CTT009

Lưu trữ dữ liệu

Lê Thị Nhàn
ltnhan@fit.hcmus.edu.vn

KHOA CÔNG NGHỆ THÔNG TIN
TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN

Nội dung

- Nhắc lại
- Nén dữ liệu
- Lỗi giao tiếp
- Hệ thống tập tin

Nhắc lại

- Số nguyên không dấu

n	Minimum	Maximum
8	0	$2^8 - 1 = \mathbf{255}$
16	0	$2^{16} - 1 = \mathbf{65,535}$
32	0	$2^{32} - 1 = \mathbf{4,294,967,295}$
64	0	$2^{64} - 1 = \mathbf{18,446,744,073,709,551,615}$

Nhắc lại

- Số nguyên có dấu

n	minimum	maximum
8	-2^7 = **-128**	$2^7 - 1$ = **+127**
16	-2^{15} = **-32,768**	$2^{15} - 1$ = **+32,767**
32	-2^{31} = **-2,147,483,648**	$2^{31} - 1$ = **+2,147,483,647**
64	-2^{63} = **-9,223,372,036,854,775,808**	$2^{63} - 1$ = **+9,223,372,036,854,775,807**

Nhắc lại

- Dấu chấm động

Precision	Min	Max
Single	1.1754×10^{-38}	3.40282×10^{38}
Double	2.2250×10^{-308}	1.7976×10^{308}

NÉN DỮ LIỆU

Tại sao phải nén dữ liệu?

- Mục đích
  - Lưu trữ dữ liệu
  - Truyền tải dữ liệu

- ***Data compression***
  - Giảm kích thước của dữ liệu nhưng vẫn giữ lại các thông tin cơ bản

Kỹ thuật

- ***Không mất (lossless)***
  - Không làm mất thông tin trong quá trình nén
- ***Mất thông tin (lossy)***
  - Có thể mất mát thông tin
  - Nén nhiều hơn lossless và lỗi nhỏ
    - Trường hợp ảnh và âm thanh

Kỹ thuật

- ***Run-length encoding***
  - Dữ liệu được nén là những chuỗi dài có cùng giá trị
  - Thay thế các chuỗi có những phần tử giống nhau bằng ***1 mã (code)***
    - Phần tử được lặp lại
    - Số lần xuất hiện trong chuỗi

Kỹ thuật

- ***Frequency-dependent encoding***
  - Chiều dài của chuỗi bits được dùng để biểu diễn cho 1 phần tử dữ liệu bằng tần suất sử dụng phần tử đó
  - Mã hóa với độ dài thay đổi
    - Phần tử dữ liệu được mã hóa với độ dài khác nhau
  - Huffman code

Kỹ thuật

- ***Relative encoding***
  - Các luồng dữ liệu (data streams) chứa nhiều đơn vị, mà mỗi đơn vị chỉ khác 1 chút so với đơn vị trước đó
    - Khung liên tiếp của một ảnh động
  - Ghi nhận lại sự khác nhau giữa các đơn vị dữ liệu liên tiếp
    - Mã hoá mối quan hệ của 1 đơn vị với đơn vị trước
  - Có thể là lossless hoặc lossy
    - Sự khác biệt giữa các đơn vị dữ liệu liên tiếp được mã hóa chính xác hay xấp xỉ

Kỹ thuật

- ***Dictionary encoding***
  - Thông điệp (message) được mã hóa thành 1 chuỗi các tham chiếu đến từ điển
  - Ví dụ : word processors
    - Có sử dụng những bộ từ điển cho mục đích kiểm tra chính tả
- Adaptive dictionary encoding
  - Biến thể của mã hóa từ điển
  - LZW encoding

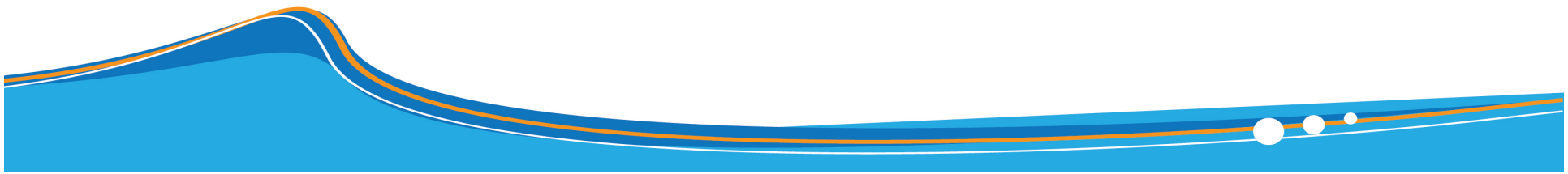
Nén ảnh

- GIF
  - Graphic Interchange Format
  - Phim hoạt hình
- JPEG
  - Joint Photographic Experts Group
  - Chụp hình
- TIFF
  - Tagged Image File Format
  - Lưu trữ hình ảnh

Nén âm thanh và video

- MPEG
  - Motion Photographic Experts Group
  - Phát sóng truyền hình HD
  - Video conferencing
- MP3
  - MPEG layer 3

LỖI GIAO TIẾP

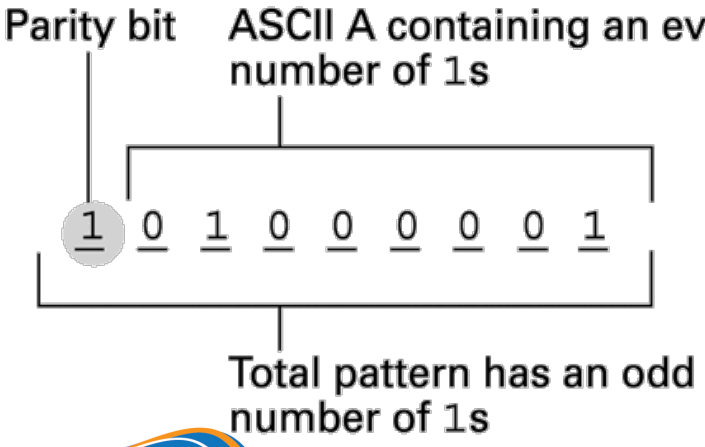


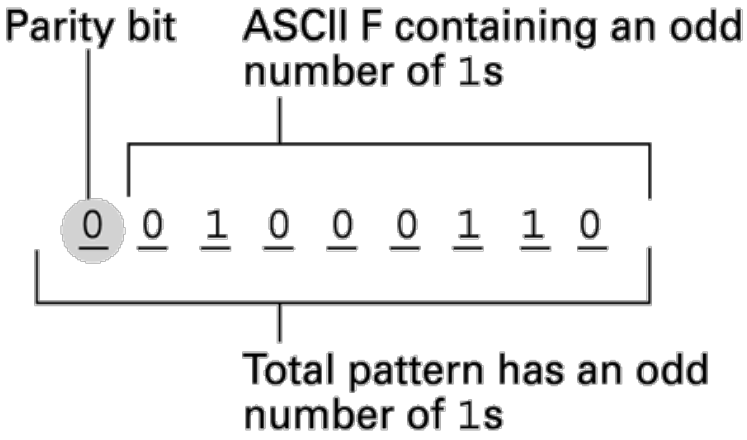
Lỗi giao tiếp là gì?

- Khi thông tin được
  - Chuyển đổi qua lại giữa các thành phần khác nhau của máy tính
  - Lưu trữ trong máy tính
- Chuỗi bit sau cùng nhận được có thể không giống với chuỗi bit ban đầu
- Nguyên nhân
  - Bụi bẩn trên bề mặt đĩa
  - Mạch bị hỏng làm cho việc đọc/ghi không chính xác
  - Đường truyền dữ liệu bị hỏng
  - Bức xạ làm thay đổi chuỗi bits trên bộ nhớ chính

Kỹ thuật

- Bit chẵn lẻ (parity bits)
  - Phát hiện sai sót dựa trên nguyên tắc: nếu 1 chuỗi bit có số lượng lẻ các bit 1 với một chuỗi bit có số lượng chẵn các bit 1 được tìm thấy, thì phải có lỗi xãy ra





Nguồn: Computer Science - An Overview, 12e

Kỹ thuật

- Checkbyte
  - Tập hợp gồm nhiều parity bits
  - Từng parity bit nằm rải rác trong chuỗi bits
    - Ví dụ, 1 parity bit liên kết với mỗi bit thứ 8 trong chuỗi bits

- Biến thể
  - Checksums and cyclic redundancy checks (CRC)

Mã sửa lỗi

- Hamming distance (2 chuỗi bits)
  - Số lượng bits khác nhau trong các chuỗi

- Ví dụ
  - Hamming(00**0000**, 00**1111**) = 4
  - Hamming(10101100, **01**100**0**100) = 3

Ví dụ

Symbol	Code
A	000000
B	001111
C	010011
D	011100
E	100110
F	101001
G	110101
H	111010

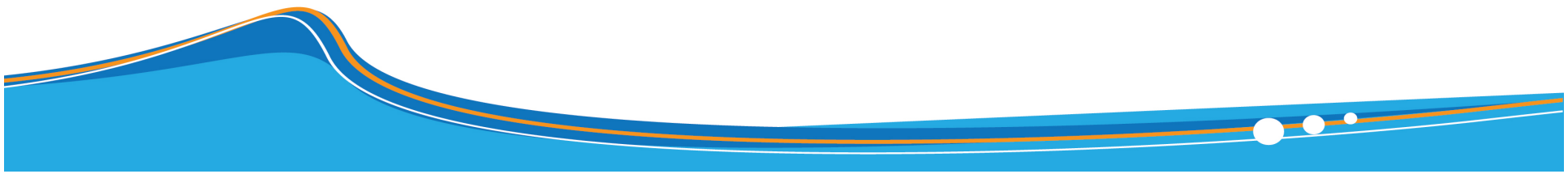
Nguồn: Computer Science - An Overview, 12e

Ví dụ

Character	Code	Pattern received	Distance between received pattern and code
A	0 0 0 0 0 0	0 **1** 0 **1** 0 0	2
B	0 0 1 1 1 1	0 **1** **0** 1 **0** **0**	4
C	0 1 0 0 1 1	0 1 0 **1** **0** **0**	3
D	0 1 1 1 0 0	0 1 **0** 1 0 0	**1** — Smallest distance
E	1 0 0 1 1 0	**0** **1** 0 1 **0** 0	3
F	1 0 1 0 0 1	**0** **1** **0** **1** 0 **0**	5
G	1 1 0 1 0 1	**0** 1 0 1 0 **0**	2
H	1 1 1 0 1 0	**0** 1 **0** **1** **0** 0	4

Nguồn: Computer Science - An Overview, 12e

HỆ THỐNG TẬP TIN



Phân loại

- Tập tin văn bản thô
  - Cấu trúc đơn giản và thông dụng
  - Có thể xem nội dung và sửa chữa bằng các lệnh của hệ điều hành hay chương trình soạn thảo đơn giản
- Tập tin nhị phân
  - Cấu trúc hóa theo một quy ước nào đó
  - Thường có phần header chứa thông tin mô tả sự bố trí và mối liên hệ của các bytes dữ liệu ở phía sau
  - Mở bằng các công cụ (phần mềm) chuyên dụng

Ví dụ tập tin văn bản thô

- Tập tin theo cấu trúc văn bản ANSI (hay ASCII)
  - Chứa các ký tự (mã từ) trong bảng mã ASCII
- Ví dụ : ma trận có 3 dòng 4 cột
  - Dòng đầu cho biết số dòng, số cột
  - 3 dòng tiếp theo mỗi dòng 4 giá trị: nội dung ma trận
- Loại tập tin văn bản cấu trúc thông dụng
  - *.RTF hoặc *.HTML

Ví dụ tập tin văn bản mở rộng

- Văn bản thô ANSI text dựa trên cơ sở các ký tự 8-bit (256 ký hiệu)
  - Bất tiện khi lưu văn bản của nhiều ngôn ngữ

- Văn bản thô dạng mở rộng cho phép lưu trữ được nhiều ngôn ngữ
  - Unicode text (lưu ký tự UTF-16)
  - UTF-8 text

Ví dụ tập tin nhị phân

- Tập tin mã thực thi
  - *.EXE, *.COM, *.DLL trên Windows
- Tập tin văn bản tích hợp văn bản, hình ảnh, bảng biểu
  - *.DOC của MS Word hay Open Office
- Tập tin multimedia
  - Ảnh: *.bmp, *.jpg, …
  - Âm thanh: *.wav, *.mp3, …
  - Video: *.avi, *.mp4, …

TÓM TẮT

Bài giảng hôm nay

- Các kỹ thuật
  - Nén dữ liệu
  - Phát hiện và sửa lỗi giao tiếp
- Dưới góc độ lập trình
  - Hệ thống tập tin

Bài giảng lần tới

- Thao tác dữ liệu (chapter 2)
  - Kiến trúc máy tính
  - Ngôn ngữ máy
  - Thực thi chương trình