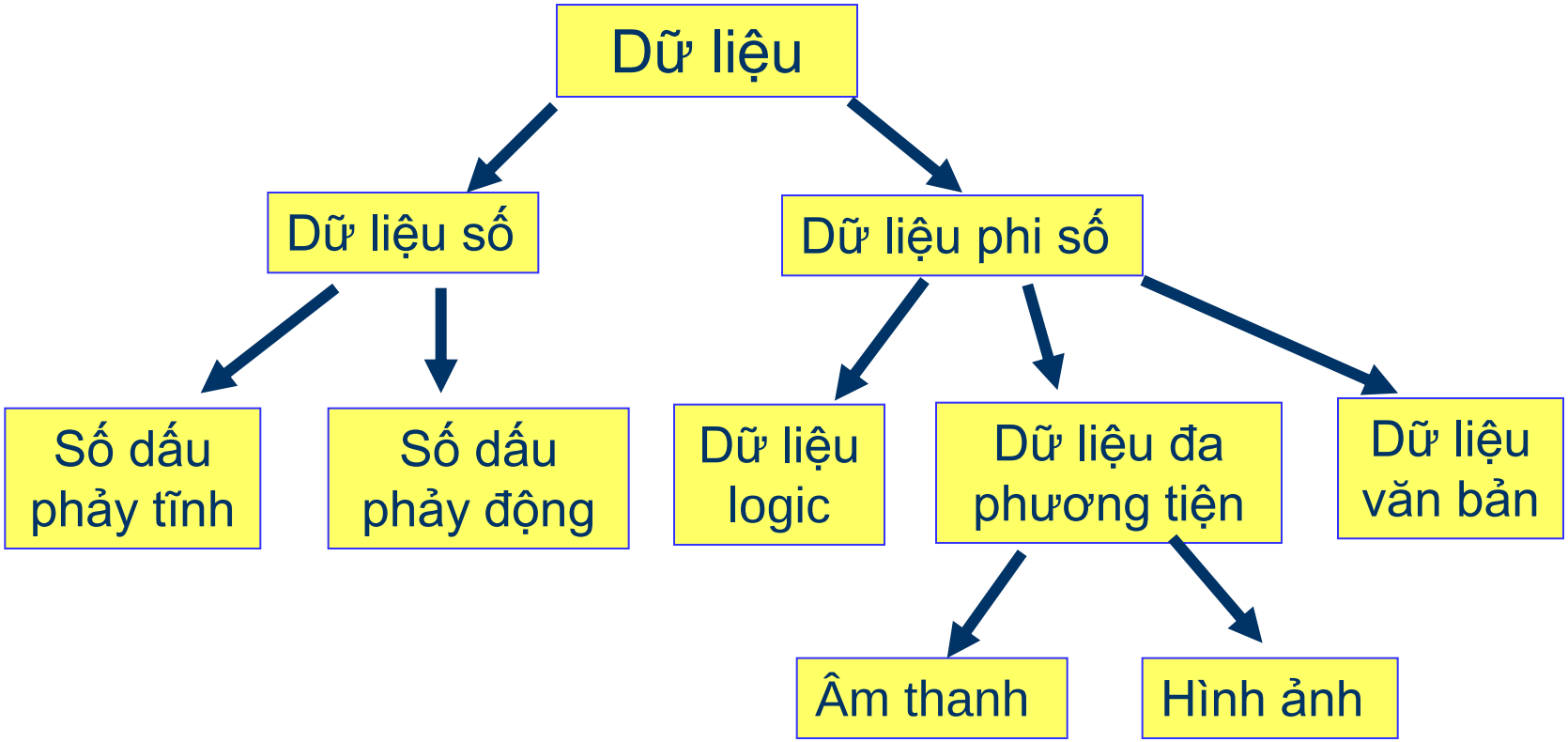
BIỂU DIỄN VÀ TRUYỀN DỮ LIỆU

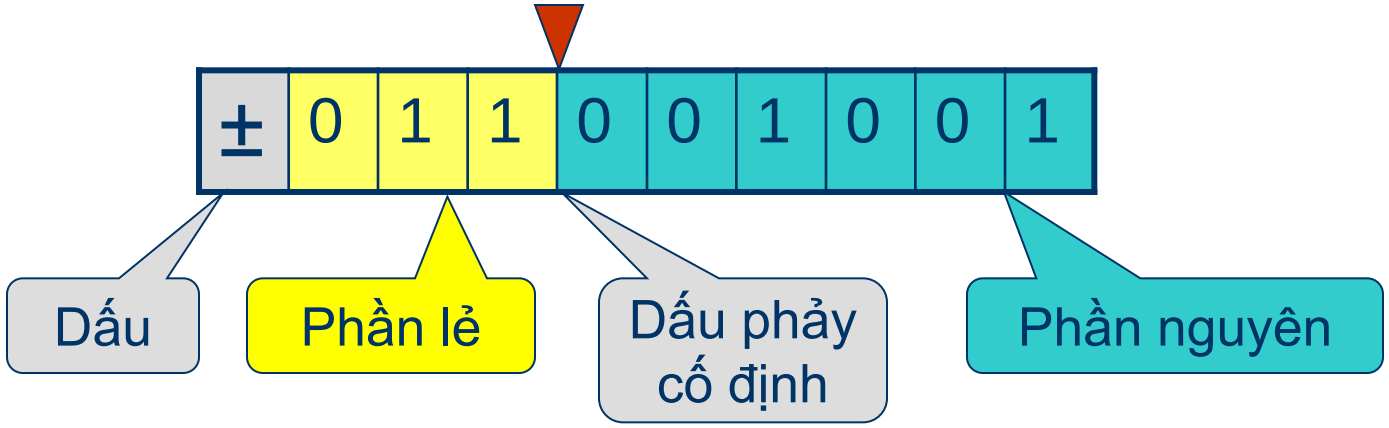
NỘI DUNG

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

PHÂN LOẠI DỮ LIỆU



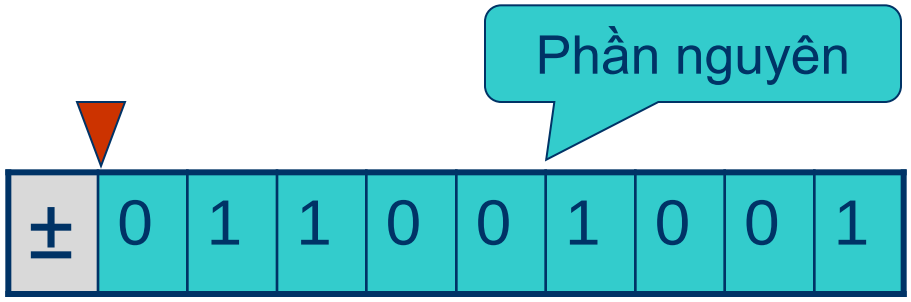
SỐ DẤU PHẢY TĨNH (fixed point number)



Ban đầu, lịch sử

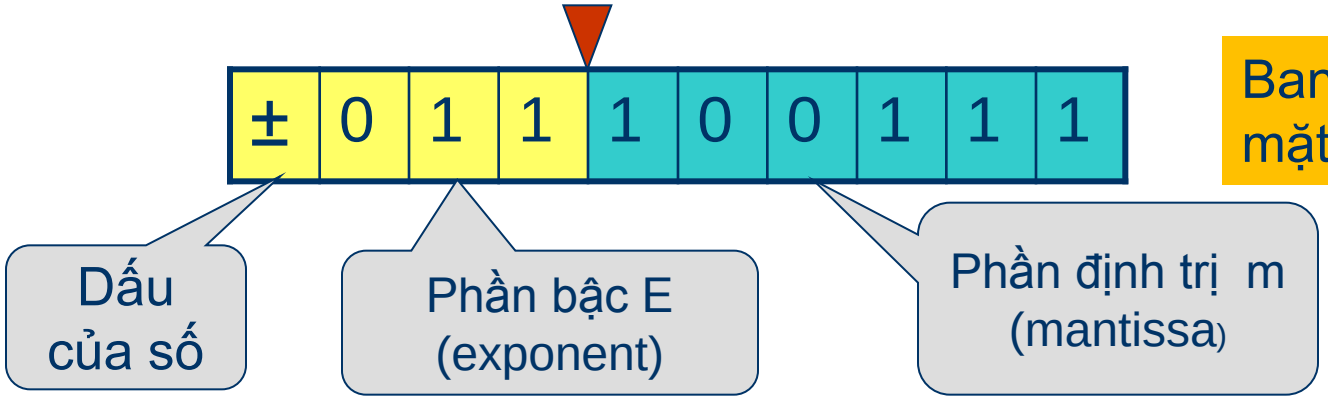
Dấu phảy tĩnh: Có một vị trí cố định ngăn cách giữa phần nguyên và phần lẻ.

Vấn đề: không quyết định được để bao nhiêu ngăn cho phần nguyên và phần lẻ là phù hợp



Thực tế: sau này chỉ dùng với số nguyên, không có phần lẻ (khai báo: byte, word, integer, long integer)

SỐ DẤU PHẢY ĐỘNG (floating point number)



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

Ví dụ $3.14 = 0.0314 \times 10^{2}$ 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.

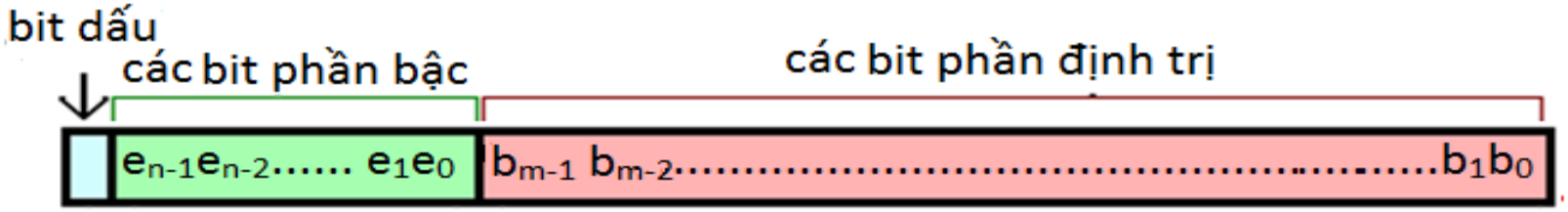
SỐ DẤU PHẢY ĐỘNG

Vấn đề

- Cách biểu diễn số không duy nhất do phần bậc có thể tùy tiện, dẫn đến phần định trị cũng biến đổi theo. Cần chuẩn hóa cách biểu diễn.

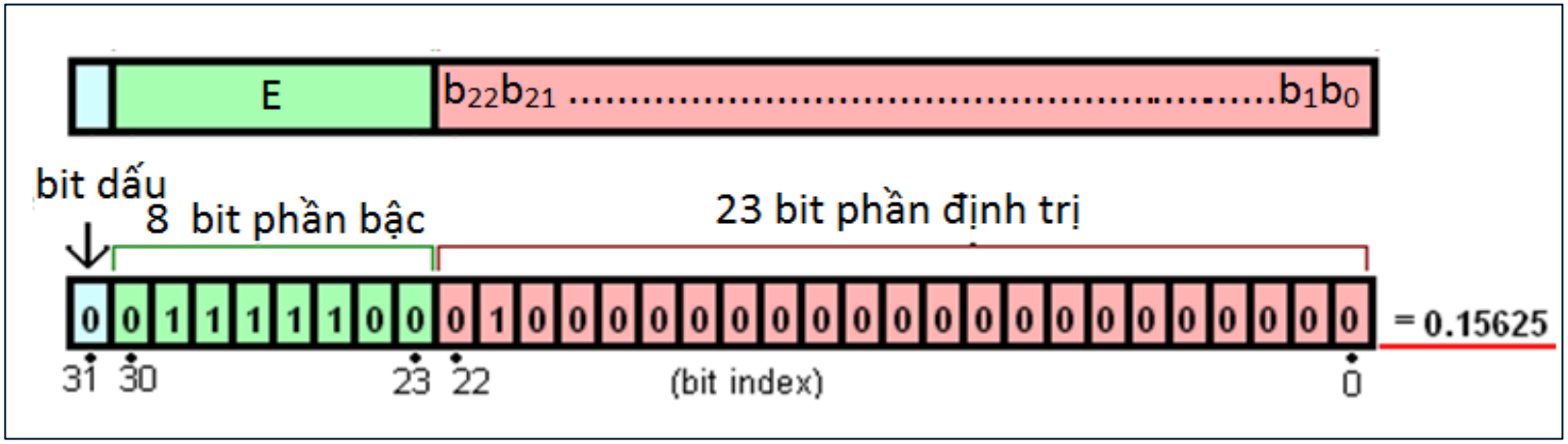
- Không xử lý được các giá trị vô hạn $+\infty$ và $-\infty$. Nên có mã cho các giá trị vô hạn.
- Không thể hiện được các giá trị không xác định
- Đôi khi cần biểu diễn một đối tượng không phải là số

MỘT GIẢI PHÁP: CHUẨN IEEE 754



Kiểu	Phần định trị bbb	Phần bậc eee
Số 0 (Zeroes)	0000…000	0
Các số phi chuẩn hoá (Denormalized numbers)	0000…000	≠ 0
Các số được chuẩn hoá (Normalized numbers)	từ 1 đến 1111...110 Mã $b_{m-1}\ b_{m-2}\ldots b_1b_0$ được hiểu có giá trị là: $1.b_{m-1}\ b_{m-2}\ldots b_1b_0$	bất kỳ
Giá trị vô hạn +∞ và - ∞	(1111...111)	0
Không phải số (NaNs)	(1111...111)	≠ 0

SỐ DẤU PHẢY ĐỘNG CHUẨN IEEE 754



Cách mã giá trị trên vùng 32 bít của chuẩn IEEE 754 đối với các số chuẩn hóa được tính như sau:

$$\text{Giá trị} = (-1)^{\text{dấu}} \times (1+\text{định trị}) \times 2^{(\text{bậc} - 127)}$$

Trong ví dụ trên, giá trị của biểu diễn số là

$$(-1)^0 \times 1.25) \times 2^{(124-127)} = 1.25 \times 2^{(-3)} = 0.15625$$

SAI SỐ LÀM TRÒN

- Khi biểu diễn một số trong chế độ dấu phảy tĩnh, phần làm tròn là phần lẻ. Còn trong chế độ dấu phảy động, phần làm tròn bị mất là phần bị bỏ qua trong phần định trị vì không đủ ngăn nhớ.
- 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 biểu diễn dấu phảy tĩnh, sai số làm tròn tuyệt đối không quá 1,
- Sai số tương đối trong chế độ dấu phảy tính là có thể lớn tuỳ theo số nhỏ hay lớn. Lớn nhất là xấp xỉ 1 khi giá trị của chính số đó xấp xỉ 2 (trường hợp số cần biểu diễn là 1.111111…1)

Đối với số dấu phảy động dùng n ngăn cho phần định trị:

- Sai số (tuyệt đối) làm tròn lớn nhất của số $x = \pm\ m_x . 2^{Ex}$ với phần định trị có n ngăn là $2^{-n} \times 2^{Ex}$
- Sai số tương đối làm tròn của số dấu phảy động luôn luôn cỡ 2^{-n} (2^{-23} đối với số IEEE 754 32 bit)

Dấu phảy động có sai số tương đối bé rất tốt cho tính toán gần đúng

KHOẢNG BIỂU DIỄN SỐ

Xét số 32 bít,

- Trong chế độ dấu phảy tĩnh
  - Số dương nhỏ nhất biểu diễn được là 1,
  - Số lớn nhất là 2^{31} khoảng 10^{9}

- Trong chế độ dấu phảy động, số IEEE 754, 32 bit
  - Số dương nhỏ nhất:

    $1.000\ldots00 \times 10^{00000000} = 2^{0-127} = 2^{-127} = 10^{-38}$
  - Số dương lớn nhất là:

    $1.1111\ldots10 \times 10^{11111111}$ xấp xỉ $2 \times 2^{255-127} = 2^{128}$ khoảng 10^{38}

Số dấu phảy động, xét về khoảng biểu diễn số tốt hơn số dấu phảy tĩnh rất nhiều

BIỂU DIỄN CHỮ VÀ VĂN BẢN

- Với k bít, có thể biểu diễn 2^k mã khác nhau. Mỗi mã dùng cho một ký tự (character – một kí hiệu của một chữ) (phân biệt với chữ - letter chỉ là một loại kí tự)

- Bộ mã EBCDIC (Extended Binary Coded Decimal Interchange Code) trong những năm 70 dùng 6 bit có thể mã được 64 ký tự: chỉ dùng cho chữ in, chữ số, các loại dấu

- Bộ mã ASCII (American Standard Codes for Information Interchange) dùng 7 bít cho phép biểu diễn 128 kí 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, chữ Slavo nhưng không thể đủ cho tiếng Trung Quốc (hơn 60.000 kí tự), thậm chí không đủ chỗ cho tiếng Việt (cần thêm 141 kí tự)

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

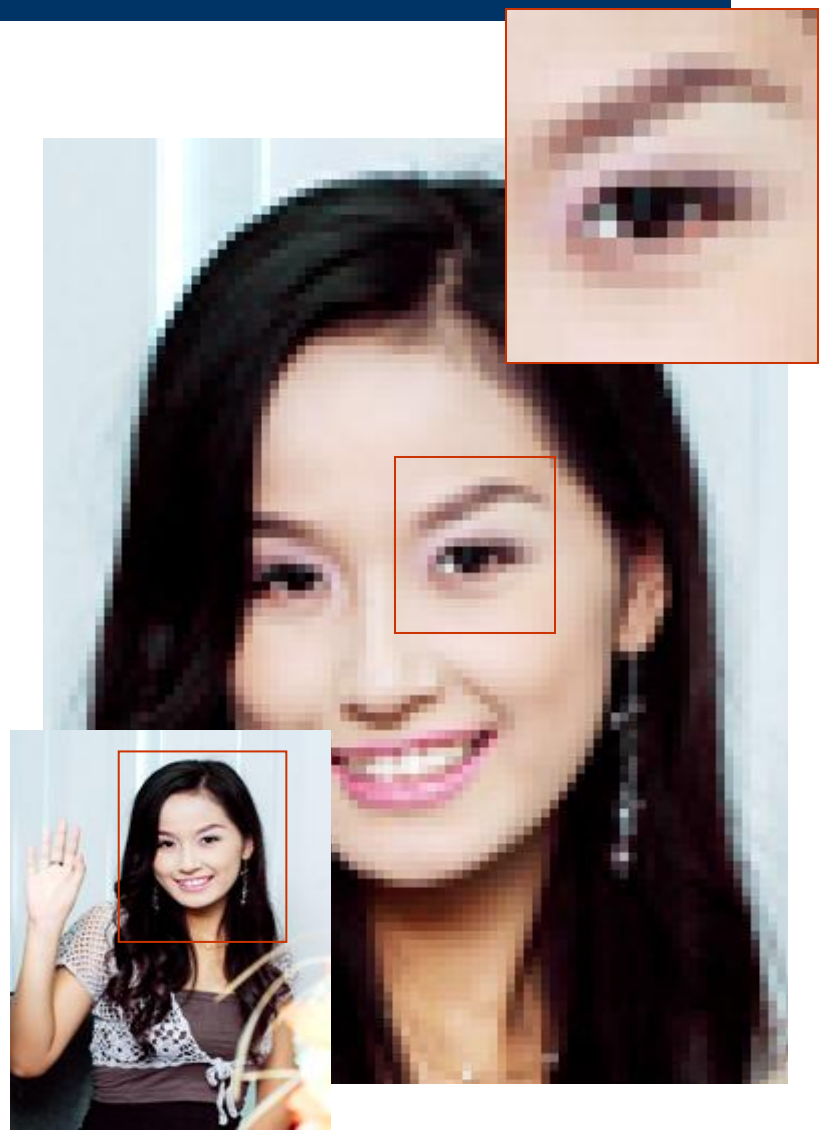
- Từng tồn tại tới 40 mã tiếng Việt 8 bít dẫn đến tình trạng loạn mã, không chia sẻ được dữ liệu. Có 141 ký tự đặc thù Việt Nam không có chỗ (vùng mở rộng chỉ có 128 chỗ)
- Bộ mã TCVN 5712/1993. Thực chất vẫn là một giải pháp chắp vá với 3 bộ mã khác nhau, không giải quyết được triệt để
  - Bộ mã 1, 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 chuối 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 Ã.
- Bộ mã tiêu chuẩn TCVN 6909/2001 là 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. Vẫn chấp nhận cả hai kiểu:
  - Mã dựng sẵn (precomposed) coi chữ dấu mũ dấu thanh là một ký tự duy nhất
  - Mã tổ hợp (Composite) thể hiện một ký tự có dấu mũ, dấu thanh qua một chuỗi ký tự khác.

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”

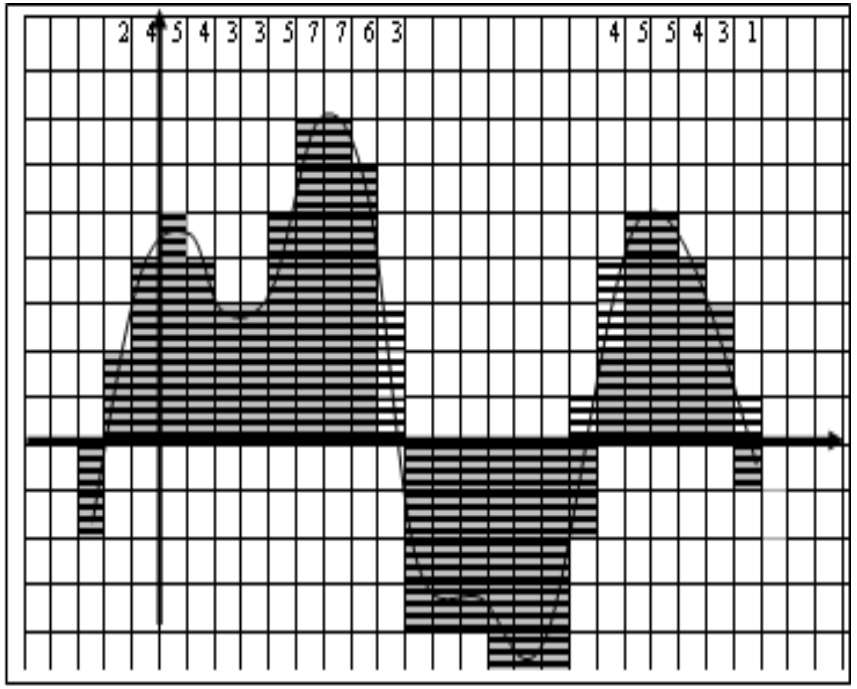
BIỂU DIỄN DỮ LIỆU HÌNH ẢNH

- 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} 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, lưu trữ các hình trên cơ sở tọa độ các điểm cơ bản. Ảnh vector tốn ít không gian lưu trữ, dễ biến đổi ảnh.



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 như mp3

TRUYỀN DỮ LIỆU

- Dữ liệu được lưu trữ dưới dạng nhị phân nhưng truyền đi bằng sóng điện từ
- Điều chế (modulation) chuyển từ tín hiệu số (digital) sang tín hiệu tương tự (analog)
- Giải điều chế (demodulation) chuyển từ tín hiệu tương tự sang tín hiệu số
- Modem: thiết bị giao tiếp hai chiều chuyển đổi tín hiệu số và tương tự
- Điều chế
  - Điều biên: biểu diễn bit 0/1 bằng các sóng có biên độ khác nhau
  - Điều tần: biểu diễn bit 0/1 bằng các sóng có tần số khác nhau
  - Điều pha: biểu diễn bit 0/1 bằng các sóng có pha khác nhau