ĐẠI HỌC QUỐC GIA HÀ NỘI
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ

BÀI GIẢNG
TIN HỌC CƠ SỞ

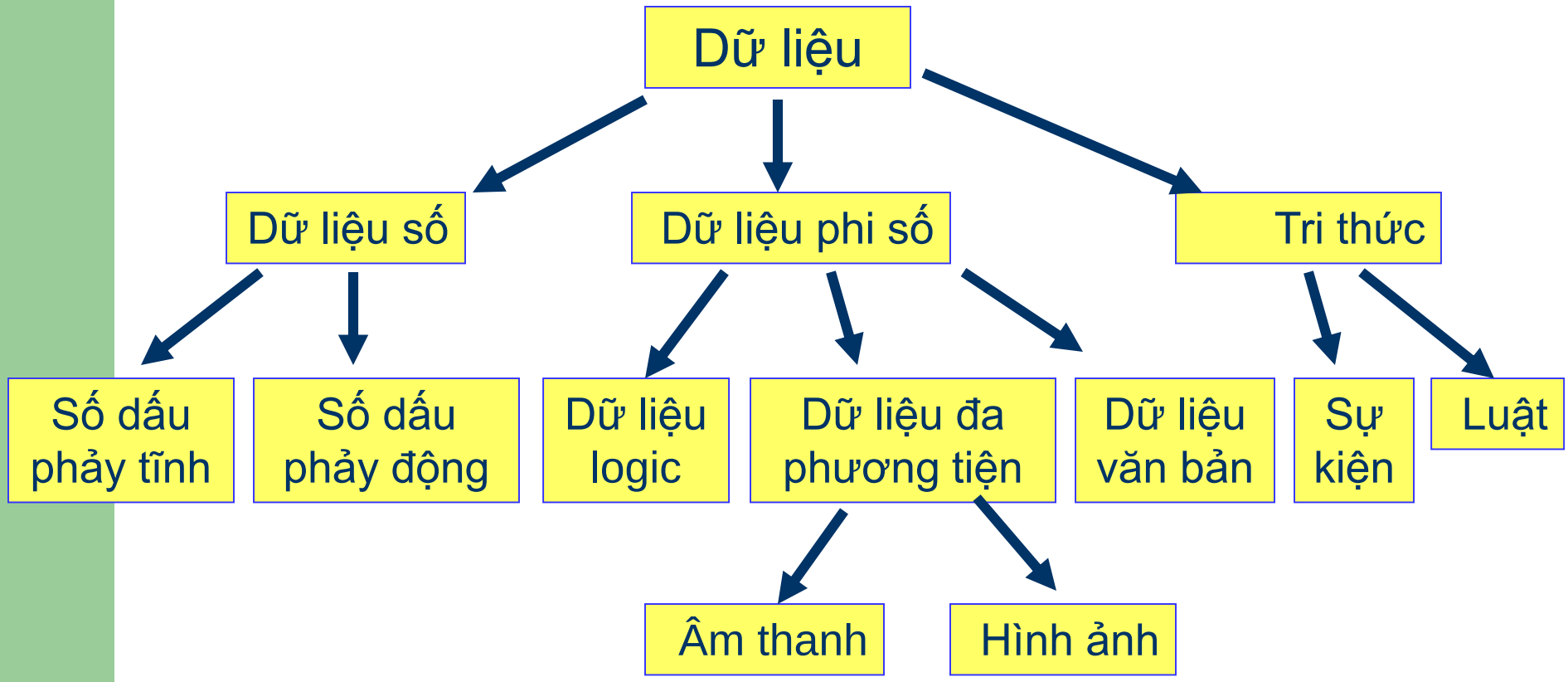
BÀI 6. BIỂU DIỄN DỮ LIỆU TRONG MÁY TÍNH

Giảng viên: ĐÀO KIẾN QUỐC
Mobile 098.91.93.980
Email: dkquoc@vnu.edu.vn

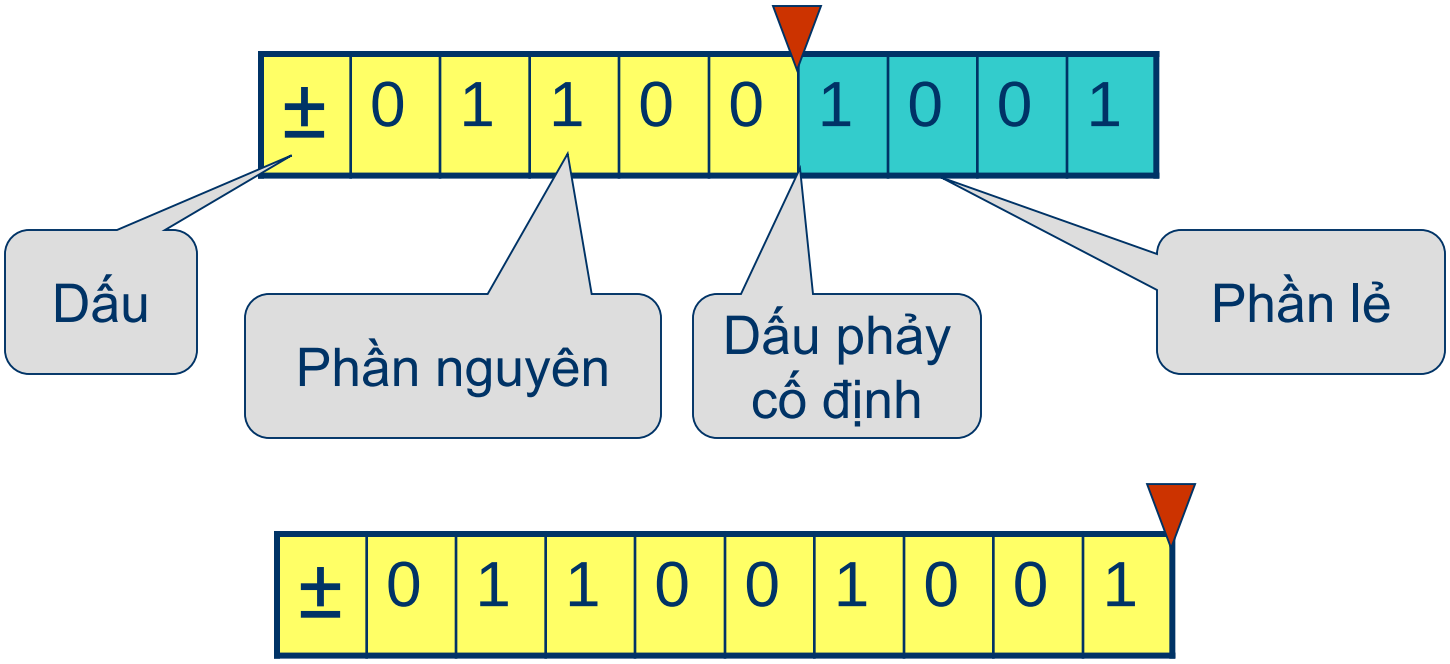
NỘI DUNG

- Phân loại dữ liệu
- Biểu diễn số (dấu phảy tĩnh và dấu phảy động)
- Biểu diễn phi số (chữ, logic, hình ảnh, âm thanh)
- Biểu diễn tri thức (sự kiện và luật)
- Truyền dữ liệu giữa các máy tính

PHÂN LOẠI DỮ LIỆU

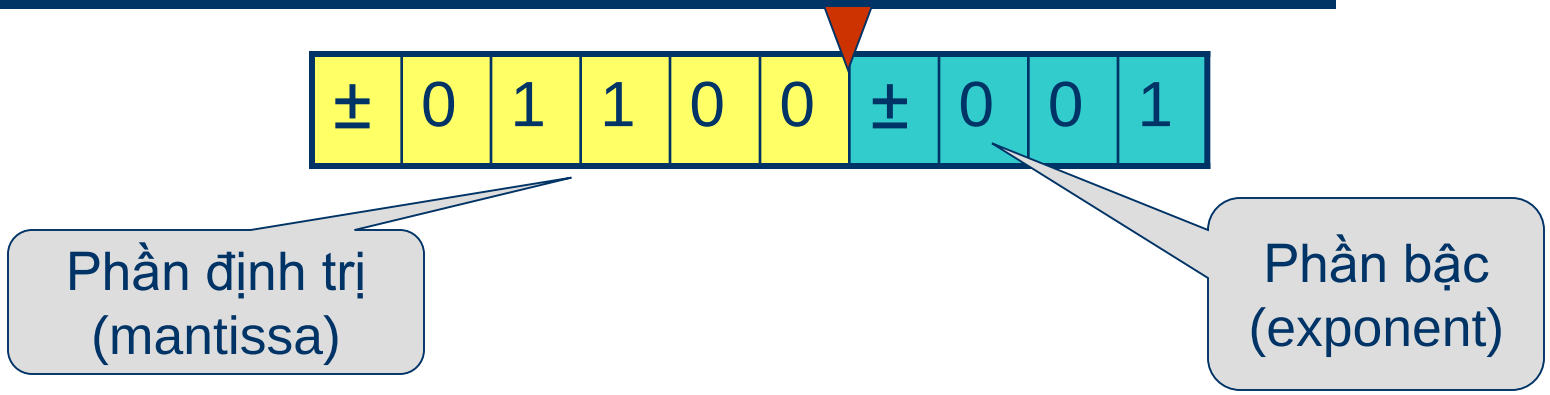


SỐ DẤU PHẢY TĨNH (fixed point number)



Có một vị trí cố định ngăn cách giữa phần nguyên và phần lẻ -> dấu phảy tĩnh

SỐ DẤU PHẢY ĐỘNG (floating point number)



Số được biểu diễn dưới dạng nửa logarit $x = \pm m_x . 10^{\pm Px}$

Ví dụ $3.14 = 0.314 \times 10^2$ hoặc $-0.0012 = -0.12 \times 10^{-2}$

Vị trí dấu phảy trong biểu diễn bình thường do phần bậc định ra trên phần định trị nên gọi là dấu phảy động. Số dấu phảy động thường được dùng với tính toán gần đúng. Trong một số ngôn ngữ lập trình nó được khai báo với kiểu là real hay double. Người ta đo tốc độ của các máy tính khoa học kỹ thuật theo Flops (floating point operations per second) hoặc Gflops

SO SÁNH KHOẢNG BIỂU DIỄN

- Về khả năng biểu diễn số. Với cùng một số ngăn nhớ, số mã khác nhau có thể biểu diễn được hoàn toàn như nhau nhưng khoảng số biểu diễn được khác nhau rất xa. Có thể xem xét qua số dương lớn nhất và số dương nhỏ nhất có thể biểu diễn đượcc. Dưới đây tất cả viết trong hệ đếm cơ số 2.
- Xét ví dụ với 4 ngăn định trị, 2 ngăn cho bậc và 2 ngăn cho dấu
- Khoảng biểu diễn được ở chế độ dấu phảy động là $0.1x10^{-11}$ đến $0.1111x10^{11}$ (tổng quát trong trường hợp m ngăn cho định trị và n ngăn cho bậc không kể dấu sẽ là từ $_{10}(10^{-111..1} - 1)$ đến $_{10}$ 111..1
- Với số dấu phảy tĩnh khoảng biểu diễn chỉ được từ 1 đến $10^{m+n} - 1$.
- Về khoảng biểu diễn, chế độ dấu phảy động tốt hơn rất nhiều

+	1	1	1	1	+	1	1
+	1	0	0	0	-	1	1

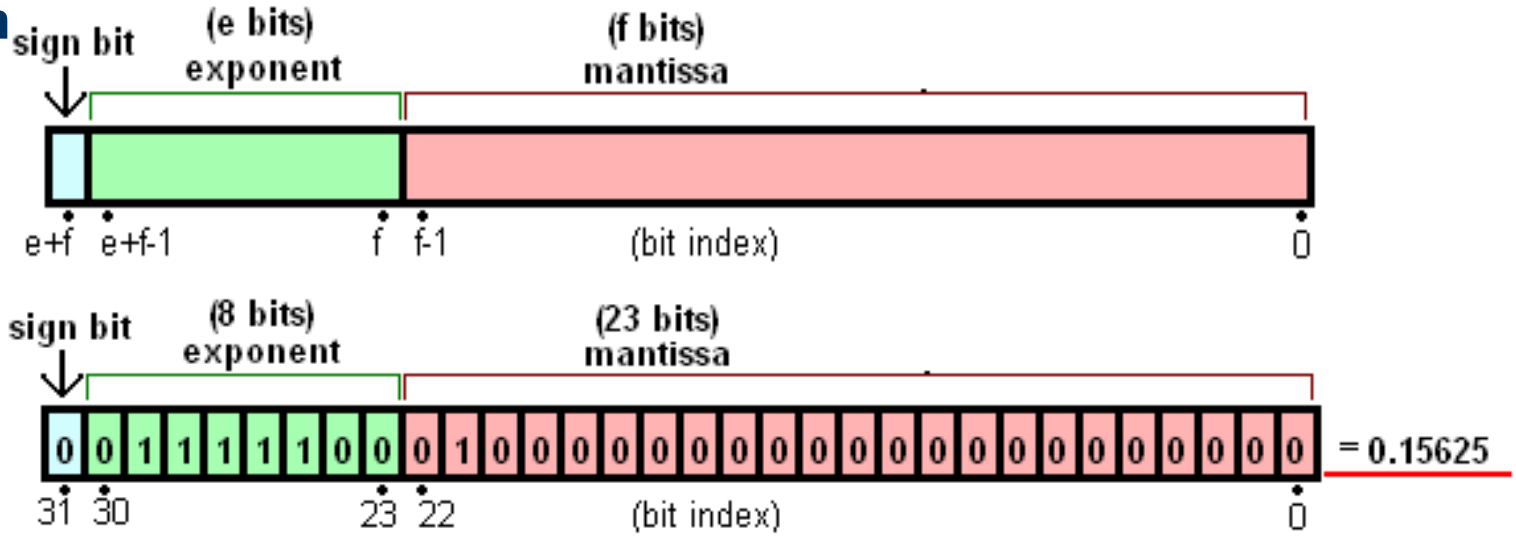
+	1	1	1	1	1	1	1
+	0	0	0	0	0	0	1

SO SÁNH ĐỘ CHÍNH XÁC

- Do số ngăn của một ô nhớ bị hạn chế nên biểu diễn sẽ mắc sai số làm tròn. Có hai loại sai số: với số x được xấp xỉ bằng x' thì |x-x'| gọi là sai số tuyệt đối, còn |(x-x')/x| được gọi là sai số tương đối
- Với dấu phảy tĩnh trong chế độ số nguyên, sai số tuyệt đối luôn là 1, còn sai số tương đối là có thể lớn tuỳ theo số nhỏ hay lơn.
- Với số dấu phảy động với m ngăn cho phần định trị và nngăn cho phần bậc sai số tương đối do làm tròn luôn luôn không quá $10^{-111..1\ (n\ so),}$, cò n sai số tương đối bị khuếch đại bới phần bậc có thể lên tới $_{10}10^{n}-1$
- Sai số tuyệt đối có thể lớn nhưng sai số tương đối thì rất tốt. Chính vì vậy trong các bài toán tính toán gần đúng, biểu diễn dấu phảy động rất phù hợp

SỐ DẤU PHẢY ĐỘNG CHUẨN IEEE 754

Chuẩn IEEE 754 là một chuẩn được sử dụng rộng rãi nhất hiện nay cho tính toán dấu phảy động. Chuẩn này định nghĩa định dạng và cách thực hiện các phép tính trên các số phảy động trong đó có cả số 0 với dấu âm, các số không chuẩn hoá, các giá trị đặc biệt như vô hạn và giá trị không phải số (NaNs). Chuẩn cũng xác định 4 kiểu làm tròn số và 5 ngoại lệ. Bit cao nhất là dấu của số, sau đó là phần bậc, cuối cùng là phần



SỐ DẤU PHẢY ĐỘNG CHUẨN IEEE 754

Kiểu	Phần bậc Exponent	Phần định trị Mantissa
Số 0 (Zeroes)	0	0
Các số không chuẩn hoá (Denormalized numbers)	0	≠ 0
Các số chuẩn hoá (Normalized numbers)	1 to $2^e - 2$ (1 -1111...110)	bất kỳ
Vô hạn (Infinities)	$2^e - 1$ (1111...111)	0
Không phải số (NaNs)	$2^e - 1$ (1111...111)	≠ 0

BIỂU DIỄN CHỮ VÀ VĂN BẢN

- Với k bít, có thể biểu diễn 2^k mã khác nhau. Ta dùng thuật ngữ ký tự (character) để chỉ một biểu diễn cho một ký hiệu phân biệt với chữ (letter) thông thường mà letter cũng chi là một loại ký tự giống như chữ số, các dấu chính tả và các dấu đặc biệt khác
- Bộ mã Mã EBCDIC (Extended Binary Coded Decimal Interchange Code) trong những năm 70 dùng 6 bit có thể mã được 64 ký tự
- Bộ mã ASCII (American Standard Codes for Information Interchange) dùng 7 bít cho phép biểu diễn 128 kì tự (32 mã đầu tiên dùng cho các mã điều khiển và truyền thông, tiếp theo là các dấu chính tả, các chữ số, các chữ thường, các chữ in và các dấu đặc biệt).
- Bộ mã ASCII mở rộng dùng 1 byte cho một ký tự nên có khả năng biểu diễn 256 ký tự. 128 chỗ vùng tiếp theo có thể cho chữ của các nước châu Âu, chữ Hy lạp hoặc bất cứ một bộ chữ nào như tiếng Việt hay ngôn ngữ Slavơ, nhưng không thể đủ cho tiếng Trung Quốc hay Nhật Bản

BẢNG CHỮ ASCII (128 ký tự đầu)

	000	001	010	011	100	101	110	111
00000	0 NUL	1 SOH	2 STX	3 EXT	4 EOT	5	6	7 BELL
00001	8 BS	9 HT	10 LF	11 VT	12 FF	13 CR	14	15
00010	16	17 DC1	18 DC2	19 DC3	20 DC4	21	22	23
00011	24	25	26	27	28	29	30	31
00100	32	33 !	34 "	35 #	36 $	37 %	38 &	39 '
00101	40 (	41)	42 *	43 +	44,	45 -	46.	47 /
00110	48 0	49 1	50 2	51 3	52 4	53 5	54 6	55 7
00111	56 8	57 9	58 :	59 ;	60 <	61 =	62 >	63 ?
01000	64 @	65 A	66 B	67 C	68 D	69 E	70 F	71 G
01001	72 H	73 I	74 J	75 K	76 L	77 M	78 N	79 O
01010	80 P	81 Q	82 R	83 S	84 T	85 U	86 V	87 W
01011	88 X	89 Y	90 Z	91 [	92 \	93]	94 ^	95 _
01100	96 `	97 a	98 b	99 c	100 d	101 e	102 f	103 g
01101	104 h	105 i	106 j	107 k	108 l	109 m	110 n	111 o
01110	112 p	113 q	114 r	115 s	116 t	117 u	118 v	119 w
01111	120 x	121 y	122 z	123 {	124 |	125 }	126 ~	127

BIỂU DIỄN CHỮ VỚI UNICODE

- Đối với quốc gia có bộ chữ lớn (như Trung quốc, Nhật bản) bộ mã 8 bít không đủ chỗ cho tất cả các chữ. Nhật Bản đã đưa ra một dự án lập bộ chữ cho toàn cầu gọi là UNICODE. Bộ chữ được chia trang cho các quốc gia. Mặt chữ nào của một nước nào đã có sẽ được dùng lại tại các phần mềm khác.
- Sau này các tổ chức chuẩn chấp nhận UNICODE dưới chuẩn ISO 10646
- Mỗi quốc gia có thể nhận các trang mã (code page), mỗi ký tự được thể hiện qua mã của trang mã và số thứ tự (code point) của ký tự đó trong trang mã - một số 2 byte). Trong bảng mã UNICODE, chữ "ơ" có điểm mã là 01A1 (so sánh với bảng mã CP1258 của Microsoft, bảng mã 8 bít, chữ "ơ" có điểm mã F5)

MÃ TIẾNG VIỆT

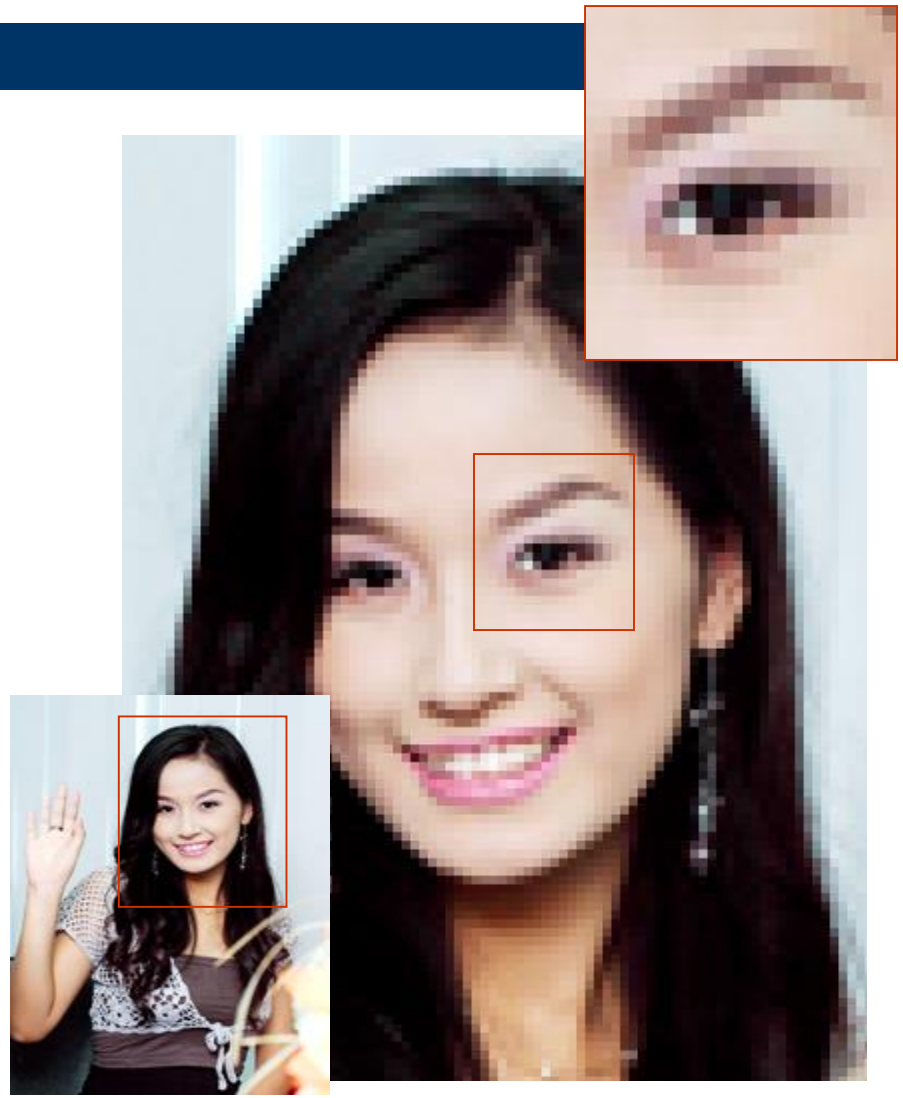
- Từng tồn tại tới 40 mã tiếng Việt 8 bít dẫn đến tình trạng loạn mã, không chia sẻ được dữ liệu. Có 141 ký tự đặc thù Việt Nam không có chỗ (vùng mở rộng chỉ có 128 chỗ)
- Năm 1993 xây dựng bộ mã TCVN 5712. Thực chất vẫn là một giải pháp chắp vá với 3 bộ mã khác nhau. Bộ mã 1, chiếm thêm một số chỗ trong vùng mã điều khiển – nguy hiểm cho truyền thông). Bộ mã 2 là bộ mã tổ hợp, dùng một chuỗi ký tự để thể hiện một mã cho các chữ thuần Việt. Bộ mã 3 hy sinh một số ký tự hoa có dấu ví dụ Ấ. Cả 3 giải pháp đều không giải quyết được triệt để
- Từ 2001, Bộ KHCN đã ban hành tiêu chuẩn TCVN 6909/2001 về việc sử dụng mã UNICODE có hiệu lực từ 1/1/2003. Các cơ quan nhà nước buộc phải dùng bộ mã này trong trao đổi dữ liệu.
- TCVN 6909 vẫn chấp nhận cả hai kiểu: mã dựng sẵn (pre-compound) với mỗi ký tự thể hiện bởi một mã 2 byte và kiểu tổ hợp cho phép dùng một chuỗi ký tự 8 bít để thể hiện một ký tự

BIỂU DIỄN CÁC GIÁ TRỊ LOGIC

- Trong đời sống, có các loại thông tin mà giá trị của nó có hai trạng thái đối lập có thể là "có/không", "đúng/sai". Dữ liệu loại này gọi là dữ liệu logic
- Các dữ liệu logic có thể tương tác với nhau thông qua các phép toán logic mệnh đề như "Và", "hoặc", "không"
- Về nguyên tắc có thể mã hoá các đại lượng logic bằng 1 bít (1 là đúng hoặc có, 0 là sai hoặc không có). Tuy nhiên người ta ít khi làm như thế vì đơn vị nhớ cơ sở là byte. Trong cài đặt cụ thể người ta có thể dung các kí tự như T (true) và F (false) để biểu diễn hai giá trị "đúng" và "sai"

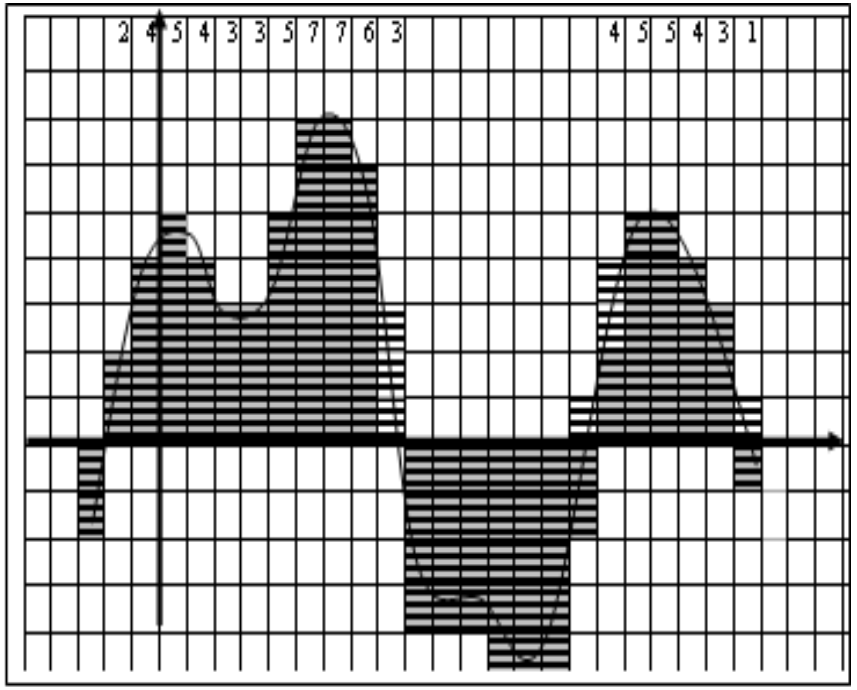
BIỂU DIỄN DỮ LIỆU HÌNH ẢNH

- Ảnh là một tập hợp các điểm ảnh (pixel), có màu sắc tạo từ 3 màu cơ bản (red, green, blue) với cường độ khác nhau.
- Ví dụ ảnh màu 24 bít, dùng mỗi byte để mã một màu với các mức từ 0 đến 255. Như vậy sẽ có 2^{24} (khoảng 19 triệu) sắc độ màu khác nhau.
- Có các chuẩn ảnh khác chủ yếu khác nhau về việc cấu trúc thông tin ảnh phù hợp với phương pháp nén ảnh và thể hiện ảnh. Một số chuẩn ảnh thông dụng là bitmap, jpeg, gif, tiff
- Ảnh trực tiếp thể hiện bằng điểm ảnh gọi là ảnh bitmap hay ảnh raster. Còn một kiểu ảnh khác là ảnh vector



BIỂU DIỄN ÂM THANH

- Cách đơn giản nhất là mã hoá bằng cách xấp xỉ dao động sóng âm bằng một chuỗi các byte thể hiện biên độ dao dộng tương ứng theo từng khoảng thời gian bằng nhau.
- Các đơn vị thời gian này cần phải đủ nhỏ để không làm nghèo âm thanh. Đơn vị thời gian này gọi là chu kỳ lấy mẫu.
- Khi phát lại, người ta dùng một mạch điện để tái tạo lại âm thanh từ các biên độ dao động của từng chu kỳ lấy mẫu



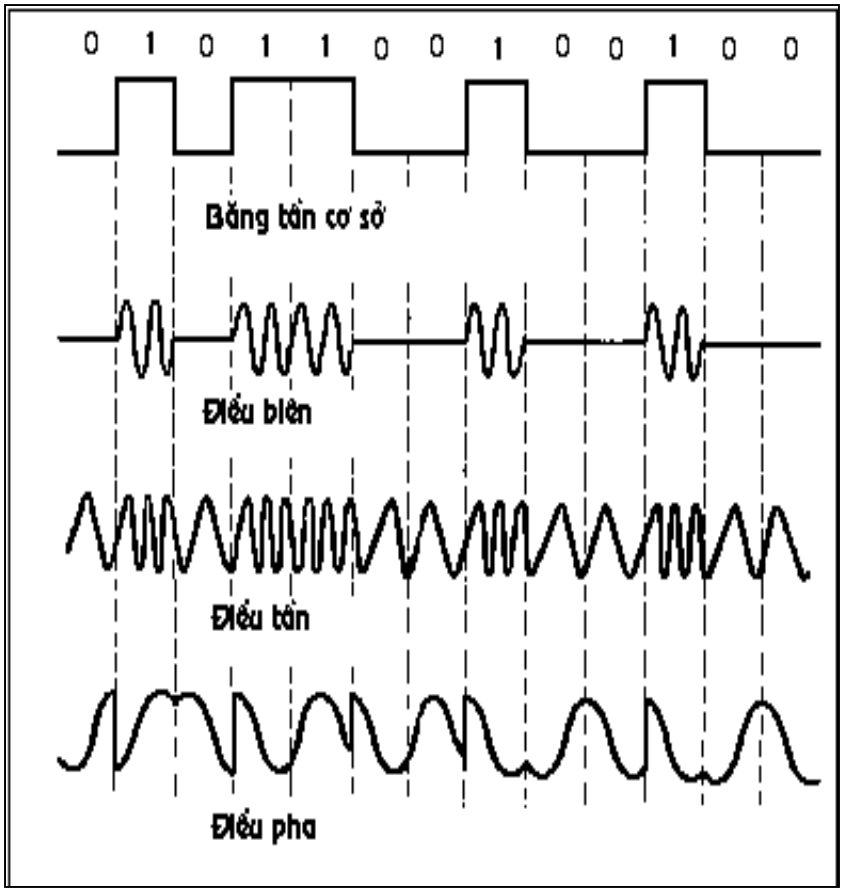
- Có một số chuẩn định dạng âm thanh như wav,một số chuẩn khác cho phép nén âm thanh cùng với các hình ảnh động

TRI THỨC = SỰ KIỆN + LUẬT

- Tri thức (knowledge) không chỉ thể hiện bằng các sự kiện (fact) mà ta có thể biểu diễn như các dữ liệu thông thường mà nó còn thể hiện cách suy luận cho bằng các luật (rule)
- VD quan hệ "Làbố" có thể cho bằng 2 chuỗi ký tự hiểu theo nghĩa tên bố và tên con. Làbố (Hùng, Cường) nghĩa là Hùng là bô của Cường.
- Quy tắc "Nếu (A là bố B) và (B là bố C) thì A là ông nội C" cho phép từ một số quan hệ này suy ra một số quan hệ khác
- Chẳng hạn từ Làbố (Bé, Cường) và Làbố (Cường, Đại) thì theo quy tắc trên sẽ rút ra Bé là ông nội của Đại

TRUYỀN DỮ LIỆU

- Dữ liệu được lưu trữ dưới dạng trạng thái nhị phân nhưng truyền đi bằng sóng điện từ
- Cần điều chế (modulation) tín hiệu trên các sóng mang trong các kênh truyền vật lý.
- Có thể điều chế theo tần số, biên độ và pha.
- Đôi khi người ta điều chế bằng cả điều pha và điều biên, cho phép truyền thông với tốc độ cao hơn cả tần số của sóng mang như trong modem 9.6 kb/s với mã hoá kiểu chòm sao (constellation)



TỔNG KẾT

- Dữ liệu là cách thể hiện thông tin với mục đích lưu trữ, xử lý và truyền tin
- Có nhiều loại dữ liệu như số, văn bản, logic, đa phương tiện và tri thức. Mỗi loại có những đặc thù riêng đi kèm với các mã hoá
- Để truyền dữ liệu, người ta phải điều chế. Đối với tín hiệu điện, thường phải gửi theo sóng mang với cơ chế mã hoá theo kiểu điều tần, điều pha, điều biên hay hỗn hợp.

CÂU HỎI VÀ BÀI TẬP

1. Người ta nói dữ liệu là hình thức biểu diễn của thông tin. Cũng có người nói dữ liệu là thông tin được xử lý bằng máy tính. Hai cách nói này có mâu thuẫn không.
2. Thế nào là dữ liệu số, thế nào là dữ liệu phi số
3. Tại sao cần các chế độ biểu diễn số khác nhau như chế độ dấu phảy động và chế độ dấu phảy tĩnh
4. Nêu các phương pháp điều chế tín hiệu để truyền dữ liệu

CẢM ƠN ĐÃ THEO DÕI

HẾT BÀI 6. HỎI VÀ ĐÁP