is a collective term referring to general business activities using the Internet. The term “dot com (.com)” is the domain name indicating the US “company.” Corporations actively doing business on the Internet are called “.com (dot com) companies” or “e-companies.”⁸

SOHO

SOHO (Small Office Home Office) is a term coined by joining the phrases **small office** and **home office**. The former is an attempt to use business resources in and out of the company effectively through networks such as the Internet. The latter refers to working at home by obtaining necessary information by accessing the company server from home and working via network communication. This is a business mode popularized with the growth of the Internet.

⁷ **Virtual company:** It is a corporate structure where a company is set up virtually on a network and is managed by multiple people.

⁸ **EC (Electronic Commerce):** It is a method of selling goods and services on a network such as the Internet instead of at a store or through conventional mail order. A business can be started with little capital, and the operating costs can be significantly reduced because there is no store and just a few people are managing the business. Also it is possible to provide different information to different customers.

Quiz

- Q1** Explain what CIO is.
- Q2** Give the general procedure of the KJ method.
- Q3** List the four principles of brainstorming.
- Q4** Explain BPR.

7.2 Corporate Accounting

Introduction

Corporate accounting is the procedure of reporting the activity status of a corporation to related parties in and out of the corporation; it can be classified into financial accounting and management accounting.

7.2.1 Financial Accounting

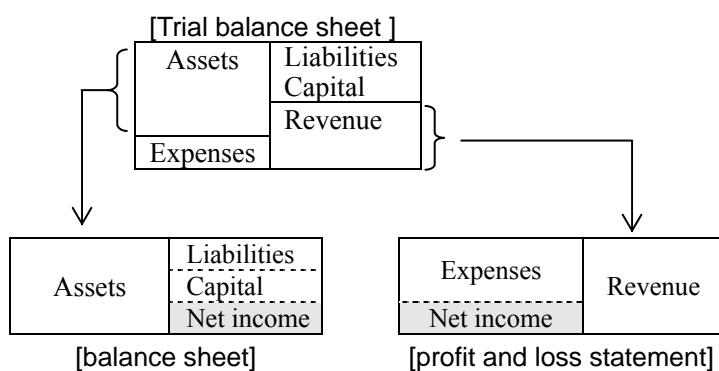
Points

- In financial accounting, the basic documents are B/S and P/L.
- For depreciation, there are declining-balance and straight-line methods.

Financial accounting is the process of reporting accounting information to related parties outside the company such as shareholders and creditors. For this process, a **balance sheet (B/S)** and a **profit and loss statement (P/L)** are necessary.

◆ B/S and P/L

A balance sheet (B/S) is a document that shows the financial status of the company at a particular point in time, indicating the relationship between assets, liabilities, and capital. A profit and loss statement (P/L) is a document that shows the business results for one accounting fiscal year, indicating the relationship between expenses and revenues (sales). Thus, the assets on the balance sheet⁹ are shown on B/S, revenues are shown on P/L, and the difference “revenues – expenses” is the net income. A net income on B/S indicates that capital has increased, and the increase equals the net income on P/L. These are related as shown in the figure below.



◆ Configuration of Account List

An element that forms asset, liability, or capital on B/S and P/L is called an **account item**. Several account items are shown in the following table.

⁹ **Trial balance sheet:** It is a table prepared to check whether or not the transaction is correct when the account is finalized. The total amounts of debit and credit will always be equal.

Assets	total of properties and rights belonging to the corporation (properties)	
Current assets	assets that will be cash in a short period of time (a year or less)	
Quick assets	those with relatively high potential for liquidation	cash, deposits, accounts receivable, notes receivable
Inventory funds	products ready to be sold, etc.	raw materials, merchandise
Fixed assets	assets that do not become cash immediately	buildings, land, patent rights
Liabilities	debts borrowed by the company (money borrowed; needs to be paid some day)	
Current liabilities	to be paid shortly	accounts payable, short-term borrowing, accounts payable
Long-term liabilities	borrowed for a long period of time	long-term borrowing
Capital¹⁰	funds necessary for business activities	
Revenue	¹¹ income of the company	
Operating income	income obtained by the business activities of the corporation	sales
Non-operating income	income from a source other than business activities	interest received, miscellaneous incomes
Expenses	expenditures needed for business activities	
Cost of sales	cost necessary to purchase merchandise	cost of purchases
Selling and general administrative expenses	expenses necessary for business activities	remuneration for board members, payroll, bonuses, welfare expenses, depreciation expenses, travel expenses, miscellaneous expenses
Non-operating expenses	expenses for things other than business activities	interest paid, commissions

◆ Depreciation¹²

Depreciation is a method of reducing the value of a fixed asset by assigning the cost for acquiring the fixed asset¹³ as an expense according to a certain method. As shown below, there are several methods including straight-line and declining-balance methods.

Method	Description
Straight-line method	<p>Find the difference between the acquisition cost and the residual value¹⁴ of the asset. Divide it by the useful life, and that fixed amount is deducted each period as depreciation.</p> $\text{Depreciation for each period} = \frac{\text{acquisition cost} - \text{residual value}}{\text{useful life}}$
Declining-balance method ¹⁵	<p>A certain fixed depreciation percentage is multiplied by the current book value (undepreciated value) of the fixed asset to obtain the depreciation expenses for the period.</p> $\text{Depreciation for each period} = \text{book value (undepreciated value)} \times \text{fixed rate}$

¹⁰ (Hints & Tips) Capital is the fund prepared by the company itself and is also called equity capital. Liabilities are capital borrowed from someone and is called borrowed capital. The capital and liabilities (equity capital and borrowed capital) are together called total capital.

¹¹ (Note) Profits include the following:

Gross profit = sales – cost of sales

Operating income = gross profit – selling and general administrative expenses

Ordinary profit = operating income + non-operating income – non-operating expenses

Generally, the ordinary profit is the corporate profit that gets evaluated.

¹² (FAQ) There will be exam questions that give an account item and an amount and ask you to calculate the operating income and ordinary profit. Know the formulas for various profits well. There are also exam questions where you have to calculate depreciation expenses. Understand well the meaning of the calculation formulas for the straight-line method and the declining-balance method.

¹³ Acquisition cost: amount paid when the asset was purchased.

¹⁴ Residual value: value of the asset anticipated at the end of its useful life. Generally, this is 10% of the cost paid to acquire the asset.

¹⁵ (Hints & Tips) The rate for the declining-balance method is determined by the depreciation duration. For example, if computers are depreciated over 6 years, the rate is 0.319.

7.2.2 Management Accounting

Points

- The break-even point is a point where the operating income is 0.
- Ways to evaluate a company include liquidity ratio and debt equity ratio.

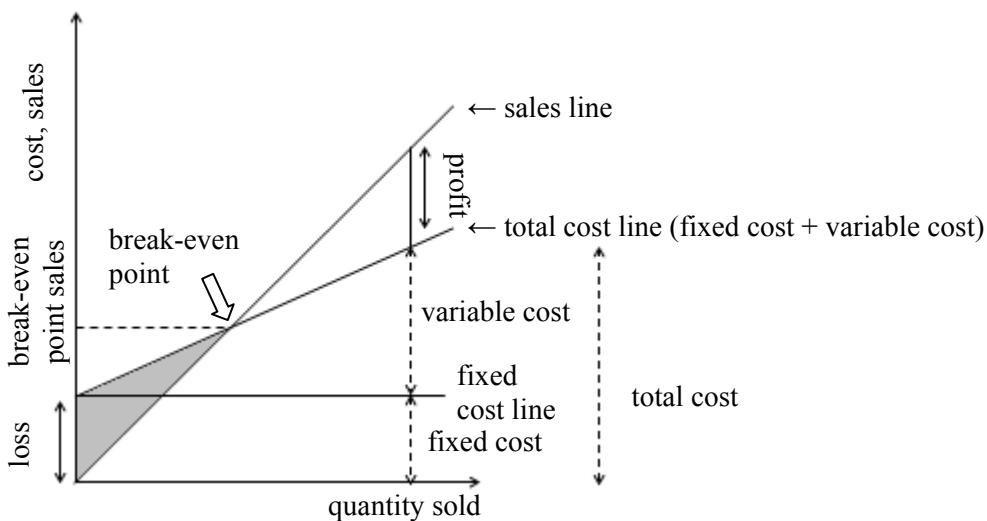
Management accounting is a procedure to provide accounting information to related parties inside the company. It includes break-even point analysis, financial indexes, cost analysis, and inventory evaluation.

◆ Break-Even Point Analysis

A **break-even point** is a point where the revenue (sales) becomes equal to the total cost (variable cost + fixed cost)¹⁶, i.e., the point where the profit becomes 0. Sales below the break-even point result in a loss (red), and sales above it result in a profit (black). By knowing the break-even point, we can identify the amount of sales necessary to avoid losses. This method of management analysis using the break-even point is called break-even point analysis.

In break-even point analysis, we graph the relation “sales = unit sale price * quantity sold” by plotting the quantity sold on the horizontal axis and the sales on the vertical axis. This graph (sales line) is normalized so that it can become a 45-degree line increasing to the right. On the other hand, since the fixed cost is constant regardless of the quantity sold, it is shown as a horizontal line. The variable cost can be indicated by the formula “variable cost = unit price of manufacturing * quantity sold,” so it becomes a line (the value is 0 if no units are sold), also increasing to the right, according to the quantity sold. The sum of the fixed cost and the variable cost is then drawn as the total cost line.

In a graph obtained by following the above procedure, the intersection point of the sales line and the total cost line is the break-even point.¹⁷



¹⁶ **Fixed cost:** It is a cost incurred regardless of the sales, including personnel expenses (payroll), rent, and utility expenses.

Variable cost: It is a cost incurred depending on the quantity sold, such as the material cost.

¹⁷ (Hints & Tips) In principle, the sales line and the total cost line do intersect. The sales line is “unit sale price * quantity sold” whereas the total cost line is “unit price of manufacturing * quantity sold.” Since the unit sale price includes the unit price of manufacturing as well as markup (profit), that is, “unit price of manufacturing + markup = unit sale price,” the slope of the sales line is greater than the slope of the total cost line.

For the break-even point sales, the following equation holds: ¹⁸

$$\begin{aligned}
 \text{Break-even point sales} &= \frac{\text{fixed cost}}{1 - \text{variable cost / sales}} \\
 &= \frac{\text{fixed cost}}{1 - \text{variable cost ratio}} \\
 &= \frac{\text{fixed cost}}{\text{contribution margin ratio}}
 \end{aligned}$$

◆ Financial Analysis

Financial analysis is conducted to evaluate the management records and financial conditions by analyzing the safety and profitability of a company. For indexes in financial analysis, relation ratios are often used.¹⁹

Ratios of safety

Ratios of safety are ratios whereby the debt-paying ability of the company is evaluated. They are shown in the following table.

Ratio	Description
Current ratio	The ratio of current assets, which has relatively high liquidity, to current liabilities that will be due shortly This indicates the short-term paying ability of the company. 200% or more is desirable.
Quick ratio	The ratio of current checking funds, which has high liquidity, to current liabilities This indicates the immediate paying ability of the company, more certain than the current ratio. 100% or more is desirable.
Fixed ratio	The ratio of fixed assets to equity capital Fixed assets are safe to procure by capital, so the smaller this ratio is, the more desirable it is. 100% or less is desirable.
Debt equity ratio	The ratio of the liability guaranteed by equity capital The less debt there is with respect to the equity capital, the safer it is; hence, a small value is desirable here. 100% or less is desirable.
Capital ratio ²⁰	The ratio of equity capital to the total capital, indicating rigidity The more equity capital there is with respect to the total capital, the better it is; therefore, a large value is desirable here. 50% or more is desirable.

¹⁸ (Hints & Tips) Note that the profit used in break-even point analysis is the operating income.

¹⁹ **Relational ratio:** It is the ratio of an account item to other account items, expressed in percentage. Comparison among the account items on B/S is called stationary analysis while comparison among account items on P/L is called dynamic analysis.

²⁰ (Note) The calculation formula for each of the ratios of safety is as follows:

$$\begin{aligned}
 \text{Current ratio} &= (\text{current assets}) / (\text{current liabilities}) \times 100 \\
 \text{Quick ratio} &= (\text{quick assets}) / (\text{current liabilities}) \times 100 \\
 \text{Fixed ratio} &= (\text{fixed assets}) / (\text{equity capital}) \times 100 \\
 \text{Debt equity ratio} &= (\text{liabilities}) / (\text{equity capital}) \times 100 \\
 \text{Capital ratio} &= (\text{equity capital}) / (\text{total capital}) \times 100
 \end{aligned}$$

Ratios of profitability

Ratios of profitability are indexes showing how much profit the company is making. They are classified as shown in the table below.

Ratio	Description
Ratio of return on equity	This shows how much profit there is with respect to the capital, i.e., the efficiency of the capital use. The larger the profit, the better it is, so a large value is desirable.
Ratio of profit to net sales	This indicates how much profit there is with respect to the sales. The larger the profit, the better it is, so a large value is desirable.
Turnover ratio	This indicates the degree to which the assets and capital are used within one accounting period. Since it is desirable to have large sales with little assets and capital, a large value is desirable.

Specifically, there are ratios as listed below. As the capital can change during one period, the mean value at the beginning and at the end of the period is commonly used.²¹

$$\text{Ratio of operating income to total capital} = \frac{\text{operating income}}{\text{total capital}} \times 100$$

$$\text{Ratio of operating income to sales} = \frac{\text{operating income}}{\text{sales}} \times 100$$

$$\text{Turnover of total capital} = \frac{\text{sales}}{\text{total capital}} \times 100$$

◆ Costs

Costs are the expenses required to manufacture and sell products. Depending on what expenses are included, they can be classified as shown in the following figure.

		Sales profit	
	General administrative cost		
	Sales expenses		
	Manufacturing indirect cost		
Direct materials cost		Total cost	Sales price
Direct labor cost	Manufacturing direct cost		
Direct expenses			

Depending on how costs are incurred, they can be classified into materials cost, labor cost, and expenses. Materials cost is the raw price for consuming materials. Labor cost is cost incurred by consuming labor. Costs besides materials cost and labor cost are the expenses.

On the other hand, if the costs are considered to be related to the product, they can be classified into direct and indirect costs. Direct costs are the costs that can be directly calculated for a specific product. Indirect costs are the costs that cannot be calculated for a specific product; they are distributed to various products according to certain criteria.

²¹ (Hints & Tips) When calculating ratios of profitability and using the total capital in the calculation, the average capital value is often used. This is because the capital (including liabilities) is different at the beginning and at the end of the period. The average capital value is the average of the total capital at the beginning and at the end of the period.

◆ Inventory Evaluation

Inventories are defined as assets that will be converted to cash by sales or be consumed for manufacturing the products. Examining the actual amount of the inventories is called inventory evaluation. There are several methods, as shown below, for inventory evaluation.

Method	Description
Last-in first-out method (LIFO)	Inventory is evaluated with the assumption that new products in the inventory go out first, leaving old products (inventory close to the beginning of the period).
Moving average method	A new unit price is calculated using the residual amount and newly delivered amount each time products are brought in. $\text{Average unit price} = \frac{\text{remaining balance} + \text{newly delivered value}}{\text{residual quantity} + \text{quantity newly delivered}}$
First-in first out method (FIFO)	Inventory is evaluated with the assumption that old products in the inventory go out first. New inventory (inventory close to the end of the period) remains.
Periodic average method	The inventory prices and quantities are totaled to find the average, regardless of the time, at the beginning or at the end of a period.

The inventory value is changed depending on which inventory evaluation method is used. Each company decides which method is to be adopted.²²

²² (Hints & Tips) The result of inventory evaluation depends on the economic conditions. For products whose purchasing unit price is gradually increasing, the unit price is higher for the units delivered into the inventory later; in this case, the FIFO method will result in the highest evaluation. If, on the other hand, the purchasing unit price is gradually dropping, the unit price is higher for those units that were delivered to inventory first. Hence, the result using the LIFO method will result in the highest evaluation. The gross-average and moving-average methods will result in intermediate values between the results of LIFO and FIFO methods.

Quiz

- Q1** When finalizing accounts at the end of a period, the following profit-and-loss statement was obtained. Calculate the operating income for the period.

Unit: million dollars

Item	Amount
Sales	150
Cost of sales	100
Selling, general and administrative expenses	20
Non-operating income	4
Non-operating expenses	3

- Q2** Give the calculation formula for current ratio.

7.3

Management Engineering

Introduction

Management engineering is a system of principles and methods for scientifically finding solutions to problems in areas such as production planning, sales planning, and inventory control. It is classified into the two main categories of IE (Industrial Engineering) and OR (Operations Research). OR includes inventory control, linear programming, schedule planning, probability, and statistics, etc.

7.3.1 IE

Points

- Common methods of IE include ABC analysis and QC seven tools.
- ABC analysis is used to find critical control points such as inventory control.

IE is a system of engineering techniques and methods for optimally designing, operating, and controlling human resources, products (machines, equipment, raw materials, auxiliary materials and energy), finances, and information, in order to set out management objectives and to achieve those objectives while taking into account a harmony with the environments (both social and natural environments). It is defined as a variety of activities related to the entire process of production management.²³

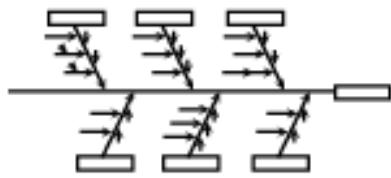
◆ Seven QC Tools

The seven QC tools are tools used to analyze mainly quantitative numerical data as shown in the following table.

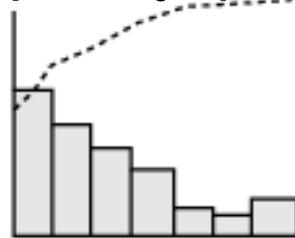
Tool	Description
Cause & effect diagram	Component elements of a certain theme are carefully analyzed, and this diagram clearly displays the structure. Due to its shape, it is sometimes called a “fishbone.”
Pareto diagram	Starting with an item with the largest quantity, the cumulative total is connected by a line while the actual quantity of each item is displayed with a bar graph. Important items and causes are then chosen from the large amount of data.
Histogram	The range of data is partitioned into subintervals, and the frequency of data in each subinterval is counted and displayed with a bar graph. The variation is then understood from the distribution condition, shape, and average.
Scatter diagram	By the look of data spread in the diagram, we can understand the existence and strength of a correlation between two attributes (such as cause and effect).
Check sheet	This is used to collect data for each item and to check for any lack of verification. It is a collective term of diagrams/tables which are easy to understand simply by checking.
Stratification	This refers to classifying obtained data and survey results into items. It is necessary to use graphs so that the difference among the items can be seen at a glance.
Control chart	This is a diagram used to study whether or not a particular process is in stable condition and to maintain the process in stable condition.

²³ (Hints & Tips) There are many definitions of IE as well as many scopes of application. As for the scope of application, the general consensus and interpretation is that it includes the analysis methods and process control centered on work research (not just to determine the efficient work methods by studying and analyzing the work methods and work conditions; this also includes a system of analysis methods for setting fair standard duration). However, some believe that, in a broad sense, IE could involve anything related to management control while, in a more limited sense, it is limited to production management.

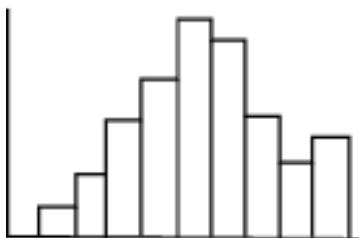
[Cause & effect diagram]



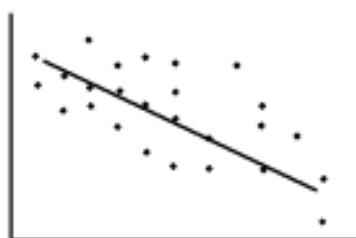
[Pareto diagram]



[Histogram]



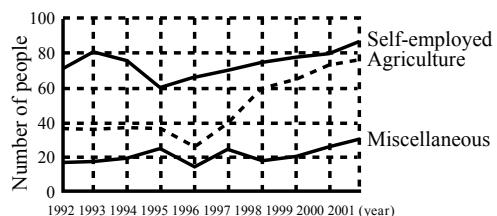
[Scatter diagram]



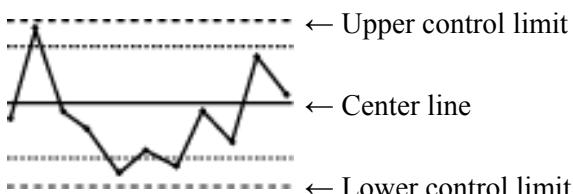
[Check sheet²⁴]

Time period	Check	Frequency
9:00- 9:59		30
10:00- 10:59		24
11:00- 11:59		17
12:00- 12:59		36
13:00- 13:59		18
Total		125

[Stratification]



[Control chart]

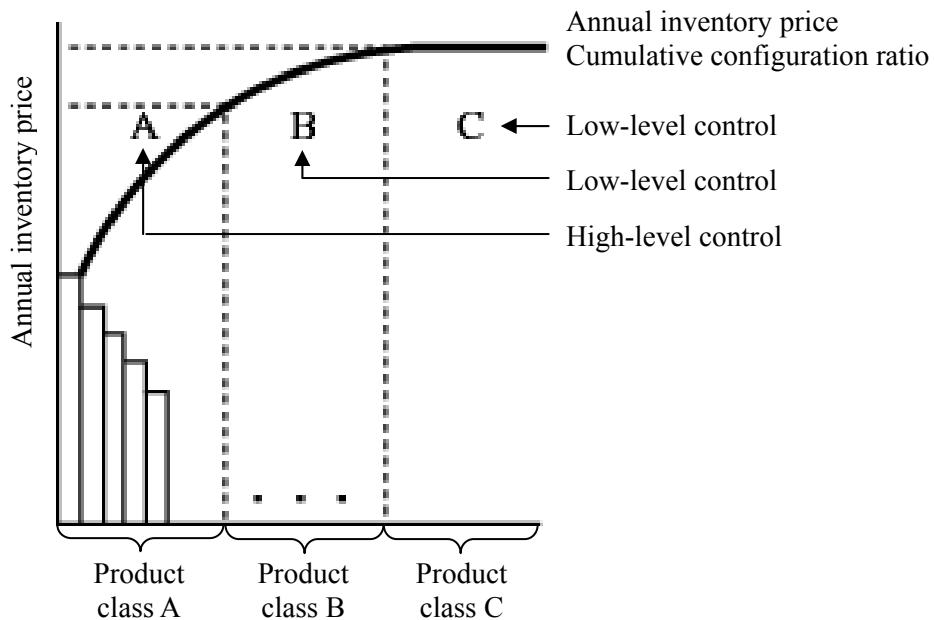


◆ ABC Analysis

Goods in an inventory can be grouped by item of goods, and then each group can be arranged in descending order by inventory price (inventory configuration ratio) or by sales revenue (sales revenue configuration ratio). Then, the cumulative sums can be shown on the same graph so that the inventory can be categorized and managed in 3 groups—Groups A, B, and C. This method is called **ABC analysis**.²⁵ For ABC analysis, we use a Pareto diagram as shown below.

²⁴ (Hints & Tips) For “check sheets” and “stratification,” there is no specific figure or graph format. We can use an appropriate diagram or graph for the objective. The figures shown here are just examples.

²⁵(FAQ) There are many exam questions on ABC analysis. Understand its viewpoint and where to apply it. In application, you may see this on programming tests. For example, as a source for discussing which programs should be managed at a high level, the number of errors for each program may be shown in a Pareto diagram.



In ABC analysis, product group A requires critical (or high-level) control while product groups B and C require relatively low-level management. Product group A represents about 70% of the cumulative configuration ratio, product group B about 70 to 90%, and product group C at 90% or higher.

ABC analysis is a technique of analysis and control based on the Pareto Principle.²⁶

◆ New Seven QC Tools

While the seven QC tools are mainly used in quantitative and numerical data analysis, the **new seven QC tools** provide a method to handle qualitative data such as language data. They are mainly used for problem-solving measurement and strategic planning. The new seven QC tools include “association diagram method,” “affinity diagram method,” “tree diagram method,” “matrix diagram method,” “matrix data analysis method,” “process decision program chart (PDPC),” and “arrow diagram method.”

²⁶ **Pareto Principle:** Only a few factors have significant impact on a certain event while most factors have very little impact. On this basis, product group A has the highest priority to be managed in ABC analysis.

7.3.2 Schedule Control (OR)

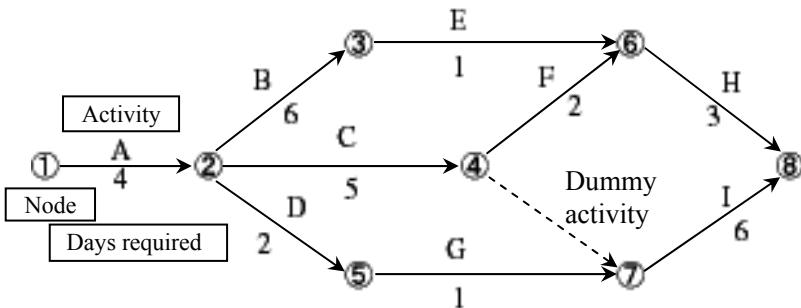
Points

- Arrow diagrams are used for schedule control.
- Mainly, the activities along a critical path are to be managed.

An analytical method called PERT (Program Evaluation and Review Technique) is used for schedule control and process control of a large-scale project. In PERT, after an arrow diagram (chart showing relationships of activities and numbers of days required) is prepared, various analyses are performed to develop an optimum schedule.

◆ Arrow Diagrams

An arrow diagram is suitable for managing large-scale projects in which multiple activities run parallel. An example of an arrow diagram is shown below.²⁷



In this arrow diagram, activity A is the initial work that begins the whole process, and its duration is 4. Following activity A, activities B, C, and D are carried out in parallel. Further, at node ⑥, activities E and F merge. This indicates that activity H cannot be initiated until both of these activities are completed.

◆ Solution by Arrow Diagrams

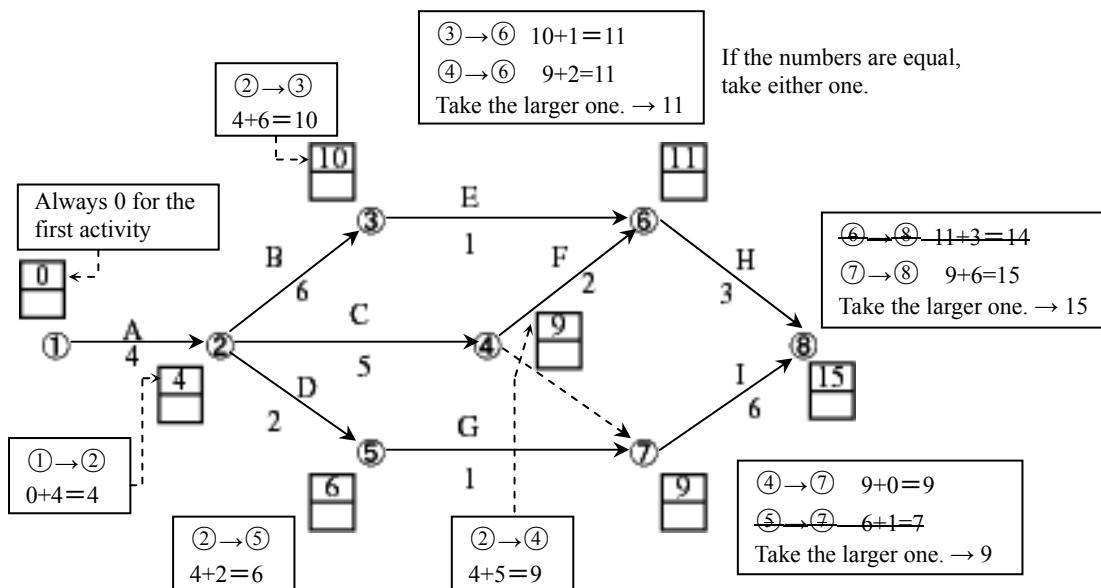
To calculate the number of days required from an arrow diagram, the forward calculation is used to find the earliest node times and the backward calculation is used to find the latest node times. Using the arrow diagram shown above, let us calculate the specific number of days required. First, draw a frame with two boxes near each of the nodes. For convenience in explanation, let us assume that the time required is expressed in “days.”

²⁷ (Note) In an arrow diagram, letters A, B, ... I, shown along the arrows are each called activities. In reality, activity titles such as systems design and programming are necessary. The numbers along the arrows refer to the numbers of days required, which represent durations required for those activities. Instead of numbers of days, we can use hours, dates, or years. However, within one arrow diagram, this unit needs to be consistent.

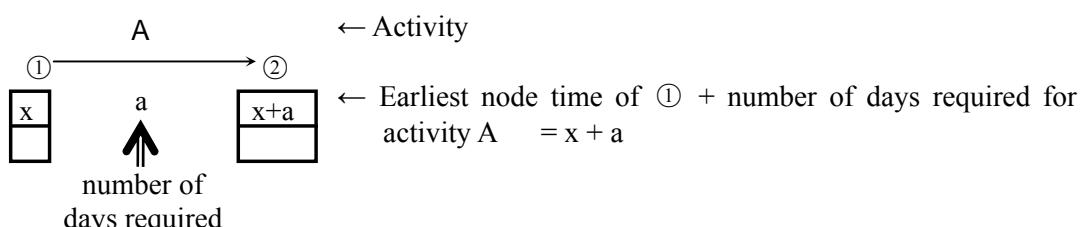
²⁸ **Dummy activity:** It is an activity without any content. A dotted line is used to indicate synchronization. When C is finished, activity F can begin immediately, but activity I cannot begin until activities C and G are both completed. The number of days required for a dummy task is considered 0.

Forward calculation

We calculate the earliest node times²⁹ as we follow the diagram starting at node ①. At nodes where multiple activities merge, we take the maximum number of days required for each path and make it the earliest node time for that merge node. This is because the next activity cannot be initiated until all the activities merging at that node are completed. The calculation result at each node is to be written in the top box at the node. Consequently, we can calculate the number of days required for the entire project.



Basically, if we take the earliest node time at one node and add the number of days required for the next activities, we get the earliest node time for the next node, unless the next node is where multiple activities merge. In that case, we should pay closer attention.

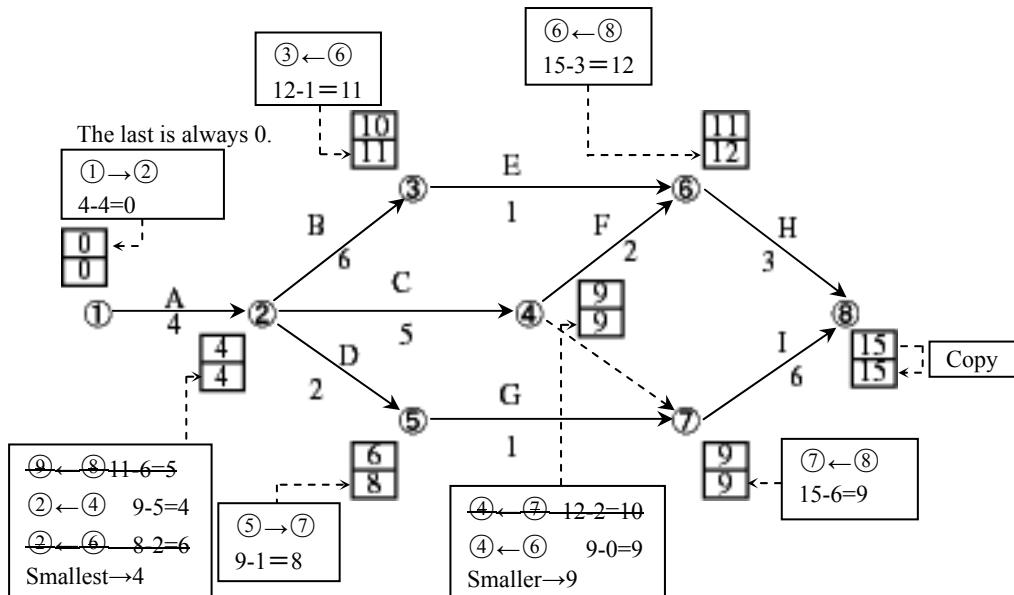


In forward calculation, we set the earliest node time of the first node to "0." Next, since activity A takes 4 days to complete, node ② can be departed 4 days later. In other words, activities B, C, and D can each begin 4 days later. However, at node ⑦, activities merge. The path "A, C, dummy activity" takes 9 days, so activity I cannot begin until 9 days later. However, the path "A, D, G" takes only 7 days, meaning that activity I can begin 7 days later. On the other hand, activity I cannot begin until activity G and the dummy activity (actually activity C) are finished. Hence, for the earliest node time at node ⑦, we need to take the larger of the two numbers.

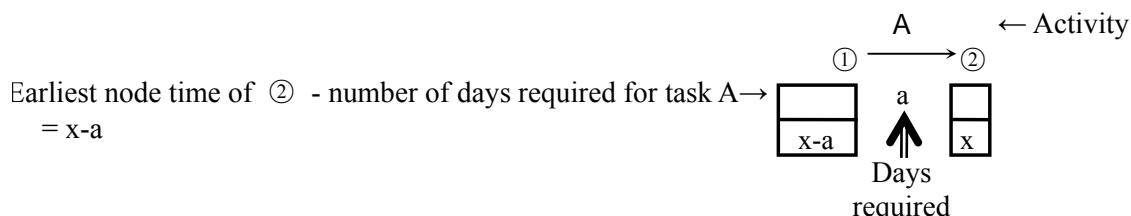
²⁹ **Earliest node time:** It is the earliest time at which an activity may be started when all preceding activities are completed as rapidly as possible.

Backward calculation

Now we calculate the latest node times,³⁰ beginning at node ⑧ and working our way backward. Where multiple activities diverge, take the minimum number of days required for each path, and that becomes the latest node time at that node. The reason is that, since activities diverge at that node, the latest node time needs to be determined by the activity that must begin earliest. The calculation result at each node is to be entered in the bottom box.



Basically, if we take the latest node time at one node and subtract the number of days required for the previous activity, we get the latest node time for the previous node, unless the node is where multiple activities diverge. In that case, we should pay closer attention.



By using forward calculation, we identified that the number of days required to complete the project is 15 days. Further, at the last node, the earliest node time and the latest node time is identical. By using backward calculation, we are now to calculate the latest day on which each activity must begin in order to reach node ⑧. For instance, take node ⑥. Note that node ⑧ needs to be reached 15 days after the project starts, and activity H takes 3 days. Hence, activity H can begin as late as 12 days after the project is initiated. Further, at node ④, if we consider the path "F, H," we see that activity H can be initiated as late as 10 ($= 15 - 3 - 2$) days after the start; however, along the path "dummy activity I," activity H must be initiated 9 ($= 15 - 6$) days after the start. Hence, at node ④, the latest node time must be set to 9.

³⁰ **Latest node time:** It is the latest time at which an activity may be started without delaying the minimum completion time of the project.

◆ Critical Paths

A path connecting nodes where the latest node time and the earliest node time are identical is called a **critical path**. Activities along a critical path have no extra time, so these activities cannot be delayed. If they are delayed, the number of days required for the entire project will change.

In the arrow diagram shown above, a critical path is “A→C→ dummy activity→ I” (①→②→④→⑦→⑧).

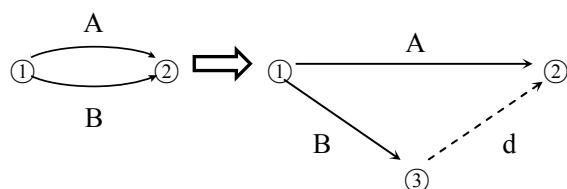
◆ Concept of Schedule Reduction

In the arrow diagram given as an example, consider node ⑤, where the earliest node time is 6 and the latest node time is 8. This means that activity G can begin 6 days after the start, but it is also acceptable to begin this activity 8 days after the start. In other words, there is a 2-day leeway. Hence, even if activity G is shortened by up to 2 days, the required time for the entire project will not be reduced.³¹

In order to shorten a project schedule, it is necessary to shorten the number of days required for activities on a critical path.

◆ Dummy Activities

A dummy activity is an activity without substance and is expressed by a dotted line. This is used only to indicate the time relation between two activities. If there are two activities between two nodes as shown below, this expression on the left can only show one activity, so we insert dummy activity d as shown on the right.³²



³¹ (Hints & Tips) If the schedule (number of days) of some activities in an arrow diagram is shortened, a critical path may change as well. There are situations where shortening an activity on a critical path by 2 days results in the entire project being shortened by only one day. When the number of days required is shortened at some node, the earliest and latest node times need to be re-calculated.

³² (FAQ) There are many exam questions that give you an arrow diagram and then ask you to calculate the number of days required or to find a critical path. If the number of days required is the only thing you need to find, forward calculation is sufficient. However, it is advisable to always perform backward calculation to verify the accuracy of the calculation. In backward calculation, the latest node time at the very first node will always be 0. If you don't get 0, you must have made a calculation error.

7.3.3 Linear Programming

Points	<ul style="list-style-type: none"> ➤ A linear programming question can be answered by reading a graph if two variables are provided. ➤ Since the solution is a production amount or some actual quantity, it cannot be negative (non-negative condition).
---------------	---

Linear programming (LP) is a method that is effective in answering questions where the condition expressions and the target equation are all linear (first-degree). For instance, it can be used when the supply of resource is limited or when the production plan at a factory or the transportation cost for distribution needs to be minimized.

◆ Constraint Conditions/ Objective Function³³

Let us consider the following situation:

“In order to manufacture 1 ton of product A, we need 4 tons and 9 tons of raw materials P and Q, respectively. For product B, we need 8 tons and 6 tons of these materials, respectively. In addition, the profits resulting from products A and B are 20,000 and 30,000 dollars per ton, respectively. However, we only have 40 tons of material P and 54 tons of material Q.”

Let us examine how much of each product should be produced to maximize the overall profit. Let x and y be the amount (in tons) of products A and B to be produced, respectively. The following table summarizes the information above.

	Product A 1 ton	Product B 1 ton	Maximum inventory	
Amount of material P needed	4 tons	8 tons	40 tons	$\rightarrow 4x + 8y \leq 40$
Amount of material Q needed	9 tons	6 tons	54 tons	$\rightarrow 9x + 6y \leq 54$
Profit	20,000 dollars	30,000 dollars	“ $2x + 3y$ ” to be maximized	

From the table, we can write down the following expressions for the constraint conditions and the objective function.

- Constraint conditions:

$$\begin{aligned} 4x + 8y &\leq 40 && (\text{for material P}) \\ 9x + 6y &\leq 54 && (\text{for material Q}) \\ x \geq 0, y \geq 0 &&& (\text{non-negative condition})^{34} \end{aligned}$$

- Objective function:

$$Z = 2x + 3y \quad (\text{to maximize } Z)$$

³³ **Constraint conditions/Objective function:** Constraint conditions are expressions regarding supply limits of the materials, etc. and are almost always expressed by inequalities. The objective function is the expression to maximize the profit, etc. Linear programming is finding the maximum value of the objective function subject to the constraint conditions.

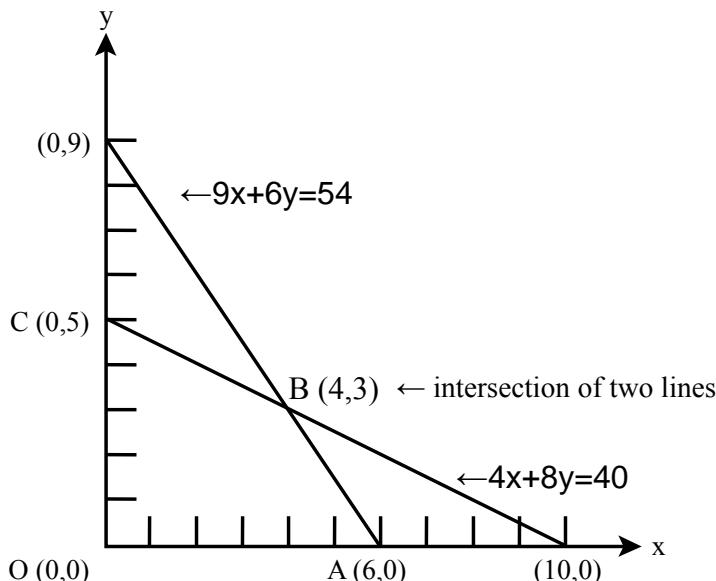
³⁴ (Hints & Tips) The non-negative condition simply means that the value cannot be negative. x and y are amounts to be produced and therefore will never be negative.

◆ Solving Method

We now change the linear inequalities expressing the constraint conditions to the corresponding equations and graph them as lines on the coordinate plane to indicate the range defined by the inequalities, using the x- and y-axes. Note the non-negative conditions: $x \geq 0, y \geq 0$.

$$4x + 8y \leq 40 \rightarrow 4x + 8y = 40 \rightarrow y = -\frac{1}{2}x + 5.$$

$$9x + 6y \leq 54 \rightarrow 9x + 6y = 54 \rightarrow y = -\frac{2}{3}x + 9.$$



The area in which the constraints are satisfied is the region on the graph surrounded by points O, A, B, and C. Regardless of the objective function, a point (x, y) where the objective function achieves its maximum value is proved to be one of the vertices of this area where the constraints are satisfied.³⁵

Hence, we now evaluate the function Z by substitution at the four vertices O, A, B, and C.

$$O : Z = 2 \times 0 + 3 \times 0 = 0$$

$$A : Z = 2 \times 6 + 3 \times 0 = 12$$

$$B : Z = 2 \times 4 + 3 \times 3 = 17$$

$$C : Z = 2 \times 0 + 3 \times 5 = 15$$

The value of Z is maximized at point B $(x, y) = (4, 3)$. Therefore, the maximum profit will be generated when 4 tons of product A and 3 tons of product B are made.³⁶

³⁵ (Note) The optimum solution in linear programming is often the intersection of the lines. You will need to know how to solve a system of linear equations.

³⁶ (FAQ) The frequency at which linear programming questions appear on the exams is rather high, but most of the Morning Exam questions will simply ask you to find the constraint expressions. The Afternoon Exam questions, however, will ask you to find the solution. Hence, you will need to know how to solve these questions by reading the graphs.

7.3.4 Inventory Control (OR)

Points	<ul style="list-style-type: none"> ➤ Methods of inventory control include periodical ordering system and fixed order quantity system. ➤ The optimum quantity to order when the demand is constant can be modeled by the EOQ formula.
---------------	--

Inventory control is a system of optimally managing the inventory of items such as products and raw materials that the company has stored. There are two ordering methods: the periodical ordering system and the fixed order quantity system. The EOQ formula is used to determine the appropriate quantity to be ordered based on balance of the inventory expense and the ordering expense.

◆ Purpose of Inventory Control

If the inventory of a product is large, the cost of storing it (inventory expense) is incurred. If, on the other hand, the ordering frequency increases, the cost of ordering (ordering expense) also rises. Hence, it is necessary to determine the appropriate amount and time for ordering, taking into account both the inventory expense and the ordering expense.

◆ Ordering Systems

The system where the orders are placed at fixed intervals and the quantity ordered varies due to changing demands and other factors is called the **periodic ordering system**. In contrast, the system where the quantity ordered stays constant and orders are placed whenever the inventory is depleted is called the **fixed order quantity system**. The characteristics of each system are shown in the following table.³⁷

		Periodic ordering system	Fixed order quantity system (Order point system)
Reorder time		Orders are placed at fixed intervals.	Ordering intervals vary (orders are placed at order points).
Quantity ordered		Demands are estimated to determine the quantity.	The quantity is constant.
Object items	Period	Cases where irregular orders cannot be placed	Cases where irregular orders can be placed
	Characteristics	High unit prices Strict control necessary	Low unit prices Large quantities in each order
	Control	Cases where there are great advantages in one-time ordering (saving in delivery cost, etc.)	Easy to check the inventory; Easy to keep records in the inventory ledger

³⁷ (Hints & Tips) The periodic ordering system is often applied to the A items in ABC analysis. The fixed order quantity system is generally applied to the B and C items in ABC analysis. A items are usually expensive, so it is necessary to keep the inventory low; hence, the demand for these must be carefully estimated.

◆ EOQ (Economic Order Quantity) Formula

In inventory control where the demand is constant, the optimum quantity to be ordered can be modeled by the EOQ formula. Consider one particular item as the object item, and define the variables as shown below.

Variable	Description
M	Firm demand for a fixed period (e.g., a year)
K	Ordering expense per order
P	Purchasing unit price
c	Inventory maintaining rate ³⁸
x	Quantity ordered per order
n	Number of orders for a fixed period (e.g., a year)

If n orders are placed in a year and the quantity ordered each time is the same, the following relation holds:

$$\text{Ordering expense} = n \times K$$

$$\text{Demand } (M) = n \times x$$

$$\text{Therefore, ordering expense} = (MK) / x$$

Now, let x be the quantity ordered at the time of delivery. Another order will be placed when the delivered units are consumed and the inventory becomes 0. Hence, the average inventory can be considered as $x / 2$. Therefore, the inventory expense can be expressed as follows:

$$\text{Inventory expense} = (x / 2) \times P \times c = (xcP) / 2$$

From the above, the total expense necessary to control the inventory (T) is as follows:

$$\begin{aligned} \text{Total expense } (T) &= \text{Ordering expense} + \text{Inventory expense} \\ &= (MK) / x + (xcP) / 2 \end{aligned}$$

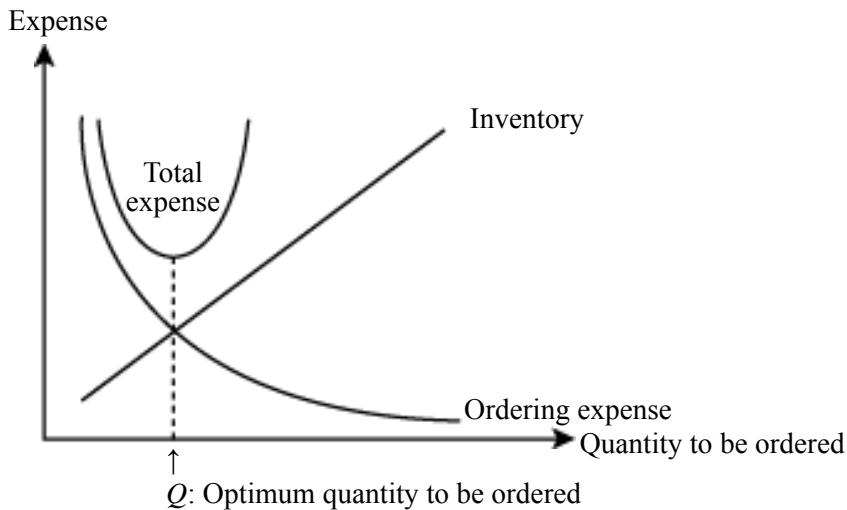
If the value of x ($= Q$) that minimizes T is identified, this value is the optimum quantity to order. As the graph shows, considering the point where the ordering expense equals the inventory expense, the optimum quantity to be ordered can be calculated as follows.³⁹

$$(MK) / x = (xcP) / 2$$

$$x = \sqrt{2MK / cP} \quad (x > 0)$$

³⁸ **Inventory maintaining rate:** It is the rate of cost necessary for maintaining the inventory: “purchasing unit price * inventory maintaining rate = inventory expense.”

³⁹ (FAQ) There will be exam questions where you are asked to calculate the difference between the periodic ordering system and the fixed order quantity system, optimum quantity to be ordered using the EOQ formula, and the number of orders to be placed. Understand the characteristics of the periodic ordering system and the fixed order quantity system. You do not need to memorize the EOQ formula as this will be given in question texts. Be sure, though, that you can use it in calculation.



7.3.5 Probability and Statistics

Points

- Variance and standard deviation are measures of dispersion of data.
- Normal distribution, because of its symmetric nature, is used for testing.

Probability is a value representing the likeliness of an event to occur. For example, the probability that we can get “1” in a roll of a die is 1/6. **Statistics** is to numerically clarify some tendency of a population from its sample. To do this, it is necessary to calculate values such as the mean and variance. When the population is large, normal distribution is used.

◆ Probability

When a die is rolled, one of the outcomes 1 through 6 will result. The numbers 1 through 6 here are called a **random variable**. The probability of each of these outcomes is 1/6, totaling 1.

Consider the following example now. Company X purchases its products from Companies A, B, and C, with percentages 50%, 30%, and 20%, respectively. Suppose that each of these companies has a defective rate of 1%, 3%, and 3%, respectively. One product purchased by X was randomly chosen, and it was defective. What is the probability that this was purchased from Company A?

For example, products from Company A make up 50% of all the products. The defective rate is 1%, meaning that the defective rate of Company A among all the products is obtained as follows:

$$\text{Defective rate of Company A over all products} = 50\% \times 1\% \quad \dots\dots(1)$$

Similar calculations can be performed for Companies B and C as follows:

$$\text{Defective rate of Company B over all products} = 30\% \times 3\% \quad \dots\dots(2)$$

$$\text{Defective rate of Company C over all products} = 20\% \times 3\% \quad \dots\dots(3)$$

The sum of (1), (2), and (3) is the defective rate over all products, and the defective rate of Company A over all products is (1), so the probability that the defective product was from Company A is calculated as follows:

$$\begin{aligned} \text{Probability that the defective product was from Company A} \\ &= (50\% \times 1\%) / (50\% \times 1\% + 30\% \times 3\% + 20\% \times 3\%) \\ &= 50 / 200 = 0.25 \end{aligned}$$

◆ Mean/ Variance/ Standard Deviation

If a sample is drawn from a population, various tendencies of the population can be estimated by calculating the sample mean, sample variance, and sample standard deviation.⁴⁰ For example, if the mean of the sample is 10, we estimate that the mean of the population is also 10.

Let \bar{x} be the mean of the sample x , V be the variance, and σ be its standard deviation.⁴¹ Then, the following equations hold:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$$

$$V = \sigma^2 = \frac{1}{n} \left\{ \sum_{i=1}^n (x_i - \bar{x})^2 \right\}$$

Since the meanings of these equations may be hard to understand, we will explain these concepts specifically. Suppose five sample values have been taken out from the population and they are as follows:

3, 2, 7, 7, 6

The mean is the sum of these sample values divided by the number of values, 5. Below, the underlined value is the number of sample values (sample size).

$$\text{Mean} = (3 + 2 + 7 + 7 + 6) \div \underline{5} = 25 \div 5 = 5$$

To calculate the variance, find the difference between each of the sample values and the mean, square each difference, add up these terms, and then divide the sum by the number of sample values. Below, the underlined value is the mean.

$$\begin{aligned} \text{Variance} &= \{(3 - \underline{5})^2 + (2 - \underline{5})^2 + (7 - \underline{5})^2 + (7 - \underline{5})^2 + (6 - \underline{5})^2\} \div 5 \\ &= (4 + 9 + 4 + 4 + 1) \div 5 \\ &= 22 \div 5 \\ &= 4.4 \end{aligned}$$

The standard deviation is the positive square root of the variance.

$$\begin{aligned} \text{Standard deviation} &= \sqrt{\text{Variance}} = \sqrt{4.4} \\ &= 2.0976 \dots \\ &\cong 2.10 \text{ (rounded to 2 decimal places)} \quad ^{42} \end{aligned}$$

In addition to the above, other measures, including the mode and the median, can be used in order to make estimates concerning the population.⁴³

⁴⁰ **Population/Sample:** In a sample study, the entire set being studied is called the population, and a subset taken from the population is called the sample. Since the population is often unknown, we make estimates about the population based on the sample.

⁴¹ **Expected value/Variance/Standard deviation:** Expected value (or mathematical expectation) is the mean value of the random variable. Variance measures the spread (variation) of the random variable; if the variance is small, the data values are relatively close to the mean value, and we say that the “variation is small.” The positive square root of the variance is called the standard deviation.

⁴² (Hints & Tips) Know the properties of standard deviation.

- The standard deviation does not change if a constant a is added to each data value.
- The standard deviation gets multiplied by a if each data value is multiplied by a .

⁴³ **Mode/Median:** Mode is the most frequently occurring value in the sample. Median is the value in the middle when the sample values are sorted in order. If there are odd sample values, the middle value is unique. If there are even values, take the

◆ Binomial Distribution

Binomial distribution is the discrete probability distribution in which $P(x)$ represents the probability that an event with probability P occurs exactly x times in n trials. Sometimes this is denoted $B(n, P)$. When two dice are rolled, the sum of the two dice gives this probability distribution. The expected value μ and the variance V in binomial distribution are expressed as follows:⁴⁴

$$\mu = nP$$

$$V = nP(1-P)$$

◆ Normal Distribution

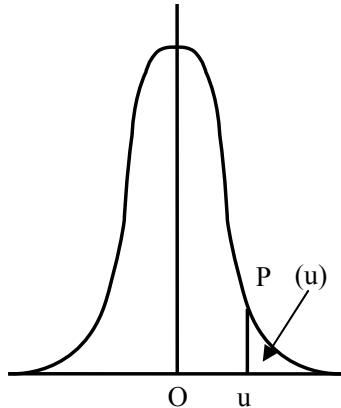
Normal distribution is a continuous probability distribution, and it approximates binomial distribution when the probability P is not small and n is large. Sometimes it is denoted $N(\mu, \sigma^2)$. μ denotes the mean, and σ is the standard deviation. However, in practice, we use the standard normal distribution.

The standard normal distribution is obtained when any normal distribution is converted by the equation $u = (x - \mu) / \sigma$, resulting in $N(0, 1)$. Here, u is the mean of the standard normal distribution, and x is a sample value.

Let us now show an example of the standard normal distribution. Note that the standard normal distribution is symmetric.

[Standard normal distribution table] [Standard normal distribution]

u	P(u)
0.0	0.5000
0.5	0.3085
1.0	0.1587
1.5	0.0668
2.0	0.0228
2.5	0.0062
3.0	0.0013



In the standard normal distribution, if $u = 2.0$, then $P(u)$ represents the area of the region under the curve satisfying " $2.0 \leq u < \infty$ ". If $u = 0.0$, then it is the area of " $0.0 \leq u < \infty$ ". Since this is exactly the right half of the standard normal distribution, the area is 0.5 (50%).

Let us now do a test using the standard normal distribution. Suppose that the dimension of a certain product manufactured at a certain fabrication process is 200mm with a standard deviation of 2mm distributed normally. Assuming that the standard is $200\text{mm} \pm 2\text{mm}$, let us calculate the probability that the product is defective.

mean of the two middle values to be the median.

⁴⁴ (Note) Binomial distribution is used for a discrete random variable while normal distribution is a continuous probability distribution. Normalizing (standardizing) normal distribution gives the standard normal distribution. A discrete random variable is a variable whose values are not a continuum; an example is the outcome of rolling a die. A continuous probability distribution describes a random variable which could take on any value in a continuous interval within a given range. An example is the probability that a defective unit gets produced in a production line.

Since this product has the mean 200mm and standard deviation 2mm, the distribution can be expressed by normal distribution $N(200, 2^2)$. Thus, the random variable 200 ± 2 (the dimension range of the product is 198mm to 202mm) can be converted to the standard normal distribution as follows:

$$u(198) = \frac{198 - 200}{2} = -1.0 \quad u(202) = \frac{2002 - 200}{2} = 1.0$$

Therefore, the product is considered good if its dimension is between -1.0 and 1.0 in the standard normal distribution, and the area of the region $-\infty$ to -1.0 as well as the region 1.0 to ∞ gives the probability that the product is defective. Searching the standard normal distribution table for $P(u)$ for $u = 1.0$, one gets $P(u) = 0.1587$. Hence, the probability we are looking for can be calculated as follows:

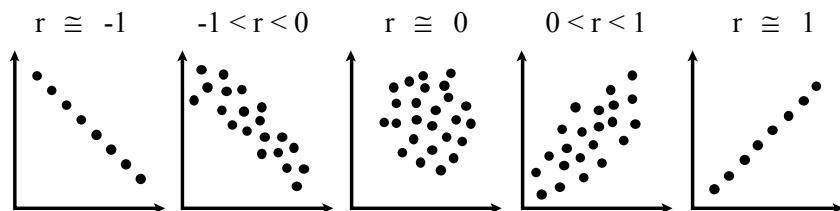
$$\begin{aligned}\text{Probability that the product is defective} &= 0.1587 \times 2 \\ &= 0.3174\end{aligned}$$

◆ Correlation Coefficient

Two quantities are said to have a correlation if there is a tendency that when one increases, so does the other, or when one increases, the other decreases. The numerical value that quantifies correlation is the correlation coefficient (r), which is interpreted in the table below.

Value of correlation coefficient	Decision	Situation
$-1 < r < 0$	Negative correlation	Opposite tendency
$0 < r < 1$	Positive correlation	Same tendency
$r \approx 0$	Weak correlation	Not closely related
$r \approx \pm 1$	Strong correlation	Closely related

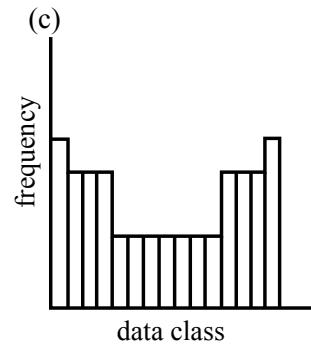
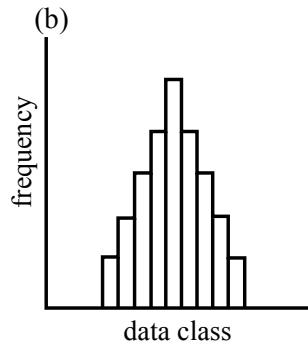
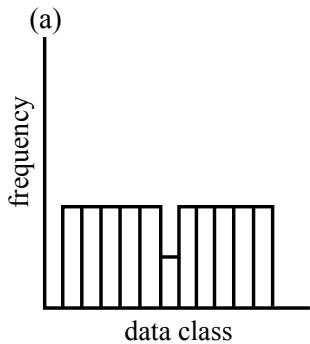
The following graphs show the difference in correlation coefficients schematically.



Quiz

Q1 Explain ABC analysis.

Q2 Which of the following histograms shows the distribution with the largest variance?



7.4 Use of Information Systems

Introduction

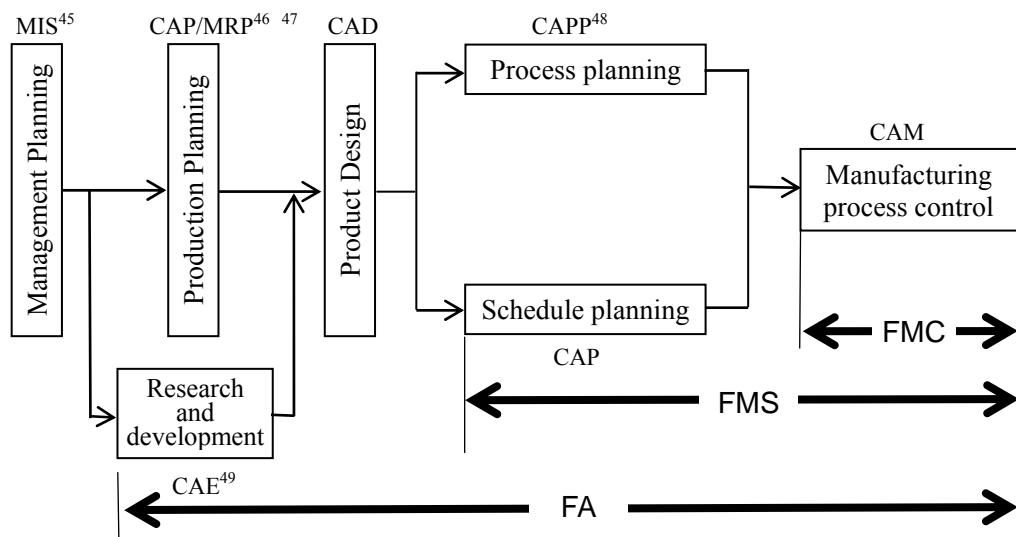
Various information systems are used in companies. These information systems can be classified into engineering systems, represented by FA, and business systems, represented by POS.

7.4.1 Engineering Systems

Points

- FA stands for factory automation.
- CIM is a system of production integrated by computer; it is a concept that includes FA.

An **engineering system** is a system for production automation. Production processes generally have a flow as shown below, and an engineering system supports these.



Let us look at the main components now.

⁴⁵ MIS: Management Information System

⁴⁶ CAP: Computer Aided Planning

⁴⁷ MRP: Material Requirement Planning

⁴⁸ CAPP: Computer Aided Process Planning

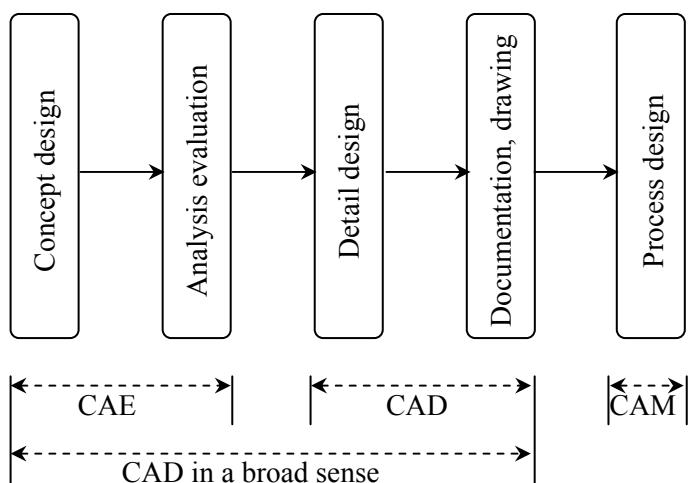
⁴⁹ CAE: Computer Aided Engineering

◆ CAD (Computer Aided Design)

CAD refers to the use of an interactive format to carry out detailed design work, including the minute shapes and dimensions of each part of a product as well as the materials. This is used in a large number of fields such as machine design, construction design, and printed circuit board design. There are various methods for designing the shape of the object being designed, including the following.

Method	Description
Wire frame model	Expression method for a 3-dimensional shape, using vertices and edges
Surface model	Faces placed between wires Expressing the intersection line of two planes and cross sections
Solid model	Expressing the interior solid below the surfaces

CAD, in a broad sense, could refer to all computer-aided processes involving design, but generally the term refers to computer support in regard to shapes and drawing of parts. To clarify this distinction, the process of concept design and analysis evaluation is referred to as CAE.^{50 51}



◆ CAM (Computer Aided Manufacturing)⁵²

CAM is manufacturing work using computers, applied over a wide range of applications, from small-scale manufacturing activities to large-scale processes using robots. Besides computers, other devices and equipment required in manufacturing processes are also considered to be within the scope of CAM.

⁵⁰ **FMS (Flexible Manufacturing System):** It means automation of assembly lines compatible with flexible and multi-model, small-quantity production. It consists of assembling machinery, robots, conveyors, unmanned transportation vehicles, and automatic storage facilities.

⁵¹ **FMC (Flexible Manufacturing Cell):** It means automation of cell processes. A cell is the smallest unit of fabrication and assembly in manufacturing.

⁵² (Hints & Tips) CAM receives data from CAD as they interact with each other; CAM then prepares manufacturing instruction data. In reality, because CAD and CAM are linked closely, they are together called CAD/CAM.

◆ FA (Factory Automation)

FA is a system that organically integrates and manages the entire production system including production planning, ordering, fabrication, assembling, testing, inspecting, transporting, storing, and delivering. Conceptually, FA contains all of CAD, CAM, and CAE, including CAT,⁵³ assembly, fabrication, and process control.

◆ CIM (Computer Integrated Manufacturing)

CIM is support and management using computers to integrate research and development, design, manufacturing, sales, and management control. Whereas FA is a system involving the actual manufacturing site, CIM includes various controls such as ordering, production, and man-hour; hence, it is a broader concept than FA.

7.4.2 Business Systems

Points	<ul style="list-style-type: none"> ➤ POS is used to study the sales tendency of a single product; EOS is used for automatic orders. ➤ Bank POS enables shopping using cash cards.
---------------	---

A **business system** is a comprehensive term referring to administrative application systems. It is used for the purpose of supporting corporate activities and enhancing business efficiency and effectiveness.

◆ POS (Point of Sales)

A POS system is a real-time data-processing system where sales data such as product codes and prices are entered at the point of sales using a terminal unit called a POS terminal located in supermarkets and retail stores. The data is expressed as OCR characters and barcodes, and the POS terminal unit is equipped with an OCR reader or a barcode reader.⁵⁴

◆ EOS (Electronic Ordering System)

A POS system can obtain the sales trend for each product, and, based on this sales information, an EOS manages the inventory properly to ensure that the store is sufficiently stocked and excessive inventory is reduced. When the inventory gets low, the inventory restocking data of the supermarket or the retail store are entered at the terminal unit, and orders are placed online, in real-time, to manufacturers or headquarters.

⁵³ **CAT (Computer Aided Testing):** It is a system where computers are used to conduct various characteristic tests on parts and products during the developing process of a product. This may also refer to a system in which computers are used to inspect products during the manufacturing process.

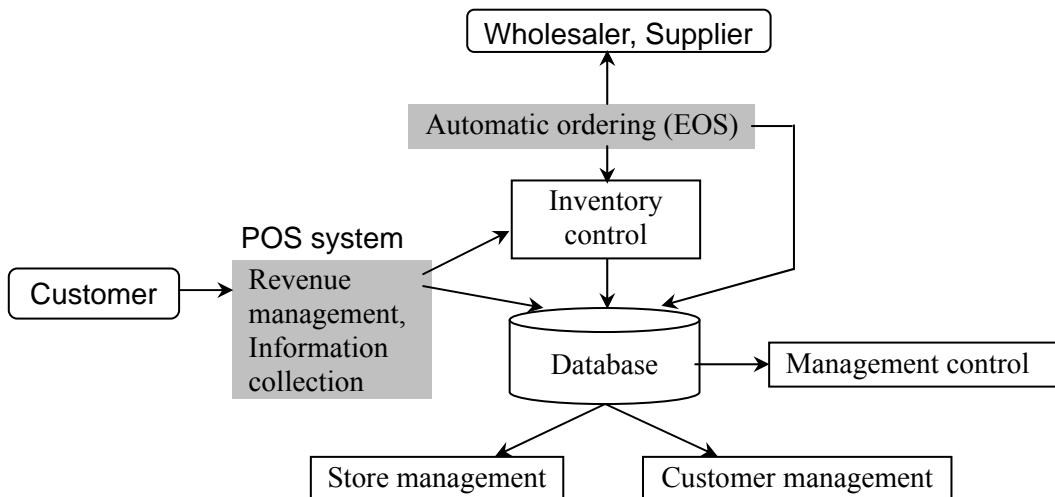
⁵⁴ (Hints & Tips) The use of a POS system could bring about the following results: shorter waiting time at checkout, certain and accurate cash register entry, improved customer service, assessment of sales promotion effects, reduction in employee training, and automation of tallying (statistical) work.

Conventional ordering management has problems including the following:

- Takes a long time from ordering to delivery
- Takes a long time to check the inventory
- Tends to incur order omissions
- Requires a high level of knowledge of the product for product inspection

EOS is said to be best-suited to solve the above problems.

The relationship between POS systems and EOS is shown below.



◆ Electronic Banking

Electronic banking is a system in which computers of financial institutions and computers or terminals of individuals and corporations are connected by communications lines whereby data such as fund transfer and balance inquiries is transferred electronically. It includes the following services.

Name	Description
Firm banking	Connecting financial institutions and corporations Fund transfer such as inter-account transfer and deposit Inquiries of deposits and withdrawals, etc.
Home banking	Connecting financial institutions and individuals Account balance inquiries Fund transfer such as inter-account transfer and deposit Application for fixed-term deposit, etc.
Internet banking	Providing banking services on the Internet PCs can be used to check balance, pay utility bills, and transfer money.

◆ Card Systems

Card systems including those listed in the table below are provided to maintain the customer base or to bring in new customers. Payment methods and the need for an ID depend on the card. Bank POS is used as debit cards.⁵⁵

Type	Point card	Prepaid card	Credit card	Bank POS card
Format	Point service	amount prepaid; no name	amount paid later; paid all at once or by installments	amount paid immediately
Purpose	keeping customers	keeping customers	absorbing customers	absorbing and keeping customers
Identification function ⁵⁶	Yes	No	Yes	Yes
Payment function	No	Yes	Yes	Yes
Record function	Yes	Yes	No	No

◆ Groupware

Groupware is a family of software used to efficiently communicate and share information within an organization such as a company.⁵⁷ Whereas a business system handles company-wide regular tasks, groupware is used for decision-making in irregular types of tasks such as schedule control of meetings.

To do joint work in a group more efficiently, it is highly effective to use PCs and networks. For instance, a group can use a PC-LAN to send and receive e-mails, manage schedules of jobs and meetings, and communicate within the group to get the joint work done smoothly. Tools of groupware include the following.

Tools of groupware	Electronic mail	Sending and receiving messages with specific persons
	Electronic bulletin board	Communication with an unlimited number of people
	Data sharing (documents, data)	Exchanging ideas with multiple members
	Schedule control	Database
	work flow	Control of document flow

⁵⁵ **Debit card:** It means a service in which a cash card issued by a bank can be used to make payment when shopping. The amount of payment is directly deducted in real-time from the bank account.

⁵⁶ (Hints & Tips) Identification function is the function to verify the identity of the person. Payment function means the ability to make payment just as cash (bills). For instance, a prepaid card does not have the ID function, so whoever has the prepaid card can make purchases. Record function refers to the capability of a card to keep a record on itself.

⁵⁷ (Hints & Tips) The original meaning of groupware was intellectual joint work as a group. However, then it would imply that the human work itself is groupware. So, in its revision, this term now refers to a system that supports joint work in an organization by the use of computers, by providing a variety of services via a network, including electronic mail, electronic bulletin boards, electronic conference, and conference room reservations.

◆ PC Communication

PC communication means connecting PCs by a communications line to a host computer providing a PC communications network so that PCs can communicate with one another and receive various services, including electronic mail, electronic bulletin boards, electronic conference rooms, and information services.⁵⁸ In information-providing services, all kinds of information are provided, such as news, weather forecasts, sports updates, market updates, corporate information, and classified advertisements.

◆ Commercial Databases

A commercial database is a database that provides, for a fee, business information such as information on a company, science/technology-related information, and patent information. Generally, this service uses databases via a communication line such as a PC communication service.⁵⁹

Quiz

Q1 Explain CAD.

Q2 What is the name of the card system whereby payments can be made by a bank-issued cash card and the money is immediately paid?

⁵⁸ (Hints & Tips) PC communication and the Internet are similar in that both provide services through communications lines. However, in PC communication, a company providing the PC communication has a host computer installed, and services are provided through this host computer. On the other hand, the Internet does not have a designated host computer.

⁵⁹ (Note) One of the means of data transfer between companies is EDI (Electronic Data Interchange). EDI is a standardized data format for electronic commerce and its procedures.

Question 1

Difficulty: **

Frequency: **

Q1. Which of the following is an appropriate description concerning the development of an overall information system plan?

- a) CIO collects all systemization requests from each user department and proceeds in sequence, starting with those that can be launched immediately.
- b) CIO makes adjustments with business plans, studies technology trends, etc. and establishes an overall plan as a mid/long-term plan. Next, CIO obtains approval and support for the plan from top management.
- c) The leaders of the individual user departments work as key persons and consolidate individual plans to form an overall plan.
- d) Specialists in telecommunications in the information systems department develop an overall plan, taking into consideration leading-edge technologies.

Answer 1

Correct Answer: b

The overall plan of an information system requires setting strategies and targets of the information system and writing down the entire subject area to be included in the information system in an outline form. Policies such as the organization of the system building, applicable tasks, and information technology are to be clarified, the entire schedule is to be established, and approximate investment effects are to be estimated.

CIO (Chief Information Officer) is the highest-ranking officer in charge of information systems. From the top management viewpoint, CIO sets a mid- to long-term plan. CIO also directs the implementation of the plan with approval and instructions from the top management (Chief Executive Officer). Typically, the officer in charge of the information systems department becomes CIO. CIO is not only required to have knowledge on information systems but also held responsible for establishing computerization strategies; therefore, he or she must have a wide range of knowledge encompassing the industry in general, the business of the company, and general administrative functions.

- a) CIO may gather systems requirements to establish computerization strategies, but this is not the main task of CIO.
- c) The overall plan is not formed by summarizing various individual plans that come from the bottom up. Rather, the plan is established top-down and is implemented.
- d) The overall plan is established by CIO. The experts on information technologies in each department implement the plan under the CIO's direction.

Question 2

Difficulty: ** Frequency: **

- Q2.** What is the sales cost in thousands of US\$ for the current term if product inventory sales at the beginning of the term were \$20,000, product purchasing costs for the term were \$100,000, and product inventory sales at the end of the term were \$30,000?
- a) 50 b) 70 c) 90 d) 110

Answer 2**Correct Answer:** c

The sales cost is the expense incurred for the sales of the product. In this case, it is the total of the product purchasing costs for the products sold.

The product inventory sales at the beginning of the term are the assessed value of the products in the inventory at the beginning of the term. The product purchasing costs for the term are the purchasing cost of the products purchased during this term. The product inventory sales at the end of the term are the assessed value of the products in the inventory at the end of the term.

Hence, the product costs for the products sold during this term are the product inventory sales at the beginning of the term plus the product purchasing costs for the term, minus the product inventory sales at the end of the term. This then is the sales cost.

$$\begin{aligned}
 \text{Sales cost} &= \text{product inventory sales at the beginning of the term} + \text{product purchasing} \\
 &\quad \text{costs for the term} - \text{product inventory sales at the end of the term} \\
 &= \$20K + \$100K - \$30K \\
 &= 120 - 30 \\
 &= 90 \text{ (thousand US dollars).}
 \end{aligned}$$

Question 3

Difficulty: **

Frequency: **

Q3. Which of the following is an appropriate description of a break-even point?

- a) If fixed costs do not change, the break-even point rises when variable cost ratio declines.
- b) If fixed costs do not change, the break-even point falls by half when variable cost ratio falls to half their original level.
- c) Sales at the break-even point are equal to the sum of fixed and variable costs.
- d) If variable cost ratio does not change, the break-even point rises when fixed costs decline.

Answer 3

Correct Answer: c

A break-even point is a point where the sales and expenses are equal to each other, indicating that the profit is 0. If the sales during a certain period are less than the sales at the break-even point, a loss will result; if they exceed the sales at the break-even point, a profit will result.

In break-even point analysis, the focus is placed on the relationship between fixed costs and variable costs. Fixed costs are expenses that are incurred with a certain fixed amount regardless of any changes in sales. These include land, lease, depreciation, insurance fees, real-estate taxes, and others. Variable costs are, on the other hand, expenses that change in correlation with the sales; they include materials costs, for example.

Let S be sales, F be fixed costs, V be variable costs, and P be the profit (target profit). The following relationship holds:

$$\text{Profit} = \text{sales} - (\text{variable costs} + \text{fixed costs}) \rightarrow \text{Sales} = \text{fixed costs} + \text{variable costs} + \text{profit}$$

$$S = F + V + P \quad \dots\dots(1)$$

Variable costs (V) are expenses directly proportional to sales, so if the constant of proportion is v, then the following equation holds:

$$V = vS \quad (v \text{ is the variable cost ratio.}) \quad \dots\dots(2)$$

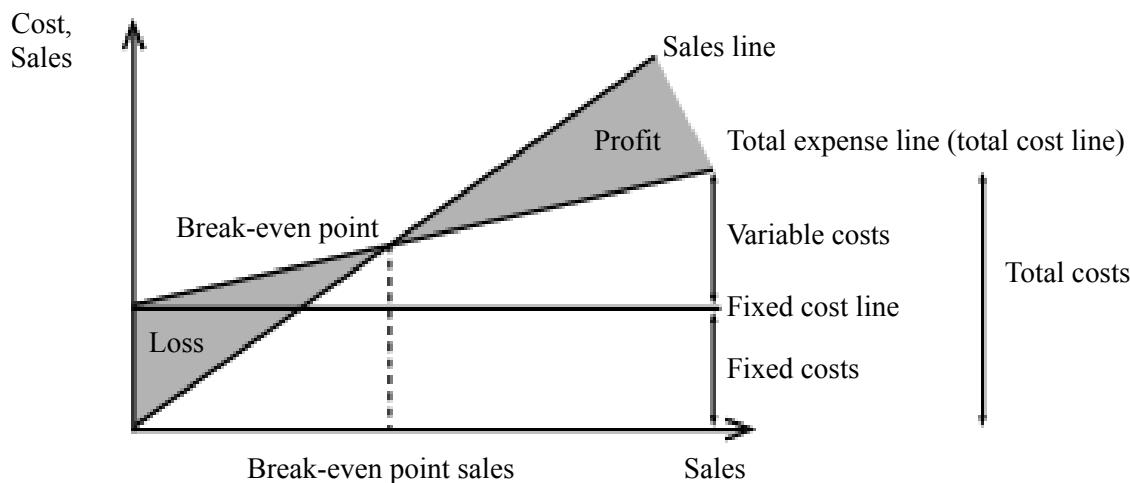
Plugging Equation (2) into Equation (1), we get the following:

$$S = F + vS + P \quad \dots\dots(3)$$

In either (1) or (3), we can solve for sales S that makes profit P equal 0, and that will be the break-even point.

$$\begin{aligned} \text{Break-even point sales} &= \frac{\text{fixed costs}}{1 - \text{variable costs / sales}} = \frac{F}{1 - V/S} \\ &= \frac{\text{fixed costs}}{1 - \text{variable cost rate}} = \frac{F}{1 - v} \quad \dots\dots(4) \end{aligned}$$

A graph showing the break-even point is called a break-even point chart. In the chart, the sales at the point where the variable cost line and the sales line intersect is the break-even point.



As shown in the chart above, if the sales are less than the break-even point sales, there is a loss; inversely, if the sales exceed the break-even point sales, there is a profit.

The break-even point sales are the sales that make the profit 0, so the profit in Equation (1) is 0. In other words, sales are equal to the sum of fixed costs and variable costs.

- a) In the formula for finding the break-even point sales, if fixed costs do not change and variable cost ratio decreases, the denominator increases ($1 - \text{variable cost ratio}$), so the break-even point sales decreases.
- b) Substitute 0.5 for v in Equation (4) for break-even point sales, and then compare that result to the result of plugging in 0.25 (which is half of 0.5, the first variable cost ratio). Note that the result is not halved.
 $\text{Break-even point sales } (v = 0.5) \quad S_{0.5} = F \div (1 - 0.5) = 2F$
 $\text{Break-even point sales } (v = 0.25) \quad S_{0.25} = F \div (1 - 0.25) \approx 1.333F$
- d) In the equation to find the break-even point sales, if the fixed costs decrease while variable cost ratio remains the same, the numerator (fixed costs) decreases, reducing the break-even point sales.

Question 4

Difficulty: **

Frequency: ***

Q4. Which of the following is an appropriate description concerning ABC analysis?

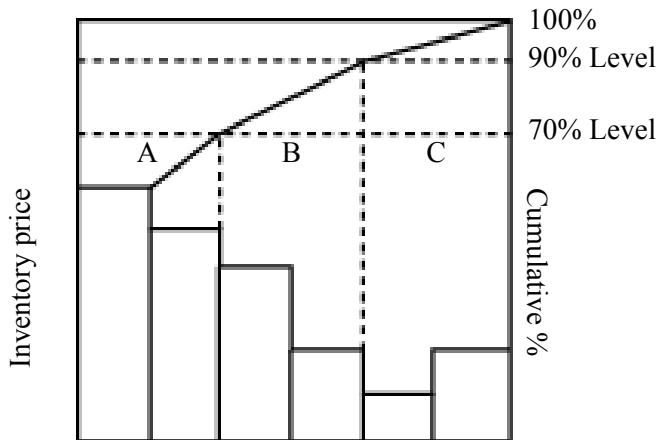
- a) Data such as combinations of products purchased by customers is analyzed, based on sales information collected by a POS system.
- b) A region which is specified by division into a grid with specific longitudes and latitudes is analyzed in great detail by collecting and analyzing various data, including population and purchasing power.
- c) For a certain objective, a region is divided into three parts and opinion leaders are selected for each part. Surveys are conducted repeatedly to identify regional trends and conditions.
- d) Products are sorted by sales or gross profit in descending order. Based upon their cumulative percentages, the products are categorized into three ranks, and product analysis is conducted to identify top-selling products.

Answer 4

Correct Answer: d

ABC analysis is a method where an inventory is grouped according to product items and then each group is sorted in descending order by the inventory price (inventory configuration ratio) or the sales revenue (sales revenue configuration ratio); the cumulative sum is then calculated so that the inventory can be managed for each product item. The result of ABC analysis is expressed with a Pareto diagram.

In ABC analysis, the inventory is categorized into 3 groups: Group A is carefully managed while Groups B and C are managed with relatively lower priority. This is based on the Pareto Principle, which states that, for many events, only a few factors have significant impact while most other factors have very little impact.



As shown in the graph, item groups are listed in descending order by price (configuration ratio). A curve is then drawn by connecting the cumulative sums. The items are grouped such that Group A makes up about 70% of the configuration ratio, Group B between 70 - 90%, and Group C the remaining items. Different management methods are applied for each of these groups.

In general, Group A receives close management attention, and the periodic ordering system is applied. For Group B, the fixed order quantity system using the EOQ formula is applied. For Group C, the fixed order quantity system where an order is placed when the inventory reaches a certain level, or the 2-bin system, is used.

- a) This is an explanation of the basket analysis (simultaneous purchase analysis). Basket analysis identifies the cross-selling opportunities by analyzing “which product and which product tend to be purchased together (i.e., there is a correlation).” For example, there is a well-known correlation: “in supermarkets, disposable diapers and beer are often sold simultaneously.” It was then discovered that men sent to the store to buy diapers often end up buying beer as well. Consequently, when a store placed diapers and beer close to each other, the sales grew.

By finding these correlations, stock of an item seemingly unrelated to some crucial product could be expanded and the sales could grow. The name comes from the idea of looking into customers' shopping baskets to find correlations.

Basket analysis is used in a variety of fields such as purchase data analysis in the retail industry and relational analysis on option requests at telephone service companies.

- b) This is an explanation of cross tabulation.
- c) This is an explanation of the Delphi method as repeated surveys are mentioned. The Delphi method is a logical projection technique used in long-term future projection and technology projection; it is classified under intuitive methods. Intuitive methods are methods of projection or prediction based on human experience and knowledge.

The Delphi method takes advantage of the feedback characteristic. In this method, opinions of a large sample of people are collected and analyzed through questionnaires, and the results of the surveys are summarized, shown to the respondents, and then the survey process is repeated. This method has many advantages. First, it is effective when projecting unpredictable and discontinuous technology changes as it employs an intuitive method. It can also help avoid being influenced by the group dynamics that tend to come from regular face-to-face meetings, etc. In addition, when a comment collected from the survey is different from the majority's opinion, invaluable new ideas can be obtained from the reasons added by the respondent. Hence, the formulation and selection of survey questions are vital to the success of this method.

Question 5

Difficulty: **

Frequency: ***

- Q5.** The table below indicates weather changes at a particular location. For example, on the day following a clear day, there is a 40% chance that the weather will be clear, a 40% chance that it will be cloudy, and a 20% chance that it will be rainy. If the change in weather is a simple Markov process, what is the probability that the weather is clear two days after it rains?

				Unit: %
		Clear next day	Cloudy next day	Rainy next day
Clear	Clear	40	40	20
	Cloudy	30	40	30
Rainy	Rainy	30	50	20

a) 15

b) 27

c) 30

d) 33

Answer 5**Correct Answer:** d

A Markov process means that the probability that an event occurs at a particular time depends on events that happened prior to that time. In a Markov process, to find the probability of an event in the future based on probabilities of past events, we need to go back a finite number of steps. In a simple Markov process, we go back only one step.

The probability that the weather is clear two days following the given rainy day is as follows:

$$\text{rainy} \rightarrow \text{rainy} \rightarrow \text{clear}: 0.2 \times 0.3 = 0.06$$

$$\text{rainy} \rightarrow \text{clear} \rightarrow \text{clear}: 0.3 \times 0.4 = 0.12$$

$$\text{rainy} \rightarrow \text{cloudy} \rightarrow \text{clear}: 0.5 \times 0.3 = 0.15$$

Hence, the correct probability is as follows:

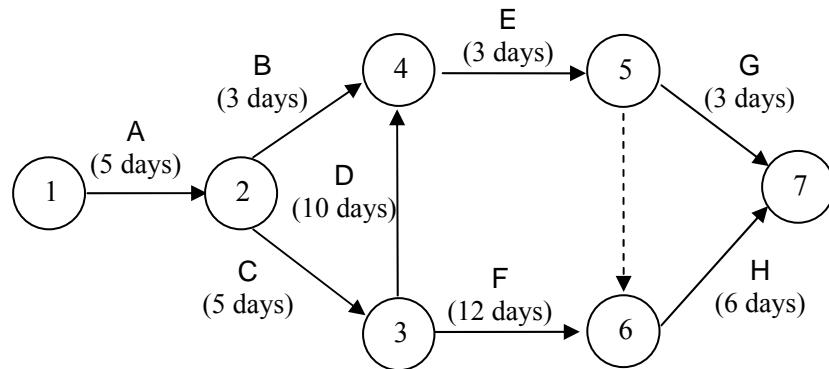
$$0.06 + 0.12 + 0.15 = 0.33 \rightarrow 33(\%)$$

Question 6

Difficulty: **

Frequency: ***

- Q6.** In the arrow diagram shown below, after each activity was reviewed, it was identified that only activity "D" can be reduced by three days. How many days can be reduced to complete all the activities ("A" through "H")? Here, a dotted-line arrow indicates a dummy activity.



- a) 0 b) 1 c) 2 d) 3

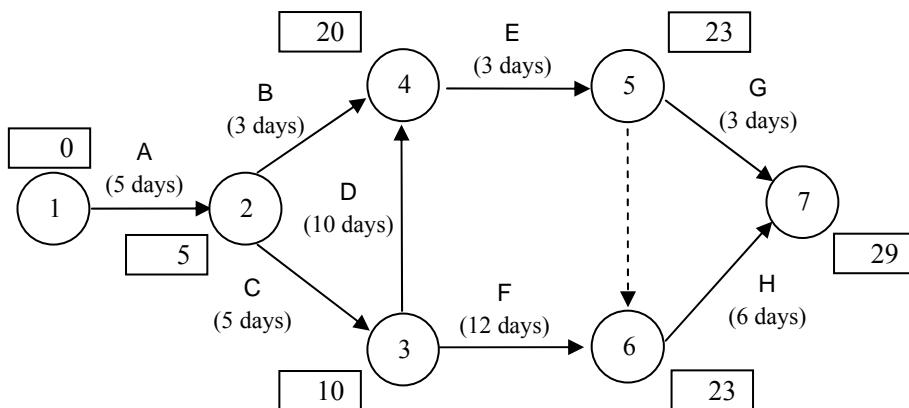
Answer 6

Correct Answer: b

We calculate the earliest node time at each node before and after the shortening. We assume that the dummy activity takes 0 days.

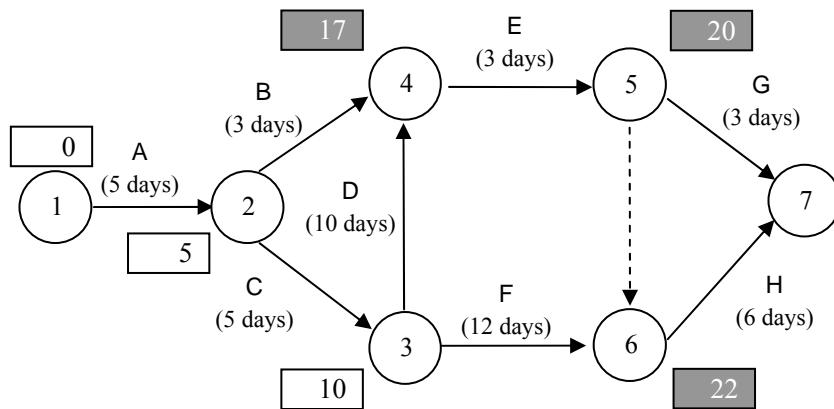
1. Earliest node times before the shortening

Each of the numbers shown below indicates the earliest node time at the respective node.



2. Earliest node times after the shortening

By reducing the number of days activity D takes by 3 days (from 10 days to 7 days), the earliest node times at the shaded nodes change.



Hence, overall, the entire work can be reduced by one day.

Question 7

Difficulty: *

Frequency: ***

Q7. Which of the following provides comprehensive support to a series of production-related tasks with the use of a computer?

- a) CIM
- b) EOS
- c) OA
- d) POS

Answer 7

Correct Answer: a

- a) CIM (Computer Integrated Manufacturing) is the concept of integrated management that uses computers in every aspect of manufacturing work including material-ordering control, production control, and process control. This is a broader concept compared to FA and includes FA and CAD/CAM/CAE as their components.
- b) EOS (Electronic Ordering System) is a system to efficiently help stock items at the store and to reduce residual inventory items. POS system analyzes the sales tendency for each item, and this sales information is used to help stock the goods at the store.
- c) OA (Office Automation) is the idea of bringing in office machines and equipment such as workstations and word processors to enhance the efficiency of information processing in the office.
- d) POS (Point Of Sales) is a system that collects sales information in real-time at the cash register and analyzes the information. Barcodes attached to or printed on the products are read by a barcode reader, and the information is automatically collected.

Question 8

Difficulty: *

Frequency: ***

Q8. Which of the following systems exchanges data between enterprises and is used in EC (Electronic Commerce)?

- a) CA
- b) EDI
- c) SET
- d) SSL

Answer 8

Correct Answer: b

EDI (Electronic Data Interchange) defines the data format for electronic data exchange on a network and its procedures so that electronic commerce can take place between different companies.

- a) CA (Certificate Authority) is an agency that certifies that a public key is valid when, for electronic commerce, etc., digital signatures based on a public-key cryptography are used.
- c) SET (Secure Electronic Transactions) is the specifications for secure processing of credit card payments on the Internet. It was developed jointly by Visa International and MasterCard International of the United States.
- d) SSL (Secure Sockets Layer) is a security protocol between a WWW server and a WWW browser. It enables authentication and encryption by combining public-key and private-key cryptography.