Bài tập Chương 2

Bài 1. Mô tả một kỹ thuật đơn giản để thực hiện giải thuật thay thế LRU trong bộ nhớ cache theo phương thức 4-way set associcative?

Bài 2. Cho một bộ nhớ cache theo phương thức ánh xạ set-associative có 64 dòng, được chia thành các tập hợp, mỗi tập hợp có 4 dòng. Bộ nhớ chính có 4K khối, mỗi khối có 128 từ. Trình bày khung dạng của địa chỉ trong bộ nhớ chính.

Bài 3. Cho một bộ nhớ cache theo phương thức ánh xạ 2-way set-associative kích thước 8KB, mỗi dòng có 16 byte. Bộ nhớ chính có dung lượng 64MB, có khả năng định vị từng byte. Trình bày khung dạng của địa chỉ trong bộ nhớ chính.

Bài 4. Cho một cache ngoại với các đặc điểm sau:

- 4-way set associative
- một dòng có 2 từ, mỗi từ 16 bit
- có khả năng chứa 4K từ (mỗi từ 32 bit) từ bộ nhớ chính
- dùng cache này cho BXL 16 bit, có 24 bit địa chỉ

Thiết kế cấu trúc cache với các thông số cho trên và giải thích cách phiên dịch địa chỉ của BXL.

Bài 5. BXL Intel 80486 có một bộ nhớ cache nội, đồng nhất (chung cho dữ liệu và lệnh). Cache này chứa 8KB và có phương thức 4-way set associative, kích thước một khối là 4 từ (mỗi từ 32 bit). Cache được tổ chức thành 128 tập hợp. Mỗi dòng cache chứa 1 bit thông tin về tính hợp lệ (line valid bit) và 3 bit B0, B1, B2 (dùng cho phương pháp thay thế LRU). Khi xảy ra cache miss, BXL đọc một khối 16 byte từ bộ nhớ chính. Vẽ sơ đồ cache và giải thích cách phiên dịch địa chỉ của BXL.

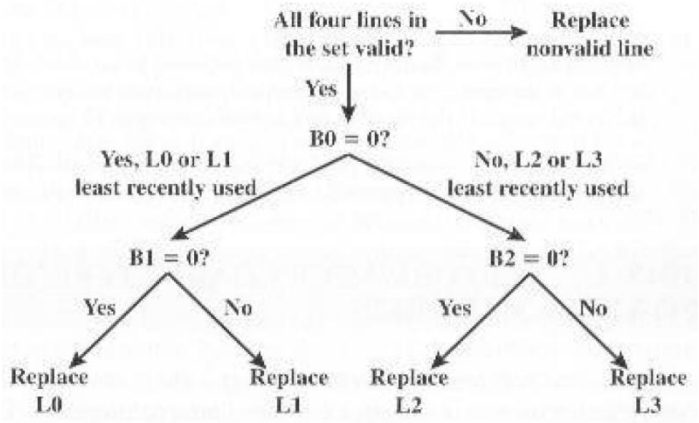
Bài 6. Cho một máy tính với bộ nhớ chính có khả năng định vị địa chỉ theo từng byte, dung lượng 64Kbyte, chia thành các khối có kích thước 8 byte. Giả sử máy tính này có cache dùng phương thức ánh xạ trực tiếp, bao gồm 32 dòng.

a) Một địa chỉ 16 bit được phân chia như thế nào?
b) Địa chỉ sau được ánh xạ vào dòng nào trong cache: 0001 0001 0001 1011, 1100 0011 0011 0100, 1101 0000 0001 1101, 1010 1010 1010 1010
c) Giả sử byte có địa chỉ 1A1A được lưu trong cache, địa chỉ byte nào cũng được lưu cùng với byte này?
d) Tổng cộng bao nhiêu byte bộ nhớ có thể được lưu trong cache?
e) Giải thích tại sao tag cũng được lưu trong cache?

Bài 7. Cho một bộ nhớ cache theo phương thức set associative có kích thước một dòng là 4 từ 16 bit và mỗi tập hợp có 2 dòng. Cache có thể chứa 4048 từ. Bộ nhớ chính có dung lượng có thể mang vào cache là 64K x 32 bit. Thiết kế cấu trúc cache và giải thích cách phiên dịch địa chỉ của BXL.

Bài 8. Cho BXL 80486 với cache nội, dùng giải thuật thay thế “pseudo LRU”. Mỗi 4 dòng (gán nhãn L0, L1, L2, L3) được nhóm thành một tập hợp và có tổng cộng 128 tập hợp. Mỗi tập hợp lưu thêm 3 bit thông tin B0, B1, B2. Giải thuật “pseudo LRU” như sau: khi phải

thay thế 1 dòng, cache sẽ xác định dòng được tham chiếu nhiều nhất trong các dòng L0, L1, L2, L3, sau đó xác định cặp dòng nào được tham chiếu ít nhất và đánh dấu thay thế.



a) Giải thích các bit B0, B1, B2 được set như thế nào và chúng được dùng như thế nào trong giải thuật thay thế trên
b) Chứng minh hiệu quả của giải thuật này xấp xỉ giải thuật LRU
c) Chứng minh rằng giải thuật LRU cần 6 bit cho 1 tập hợp

Bài 9. Cho hệ thống bộ nhớ dùng 32 bit để đánh địa chỉ theo từng byte và một cache có kích thước mỗi dòng là 64 byte.

a) Giả sử cache dùng phương thức ánh xạ trực tiếp với trường tag trong địa chỉ là 20 bit. Trình bày khung dạng địa chỉ và xác định các thông số sau: số đơn vị có khả năng định vị địa chỉ, số khối trong bộ nhớ chính, số dòng trong cache
b) Giả sử cache dùng phương thức fully associative. Trình bày khung dạng địa chỉ và xác định các thông số sau: số đơn vị có khả năng định vị địa chỉ, số khối trong bộ nhớ chính, số dòng trong cache và kích thước của tag
c) Giả sử cache dùng phương thức 4-way set associative với trường tag trong địa chỉ là 9 bit. Trình bày khung dạng địa chỉ và xác định các thông số sau: số đơn vị có khả năng định vị địa chỉ, số khối trong bộ nhớ chính, số dòng trong cache, số tập hợp trong cache

Bài 10. Xét một máy tính với các đặc điểm sau

- Tổng kích thước của bộ nhớ chính là 1MB
- Kích thước từ là 1 byte
- Mỗi khối bao gồm 16 byte
- Kích thước cahe là 64KB

a) Theo phương thức ánh xạ trực tiếp của cache, các địa chỉ sau được phiên dịch như thế nào: F0010, 01234, CABBE
b) Cho ví dụ 2 địa chỉ bộ nhớ chính được ánh xạ vào cùng dòng trong cache (phương thức ánh xạ trực tiếp)
c) Theo phương thức fully associative của cache, các địa chỉ sau được phiên dịch như thế nào: F0010, CABBE
d) Theo phương thức 2-way set associative của cache, các địa chỉ sau được phiên dịch như thế nào: F0010, CABBE

Bài 11. Cho bộ nhớ cache có kích thước một dòng là 32 byte, một từ có 4 byte. Khi một khối dữ liệu được mang từ bộ nhớ chính vào cache, thời gian truy xuất cần cho từ đầu tiên là 30ns và các từ sau mỗi từ cần 5ns. Đối với các dòng được cập nhật ghi tối thiểu 1 lần trước khi được thay thế bởi một khối dữ liệu khác, xác định số lần trung bình cần thiết một dòng được cập nhật ghi trước khi bị thay thế để cho phương pháp write-back hiệu quả hơn write-through.

Bài 12. Cho máy tính có dung lượng bộ nhớ chính là 32K x 16 bit. Nó có bộ nhớ cache 4K từ, được chia thành các tập hợp 4 dòng, mỗi dòng chứa 64 từ. Giả sử ban đầu cache không chứa gì và nó dùng giải thuật LRU cho việc thay thế. BXL nạp các từ từ các ô nhớ 0, 1, 2, …, 4351 theo thứ tự. Nó lặp lại quá trình nạp này thêm 9 lần nữa. Cho biết cache nhanh hơn 10 lần so với bộ nhớ chính. Ước tính độ cải thiện hiệu suất khi dùng cache

Bài 13. Cho hệ thống bộ nhớ với các thông số sau

- T_{cache} = 100ns $C_{cache} = 10^{-4}$ \$/bit
- T_{memory} = 1200ns $C_{memory} = 10^{-5}$ \$/bit

a) Cho biết chi phí 1Mbyte bộ nhớ chính
b) Cho biết chi phí 1Mbyte bộ nhớ chính có dùng cache
c) Nếu thời gian truy xuất thực tế lớn hơn 10% so với thời gian truy xuất cache, xác định hit rate H

Bài 14. Một hệ thống có cache, bộ nhớ chính và đĩa cứng dùng cho bộ nhớ ảo. Nếu ô nhớ cần tham chiếu nằm trong cache, thời gian truy xuất cần thiết là 20ns. Nếu ô nhớ cần tham chiếu nằm trong bộ nhớ chính, thời gian cần để mang nó vào cache là 60ns và sau đó quá trình tham chiếu bắt đầu lại. Nếu ô nhớ cần tham chiếu không nằm trong bộ nhớ chính, thời gian cần để mang nó vào bộ nhớ chính là 12ms, thời gian cần để mang nó vào cache là 60ns và sau đó quá trình tham chiếu bắt đầu lại. Cho xác xuất tìm thấy trong cache là 90% và xác xuất tìm thấy trong bộ nhớ chính là 60%. Xác định thời gian trung bình cần thiết (theo ns) để tham chiếu một ô nhớ trong hệ thống này

Bài 15. Khảo sát tính cục bộ của tham chiếu trong việc tính toán ma trận. Cho đoạn mã C, trong đó các phần tử trên cùng một hàng ma trận được lưu liên tục trong bộ nhớ

```
for (i=0; i<8000; i++)
  for (j=0; j<8; j++)
    A[i][j] = B[j][0] + A[j][i];
```

a) Bao nhiêu số nguyên 32 bit có thể được lưu trong một dòng 16 byte của cache
b) Biến nào trong đoạn chương trình có tính tham chiếu cục bộ về mặt thời gian ?
c) Biến nào trong đoạn chương trình có tính tham chiếu cục bộ về mặt không gian ?

Đoạn chương trình trên có thể được viết bằng Matlab, trong đó, khác với C, các phần tử ma trận trên cùng một cột sẽ được lưu trữ liên tục trong bộ nhớ.

```
for i=1:8000
  for j=1:8
    A(i,j) = B(j,0) + A(j,i);
  end
end
```

d) Cần bao nhiêu dòng cache 16 byte để lưu trữ toàn bộ các phần tử ma trận (mỗi phần tử 32 bit) đang được tham chiếu
e) Biến nào trong đoạn chương trình trên có tính tham chiếu cục bộ về mặt thời gian ?
f) Biến nào trong đoạn chương trình trên có tính tham chiếu cục bộ về mặt không gian ?

Bài 16. Cho danh sách tham chiếu của các từ trong bộ nhớ (địa chỉ bộ nhớ tham chiếu 32 bit):
1, 134, 212, 1, 135, 213, 162, 161, 2, 44, 41, 221

a) Giả sử cache dùng phương thức ánh xạ trực tiếp có 16 dòng (mỗi dòng 1 từ). Giả sử ban đầu cache không chứa gì. Xác định thứ tự dòng và tag mà mỗi từ bộ nhớ được tham chiếu. Xác định rõ mỗi tham chiếu là hit hoặc miss.
b) Giả sử cache dùng phương thức ánh xạ trực tiếp có 8 dòng (mỗi dòng 2 từ). Giả sử ban đầu cache không chứa gì. Xác định thứ tự dòng và tag mà mỗi từ bộ nhớ được tham chiếu. Xác định rõ mỗi tham chiếu là hit hoặc miss.
c) Giả sử ta phải tối ưu cache cho trường hợp tham chiếu trên. Có 3 khả năng thiết kế cache với phương thức ánh xạ trực tiếp: C1) mỗi dòng 1 từ, C2) mỗi dòng 2 từ và C3) mỗi dòng 4 từ. Căn cứ vào miss rate, cách thiết kế nào tốt nhất? Nếu thời gian chờ trong trường hợp miss là 25 chu kỳ, và C1 có thời gian truy xuất 2 chu kỳ, C2 có thời gian truy xuất 3 chu kỳ, C3 có thời gian truy xuất 5 chu kỳ, cách thiết kế nào tốt nhất?

Giả sử kích thước dữ liệu của cache là 64KB, kích thước một dòng cache là 1 từ, thời gian truy xuất cache là 1 chu kỳ

d) Tính tổng số bit của cache trên. Cho cache với tổng kích thước này, tím kích thước của cache ánh xạ trực tiếp gần nhất (bằng hoặc lớn hơn), với kích thước một dòng là 16 từ. Giải thích tại sao với cache thứ 2 này, mặc dù kích thước dữ liệu tăng, nhưng hiệu suất có thể thấp hơn so với cache thứ 1?
e) Tạo ra một chuỗi các yêu cầu tham chiếu bộ nhớ trên một cache (kích thước 2 KB, 2-way set associative) sao cho nó có miss rate nhỏ hơn so với cache thứ 1 trong câu trên? Chỉ ra một giải pháp khả thi sao cho cache thứ 1 trong câu trên có miss rate bằng hoặc nhỏ hơn cache 2 KB này? Phân tích ưu và nhược điểm của giải pháp này.
f) Theo phương thức ánh xạ trực tiếp, số thứ tự dòng trong cache (hàm chỉ mục) sẽ bằng (Block address MODULO Number of cache blocks). Giả sử địa chỉ bao gồm 32 bit và có 1024 khối trong cache, xét một hàm chỉ mục khác như sau: (Block address[31:27] XOR Block address[26:22]). Có thể dùng hàm chỉ mục này cho phương pháp ánh xạ trực tiếp hay không? Nếu được, giải thích tại sao và phân tích các thay đổi cần thiết để thực hiện. Nếu không được, giải thích tại sao?

Bài 17. Cho bộ nhớ cache theo phương thức ánh xạ trực tiếp với 32 bit địa chỉ, được phiên thành các trường sau để truy cập cache

Tag	Line	Word
22	6	4

a) Xác định kích thước một dòng trong cache (theo word)
b) Xác định số dòng trong cache
c) Xác định tỉ số giữa tổng số bit của cache này so với các bit dữ liệu
d) Khi bật nguồn hệ thống, cho chuỗi tham chiếu đến các địa chỉ byte như sau: 0, 4, 16, 132, 232, 160, 1024, 30, 140, 3100, 180, 2180. Xác định số khối bị thay thế? Xác định hit rate? Xác định trạng thái cuối cùng của cache, trong đó mỗi dòng được trình bày theo định dạng <line, tag, data>

Bài 18. Cho cache với write policy được cho trong bảng sau

L1 cache	L2 cache
Write-back, write allocate	Write-through, non write allocate

a) Người ta dùng buffer các lớp bộ nhớ khác mức để giảm latency. Liệt kê các buffer cần thiết giữa cache L1 và L2, và giữa cache L2 và bộ nhớ chính
b) Mô tả qui trình xử lý trong trường hợp write misss ở cache L1
c) Đối với multilevel exclusive cache (nghĩa là, 1 khối chỉ tồn tại trên 1 trong các cache L1 và L2), mô tả qui trình xử lý trong trường hợp write misss ở cache L1
d) Cho thông tin về chương trình và hành vi của cache như sau

Data reads per 1000 instructions	Data writes per 1000 instructions	Instruction cache miss rate	Data cache miss rate	Block size (byte)
200	160	0.20%	2%	8

e) Đối với bộ nhớ cache write-through, write-allocate, băng thông read và write tối thiểu (đo bằng byte/cycle) cần thiết để đạt được CPI = 2?
f) Đối với bộ nhớ cache write-back, write-allocate, giả sử 30% các khối dữ liệu bị thay thế là cần thiết phải cập nhật lên bộ nhớ chính. Xác định băng thông read và write tối thiểu cần thiết để đạt được CPI = 2?
g) Xác định băng thông tối thiểu cần thiết để đạt được CPI = 1.5?

Bài 19. Giả sử thời gian truy xuất bộ nhớ chính là 70ns và 36% truy xuất bộ nhớ này là các lệnh. Bảng sau trình bày dữ kiện của cache L1 được nối với 2 BXL P1 và P2.

	L1 size	L1 miss rate	L1 hit time
P1	1 KB	11.4%	0.62ns
P2	2 KB	8%	0.66ns

a) Giả sử hit time của L1 xác định thời gian chu kỳ cho các BXL. Xác định tốc độ của các BXL?
b) Xác định AMAT cho các BXL
c) Giả sử CPI_{ideal} = 1, xác định CPI thực tế của các BXL P1 và P2. BXL nào chạy nhanh hơn?

Giả sử thêm cache L2 với các đặc tính được cho trong bảng sau vào BXL P1. Miss rate của L2 cho trong bảng là miss rate cục bộ của riêng nó.

L2 size	L2 miss rate	L2 hit time
512 KB	98%	3.22ns

d) Xác định lại AMAT của BXL P1 sau khi thêm cache L2. AMAT này tốt hơn hay xấu hơn?
e) Giả sử CPI_{ideal} = 1, xác định CPI thực tế của P1
f) Với BXL P1 được thêm cache L2, P1 hay P2 nhanh hơn? Nếu P1 nhanh hơn, miss rate của cache L1 trong BXL P2 phải bằng bao nhiêu để hiệu suất của P2 bằng với hiệu suất P1? Nếu P2 nhanh hơn, miss rate của cache L1 trong BXL P1 phải bằng bao nhiêu để hiệu suất của P1 bằng với hiệu suất P2?

Bài 20. BXL Barcelona và Nehalem là các bộ đa xử lý, có nhiều lõi và cache trên cùng một chip. Việc thiết kế cache nội L2 có một số vấn đề cần quan tâm. Bảng sau trình bày miss rate và hit latency cho 2 benchmark với thiết kế L2 dùng chung hoặc riêng. Giả sử cache L1 miss mỗi 32 lệnh.

	Private	Shared
Benchmark A (misses per instruction)	0.30%	0.12%
Benchmark B (misses per instruction)	0.06%	0.03%

Bảng sau thể hiện hit latency

Private cache	Shared cache	Memory
6	12	120

a) Cách thiết kế nào tốt hơn đối với mỗi benchmark?
b) Cache dùng chung có latency tăng khi kích thức BXL tăng. Chọn cách thiết kế tốt nhất nếu latency của cache dùng chung tăng gấp đôi? Băng thông trao đổi dữ liệu bên ngoài chip sẽ trở thành thắt cổ chai khi số lõi BXL tăng. Chọn cách thiết kế tốt nhất nếu latency bộ nhớ ngoại tăng gấp đôi?