Computer Organization

Douglas Comer

Computer Science Department
Purdue University
250 N. University Street
West Lafayette, IN 47907-2066

http://www.cs.purdue.edu/people/comer

X

Physical Memory And Physical Addressing

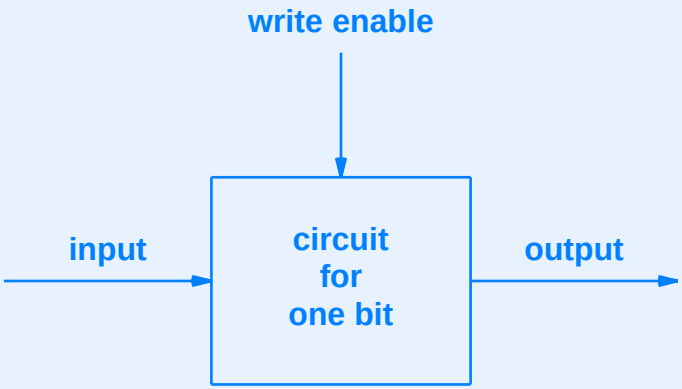
Computer Memory

- Main memory known as *Random Access Memory* (*RAM*)
- Usually volatile
- Two basic technologies available
  - Static RAM
  - Dynamic RAM

Static RAM (SRAM)

- Easiest to understand
- Similar to flip-flop

Illustration Of Static RAM



- When *enable* is high, output is same as input
- Otherwise, output holds last value

Advantages And Disadvantages Of SRAM

- Chief advantage: high speed
- Chief disadvantage: power consumption and heat

Dynamic RAM (DRAM)

- Alternative to SRAM
- Consumes less power
- Acts like a *capacitor*
  - Stores charge

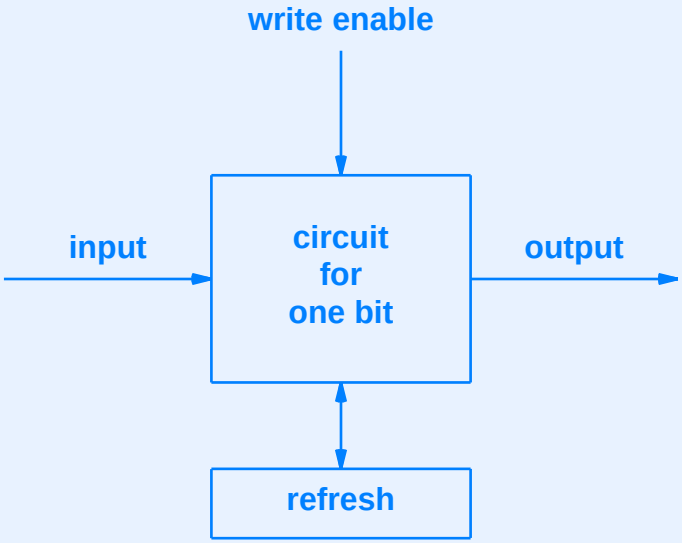
The Facts Of Electronic Life

- Capacitor gradually loses charge
- When left for a long time, logical 1 changes to logical 0
- Time to discharge can be under a second
- Although it is inexpensive, DRAM is an imperfect memory device!

Making DRAM Work

- Invent extra hardware that operates independently
- Repeatedly step through each location of DRAM
- Read value from location in DRAM
- Write value back into same location
- Extra hardware known as *refresh circuit*

Illustration Of Bit In DRAM



DRAM Refresh Circuit

- More complex than figure implies
- Must coordinate with normal *read* and *write* operations

Measures Of Memory Technology

- Density
- Latency and cycle time

Memory Density

- Refers to memory cells per square area of silicon
- Usually stated as number of bits on standard size chip
- Examples
  - *1 meg chip* holds one megabit of memory
  - *4 meg chip* holds four megabits of memory
- Note: higher density chip generates more heat

Separation Of Read And Write Performance

In many memory technologies, the time required to fetch information from memory differs from the time required to store information in memory, and the difference can be dramatic. Therefore, any measure of memory performance must give two values: the performance of read *operations and the performance of* write *operations.*

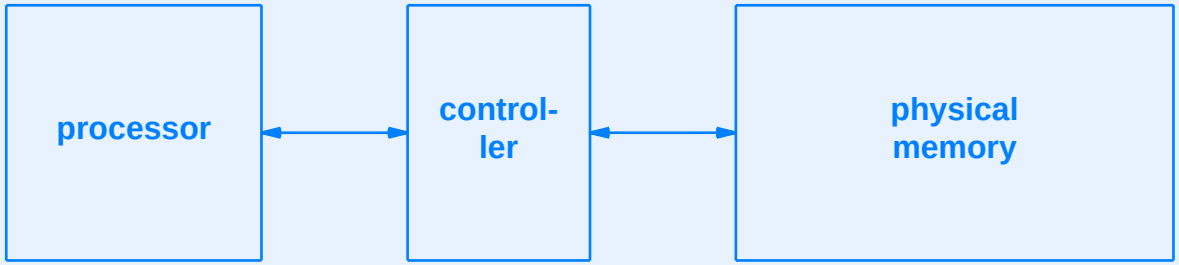
Latency

- Time that elapses between the start of an operation and the completion of the operation
- Not a constant

Memory Organization

- Hardware unit connects computer to physical memory chips
- Called a *memory controller*

Illustration Of Memory Organization



Honoring A Memory Request

- Computer
  - Presents request to controller
  - Waits for response
- Controller
  - Translates request into signals for physical memory chips
  - Returns answer to computer immediately
  - Sends signals to reset physical memory for next request

Consequence Of The Need To Reset Memory

Because a memory controller may need extra time between operations to reset the underlying physical memory, latency is an insufficient measure of performance; a performance measure needs to measure the time required for successive operations.

Memory Cycle Time

- Time that must elapse between two successive memory operations
- More accurate measure than latency
- Two separate measures
  - Read cycle time (tRC)
  - Write cycle time (tWC)

The Point About Cycle Times

The read cycle time and write cycle time are used as measures of memory system performance because they measure how quickly the memory system can handle successive requests.

Synchronized Memory Technologies

- Use same hardware clock as CPU
- Avoid unnecessary delays
- Can be used with SRAM or DRAM
- Terminology
  - *Synchronized Static Random Access Memory* (*SSRAM*)
  - *Synchronized Dynamic Random Access Memory* (*SDRAM*)
- Note: the RAM in many computers is now SDRAM

Multiple Data Rate Memory Technologies

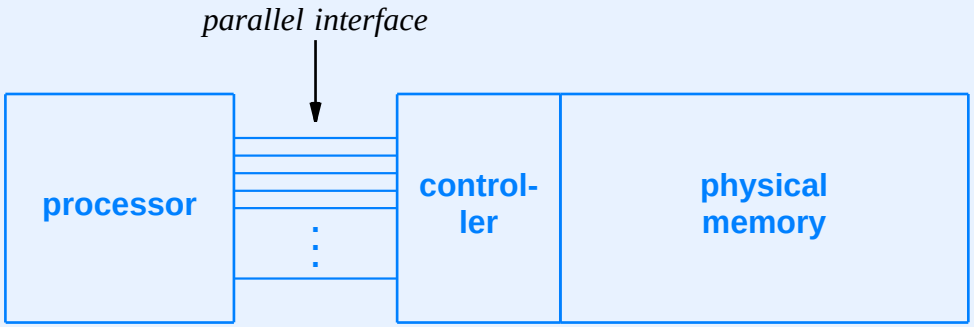
- Technique to improve memory performance
- Avoids a memory bottleneck
- Memory hardware runs at a multiple of CPU clock
- Examples
  - Double Data Rate SDRAM (DDR-SDRAM)
  - Quad Data Rate SRAM (QDR-SRAM)

Example Memory Technologies

Technology	Description
DDR-DRAM	Double Data Rate Dynamic RAM
DDR-SDRAM	Double Data Rate Synchronized Dynamic RAM
FCRAM	Fast Cycle RAM
FPM-DRAM	Fast Page Mode Dynamic RAM
QDR-DRAM	Quad Data Rate Dynamic RAM
QDR-SRAM	Quad Data Rate Static RAM
SDRAM	Synchronized Dynamic RAM
SSRAM	Synchronized Static RAM
ZBT-SRAM	Zero Bus Turnaround Static RAM
RDRAM	Rambus Dynamic RAM
RLDRAM	Reduced Latency Dynamic RAM

- Many others exist

Memory Organization



- Parallel interface used between computer and memory
- Called a *bus* (more later in the course)

Memory Transfer Size

- Amount of memory that can be transferred to computer simultaneously
- Determined by bus between computer and controller
- Example memory transfer sizes
  - 16 bits
  - 32 bits
  - 64 bits
- Important to programmers

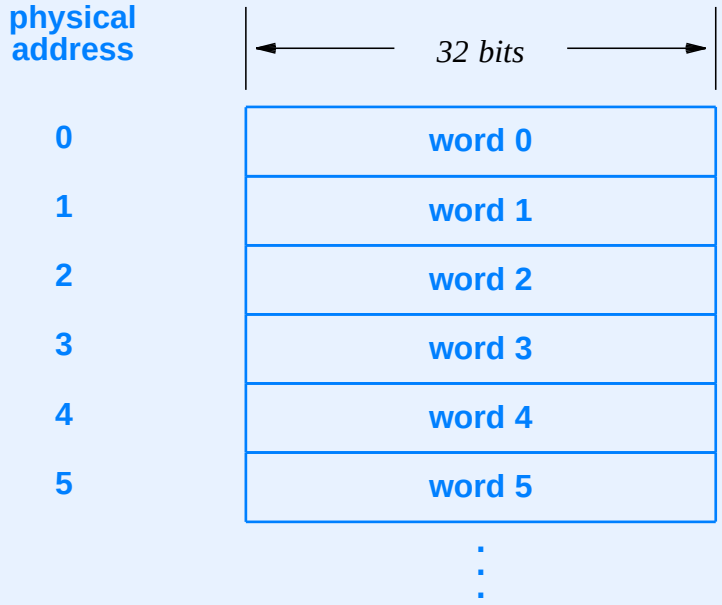
Physical Memory And Word Size

- Bits of physical memory are divided into blocks of *N* bits each
- Terminology
  - Group of *N* bits is called a *word*
  - *N* is known as the *width* of a word or the *word size*

Physical Memory Addresses

- Each word of memory is assigned a unique number known as a *physical memory address*
- Programmer imagines physical memory to be an array of words
- Note: entire word must be transferred

Illustration Of Physical Memory



- Illustration depicts a 32-bit word size

Summary of Physical Memory Organization

Physical memory is organized into words, where a word is equal to the memory transfer size. Each read *or* write *operation applies to an entire word.*

Choosing A Word Size

- Larger word size
  - Implemented with more parallel wires
  - Results in higher performance
  - Has higher cost
- Note: architect usually designs all parts of computer to use one size for:
  - Memory word
  - Integer (general-purpose registers)
  - Floating point number

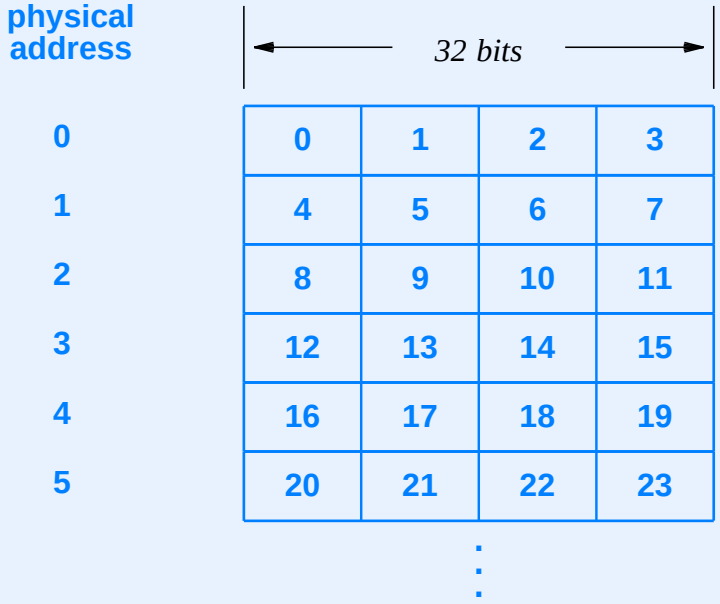
Byte Addressing

- View of memory presented to processor
- Each byte of memory assigned an address
- Convenient for programmers
- Underlying memory can still use word addressing

Translation Between Byte And Word Addresses

- Performed by *intelligent memory controller*
- CPU can use byte addresses (convenient)
- Physical memory can use word addresses (efficient)

Illustration Of Address Translation



Mathematics Of Translation

- Word address given by:

$$W = \left\lfloor \frac{B}{N} \right\rfloor$$

- Offest given by:

$$O = B \bmod N$$

- Example
  - $N = 4$
  - Byte address 11
  - Found in word 2 at offset 3

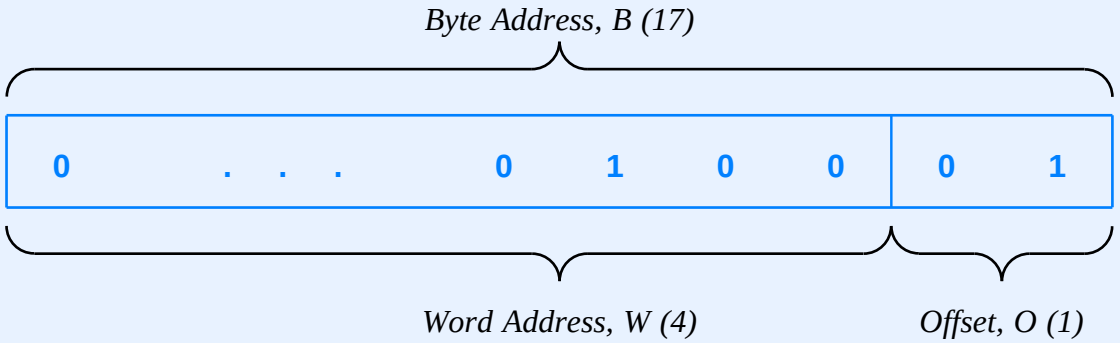
Efficient Translation

- Choose word size as power of 2
- Word address computed by extracting high-order bits
- Offset computed by extracting low-order bits

The Important Point

To avoid arithmetic calculations such as division or remainder, physical memory is organized such that the number of bytes per word is a power of two, which means the translation from a byte address to word address and offset can be performed by extracting bits.

Example Of Byte-To-Word Translation



Byte Alignment

- Refers to integer storage in memory
- In some architectures
  - Integer in memory must correspond to word in underlying physical memory
- In other architectures
  - Integer can be unaligned, but *fetch* and *store* operations are much slower

The Point For Programmers

The organization of physical memory affects programming: even if a processor allows unaligned memory access, aligning data on boundaries that correspond to the physical word size can improve program performance.

Memory Size And Address Space

- Size of address limits maximum memory
- Example: 32-bit address can represent

$$2^{32} = 4{,}294{,}967{,}296$$

  unique addresses
- Known as *address space*
- Note: word addressing allows larger memory than byte addressing

Programming With Word Addressing

- To obtain a single byte
  - Fetch word from memory
  - Extract byte from word
- To store a single byte
  - Fetch word from memory
  - Replace byte in word
  - Write entire word back to memory
- Performed by software

Measures Of Physical Memory Size

Physical memory is organized into a set of M words that each contain N bytes; to make controller hardware efficient, M and N are each chosen to be powers of two.

- Note
  - Memory sizes expressed as powers of two
  - Kilobyte defined to be 2^{10} bytes
  - Megabyte defined to be 2^{20} bytes

Consequence To Programmers

- Speed of computer network or other I/O device usually expressed in powers of ten
  - Example: *megabits per second* is 10^6 bits per second
- Programmer must accommodate differences between measures for storage and transmission

C Programming And Memory Addressability

- C has a heritage of both byte and word addressing
- Example of byte pointer declaration

*char *iptr;*

- Example of integer pointer declaration

*int *iptr;*

5.If integer size is four bytes, iptr++ increments by four

Memory Dump

- Used for debugging
- Printable representation of bytes in memory
- Each line of output specifies memory address and bytes starting at that address

Example Memory Dump

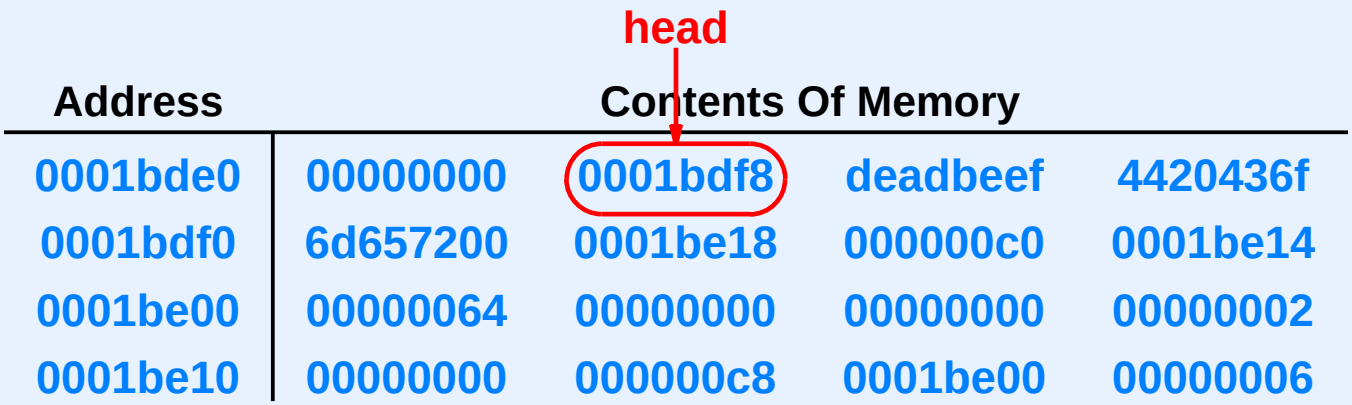
- Assume linked list in memory
- Head consists of pointer
- Each node has the following structure:

```
struct  node  {
        int count;
        struct node *next;
}
```

Example Memory Dump

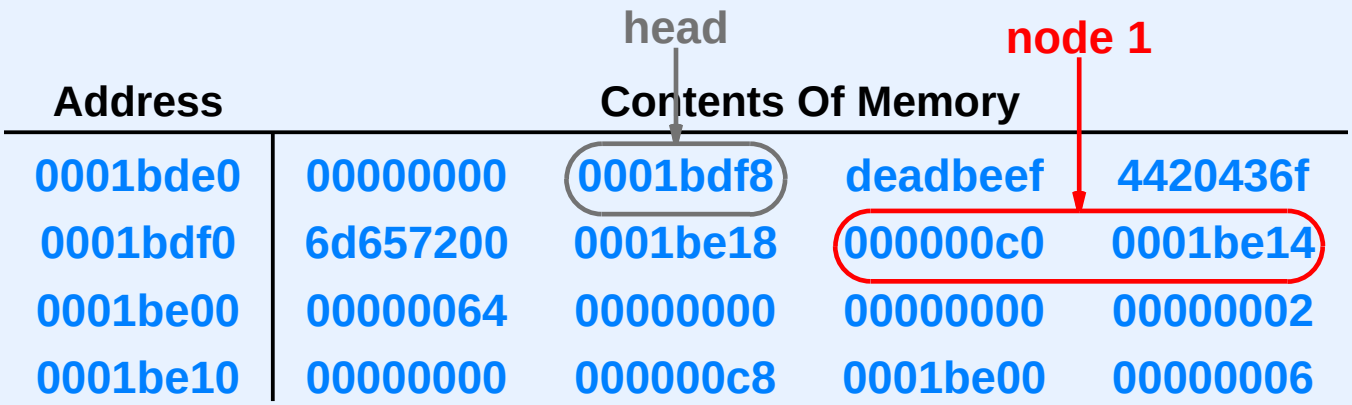
Address	Contents Of Memory			
0001bde0	00000000	0001bdf8	deadbeef	4420436f
0001bdf0	6d657200	0001be18	000000c0	0001be14
0001be00	00000064	00000000	00000000	00000002
0001be10	00000000	000000c8	0001be00	00000006

Example Memory Dump



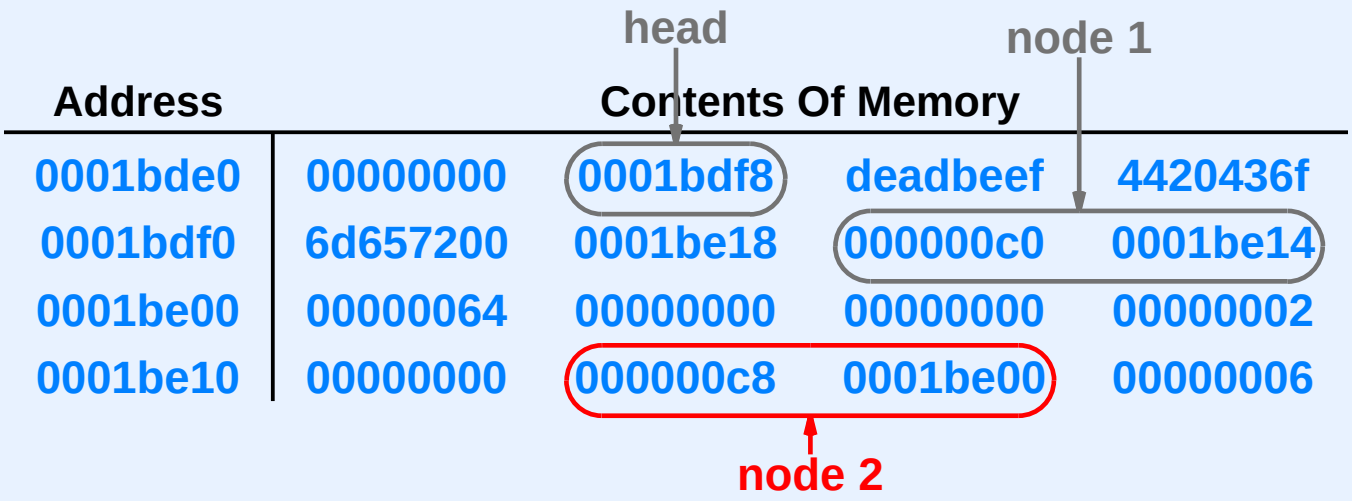
- Head found at address 0x0001bde4

Example Memory Dump



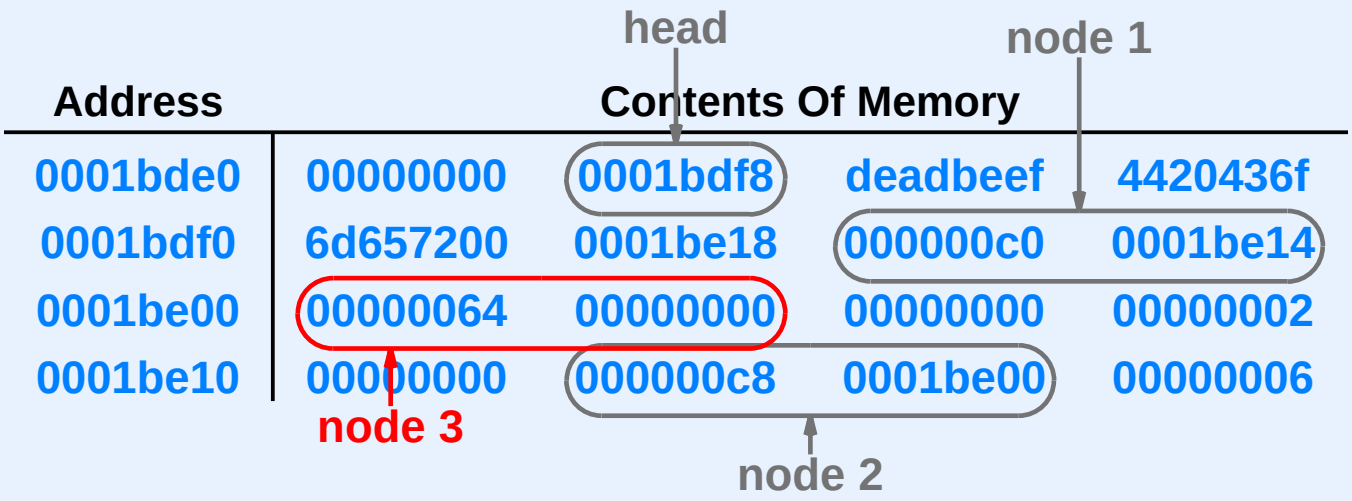
- Head found at address 0x0001bde4
- First node at 0x0001bdf8 contains 192 (0xc0)

Example Memory Dump



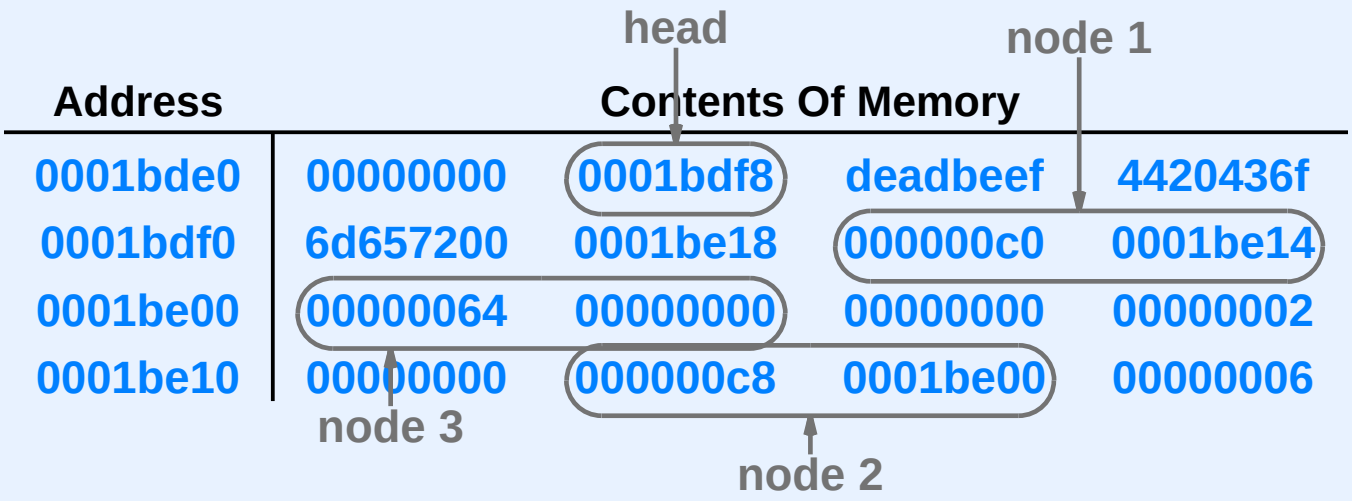
- Head found at address 0x0001bde4
- First node at 0x0001bdf8 contains 192 (0xc0)
- Second node at 0x0001be14 contains 200 (0xc8)

Example Memory Dump



- Head found at address 0x0001bde4
- First node at 0x0001bdf8 contains 192 (0xc0)
- Second node at 0x0001be14 contains 200 (0xc8)
- Last node at 0x001be00 contains 100 (0x64)

Example Memory Dump



- Head found at address 0x0001bde4
- First node at 0x0001bdf8 contains 192 (0xc0)
- Second node at 0x0001be14 contains 200 (0xc8)
- Last node at 0x001be00 contains 100 (0x64)

Example Memory Dump

Address	Contents Of Memory			
0001bde0	00000000	0001bdf8	deadbeef	4420436f
0001bdf0	6d657200	0001be18	000000c0	0001be14
0001be00	00000064	00000000	00000000	00000002
0001be10	00000000	000000c8	0001be00	00000006

- Head found at address 0x0001bde4
- First node at 0x0001bdf8 contains 192 (0xc0)
- Second node at 0x0001be14 contains 200 (0xc8)
- Last node at 0x001be00 contains 100 (0x64)

Memory Banks And Interleaving

- Two techniques used to increase memory performance
- Use parallel hardware

Memory Banks

- Alternative to single memory and single memory controller
- Processor connects to multiple controllers
- Each controller connects to separate physical memory
- Controllers and memories can all operate simultaneously

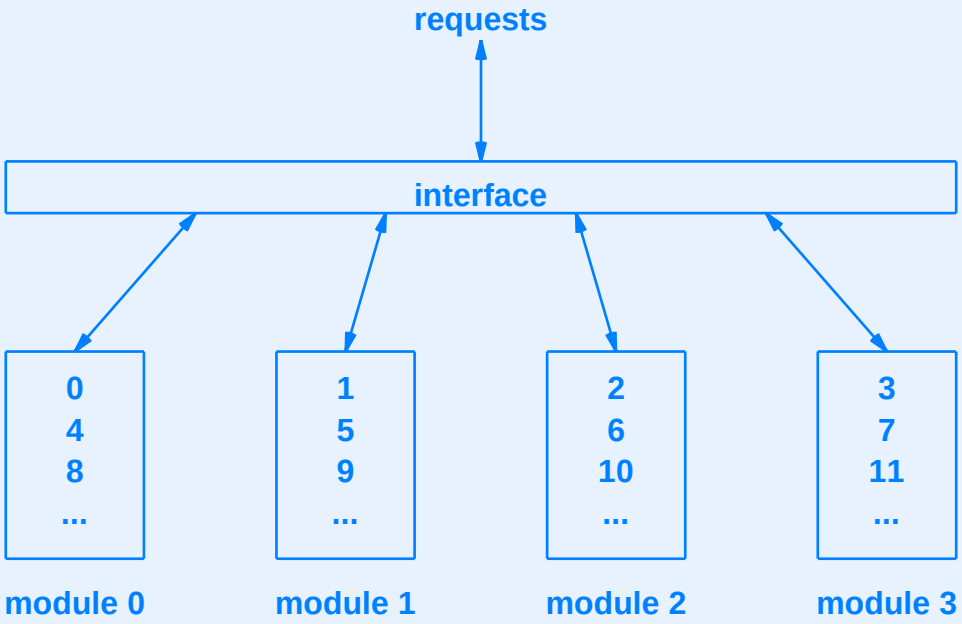
Programming With Memory Banks

- Two approaches
- Transparent
  - Programmer is not concerned with banks
  - Hardware automatically finds and exploits parallelism
- Opaque
  - Banks visible to programmer
  - To optimize performance, programmer must place items that will be accessed simultaneously in separate banks

Interleaving

- Related to memory banks
- Transparent to programmer
- Places consecutive bytes in separate physical memory
- Uses low-order bits of address to choose module
- Known as *N-way interleaving* (*N* is number of physical memories)

Illustration Of 4-Way Interleaving



- Consecutive bytes stored in separate physical memory

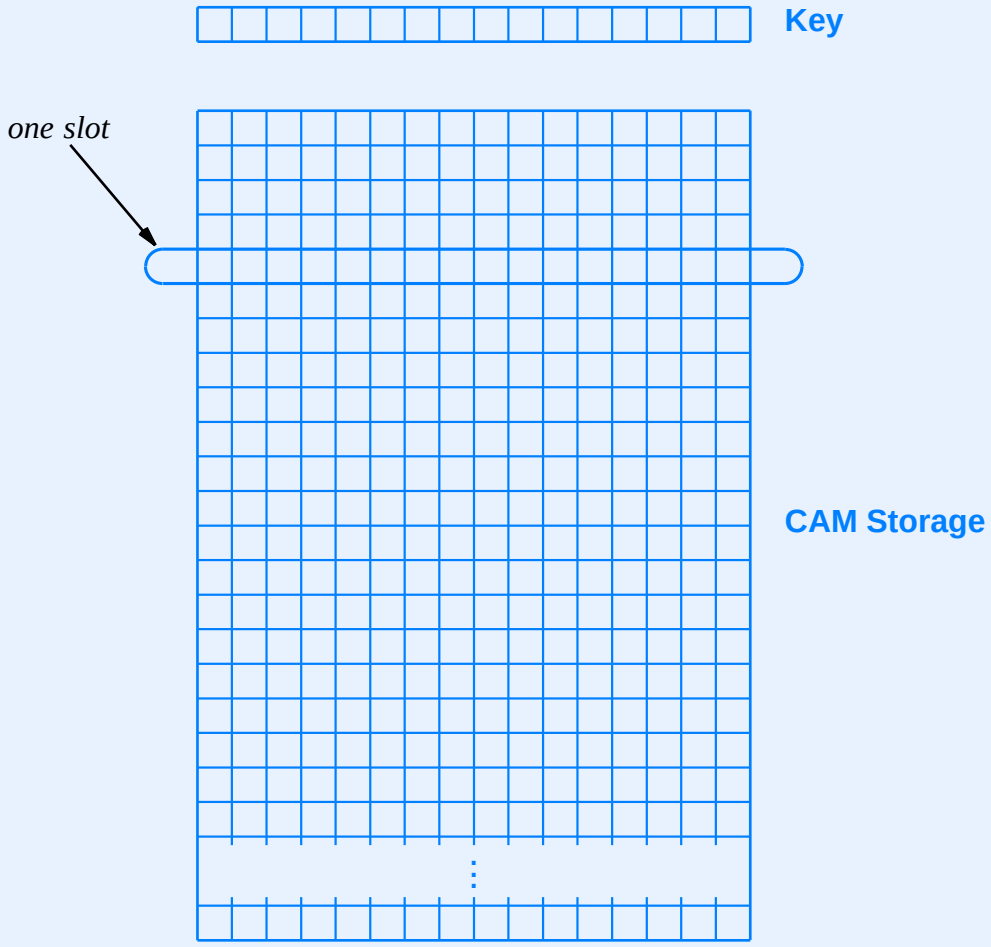
Content Addressable Memory (CAM)

- Blends two key ideas
  - Memory technology
  - Memory organization
- Includes parallel hardware for high-speed search

CAM

- Think of memory as a two-dimensional array
- Row in the array is called a *slot*
- Lookup hardware can answer the question: ‘‘is *X* stored in any row of the CAM?’’

Illustration Of CAM



Lookup In A CAM

- CAM presented with key for lookup
- Hardware searches slots to determine whether key is present
  - Search operation performed in parallel on all slots
  - Result is index of slot where value found
- Note: parallel search hardware makes CAM expensive

Ternary CAM (T-CAM)

- Variation of CAM
- Extends CAM to use *partial match searching*
- Each bit in slot can have one of three possible values:
  - Zero
  - One
  - Don’t care
- CAM either reports
  - First match
  - All matches (bit vector)

Summary

- Physical memory
  - Organized into fixed-size words
  - Accessed through a controller
- Controller can use
  - Byte addressing when communicating with a processor
  - Word addressing when communicating with a physical memory
- To avoid arithmetic, use powers of two for
  - Address space size
  - Bytes per word

Summary
(continued)

- Many memory technologies exist
- A memory dump that shows contents of memory in a printable form can be an invaluable tool
- Two techniques used to optimize memory access
  - Separate memory banks
  - Interleaving
- Content Addressable Memory (CAM) permits parallel search; variation of CAM known as Ternary Content Addressable Memory (T-CAM) allows partial match retrieval


Questions?