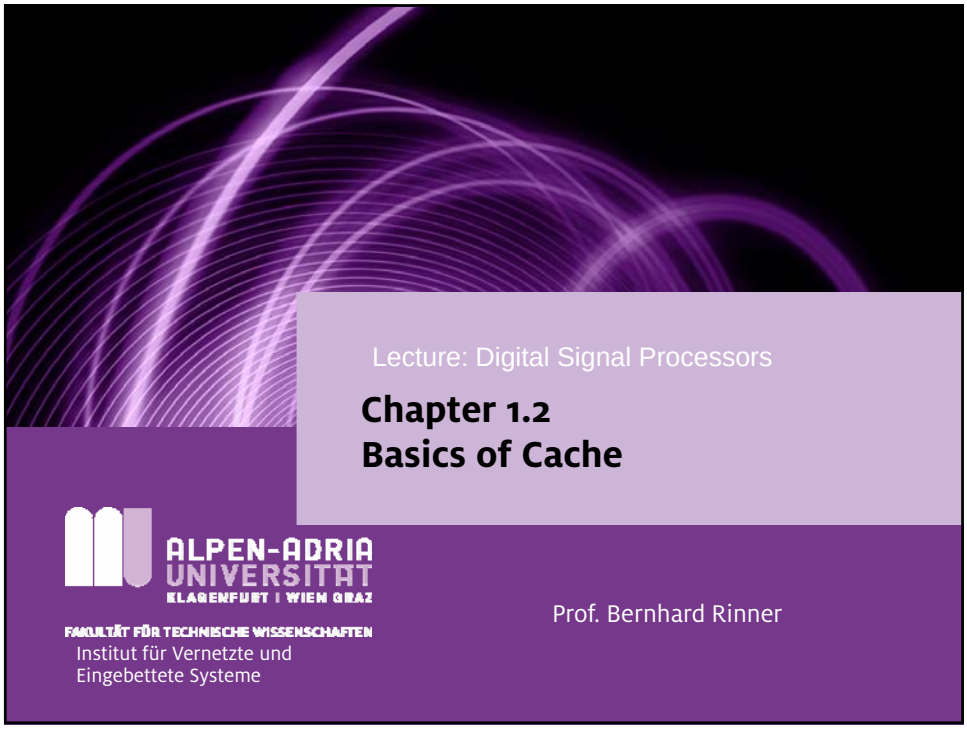


Lecture: Digital Signal Processors

Chapter 1.2
Basics of Cache

ALPEN-ADRIA UNIVERSITÄT
KLAGENFURT | WIEN GRAZ
FAKULTÄT FÜR TECHNISCHE WISSENSCHAFTEN
Institut für Vernetzte und Eingebettete Systeme

Prof. Bernhard Rinner

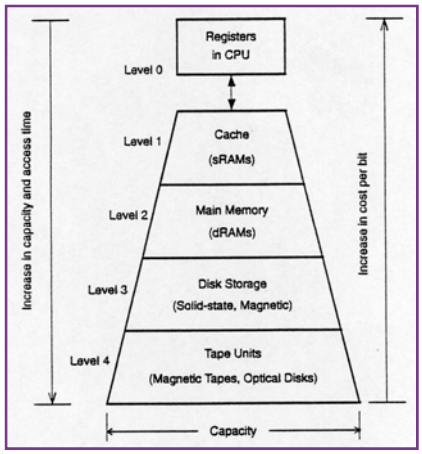
Agenda

PART 2

1. Memory System
2. Cache Memory
   - Cache Architecture
   - Cache Organization
3. Direct Memory Transfer

Memory System

- Memory hierarchy



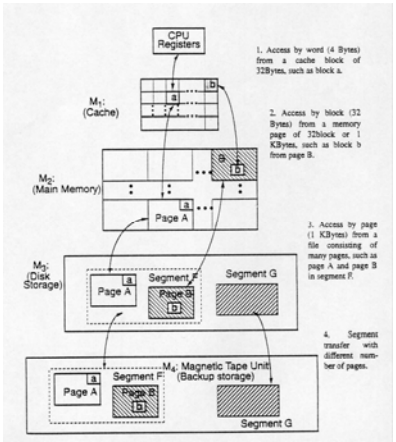
- Parameters
  - Access time
  - Memory size
  - Cost/Byte
  - Transmission bandwidth
  - Transmission unit

B.Rinner: Digital Signal Processors (Chapter 1.2)

3

Memory Management (2)

- Inclusion



M1 ⊂ M2 ⊂ … ⊂ Mn

Whole information is originally stored in memory Mn

B.Rinner: Digital Signal Processors (Chapter 1.2)

4

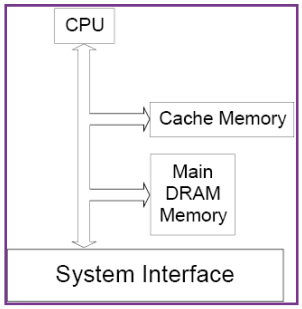
Memory Management (3)

- Coherence
  - Copies of same data must remain consistent at higher layers
  - Example: Modifications of cached data
    ➔update modified data on higher layers
- Cache Strategies
  - Write-through (WT): immediate update
  - Write-back (WB): delayed update

Cache Memory

Cache

- Cache is a small high speed memory
- Cache memory helps to reduce the time for accessing data (between processor and main memory).
- Reducing the access time based on "locality of reference".
  - Typically, the processor accesses memory in a small or localized region.
  - Keep the localized memory region in cache to reduce access time



B.Rinner: Digital Signal Processors (Chapter 1.2)

7

Cache Cycles

- CPU performs a read or write, the cache may intercept the bus transaction for decrease the response time
- Cache Hits : When ever the cache contains the information requested
- Cache Miss : When ever the cache does not contain the information requested
- Cache Consistency : Cache always reflects what is in main memory
  - Snoop : A cache is watching the address lines for transaction
  - Snarf : When a cache takes the information from the data lines
  - Dirty Data : When data is modified within cache but not modified in main memory
  - Stale Data : When data is modified within main memory but not modified in cache

B.Rinner: Digital Signal Processors (Chapter 1.2)

8

Cache Architecture

- Read architectures: “Look Aside” or “Look Through”
- Write policies: “Write-Back” or “Write-Through”
- Look Aside: The discerning feature of this cache unit is that it sits in parallel with main memory.
  - Both main memory and cache see a bus cycle at the same time.
  - Look aside caches are less complex and less expensive.
  - This architecture provides better response to a cache miss since both the DRAM and the cache see the bus cycle at the same time.
  - The draw back is the processor cannot access cache while another bus master is accessing main memory.

Cache Architecture (continued)

- Look Through: sits between processor and main memory
  - The cache sees the processors bus cycle before allowing it to pass on to the system bus.
  - This architecture allows the processor to read out of cache while another bus master is accessing main memory
  - This cache architecture is more complex and more costly.
  - Another down side is that memory accesses on cache misses are slower because main memory is not accessed until after the cache is checked.

Cache Architecture (continued)

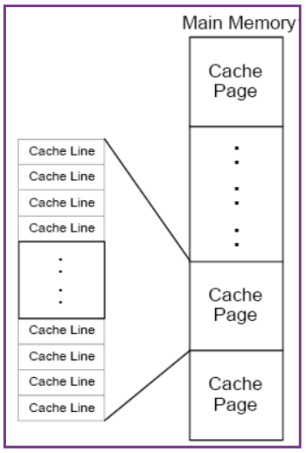
- Write policy: What to do when data is modified?
- Write-back policy: Cache acts like a buffer.
  - Write cycle updates the data in the cache
  - Update of main memory is performed later (cp. “dirty data”)
  - Might reduce main memory updates, but increases complexity.
- Write-through policy: The processor writes through the cache to main memory.
  - Write cycle updates both cache and main memory
  - Access to main memory at each write (no “dirty data”).
  - Simpler but less effective.

Cache Components

- Cache Sub-Systems: SRAM, Tag RAM, and Cache Controller.
  - In actual designs, these blocks may be implemented by multiple chips or all may be combined into a single chip.
- SRAM: Static Random Access Memory (SRAM) is the memory block which holds the data. The size of the SRAM determines the size of the cache.
- Tag RAM: Tag RAM (TRAM) is a small piece of memory that stores the addresses of the data that is stored in the SRAM.
- Cache Controller: The cache controller implements the cache policies.
  - Updates the SRAM and TRAM and implementing the write policy
  - Performs the “snoops” and “snarfs”

Cache Organization

- Caches organized into two terms cache page and cache line.
- Cache page: Main memory is divided into equal pieces
  - The size of a page is dependent on the size of the cache
  - Cache page is broken into smaller pieces, called a cache line.
  - Cache line is determined by processor and cache design.
- Determination of cache size

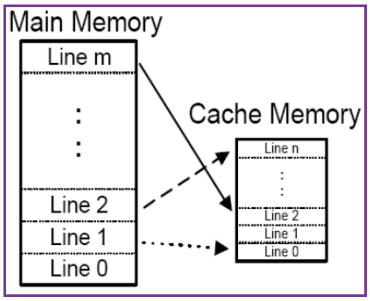


B.Rinner: Digital Signal Processors (Chapter 1.2)

13

Fully-Associative Cache

- Any line in main memory is allowed to store in any location in the cache
  - The disadvantage is the complexity of implementing.
  - TRAM access time is critical for overall performance

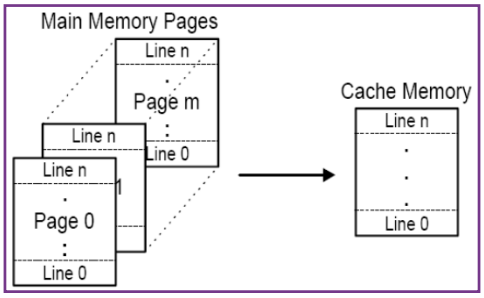


B.Rinner: Digital Signal Processors (Chapter 1.2)

14

Direct-Mapped Cache

- Main memory is divided into cache pages. The size of each page is equal to the size of the cache.
  - Unique cache line for line i of all pages
  - It is the least complex of all three caching schemes

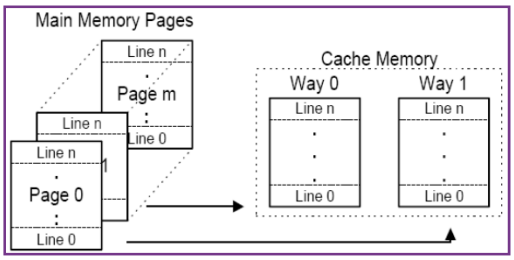


B.Rinner: Digital Signal Processors (Chapter 1.2)

15

Set-Associative Cache

- Cache SRAM is divided into equal sections called cache ways.
  - The cache page size is equal to the size of the cache way.
  - Each cache way is treated like a small direct mapped cache.
  - In this scheme, two lines of memory may be stored at any time.
  - This helps to reduce trashing (loading/replacing the same line)



B.Rinner: Digital Signal Processors (Chapter 1.2)

16

Direct Memory Access

Direct Memory Access (DMA)

- Data transfer mechanisms
  - Polling
    - Processor is dedicated to acquire incoming data, often by waiting in a loop
    - Not an efficient method of data acquisition because processor cannot do other task until data transfer is completed
  - Interrupts
    - Processor is interrupt to acquire incoming data
  - Direct Memory Access
    - Dedicated device for data acquisition, that reads incoming data and stores the data in system memory for later retrieval by the processor
    - Processor can work on other tasks in parallel
    - Improved efficiency (reduced CPU overhead)!

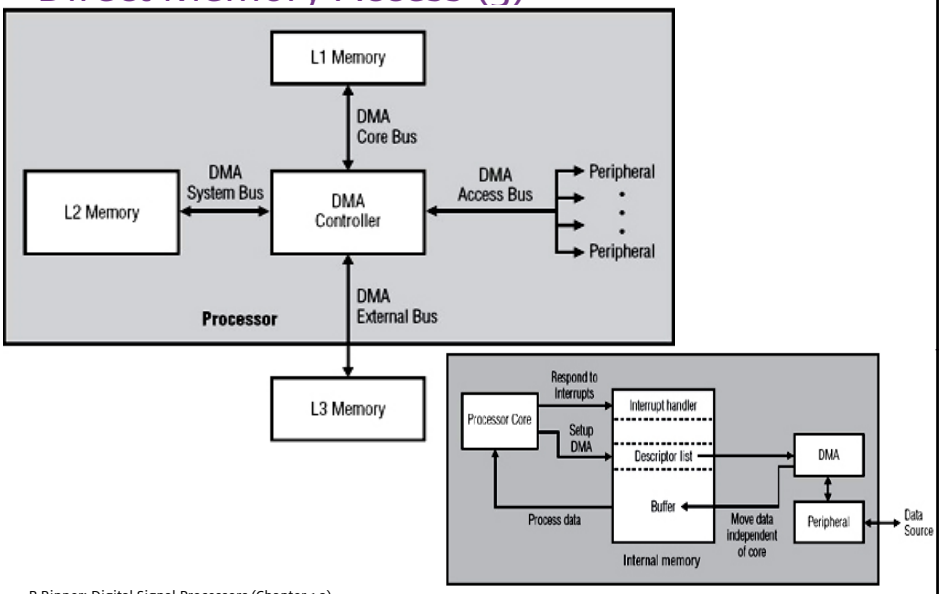
Direct Memory Access (2)

- External peripherals communicate with DMA Controller for data transfer by asserting hardware signal called DMA request signal
- DMA controller manages several channels for data transfer and DMA request signal is issued in specific to one channel
- DMA request signal is monitored and responded in a similar way the processor responds to interrupts
- Channels must be enabled by the processor for DMA controller to respond to DMA requests i.e. processor must initiate the communication

B.Rinner: Digital Signal Processors (Chapter 1.2)

19

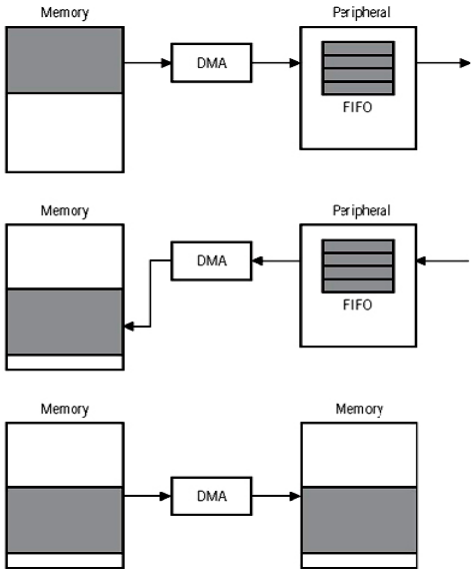
Direct Memory Access (3)



B.Rinner: Digital Signal Processors (Chapter 1.2)

20

DMA Data Flow



B.Rinner: Digital Signal Processors (Chapter 1.2)

21

DMA Transfer Types

- Fly-by DMA Transfer
  - Fastest data transfer also referred as single cycle or single address transfer (single memory cycle per transfer)
  - DMA controller supplies the address and device reads or writes the data
  - Memory-to-memory transfers are not possible!
- Fetch-and-deposit DMA Transfer
  - Also referred as dual-cycle, dual address or flow through transfer
  - Data is read from the device into the temporary register of DMA during the first cycle and written to memory or other device in the next cycle
  - Though inefficient, generally used for interface devices with different bus widths
  - DMA controller of this type has two address registers per channel

B.Rinner: Digital Signal Processors (Chapter 1.2)

22

DMA Controller Operation

- DMA channel is enabled or disabled via DMA Mask Register
- Count register determines the number of pending transfers and is decremented after a transfer. Terminal count signal (value goes from o to -1) signifies completion of DMA transfer sequence
- Value from the address register is driven onto address bus and is automatically incremented or decremented

DMA Applications

- Network cards
- Intra-chip data transfer in multi-core processors
- Graphic cards
- Disk drive controllers
- Sound cards
- .....