1. Khái niệm cơ bản

* Bộ nhớ đệm ( Cache) là thành phần nhớ trong sơ đồ phân cấp bộ nhớ máy tính. Nó hoạt động như thành phần trung gian giữa CPU và bộ nhớ chính
* Cache có thể ở ngoài CPU ( trong các hệ thống cũ) hoặc trong CPU (các hệ thống hiện đại)
* Cache thường có kích thước nhỏ: 16K, 32K,..., 128K đối với các hệ thống cũ và 256K, 512K, 1MB, 2MB,... với các hệ thống mới
* Tốc độ của cache nhanh hơn so với tốc độ bộ nhớ chính
* Chi phí tính trên bit của cache đắt hơn so với bộ nhớ chính

1. Vai trò của Cache

* Nâng cao hiệu năng hệ thống:
* Cân bằng giữa tốc độ nhanh của CPU và tốc độ chậm của bộ nhớ chính (giảm số lượng truy cập trực tiếp của CPU vào bộ nhớ chính)
* Thời gian truy cập trung bình của hệ thống bộ nhớ gần bằng thời gian truy cập cache
* Giảm chi phí sản xuất:
* Nếu 2 hệ thống có cùng hiệu năng thì hệ thống có cache sẽ rẻ hơn
* Nếu 2 hệ thống cùng chi phí, hệ thống có cache sẽ nhanh hơn

1. Các nguyên lý cơ bản của Cache

* Cục bộ về không gian:
* Nếu một vị trí bộ nhớ được truy cập, thì khả năng các vị trí gần đó được truy cập trong thời gian gần tới là cao
* Áp dụng với các mục dữ liệu và các lệnh có thứ tự tuần tự theo chương trình
* Hầu hết các lệnh trong chương trình có thứ tự tuần tự, do đó cache đọc một khối dữ liệu trong bộ nhớ, mà bao gồm cả các phần tử xung quanh vị trí phần tử hiện tại được truy cập
* Cục bộ về thời gian:
* Nếu một vị trí bộ nhớ được truy cập, thì khả năng nó sẽ được truy cập trong thời gian gần tới là cao
* Áp dụng với các mục dữ liệu và các lệnh trong vòng lặp
* Cache đọc khối dữ liệu trong bộ nhớ bao gồm tất cả các thành phần trong vòng lặp

1. Các kỹ thuật tổ chức Cache

* Ánh xạ trực tiếp:
* Đơn giản, nhanh
* Ánh xạ cố định
* Ánh xạ liên kết đầy đủ (associative):
* Phức tạp, chậm
* Ánh xạ linh hoạt
* Ánh xạ liên kết tập hợp/ theo bộ (set):
  + Phức tạp
  + Nhanh, ánh xạ linh hoạt