

# Computer Systems Architecture CMSC 411 Unit 5 – Memory Hierarchy

Alan Sussman  
March 2, 2006

## Administrivia

- Quiz 1 returned Tuesday
- HW #3 due Tuesday – questions?
- Read Chapter 5
  - except 5.11-5.15
- Project 1 posted

CMSC 411 - Alan Sussman

2

## Last time

- Pipelining
  - detect hazards in ID stage, delay execute cycle if structural or data hazard detected
  - precise exception handling harder with long instructions
    - delay commit of instruction until earlier insts complete
    - or, history or future register file
    - or, detect all exceptions early in pipeline
- MIPS R4000 8-stage pipeline
  - extra stages for memory accesses (insts and data)
  - 2 cycle load delay
  - 3 cycle branch delay w/1 delay slot + 2 cycle stall for *taken* branch
  - complex 8 stage FP unit, combined in different ways for different FP ops

CMSC 411 - Alan Sussman

3

## Memory Hierarchy

- From fast and small, to slow and large
  - Registers
  - Cache
    - maybe multiple levels
  - Main memory
  - I/O devices
    - disk
    - tape
    - CD/DVD/...

CMSC 411 - Alan Sussman

4

## Cache Memory

## Issues to consider

- How big should the fastest memory (cache memory) be?
- How do we decide what to put in cache memory?
- If the cache is full, how do we decide what to remove?
- How do we find something in cache?
- How do we handle writes?

CMSC 411 - Alan Sussman

6

## First, there is main memory

- Jargon:
  - frame address – which page?
  - block number – which cache block?
  - contents – the data

CMSC 411 - Alan Sussman

7

## Then add a cache

- Jargon: Each address of a memory location is partitioned into
  - block address
    - tag
    - index
  - block offset

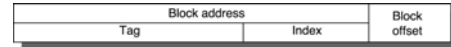


Fig. 5.5

© 2003 Elsevier Science (USA). All rights reserved.  
CMSC 411 - Alan Sussman

8

## How does cache memory work?

- The following slides discuss:
  - what cache memory is
  - three organizations for cache memory
    - direct mapped.
    - set associative
    - fully associative
  - how the bookkeeping is done
- **Important note:** All addresses shown are in octal. Addresses in the book are usually decimal.

CMSC 411 - Alan Sussman

9

## What is cache memory? Main memory first

Main memory is divided into (cache) blocks.  
Each block contains many words (16-64 common now).

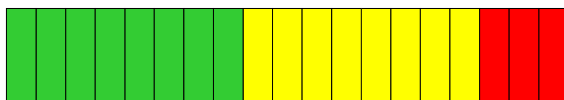


CMSC 411 - Alan Sussman

10

## Main memory

Blocks are grouped into frames (pages), 3 frames in this picture.

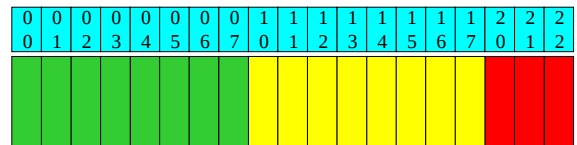


CMSC 411 - Alan Sussman

11

## Main memory (cont.)

Blocks are addressed by their frame number, and their block number within the frame.



CMSC 411 - Alan Sussman

12

## Cache memory

Cache has many, MANY fewer blocks than main memory, each with

a **block number**, 

|   |   |   |   |   |   |   |   |
|---|---|---|---|---|---|---|---|
| 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|---|

a **memory address**, 

|    |    |    |    |    |    |    |    |
|----|----|----|----|----|----|----|----|
| 10 | 21 | 42 | 53 | 74 | 25 | 16 | 77 |
|----|----|----|----|----|----|----|----|

**data**, 

|  |  |  |  |  |  |  |  |
|--|--|--|--|--|--|--|--|
|  |  |  |  |  |  |  |  |
|--|--|--|--|--|--|--|--|

a **valid bit**, 

|   |   |   |   |   |   |   |   |
|---|---|---|---|---|---|---|---|
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
|---|---|---|---|---|---|---|---|

a **dirty bit**, 

|   |   |   |   |   |   |   |   |
|---|---|---|---|---|---|---|---|
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
|---|---|---|---|---|---|---|---|

CMSC 411 - Alan Sussman

13

## Cache memory (cont.)

Initially, all the **valid** bits set to zero.

|    |    |    |    |    |    |    |    |
|----|----|----|----|----|----|----|----|
| 0  | 1  | 2  | 3  | 4  | 5  | 6  | 7  |
| 10 | 21 | 42 | 53 | 74 | 25 | 16 | 77 |
|    |    |    |    |    |    |    |    |
| 0  | 0  | 0  | 0  | 0  | 0  | 0  | 0  |
| 0  | 0  | 0  | 0  | 0  | 0  | 0  | 0  |

CMSC 411 - Alan Sussman

14

## Cache memory (cont.)

Suppose want to load block 14 (octal) from memory into cache.

|    |    |    |    |    |    |    |    |
|----|----|----|----|----|----|----|----|
| 0  | 1  | 2  | 3  | 4  | 5  | 6  | 7  |
| 10 | 21 | 42 | 53 | 74 | 25 | 16 | 77 |
|    |    |    |    |    |    |    |    |
| 0  | 0  | 0  | 0  | 0  | 0  | 0  | 0  |
| 0  | 0  | 0  | 0  | 0  | 0  | 0  | 0  |

Three ways to organize cache

- direct mapped
- set associative
- fully associative

CMSC 411 - Alan Sussman

15

## Direct mapped cache

In **direct mapped** cache, block 14 can only be put in the cache block with address 4.

|    |    |    |    |    |    |    |    |
|----|----|----|----|----|----|----|----|
| 0  | 1  | 2  | 3  | 4  | 5  | 6  | 7  |
| 10 | 21 | 42 | 53 | 74 | 25 | 16 | 77 |
|    |    |    |    |    |    |    |    |
| 0  | 0  | 0  | 0  | 0  | 0  | 0  | 0  |
| 0  | 0  | 0  | 0  | 0  | 0  | 0  | 0  |

So the cache will no longer hold the block with memory address 74.

CMSC 411 - Alan Sussman

16

## Direct mapped cache (cont.)

After the load, the contents look like this.

|    |    |    |    |    |    |    |    |
|----|----|----|----|----|----|----|----|
| 0  | 1  | 2  | 3  | 4  | 5  | 6  | 7  |
| 10 | 21 | 42 | 53 | 14 | 25 | 16 | 77 |
|    |    |    |    |    |    |    |    |
| 0  | 0  | 0  | 0  | 1  | 0  | 0  | 0  |
| 0  | 0  | 0  | 0  | 0  | 0  | 0  | 0  |

CMSC 411 - Alan Sussman

17

## Set associative cache

In **set associative** cache, each memory block can be put in any of a set of possible blocks in cache.

For example, if divide cache into 4 sets, block 14 can be put in any block in Set 0 (since last two bits of 14 octal are zero).

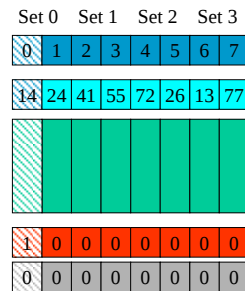
| Set 0 | Set 1 | Set 2 | Set 3 |
|-------|-------|-------|-------|
| 0     | 1     | 2     | 3     |
| 4     | 5     | 6     | 7     |
| 10    | 24    | 41    | 55    |
| 72    | 26    | 13    | 77    |
|       |       |       |       |
| 0     | 0     | 0     | 0     |
| 0     | 0     | 0     | 0     |

CMSC 411 - Alan Sussman

18

## Set associative cache (cont.)

So after loading the block, cache memory might look like this.

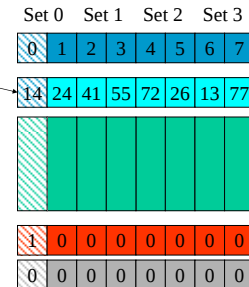


CMSC 411 - Alan Sussman

19

## Set associative cache (cont.)

Note that the last two bits of the memory block's address always match the set number, so do not need to be stored. This part of the address is called the **index**. The higher order bits are stored, and are called the **tag**. In these pictures, both index and tag shown.



CMSC 411 - Alan Sussman

20

## Set associative cache replacement

- Which entry in the set to replace?
- Three common choices:
  - Replace an eligible *random* block
  - Replace the least recently used (LRU) block
    - can be hard to keep track of, so often only approximated
  - Replace the oldest eligible block (First In, First Out, or FIFO)

CMSC 411 - Alan Sussman

21

## Data cache replacement – Fig. 5.6

SPEC2000, in misses per 1000 instructions

Set associativity

| Size  | Two-way |        |       | Four-way |        |       | Eight-Way |        |       |
|-------|---------|--------|-------|----------|--------|-------|-----------|--------|-------|
|       | LRU     | Random | FIFO  | LRU      | Random | FIFO  | LRU       | Random | FIFO  |
| 16KB  | 114.1   | 117.3  | 115.5 | 111.7    | 115.1  | 113.3 | 109.0     | 111.8  | 110.4 |
| 64KB  | 103.4   | 104.3  | 103.9 | 102.4    | 102.3  | 103.1 | 99.7      | 100.5  | 100.3 |
| 256KB | 92.2    | 92.1   | 92.5  | 92.1     | 92.1   | 92.5  | 92.1      | 92.1   | 92.5  |

CMSC 411 - Alan Sussman

22

## Computer Systems Architecture CMSC 411 Unit 5 – Memory Hierarchy

Alan Sussman  
March 7, 2006

## Administrivia

- Quiz 1 returned today
  - Average: 75
  - Median: 70 25%: 60 75%: 90
  - questions?
- Midterm next Thursday?
- HW #3 due today
- Project questions?

CMSC 411 - Alan Sussman

24

## Last time

- Main memory
  - frame address – page
  - block number – which cache block
- Cache memory
  - block address – tag & index
  - block offset
- Cache organization
  - block number, memory address, data, valid bit, dirty bit
  - direct mapped – memory block maps to 1 cache block
  - set associative – block maps to set of cache blocks, a *set* – use the index to pick the set, rest of address bits are the tag
- Cache replacement
  - random, LRU, FIFO policies
  - higher set associativity gets higher hit rates, LRU a little better than FIFO, on SPEC2000 benchmarks

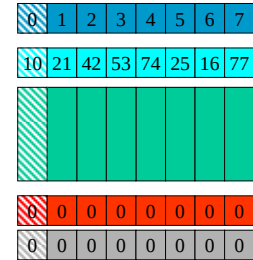
CMSC 411 - Alan Sussman

25

## Fully associative cache

In **fully associative** cache, memory blocks may be stored anywhere.

So block 14 might be put in the first available block -- one with **valid** = 0.

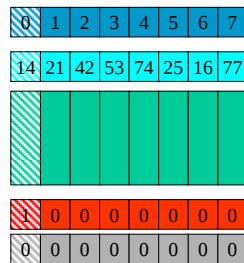


CMSC 411 - Alan Sussman

26

## Fully associative cache (cont.)

With this result.



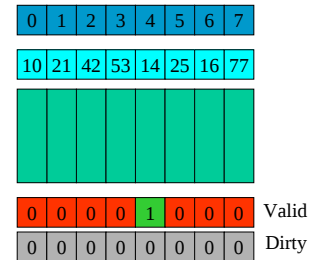
CMSC 411 - Alan Sussman

27

## Managing cache

Use direct mapped cache as an example.

After first read operation, cache memory looked like this.



Valid

Dirty

CMSC 411 - Alan Sussman

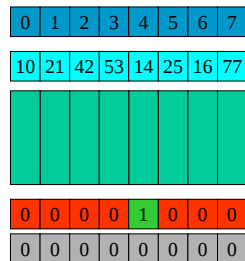
28

## Managing cache (cont.)

If all other memory references involved block 14, no other blocks would need to be fetched from memory.

But suppose eventually need to fetch blocks 10, 31 and 66.

Need to fetch all three, because don't have valid versions of them.



Valid

Dirty

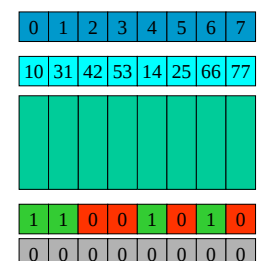
CMSC 411 - Alan Sussman

29

## Managing cache (cont.)

The result looks like this.

Now suppose **write** to block 66.



Valid

Dirty

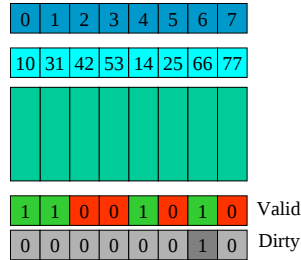
CMSC 411 - Alan Sussman

30

## Managing cache (cont.)

The block is valid in cache, so don't need to fetch.

But the **write** operation sets the **dirty** bit for that block.



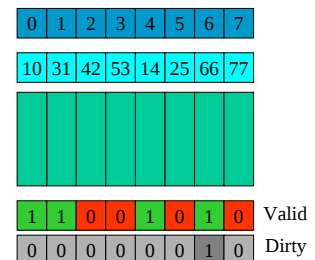
CMSC 411 - Alan Sussman

31

## Managing cache (cont.)

The **write** operation sets the **dirty** bit for that block.

That means that the cached block is different from the memory block, so must eventually be written back.



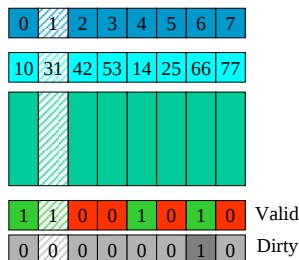
CMSC 411 - Alan Sussman

32

## Managing cache (cont.)

Now suppose need to **read** from a block not in cache.

If it is block 41, then must overwrite block 31.



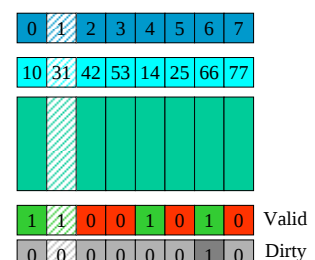
CMSC 411 - Alan Sussman

33

## Write-through cache

In **write-through** caches, every **write** causes an *immediate* change both to cache and to main memory.

So the **read** just involves fetching the block.



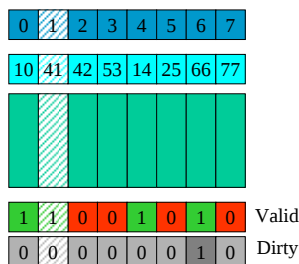
CMSC 411 - Alan Sussman

34

## Write-back cache

In **write-back** caches, every **write** causes a change only to cache.

So the **read** involves writing block 31 back to memory if its **dirty** bit is set, then fetching block 41.



CMSC 411 - Alan Sussman

35

## Reads easy, writes are not

- Most memory access is read, not write, because read both data and instructions but write only data
- If the data requested is not in cache, call that a *cache miss*
- It's easy to make most reads from cache fast: just pull the data into a register as soon as it is accessed, while checking whether the address matches the tag. If not, that is a cache miss, so load a block from main memory to cache.
- Can't do this with write:
  - must verify the address *before* changing the value of the cache location

CMSC 411 - Alan Sussman

36

## Write through vs. write back

- Which is better?
  - Write back gives faster writes, since don't have to wait for main memory
  - Write back is very efficient if want to modify many bytes in a given block
  - But write back can slow down some reads, since a cache miss might cause a write back
  - In multiprocessors, write through might be the only correct solution. Why?

CMSC 411 - Alan Sussman

37

## Cache summary

- Cache memory can be organized as direct mapped, set associative, or fully associative
- Can be write-through or write-back
- Extra bits such as **valid** and **dirty** bits help keep track of the status of the cache

CMSC 411 - Alan Sussman

38

## How much do memory stalls slow down a machine?

- Suppose that on pipelined MIPS, each instruction takes, on average, 2 clock cycles, not counting cache faults/misses
- Suppose, on average, there are 1.33 memory references per instruction, memory access time is 50 cycles, and the miss rate is 2%
- Then each instruction takes, on average:  
 $2 + (0 \times .98) + (1.33 \times .02 \times 50) = 3.33$  clock cycles

CMSC 411 - Alan Sussman

39

## Memory stalls (cont.)

- To reduce the impact of cache misses, can reduce any of three parameters:
  - main memory access time (miss penalty)
  - miss rate
  - cache access (hit) time

CMSC 411 - Alan Sussman

40

## Reducing cache miss penalty

- 5 strategies:
  - Give priority to *read* misses over *write* misses
  - Don't wait for the whole block
  - Use a nonblocking cache
  - Multi-level cache
  - Victim caches
- First 4 used in most desktop and server machines

CMSC 411 - Alan Sussman

41

## Give priority to read misses over write misses

- But need to be careful
- Example:
  - Suppose have a direct mapped cache, with room for 8 blocks of 16 bytes each
  - Then **M[512]** and **M[1024]** both get stored in block 0, so can't be in cache at the same time
- Consider the following instructions:  
SD R3, 512(R0)  
LD R1, 1024(R0)  
LD R2, 512(R0)

CMSC 411 - Alan Sussman

42

### Example (cont.)

- If the cache is write-through, the **SD** will cause memory location 512 to be changed
- The first **LW** will cause block 0 to be replaced, so that the contents **M[512]** are no longer available in cache
  - If the system is write-back, this is when memory location 512 will be changed
- Physically, the contents of block 0 will be put into temporary storage (a *write buffer*) while the new block is loaded, then the write back proceeds
- The second **LW** again replaces block 0, but this time no write-back is necessary
- But get a RAW hazard if don't ensure that the write-through or write-back completes before the second **LW** reads memory

CMSC 411 - Alan Sussman

43

### Example (cont.)

- To avoid such RAW hazards:
  - Can force the read miss to always wait until the write buffer is empty
  - Or can force the hardware to check the write buffer before read and only wait if there is a potential hazard

CMSC 411 - Alan Sussman

44

### Another write buffer optimization

- Write buffer mechanics, with *merging*
  - An entry may contain multiple words (maybe even a whole cache block)
  - If there's an empty entry, the data and address are written to the buffer, and the CPU is done with the write
  - If buffer contains other modified blocks, check to see if new address matches one already in the buffer – if so, combine the new data with that entry
  - If buffer full and no address match, cache and CPU wait for an empty entry to appear (meaning some entry has been written to main memory)
  - Merging improves memory efficiency, since multi-word writes usually faster than one word at a time

CMSC 411 - Alan Sussman

45

### Don't wait for the whole block

- Two ways to do this – suppose need the 10th word in a block:
  - *Early restart*: access the required word as soon as it is fetched, instead of waiting for the whole block
  - *Critical word first*: start the fetch with word 10, and fill in the first few later

CMSC 411 - Alan Sussman

46

### Use a nonblocking cache

- With this optimization, the cache doesn't stop for a miss, but continues to process later requests if possible, even though an earlier one is not yet fulfilled
  - Introduces significant complexity into cache architecture – have to allow multiple outstanding cache requests (maybe even multiple misses)

CMSC 411 - Alan Sussman

47

## Computer Systems Architecture CMSC 411 Unit 5 – Memory Hierarchy

Alan Sussman  
March 9, 2006

## Administrivia

- HW #4 posted
  - questions not from book
- Midterm likely April 4
- Project questions?

CMSC 411 - Alan Sussman

49

## Last time

- Fully associative cache
  - memory block can be stored in any cache block
- Write-through cache
  - write (store) changes both cache and main memory right away
  - reads only require getting block on cache miss
- Write-back cache
  - write changes only cache
  - read causes write of dirty block to memory on a replace
- Reads easy to implement, writes harder
  - read data from cache in parallel with checking address against tag of cache block
  - write must verify address against tag before update
- Reducing memory stalls
  - reduce miss penalty, miss rate, cache hit time
- Reducing miss penalty
  - give priority to read over write misses
  - don't wait for the whole block
  - use a non-blocking cache

CMSC 411 - Alan Sussman

50

## Multi-level cache

- For example, if cache takes 1 clock cycle, and memory takes 50, might be a good idea to add a larger (but necessarily slower) secondary cache in between, perhaps capable of 10 clock cycle access
- Complicates performance analysis (see H&P), but 2<sup>nd</sup> level cache captures many of 1<sup>st</sup> level cache misses, lowering effective miss penalty

CMSC 411 - Alan Sussman

51

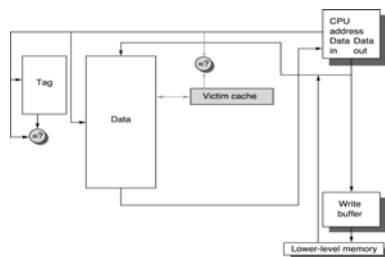
## Victim caches

- To remember a cache block that has recently been replaced (*evicted*)
  - use a small, *fully associative* cache between a cache and where it gets data from
  - check the victim cache on a cache miss, before going to next lower-level memory
    - if found, swap victim block and cache block
  - reduces *conflict* misses, soon to be discussed

CMSC 411 - Alan Sussman

52

## Victim caches (cont.) – Fig. 5.13



© 2003 Elsevier Science (USA). All rights reserved.  
CMSC 411 - Alan Sussman

53

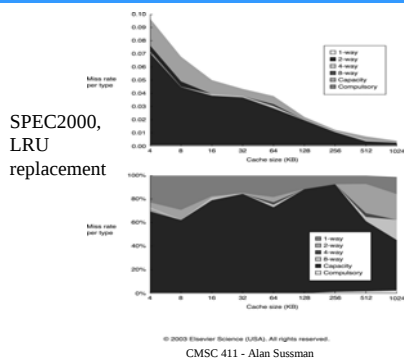
## Reducing the miss rate

- Sometimes cache misses are inevitable:
  - The first time a block is used, need to bring it into cache (a *compulsory miss*)
  - If need to use more blocks at once than can fit into cache, some will bounce in and out (*capacity miss*)
  - In direct mapped or set associative caches, there are certain combinations of addresses that cannot be in cache at the same time (*conflict miss*)

CMSC 411 - Alan Sussman

54

## Miss rate – Fig. 5.15



55

## How to reduce the miss rate?

- Use larger blocks
- Use more associativity, to reduce conflict misses
- Victim cache
- Pseudo-associative caches (won't talk about this)
- Prefetch (hardware controlled)
- Prefetch (compiler controlled)
- Compiler optimizations

CMSC 411 - Alan Sussman

56

## Increasing block size

- Want the block size **large** so don't have to stop so often to load blocks
- Want the block size **small** so that blocks load quickly

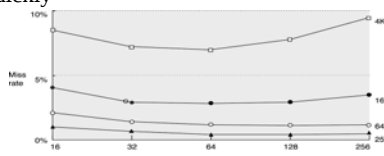


Fig. 5.16 – SPEC92

© 2003 Elsevier Science (USA). All rights reserved.  
CMSC 411 - Alan Sussman

57

## Increasing block size (cont.)

- So large block size reduces miss rates, but...
- Example:
  - Suppose that loading a block takes 80 cycles (overhead) plus 2 clock cycles for each 16 bytes
  - A block of size 64 bytes can be loaded in  $80 + 2 \cdot 64 / 16$  cycles = 88 cycles (miss penalty)
  - If the miss rate is 7%, then the average memory access time is  $1 + .07 \cdot 88 = 7.16$  cycles

CMSC 411 - Alan Sussman

58

## Memory Access Times – Fig. 5.18

SPEC92 benchmarks on DEC workstation

| Block size | Miss penalty | Cache size   |              |              |              |
|------------|--------------|--------------|--------------|--------------|--------------|
|            |              | 4K           | 16K          | 64K          | 256K         |
| 16         | 82           | 8.027        | 4.231        | 2.673        | 1.894        |
| 32         | 84           | <b>7.082</b> | 3.411        | 2.134        | 1.588        |
| 64         | 88           | 7.160        | <b>3.323</b> | <b>1.933</b> | <b>1.449</b> |
| 128        | 96           | 8.469        | 3.659        | 1.979        | 1.470        |
| 256        | 112          | 11.651       | 4.685        | 2.288        | 1.549        |

CMSC 411 - Alan Sussman

59

## Higher associativity

- A direct-mapped cache of size  $N$  has about the same miss rate as a 2-way set-associative cache of size  $N/2$ 
  - 2:1 cache rule of thumb (seems to work up to 128KB caches)
- But associative cache is slower than direct-mapped, so the clock may need to run slower
- Example:
  - Suppose that the clock for 2-way memory needs to run at a factor of 1.1 times the clock for 1-way memory
    - the hit time increases with higher associativity
  - Then the average memory access time for 2-way is  $1.10 + \text{miss rate} \times 50$  (assuming that the miss penalty is 50)

CMSC 411 - Alan Sussman

60

## Memory access time – Fig. 5.19

| Cache size<br>(KB) | Associativity |             |             |             |
|--------------------|---------------|-------------|-------------|-------------|
|                    | One-way       | Two-way     | Four-way    | Eight-way   |
| 4                  | 3.44          | 3.25        | 3.22        | <b>3.28</b> |
| 8                  | 2.69          | 2.58        | 2.55        | <b>2.62</b> |
| 16                 | 2.23          | <b>2.40</b> | <b>2.46</b> | <b>2.53</b> |
| 32                 | 2.06          | <b>2.30</b> | <b>2.37</b> | <b>2.45</b> |
| 64                 | 1.92          | <b>2.14</b> | <b>2.18</b> | <b>2.25</b> |
| 128                | 1.52          | <b>1.84</b> | <b>1.92</b> | <b>2.00</b> |
| 256                | 1.32          | <b>1.66</b> | <b>1.74</b> | <b>1.82</b> |
| 512                | 1.20          | <b>1.55</b> | <b>1.59</b> | <b>1.66</b> |

CMSC 411 - Alan Sussman

61

## Pseudo-associative cache

- Uses the technique of *chaining*, with a series of cache locations to check if the block is not found in the first location
  - e.g., invert most significant bit of index part of address (as if it were a set associative cache)
- The idea:
  - Check the direct mapped address
  - Until the block is found or the chain of addresses ends, check the next alternate address
  - If the block has not been found, bring it in from memory
- Three different delays generated, depending on which step succeeds

CMSC 411 - Alan Sussman

62

## Hardware prefetch

- Idea: If read page  $k$  of a book, the next page read is most likely page  $k+1$
- So, when a block is read from memory, read the next block too
  - maybe into a separate buffer that is accessed on a cache miss before going to memory
- Advantage:
  - if use blocks sequentially, will need to fetch only half as often from memory
- Disadvantages:
  - more information to move
  - may fill the cache with useless blocks
  - may compete with demand misses for memory bandwidth

CMSC 411 - Alan Sussman

63

## Compiler-controlled prefetch

- Idea: The compiler has a better idea than the hardware does of when blocks are being use sequentially
- Want the prefetch to be nonblocking:
  - don't slow the pipeline waiting for it
- Usually want the prefetch to fail quietly:
  - if ask for an illegal block (one that generates a page fault or protection exception), don't generate an exception; just continue as if the fetch wasn't requested
  - called a *non-binding* cache prefetch

CMSC 411 - Alan Sussman

64

## Compiler optimizations to reduce cache miss rate

## Four compiler techniques

- 4 techniques to improve cache locality:
  - merging arrays
  - loop interchange
  - loop fusion
  - blocking

CMSC 411 - Alan Sussman

66

## Technique 1: merging arrays

- Suppose have two arrays:

```
int val[size];  
int key[size];
```

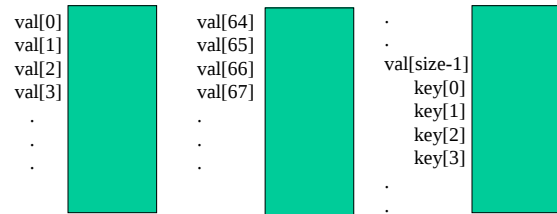
- and that usually use both of them together

CMSC 411 - Alan Sussman

67

## Merging arrays (cont.)

This is how they would be stored if cache blocksize is 64 words:

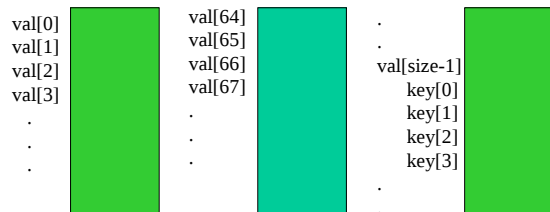


CMSC 411 - Alan Sussman

68

## Merging arrays (cont.)

Means that at least 2 blocks must be in cache to begin using the arrays.

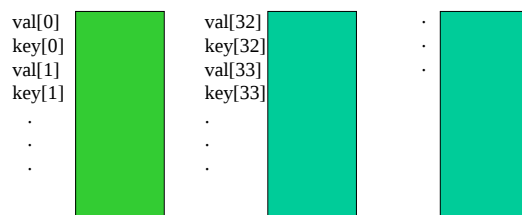


CMSC 411 - Alan Sussman

69

## Merging arrays (cont.)

More efficient, especially if more than two arrays are coupled this way, to store them together.

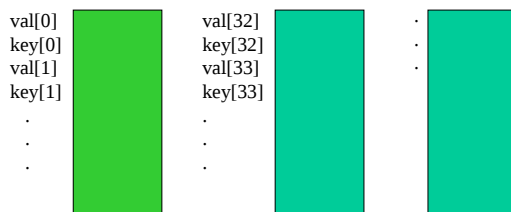


CMSC 411 - Alan Sussman

70

## Merging arrays (cont.)

Can do this by making the two arrays part of a structure.

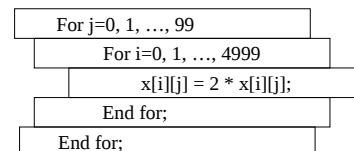


CMSC 411 - Alan Sussman

71

## Technique 2: interchanging loops

Example:



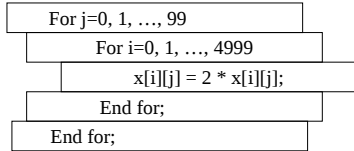
CMSC 411 - Alan Sussman

72

## Interchanging loops (cont.)

Notice that accesses are by columns, so the elements are spaced 100 words apart.

Blocks are bouncing in and out of cache.

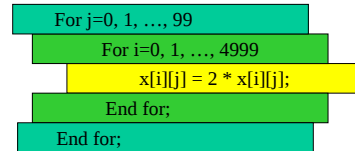


CMSC 411 - Alan Sussman

73

## Interchanging loops (cont.)

First color the loops:

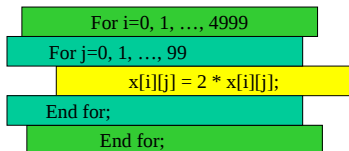


CMSC 411 - Alan Sussman

74

## Interchanging loops (cont.)

Notice that the program has the same effect if the two loops are interchanged:

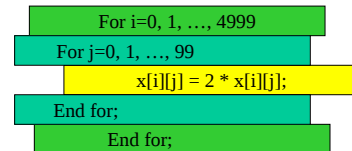


CMSC 411 - Alan Sussman

75

## Interchanging loops (cont.)

But with this ordering, use every element in a cache block before needing another block!



CMSC 411 - Alan Sussman

76

## Computer Systems Architecture CMSC 411 Unit 5 – Memory Hierarchy

Alan Sussman  
March 14, 2006

## Administrivia

- HW #4 posted
  - questions?
- Midterm April 4
- Project questions?

CMSC 411 - Alan Sussman

78

## Last time

- Reducing cache miss penalty
  - multilevel cache – to lower *effective* miss penalty, by capturing misses
  - victim cache – small, fully associative cache to reduce *conflict* misses
- Reducing cache miss rate
  - compulsory vs. capacity vs. conflict misses
  - increase block size – load bigger, but fewer, cache blocks from memory
    - helps up to some block size (64-128 bytes)
  - increase associativity – at cost of slower clock
    - direct mapped generally best for larger caches
  - hardware prefetch – read next sequential block on a miss – at cost of maybe moving extra data from memory
  - compiler controlled prefetch – a non-blocking load instruction – and *non-binding* (fails silently)
- Compiler optimizations
  - merging arrays – turn multiple arrays used together into an array of structs, to access fewer cache blocks
  - interchanging loops – to change the *order* of access to array elements, to improve locality (spatial and temporal)

CMSC 411 - Alan Sussman

79

## Compiler optimizations for reducing cache misses, continued

### Technique 3: loop fusion

Example:

```

For j=0, 1, ..., 99
  For i=0, 1, ..., 4999
    x[i][j] = 2 * x[i][j];
  End for;
End for;

For j=0, 1, ..., 99
  For i=0, 1, ..., 4999
    y[i][j] = x[i][j] * a[i][j];
  End for;
End for;
    
```

CMSC 411 - Alan Sussman

81

### Loop fusion (cont.)

Note that the loop control is the same for both sets of loops.

```

For j=0, 1, ..., 99
  For i=0, 1, ..., 4999
    x[i][j] = 2 * x[i][j];
  End for;
End for;

For j=0, 1, ..., 99
  For i=0, 1, ..., 4999
    y[i][j] = x[i][j] * a[i][j];
  End for;
End for;
    
```

CMSC 411 - Alan Sussman

82

### Loop fusion (cont.)

And note that the array *x* is used in each, so probably needs to be loaded into cache twice, which wastes cycles.

```

For j=0, 1, ..., 99
  For i=0, 1, ..., 4999
    x[i][j] = 2 * x[i][j];
  End for;
End for;

For j=0, 1, ..., 99
  For i=0, 1, ..., 4999
    y[i][j] = x[i][j] * a[i][j];
  End for;
End for;
    
```

CMSC 411 - Alan Sussman

83

### Loop fusion (cont.)

So combine, or **fuse**, the loops to improve efficiency.

```

For j=0, 1, ..., 99
  For i=0, 1, ..., 4999
    x[i][j] = 2 * x[i][j];
    y[i][j] = x[i][j] * a[i][j];
  End for;
End for;
    
```

CMSC 411 - Alan Sussman

84

## Technique 4: blocking access to arrays

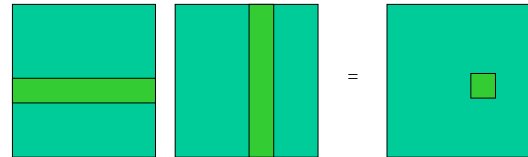
- *Blocking* does not mean *preventing*. Means gathering the accesses into blocks that optimize cache use (access order, block size, etc.)

CMSC 411 - Alan Sussman

85

## Blocking access to arrays (cont.)

Example: Matrix-matrix multiplication

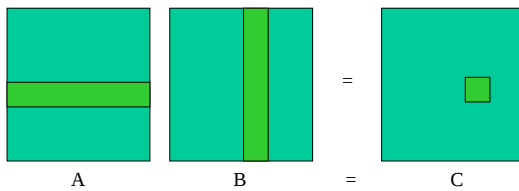


CMSC 411 - Alan Sussman

86

## Blocking access to arrays (cont.)

Trouble: Easy to get rows of A;  
not so efficient to get columns of B.



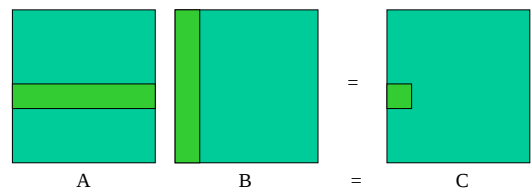
CMSC 411 - Alan Sussman

87

## Blocking access to arrays (cont.)

And if cycle through rows of A, end up loading all of B  $m$  times, where  $m$  is the number of rows of A.

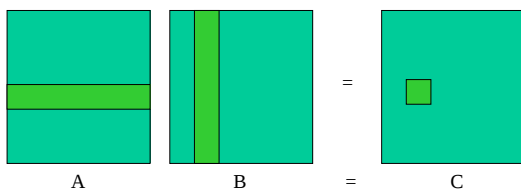
Computing the elements in the columns of C:



CMSC 411 - Alan Sussman

88

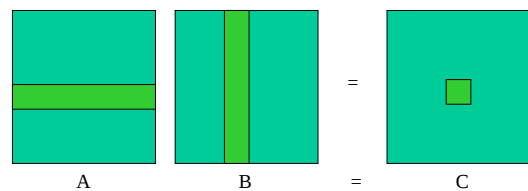
## Blocking access to arrays (cont.)



CMSC 411 - Alan Sussman

89

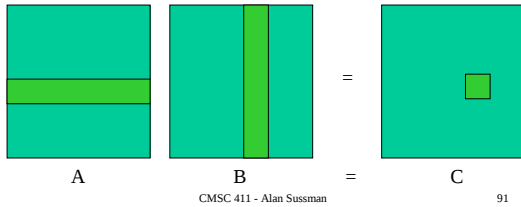
## Blocking access to arrays (cont.)



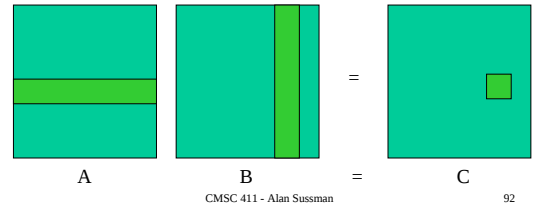
CMSC 411 - Alan Sussman

90

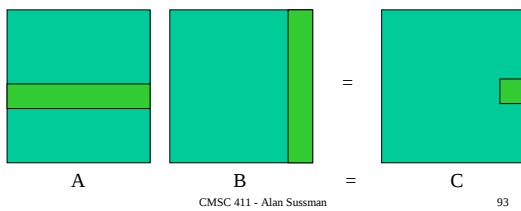
## Blocking access to arrays (cont.)



## Blocking access to arrays (cont.)

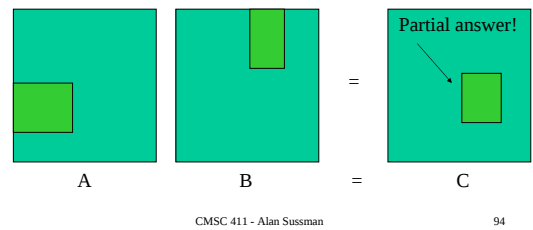


## Blocking access to arrays (cont.)



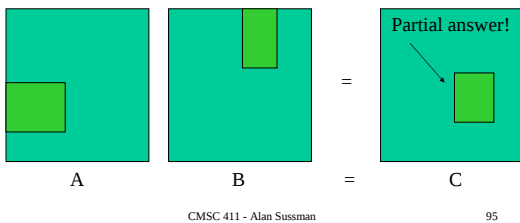
## Blocking access to arrays (cont.)

Instead, order the computation using rectangular blocks of A and B.



## Blocking access to arrays (cont.)

If the block of A has  $k$  rows, then only need to load B  $m/k$  times.



## Blocking access to arrays (cont.)

Improves *temporal locality*

```
/* Before */
for (i=0; i<N; i++)
  for (j=0; j<N; j++) {
    r=0;
    for (k=0; k<N; k++)
      r=r+y[i][k]*z[k][j];
    x[i][j]=r;
  }

/* After */
for (jj=0; jj<N; jj=jj+B)
  for (kk=0; kk<N; kk=kk+B)
    for (i=0; i<N; i++)
      for (j=jj; j<min(jj+B,N); j++) {
        r=0;
        for (k=kk; k<min(kk+B,N); k++)
          r=r+y[i][k]*z[k][j];
        x[i][j]=x[i][j]+r;
      }
```

CMSC 411 - Alan Sussman

96

## Reducing the time for cache hits

- K.I.S.S.
- Use *virtual addresses* rather than *physical addresses* in the cache.
- Pipeline cache accesses
- Trace caches (won't talk about these)

CMSC 411 - Alan Sussman

97

## K.I.S.S.

- Cache should be small enough to fit on the processor chip
- Direct mapped is faster than associative, especially on *read*
  - overlap tag check with transmitting data
- For current processors, small L1 caches to keep fast clock cycle time, hide L1 misses with dynamic scheduling, and use L2 caches to avoid main memory accesses

CMSC 411 - Alan Sussman

98

## Simulated cache access times

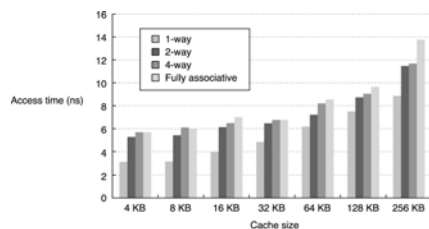


Fig. 5.24 – 0.8-micron feature size, 1 R/W port, 32 address & 64 data bits, 32-byte blocks

© 2003 Elsevier Science (USA). All rights reserved.

CMSC 411 - Alan Sussman

99

## Use virtual addresses

- Each user has his/her own *address space*, and no addresses outside that space can be accessed
- To keep address length small, each user addresses by offsets relative to some physical address in memory (pages)
- For example:

| Physical address | Virtual address |
|------------------|-----------------|
| 5400             | 00              |
| 5412             | 12              |
| 5500             | 100             |

CMSC 411 - Alan Sussman

100

## Virtual addresses (cont.)

- Since instructions use virtual addresses, use them for index and tag in cache, to save the time of translating to physical address space (the subject of the next part of this unit)
- Note that it is important to flush the cache and set all blocks invalid when switch to a new user in the OS (a *context switch*), since the same virtual address then may refer to a different physical address
  - or use the process/user ID as part of the tag in cache
- Aliases are another problem
  - when two different virtual addresses map to the same physical address – can get 2 copies in cache
    - what happens when one copy is modified?

CMSC 411 - Alan Sussman

101

## Pipelined cache access

- Latency to first level cache is more than one cycle
  - we've already seen this in Unit 3
- Benefit is fast cycle time
- Penalty is slower hits
  - also more clock cycles between a load and the use of the data (maybe more pipeline stalls)

CMSC 411 - Alan Sussman

102

## Trace cache

- Find a dynamic sequence of instructions to load into a cache block, including *taken* branches
  - instead of statically, from how the instructions are laid out in memory
  - branch prediction needed for loading cache
- One penalty is complicated address mapping, since addresses not always aligned to cache block size
  - can also end up storing same instructions multiple times
- Benefit is only caching instructions that will actually be used (if branch prediction is right), not all instructions that happen to be in the same cache block

CMSC 411 - Alan Sussman

103

## Main Memory

## Main memory management

- Questions:
  - How big should main memory be?
  - How to handle reads and writes?
  - How to find something in main memory?
  - How to decide what to put in main memory?
  - If main memory is full, how to decide what to replace?

CMSC 411 - Alan Sussman

105

## The scale of things

- Typically (as of 2000):
  - *Registers*: < 1 KB, access time .25 - .5 ns
  - *Cache*: < 8 MB, access time .5 - 25 ns
  - *Main Memory*: < 4 GB, access time 150 - 250 ns
  - *Disk Storage*: > 30 GB, access time 5,000,000 ns (5ms)
- Memory Technology: CMOS (Complementary Metal Oxide Semiconductor)
  - uses a combination of n- and p-doped semiconductor material to achieve low power dissipation.

CMSC 411 - Alan Sussman

106

## Memory hardware

- *DRAM*: dynamic random access memory, typically used for main memory
  - one transistor per data bit
  - each bit must be refreshed periodically (e.g., every 8 milliseconds), so maybe 5% of time is spent in refresh
  - access time < cycle time
  - address sent in two halves so that fewer pins are needed on chip (row and column access)

CMSC 411 - Alan Sussman

107

## Memory hardware (cont.)

- *SRAM*: static random access, typically used for cache memory
  - 4-6 transistors per data bit
  - no need for refresh
  - access time = cycle time
  - address sent all at once, for speed

CMSC 411 - Alan Sussman

108

# Computer Systems Architecture CMSC 411 Unit 5 – Memory Hierarchy

Alan Sussman  
March 16, 2006

## Administrivia

- HW #4 questions?
- Midterm April 4
- Project questions?

CMSC 411 - Alan Sussman

110

## Last time

- Compiler optimizations to reduce cache miss rate
  - loop fusion – to improve temporal locality
  - blocking array accesses – rearrange loop iterations to work on a block of each array in a loop at a time – example is matrix multiply
    - also improves temporal locality
- Reducing time for cache hits
  - Small, direct mapped, and multi-level caches
  - Use virtual, instead of physical addresses
  - Pipeline cache accesses (multi-cycle)
- Main memory
  - DRAM, vs. SRAM for cache (and registers)

CMSC 411 - Alan Sussman

111

## Bottleneck

- Main memory access will slow down the CPU unless the hardware designer is careful
- Some techniques can improve *memory bandwidth*, the amount of data that can be delivered from memory in a given amount of time:
  - wider main memory
  - interleaved memory
  - independent memory banks
  - avoiding memory bank conflicts

CMSC 411 - Alan Sussman

112

## Wider main memory

- Cache miss: If a cache block contains  $k$  words, then each cache miss involves these steps repeated  $k$  times:
  - Send the address to main memory
  - Access the word (i.e., locate it)
  - Send the word to cache, with the bits transmitted in parallel
- Idea behind wider memory: the user thinks about 32 bit words, but physical memory can have longer words
- Then the operations above are done only  $k/n$  times, where  $n$  is the number of 32 bit words in a physical word

CMSC 411 - Alan Sussman

113

## Wider main memory (cont.)

- Extra costs:
  - a wider *memory bus*: hardware to deliver  $32n$  bits in parallel, instead of 32 bits
  - a multiplexor to choose the correct 32 bits to transmit from the cache to the CPU

CMSC 411 - Alan Sussman

114

## Interleaved memory

- Partition memory into banks, with each bank able to access a word and send it to cache in parallel
- Organize address space so that adjacent words live in different banks - called *interleaving*
- For example, 4 banks might have words with the following octal addresses:

| Bank 0 | Bank 1 | Bank 2 | Bank 3 |
|--------|--------|--------|--------|
| 00     | 01     | 02     | 03     |
| 04     | 05     | 06     | 07     |
| 10     | 11     | 12     | 13     |
| ...    | ...    | ...    | ...    |

CMSC 411 - Alan Sussman

115

## Interleaved memory (cont.)

- Note how nice interleaving is for write-through
- Also helps speed read and write-back
- *Note:* Interleaved memory acts like wide memory, except that words are transmitted through the bus sequentially, not in parallel

CMSC 411 - Alan Sussman

116

## Independent memory banks

- Each *bank* of memory has its own address lines and (usually) a bus
- Can have several independent banks: perhaps
  - one for instructions
  - one for data
  - this is called a *Harvard architecture*
- Banks can operate independently without slowing others

CMSC 411 - Alan Sussman

117

## Avoid memory bank conflicts

- By having a prime number of memory banks
- Since arrays frequently have even dimension sizes - and often dimension sizes that are a power of 2 - strides that match the number of banks (or a multiple) give very slow access

CMSC 411 - Alan Sussman

118

## Example

- ```
int x[256][512];
for (j=0; j<512; j=j+1)
  for (i=0; i<256; i=i+1)
    x[i][j] = 2 * x[i][j];
```
- First access the first column of **x**:
    - $x[0][0], x[1][0], x[2][0], \dots, x[255][0]$
  - with addresses
    - $K, K+512*4, K+512*8, \dots, K+512*4*255$
  - With 4 memory banks, all of the elements live in the same memory bank, so the CPU will stall in the worst possible way

CMSC 411 - Alan Sussman

119

## Number theory to the rescue!

- Subtitle: One reason why computer scientists need math
- *Fact 1:* It is easy to compute **mod**, if the base is a prime number that is one less than a power of 2
- *Fact 2:* The Chinese remainder theorem says that it is safe to do a rather bizarre, but convenient, mapping of words to memory banks
- Idea: Suppose have 3 memory banks with 8 words each
  - Map word  $k$  to memory bank  $k \bmod 3$ , word  $k \bmod 8$
  - For example, word 17 (decimal) goes to bank 2, address 1
  - Word 9 goes to bank 0, word 1

CMSC 411 - Alan Sussman

120

## Number theory (cont.)

- That mapping gives the following arrangement of words:

| Address within bank | Memory bank              |    |    |                    |    |    |
|---------------------|--------------------------|----|----|--------------------|----|----|
|                     | Sequentially interleaved |    |    | Modulo interleaved |    |    |
|                     | 0                        | 1  | 2  | 0                  | 1  | 2  |
| 0                   | 0                        | 1  | 2  | 0                  | 16 | 8  |
| 1                   | 3                        | 4  | 5  | 9                  | 1  | 17 |
| 2                   | 6                        | 7  | 8  | 18                 | 10 | 2  |
| 3                   | 9                        | 10 | 11 | 3                  | 19 | 11 |
| 4                   | 12                       | 13 | 14 | 12                 | 4  | 20 |
| 5                   | 15                       | 16 | 17 | 21                 | 13 | 5  |
| 6                   | 18                       | 19 | 20 | 6                  | 22 | 14 |
| 7                   | 21                       | 22 | 23 | 15                 | 7  | 23 |

CMSC 411 - Alan Sussman

121

## How much good do these techniques do?

- Example: Assume a cache block of 4 words, and
  - 4 cycles to send address to main memory
  - 24 cycles to access a word, once the address arrives
  - 4 cycles to send a word back to cache
- Basic miss penalty:  $4 \times 32 = 128$  cycles, since each of 4 words has the full 32 cycle penalty
- Memory with a 2-word width:  $2 \times 32 = 64$  cycle miss penalty
- Simple interleaved memory: address can be sent to each bank simultaneously, so miss penalty is  $4 + 24 + 4 \times 4$  (for sending words) = 44 cycles
- Independent memory banks: 32 cycle miss penalty, as long as the words are in different banks, since each has its own address lines and bus

CMSC 411 - Alan Sussman

122

## A confession: We've been lying

- The lie: User's don't really use physical addresses in their programs
- Instead, they use *virtual addresses*, where *virtual* means just what it does in virtual reality!
- This idea is over 40 years old, invented by the designers of the Atlas computer (Section 5.18)
  - Cache memory is just a little newer, discussed in print in 1965
- So when the user addresses word 450, the hardware/OS provides the illusion that the data is actually in this address, when really the data is somewhere else
- This *address translation* or *memory mapping* is invisible to the user

CMSC 411 - Alan Sussman

123

## Why virtual addressing

- Computers are designed so that multiple programs can be active at the same time
- At the time a program is compiled, the compiler has to assign addresses to each data item. But how can it know what memory addresses are being used by other programs?
- Instead, the compiler assigns virtual addresses, and expects the loader/OS to provide the means to map these into physical addresses

CMSC 411 - Alan Sussman

124

## In the olden days ...

- The loader would locate an unused set of main memory addresses and load the program and data there
- There would be a special register called the *relocation register*, and all addresses that the program used would be interpreted as addresses relative to the base address in that register
- So if the program jumped to location 54, the jump would really be to  $54 + \text{contents of relocation register}$ . A similar thing, perhaps with a second register, would happen for data references

CMSC 411 - Alan Sussman

125

## In the less-older days ...

- It became difficult to find a contiguous segment of memory big enough to hold program and data, so the program was divided into *pages*, with each page stored contiguously, but different pages in any available spot, either in main memory or on *disk*
- This is the *virtual addressing* scheme
  - to the program, memory looks like a contiguous segment, but actually, data is scattered in main memory and perhaps on disk

CMSC 411 - Alan Sussman

126

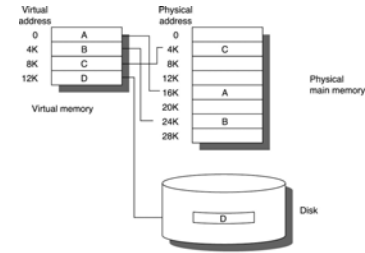
## But we know all about this!

- Already know that a program and data can be scattered between cache memory and main memory
- Now add the reality that its location in main memory is also determined in a scattered way, and some pages may also be located on disk
- So each page has its own relocation value

CMSC 411 - Alan Sussman

127

## Virtual Memory – Fig. 5.31



© 2003 Elsevier Science (USA). All rights reserved.

CMSC 411 - Alan Sussman

128

## Parameters – Fig. 5.32

| Parameter         | First-level cache                                     | Virtual memory                                          |
|-------------------|-------------------------------------------------------|---------------------------------------------------------|
| Block (page) size | 16-128 bytes                                          | 4096-65,536 bytes                                       |
| Hit time          | 1-3 clock cycles                                      | 50-150 cc                                               |
| Miss penalty      | 8-150 cc                                              | $10^6$ - $10^7$ cc                                      |
| (access time)     | (6-130 cc)                                            | (.8- $8 \times 10^6$ cc)                                |
| (transfer time)   | (2-20 cc)                                             | (.2- $2 \times 10^6$ cc)                                |
| Miss rate         | 0.1-10%                                               | 0.00001-0.001%                                          |
| Address mapping   | 25-45 bit physical address to 14-20 bit cache address | 32-64 bit virtual address to 25-45 bit physical address |

CMSC 411 - Alan Sussman

129

## Cache vs. virtual memory

| Cache                                     | Virtual memory                                                       |
|-------------------------------------------|----------------------------------------------------------------------|
| Cache miss handled by hardware            | Page faults handled by operating system                              |
| Cache size fixed for a particular machine | Virtual memory size fixed for a particular program                   |
| Fundamental unit is a block               | Fundamental unit is a fixed-length page or a variable-length segment |
| cache fault                               | page fault                                                           |

CMSC 411 - Alan Sussman

130

## Paging vs. segmentation – Fig. 5.34

|                         | Page                                 | Segment                                           |
|-------------------------|--------------------------------------|---------------------------------------------------|
| Words per address       | One                                  | Two (segment/offset)                              |
| Programmer visible?     | Invisible to app programmer          | May be visible to app programmer                  |
| Replacing a block       | Trivial (all blocks same size)       | Hard (must find contiguous, variable-sized chunk) |
| Memory use inefficiency | Internal fragmentation (within page) | External fragmentation (in unused memory)         |
| Efficient disk traffic  | Yes (can adjust page size)           | Not always (small segment problem)                |

CMSC 411 - Alan Sussman

131

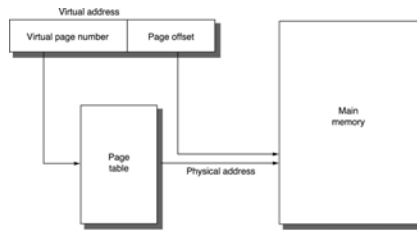
## Managing main memory

- Fully-associative mapping, because page faults are *really*, *really* expensive
- Page is located using a *page table*, one entry per page in the virtual address space
  - Size is sometimes reduced by hashing, to make one entry per physical page in main memory – an *inverted page table*
- Since locality says that a page will be used multiple times, address translation usually tests the addresses of the recently referenced pages before looking in other places
- So address translation information is held in the *translation look-aside buffer* (TLB)

CMSC 411 - Alan Sussman

132

## Page tables



© 2003 Elsevier Science (USA). All rights reserved.

CMSC 411 - Alan Sussman

133

## Managing main memory (cont.)

- Most machines replace the LRU page
  - moving pages between memory and disk is so slow that it's worth doing something close to real LRU
- Disks are so slow that machines use write-back, not write-through, and keep a *dirty bit* for each page

CMSC 411 - Alan Sussman

134

## Choosing page size

- A large page size
  - keeps page table small.
  - reduces cache miss times, if accesses have locality
  - reduces start-up overhead in moving data from disk to memory
  - means fewer TLB misses
- but also
  - wastes memory (internal fragmentation)
  - increases the time to start up a program

CMSC 411 - Alan Sussman

135

## Computer Systems Architecture CMSC 411 Unit 5 – Memory Hierarchy

Alan Sussman  
March 28, 2006

## Administrivia

- HW #4 due Friday, March 31, in my office, by 6PM
- Midterm next Tuesday, April 4
  - practice exam posted by tomorrow
  - given by my grad student
- Project questions?
  - output from mult.s in my linuxlab account

CMSC 411 - Alan Sussman

137

## Last time

- Improving memory bandwidth
  - wider main memory – transfer  $k$  words at a time
  - interleaved memory – separate banks, but transfer back 1 word at a time
  - independent memory banks – separate address lines and bus for each bank
    - try to avoid memory bank conflicts
- Virtual addressing
  - map VM pages to physical memory page frames (in memory or on disk), with a fully associative mapping
    - page table, also can use *inverted page table* to reduce size
  - cache address translation info in TLB
  - page faults handled by OS
  - LRU replacement, with write-back

CMSC 411 - Alan Sussman

138

## Memory protection

- Each program “lives” in its own virtual space, called its *process*
- When the CPU is working on one process, others may be partially completed or waiting for attention
- The CPU is *time shared* among the processes, working on each in turn
- And main memory is also shared among processes

CMSC 411 - Alan Sussman

139

## Protection modes

- User processes need to be protected from each other
- Two registers, *base* and *bound* test whether this virtual address belongs to this process
- If not, a *memory protection violation* exception is raised
- Users cannot change the base and bound registers

CMSC 411 - Alan Sussman

140

## Who can change them?

- The operating system needs access to the base and bound registers
- So a process that is labeled *kernel* (also called *supervisor* or *executive*) can access any memory location and change the registers
- Kernel processes are accessed through *system calls*, and a return to user mode is like a subroutine return, restoring the state of the user process

CMSC 411 - Alan Sussman

141

## Memory protection - bookkeeping

- Each process has
  - base and bound values, denoting its virtual address space
  - page tables, giving mapping to accessible pages in memory
  - mode bit (user or kernel)
- A user process cannot modify any of this information

CMSC 411 - Alan Sussman

142

## Examples

- The book uses examples to illustrate these concepts:
  - Alpha and Intel Pentium
- Read them if you are curious
  - and to see how ugly Pentium segments are

CMSC 411 - Alan Sussman

143

## Fallacies & Pitfalls

- *Biggest pitfall in computer design*: If the addresses are too short, the length of programs and the amount of data they can use will be too limited
- *Other pitfalls/fallacies*:
  - using benchmarks to predict memory performance
  - emphasizing DRAM bandwidth vs. latency
  - reducing memory efficiency by a poorly designed operating system
  - making the write buffer too small

CMSC 411 - Alan Sussman

144