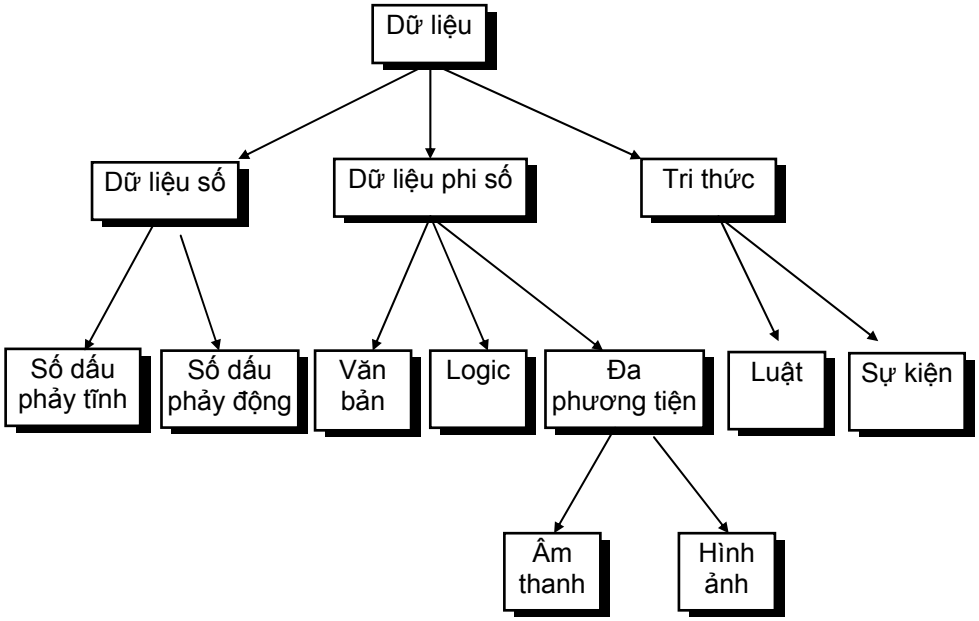
MODULE 6. BIỂU DIỄN THÔNG TIN TRONG MÁY TÍNH

6.1. Dữ liệu

Để đạt được hiệu quả cao khi xử lý, lưu trữ và truyền thông tin điều cần thiết là phải tìm cách tổ chức và biểu diễn (thể hiện) thông tin trong MTĐT một cách hợp lý. Như đã biết, dữ liệu là hình thức biểu diễn thông tin. Vậy đối với máy tính dữ liệu chính là các thông tin đã được mã hoá dưới dạng nhị phân. Dữ liệu - thông tin được máy tính xử lý có thể có các dạng khác nhau.

Máy tính có thể tính toán trên các số, có thể xử lý thông tin chữ hay thông tin logic, có thể xử lý những thông tin đa phương tiện (multimedia) như âm thanh và hình ảnh. Máy tính còn có thể xử lý tri thức (knowledge).



Hình 6.1. Phân loại các dạng dữ liệu cơ bản.

Thông tin về một đối tượng có thể rất phức tạp và có thể được thể hiện bằng nhiều dữ liệu có kiểu khác nhau. Ví dụ thông tin về một cán bộ có thể có tên, nơi sinh là văn bản; ngày sinh, lương là số; ảnh chân dung là ảnh...

Để lưu trữ trong MTĐT cả dữ liệu số, phi số và tri thức đều được mã hóa bằng các mã nhị phân. Theo nghĩa đó mọi dữ liệu dù là bản chất có khác nhau nhưng đều được số hoá. Sự phân biệt theo sơ đồ trên nặng về ý nghĩa sử dụng hơn là cách biểu diễn. Dưới đây ta sẽ trình bày chi tiết hơn các lớp dữ liệu. Trong trường hợp biểu diễn thông tin không quá phức tạp ta sẽ trình bày một chút về cách biểu diễn.

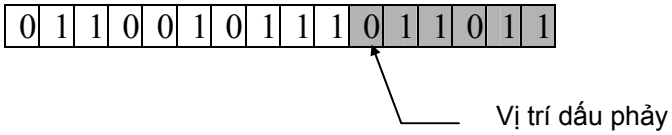
6.2. Dữ liệu kiểu số

Người ta thường dùng hai cách biểu diễn số là số dấu phảy tĩnh và số dấu phảy động.

6.2.1. Biểu diễn số dấu phảy tĩnh

(fixed point number)

Với kiểu biểu diễn số dấu phảy tĩnh, người ta chọn một độ rộng n bít nào đó cho một số. Trong n bit này, bít đầu tiên dùng để mã dấu của số theo cách bít 0 dùng để mã dấu dương, bít 1 dùng để mã dấu âm. Trong n-1 bít còn lại, lấy một số bít cho phần nguyên và phần còn lại cho phần lẻ. Ví dụ trong dãy 16 bít sau nếu ta dùng 7 bít cho phần nguyên và 8 bít cho phần lẻ và một bít cho dấu thì biểu diễn sau thể hiện số 1100101,11011011



Hình 6.2. Biểu diễn dấu phảy tĩnh

Do với mỗi kiểu biểu diễn đã chọn, vị trí dấu phảy mang tính quy ước nằm ở một vị trí cố định nên kiểu biểu diễn này gọi là kiểu dấu phảy tĩnh.

Trên thực tế đa số các môi trường xử lý quy ước dấu phảy đứng sau ô cuối cùng có nghĩa là chỉ áp dụng chế độ dấu phảy tĩnh cho số nguyên. Độ dài của biểu diễn tuỳ thuộc vào nhu cầu. Các số nguyên thường dùng chủ yếu có các loại độ dài 8 bit, 16 bít và 32 bit.

Mã số nguyên trình bày trên đây được gọi là mã thuận. Thực ra để tiện cho việc thực hiện các phép tính đại số, người ta còn sử dụng nhiều loại mã số nguyên khác như mã ngược, mã bù...mà ta sẽ không trình bày ở đây.

6.2.2. Biểu diễn số dấu phảy động

(floating point number)

Biểu diễn dấu phảy tĩnh không đáp ứng được một số nhu cầu, đặc biệt trong tính toán gần đúng. Đối với các bài toán tính gần đúng người ta có thể chấp nhận những sai số là lớn về tuyệt đối nhưng tỉ số của sai số trên giá trị thực của số là nhỏ (sai số tương đối). Mặt khác cách biểu diễn số trong dấu phảy tĩnh không đủ mềm dẻo để thể hiện các số quá lớn hoặc quá bé. Đã từ lâu, khi có nhu cầu tính toán gần đúng trên máy tính người ta thường dùng một loại biểu diễn số khác là biểu dấu phảy động. Trong dạng này số phải được phân tích trong dạng mũ hay còn là dạng nửa logarit như sau:

$x = \pm m_x.10^{\pm Px}$

trong đó m_x gọi là phần định trị còn $\pm P_x$ gọi là phần bậc.

Ví dụ

$3,14 = 0,314 \times 10^{1}$

$- 0.0012 = - 0.12 \times 10^{-2}$

Như vậy phân tích của một số ra dạng mũ là không duy nhất nhưng nếu kèm thêm điều kiện phần định trị phải nằm giữa 1 và 10^{-1} thì phân tích luôn duy nhất. Phân tích như thế gọi là dạng chuẩn. Với dạng chuẩn, phần định trị không có phần nguyên nên chỉ phải biểu

diễn phần lẻ của nó. Hơn nữa trong dạng chuẩn nếu một số khác 0 thì chữ số đầu tiên của phần định trị phải khác 0. Trong hệ đếm cơ số 2 nó phải là 1.

Để biểu diễn một số trong một vùng nhớ n bít người ta sẽ dành một phần biểu diễn phần định trị và một vùng biểu diễn phần bậc.

±	Phần định trị	±	Phần bậc

Hình 6.3. Biểu diễn dấu phảy động

Nguyên tắc mã dấu của phần định trị và phần bậc cũng giống như trong trường hợp số dấu phảy tĩnh. Như vậy ví trí dấu phảy thực sự của số là do phần bậc định ra trên phần định trị. Chính vì vậy người ta gọi kiểu biểu diễn số này là biểu diễn dấu phảy động.

Biểu diễn dấu phảy động có hai ưu điểm so với biểu diễn dấu phảy tĩnh là:

- Khoảng biểu diễn số rất lớn theo nghĩa với cùng một số vị trí (tương ứng là các ngăn lưu trữ trong một ô nhớ), kiểu dấu phảy động có thể biểu diễn các số có giá trị tuyệt đối rất lớn hoặc rất nhỏ so với biểu diễn dấu phảy tĩnh.

- Một khi số lượng các ngăn chứa các chữ số của một đã được xác định thì cũng có nghĩa là ta phải chấp nhận sai số làm tròn. Giả sử có số x mà do làm trong ta chỉ biểu diễn được giá trị x'. Khi đó |x-x'| được gọi là sai số tuyệt đối và |x-x'| /|x| được gọi là sai số tương đối. Kiểu biểu diễn số dấu phảy động có sai số tương đối khá tốt.

Để dễ hình dung, ta minh họa *trên hệ thập phân* với ô nhớ gồm 10 ngăn, mỗi ngăn chứa được một chữ số thập phân.

Giả sử trong chế độ dấu phảy tĩnh ta dùng 1 ngăn dấu của số, 6 số cho phần nguyên, 3 số cho phần lẻ.

±		
	phần nguyên	phần lẻ

Số dương lớn nhất biểu diễn được là 999999,99. Số dương nhỏ nhất là 0,001.

Một cách tổng quát khi dùng m ngăn cho phần nguyên, n ngăn cho phần lẻ thì số dương lớn nhất biểu diễn được là 10^m-10^{-n}, còn số dương nhỏ nhất biểu diễn được là 10^{-n}

Nếu cũng vẫn 10 ngăn đó ta dùng 6 ngăn cho phần định trị và 2 ngăn cho phần bậc, một ngăn cho dấu của bậc.

±		±	
	phần định trị		phần bậc

Số dương lớn nhất biểu diễn được là $0.999999.10^{99}$. Số dương nhỏ nhất biểu diễn được

là 0,1. $10^{-99} = 10^{-100}$.

Một cách tổng quát, nếu dùng m ngăn cho phần định trị và n ngăn cho phần bậc thì số dương lớn nhất có thể biểu diễn được là

$$(1-10^{-m})\ 10^{10^{m}-1}$$

Còn số dương nhỏ nhất có thể biểu diễn được là

$$10^{10^{-m}}$$

Rõ ràng là khoảng số biểu diễn được của chế độ dấu phảy động tốt hơn rất nhiều so với chế độ biểu diễn dấu phảy tĩnh.

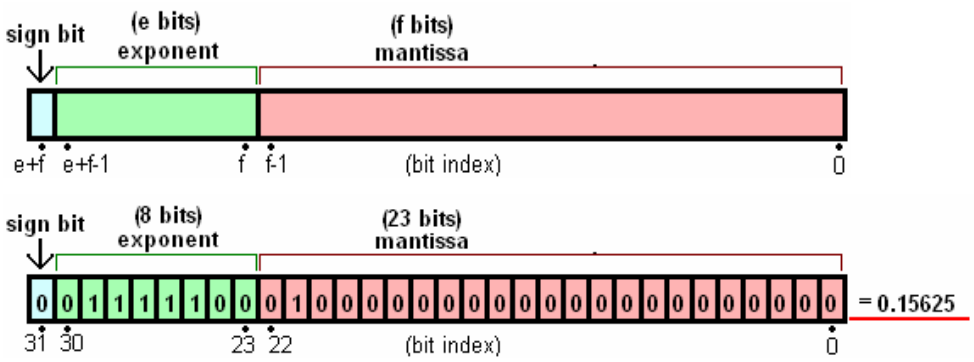
Trong ví dụ trên, với kiểu biểu diễn dấu phảy tĩnh, sai số tuyệt đối là 10^{-3} (0,001) và sai số tương đối xấp xỉ 10^{-9} .(0,001/999999,999). Một cách tổng quát nếu ta dùng m ngăn cho phần lẻ và n ngăn cho phần nguyên thì chế độ dấu phảy tĩnh sai số tuyệt đối gây ra do làm tròn là 10^{-m} còn sai số tương đối có thể lên tới 10^{-m-n}.

Với chế độ dấu phảy động, trong ví dụ trên, sai số tuyệt đối có thể khá lớn do bị khuyếch đại theo hệ số 10^{Px} như $0{,}000001.10^{99}$. Còn sai số tương đối thì rất tốt, luôn luôn đạt mức 10^{-m-n} .

Chuẩn dấu phảy động IEEE 754

Chuẩn dữ liệu dấu phảy động được dùng rộng rãi hiện nay có khác chút ít với chuẩn dữ liệu dấu phảy động nêu trên. Với chuẩn IEEE 754, người ta đưa vào những yếu tố mới để việc xử lý số liệu đa dạng hơn, phản ánh được đầy đủ thực tiễn tính toán với các số gần đúng.

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 mode làm tròn số và 5 ngoại lệ. Bit đầu tiên là dấu của số, sau đó là phần bậc, cuối cùng là phần định trị.



Bảng dưới đây mô tả định dạng cho chuẩn dấu phảy động IEEE 754

Kiểu	Exponent	Mantissa
Zeroes (Số 0)	0	0
Denormalized numbers (Số không được chuẩn hoá)	0	khác 0
Normalized numbers (Số trong dạng chuẩn hoá)	1 to $2^e - 2$ Khác 1…11	bất kỳ
Infinities (Vô hạn)	$2^e - 1$ 11111…11	0
NaNs (Phi số, định dạng không phải số)	$2^e - 1$ 11111…11	khác 0

6.3. Dữ liệu phi số

6.3.1. Mã hoá chữ và dữ liệu kiểu văn bản.

Đơn vị cơ sở của dữ liệu văn bản là chữ. Ở đây khái niệm chữ cần được hiểu theo nghĩa rộng, không chỉ là các chữ cái la tinh mà kể cả chữ số, các dấu chính tả, các dấu toán học, các kí hiệu để trình bày. Mặt khác không phải dân tộc nào cũng dùng chữ latinh nên đối với một số dân tộc có thể có những chữ riêng. Ví dụ bộ chữ Trung hoa có đến hơn 60 nghìn chữ.

Để đỡ gây nhầm lẫn giữa khái niệm chữ theo nghĩa chữ cái thông thường (letter) với "chữ" dùng trong văn bản nói chung kể cả văn bản máy tính, từ đây trở đi chúng ta sẽ dùng thuật ngữ *ký tự* (character) với ý nghĩa là một ký hiệu dùng trong văn bản.

Nếu dùng một vùng nhớ k bit để mã hoá một chữ thì chỉ có thể biểu diễn được tối đa là 2^k kí tự vì chỉ có thể tạo được đúng 2^K các mã nhị phân khác nhau. Điều này giải thích tại sao người Mỹ chỉ cần 7 bit để mã cho các chữ của họ; để có thêm các mặt chữ châu Âu, chữ Hy lạp và một số ký hiệu trình bày cũng chỉ cần 8 bít; trong khi đó người Trung hoa hay người Nhật phải dùng các mã 16 bít.

Các văn bản được hình dung như một chuỗi kí tự. Nội dung một cuốn sách, một bài thơ được đưa vào máy tính là những ví dụ cụ thể về thông tin văn bản. Hầu hết các máy tính và môi trường lập trình hiện nay đều sử dụng một byte để mã hoá một chữ.

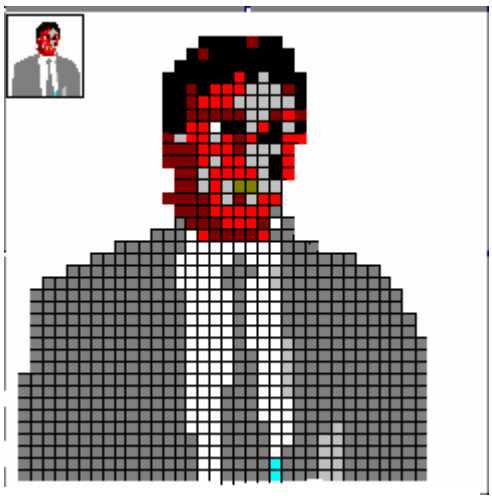
6.3.2. Các dữ liệu logic

Dữ liệu loại logic chỉ thể hiện một trong hai trạng thái đối lập là đúng/sai, hoặc có/không. Điều này ta thường thấy trong rất nhiều loại hồ sơ. Ví dụ trong lý lịch cá nhân: họ tên, quê quán là dữ liệu kiểu văn bản, ngày tháng năm sinh, lương có thể thể hiện bằng số, còn các thông tin như có là đoàn viên không, có gia đình hay không là các thông tin có

kiểu logic. Các thông tin kiểu logic chịu tác động của các phép toán đặc trưng như các phép toán so sánh, các phép toán nhân logic "và", cộng logic "hoặc" hay phủ định logic "không" ví dụ trong một hệ thống quản lý sinh viên người ta có thể phải đặt các câu hỏi như: In ra danh sách các sinh viên (mà) tuổi < 20 hoặc tuổi <21 và là nữ và là đoàn viên và không phải nơi sinh là Hà Nội.

Về nguyên tắc có thể mã giá trị sai hay không bởi bít 0, giá trị đúng hay có bởi bít 1. Tuy nhiên ít khi người ta sử dụng tới mức bít vì cơ chế địa chỉ hoá thường ít nhất ở mức byte. Khi đó người ta vẫn dùng một byte để mã hoá các giá trị logic

6.3.3. Hình ảnh



Hình 6.4. Ảnh bitmap

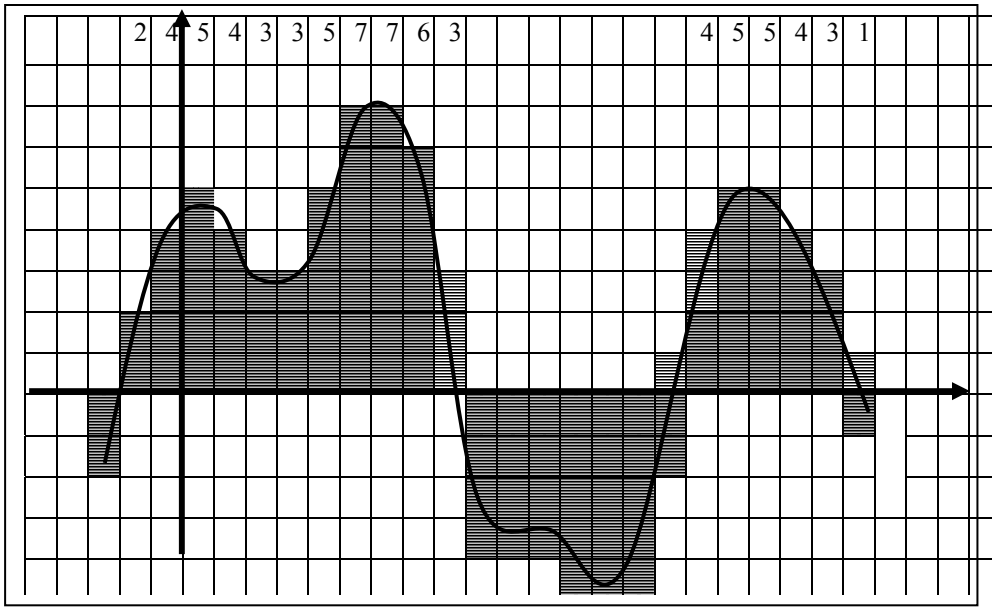
Hình ảnh cũng có thể xử lý bằng máy tính. Khác với hình ảnh thông thường, hình ảnh trong máy tính được mã hoá dưới dạng nhị phân. Có rất nhiều kiểu mã hoá ảnh trong đó hai kiểu thông dụng nhất là.

Ảnh bitmap (nghĩa là bản đồ các bít) thể hiện ảnh như một lưới điểm. Như vậy mỗi điểm sẽ phải nằm trong một hàng và một cột nào đó trong lưới, ngoài ra màu của điểm cũng được mã hoá. Các ảnh khí tượng do các vệ tinh chụp gửi về, ảnh phong cảnh, chân dung đều có thể thể hiện theo kiểu này. Ta cũng có thể đưa một ảnh bất kỳ vào máy dưới dạng bitmap bằng máy quét ảnh (scanner), máy quay video số (digital video camera) hay máy chụp ảnh số (digital camera)... Nói chung dữ liệu ảnh này là dữ liệu lớn. Vì vậy, người ta thường sử dụng các kỹ thuật nén ảnh trước khi đưa vào máy lưu trữ và khôi phục ảnh khi trình bày. Có rất nhiều chuẩn ảnh khác nhau, chủ yếu khác nhau ở cách tổ chức để nén được ảnh mà vẫn giữ được chất lượng và thể hiện được các hiệu ứng ảnh. Còn lúc hiển thị để xem thì ảnh sẽ được khôi phục dưới dạng bitmap. Ảnh thể hiện theo từng điểm còn gọi là ảnh raster.

Kiểu thứ 2 thể hiện ảnh theo cách vẽ. Kiểu này chỉ phù hợp với các ảnh có thành phần là các điểm rời rạc, các đường hoặc hình thể hiện bằng các đường biên như bản vẽ kiến trúc, các bản vẽ kỹ thuật, bản đồ. Cách lưu trữ là lưu thông tin về các thành phần của ảnh. Đối với một đoạn thẳng thì chỉ lưu toạ độ các đầu mút, đối với một hình tròn thì chỉ lưu toạ độ tâm và bán kính... Vì thế các ảnh này thường gọn gàng và dễ phóng to thu nhỏ (vì chỉ dùng các phép biến đổi toạ độ). Các ảnh kiểu này gọi là ảnh vector.

6.3.4. Âm thanh

Âm thanh cũng có thể được xử lý bằng máy tính. Cũng có nhiều phương pháp mã hoá 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. Dĩ nhiên 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. Hình vẽ dưới đây minh hoạ cách lưu trữ xấp xỉ sóng âm, theo đó sẽ lưu lại dãy các giá trị sau:



Hình 6.5. Số hoá âm thanh

Khi phát, một mạch điện sẽ khôi phục lại sóng âm với một sai lệch chấp nhận được.

Một cách khác là phân tích dao động âm thanh thành tổng các dao động điều hoà (các dao động hình sinh với tần số và biên độ khác nhau) và chỉ lưu lại các đặc trưng về tần số, và biên độ.

Còn có nhiều cách mã hoá âm thanh dựa theo những nguyên lý nén dữ liệu rất hiệu quả

Việc số hoá âm thanh cũng được thực hiện nhờ các thiết bị chuyên dụng

Xử lý âm thanh trên máy tính gồm những việc sau:

- Thu và mã hoá âm thanh,
- Biên tập (sửa chữa, ghép, cắt),
- Phân tích (tìm các đặc trưng để nhận dạng tiếng nói). Một số máy tính đã có thể nghe được các lệnh đơn giản. Các máy điện thoại di động hiện nay đã có khả năng nhận dạng tiếng nói,
- Tổng hợp tiếng nói. Ở mức độ đơn giản máy tính có thể đọc văn bản thành lời.

6.4. Biểu diễn vật lý của thông tin trong máy tính

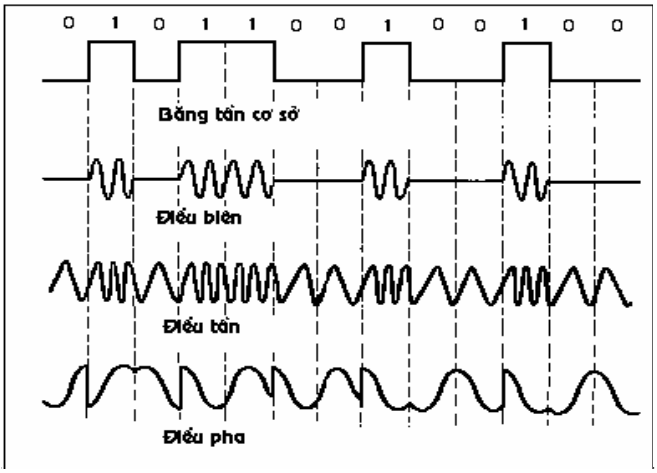
Đối với bộ nhớ trong, các thông tin sau khi mã hoá dưới dạng nhị phân được đưa vào bộ nhớ theo quy ước. Mỗi ngăn của ô nhớ sẽ lưu giữ một trong hai trạng thái được quy ước là một trong hai bit 0 hoặc 1.

Với đĩa từ, người ta không ghi mã hoá bit 0 hay 1 theo chiều của từ thông của các vùng nhiễm từ. Vấn đề là khi đọc, đầu đọc trên nguyên tắc cảm ứng từ chỉ có thể cảm nhận được các trạng thái khác nhau qua sự biến thiên của từ trường. Nếu dùng chiều của từ

thông để mã hoá thì không thể phân biệt được các bít giống nhau đứng liền nhau. Thông thường các bít được ghi theo kiểu điều tần. Các bít được thể hiện qua các kiểu biến thiên của từ trường chứ không phải chiều của từ thông một vùng nhiễm từ trên đĩa. Thực ra cách ghi trên điã từ khá phức tạp vì người ta không những chỉ ghi dữ liệu mà còn có các thông tin về địa chỉ và các thông tin đồng bộ giúp cho việc đọc thông tin được chính xác.

6.5. Truyền tin giữa các máy tính.

Người ta muốn truyền các bít từ máy này đến máy khác và cách đơn giản nhất là phân biệt các bit bằng điện áp, ví dụ điện áp 5 v để thể hiện bít 1, điện áp -5 v thể hiện bit 0. Thực tế không bao giờ có thể truyền và nhận các tín hiệu dưới dạng các xung vuông mà bao giờ khi chuyển từ mức điện áp nọ đến mức điện áp kia cũng có những giai đoạn điện áp nhận những giá trị trung gian. Không những thế còn rất nhiều ảnh hưởng khác làm tín hiệu lúc nhận khác với tín hiệu mức thu như sự suy yếu tín hiệu làm biên độ thay đổi, hình dạng tín hiệu thay đổi - bị méo và đương cong tín hiệu không trơn như lúc đầu do nhiễu. Những vấn đề này cần được khắc phục bằng các mạch vật lý.



Hình 6.6. Điều chế tín hiệu

a) Tín hiệu cơ sở (nhị phân);b) Điều biên 1: biên độ khác 0, 0: biên độ 0; c) Điều tần 1: tần số cao, 0: tần số thấp. d) Điều pha 1: pha $\pi/2$, 0: pha $-\pi/2$

Thể hiện các bít bằng mức điện áp chỉ là một cách điều chế tín hiệu mà ta gọi là điều biên. Còn có các phương pháp điều chế tín hiệu khác gọi là điều tần theo đó, các bít thể hiện bằng sự thay đổi tần số của tín hiệu và phương pháp điều pha theo đó các bít được thể hiện bằng pha của tín hiệu hình sin. Người ta cũng kết hợp cả nhiều cách điều chế để có thể tăng tốc độ truyền tin. Một trong các thiết bị cho phép truyền tin giữa các máy qua mạng điện thoại là modem. Tên "modem" có nguồn gốc từ cặp từ "modulation - demodulation" nghĩa là điều chế và giải điều chế. Các dữ liệu truyền ra từ một máy tính sẽ được modem điều chế thành tín hiệu tương tự và gửi đi theo đường điện thoại. Modem nhận sẽ giải điều chế từ tín hiệu tương tự thành các bít chuyển cho máy tính nhận. Một cách mã hoá để truyền trong modem là kết hợp điều biên và điều pha cho phép có thể truyền tin với tốc độ cao hơn tần số của sóng mang.

Còn rất nhiều kiểu truyền thông khác như truyền thông nhờ các môi trường không dây như sóng điện từ hay tia hồng ngoại. Đối với mỗi kiểu truyền thông đó đều có một cách điều chế tín hiệu riêng.