2009-06-11

計算機アーキテクチャ 第一 (E)

6. メモリ2:
半導体メモリシステム, ファイルメモリシステム

吉瀬 謙二 計算工学専攻
kise_at_cs.titech.ac.jp
W641講義室 木曜日13:20 ー 14:50

Acknowledgement

- Lecture slides for Computer Organization and Design, Third Edition, courtesy of **Professor Mary Jane Irwin**, Penn State University
- Lecture slides for Computer Organization and Design, third edition, Chapters 1-9, courtesy of **Professor Tod Amon,** Southern Utah University.

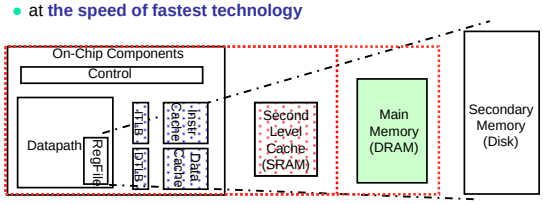
Adapted from *Computer Organization and Design*, Patterson & Hennessy, © 2005

The Memory Hierarchy **Goal**

- **Fact**:
  Large memories are slow and
  fast memories are small
- How do we create a memory that gives the illusion of being large, cheap and fast ?
  - With **hierarchy**（階層）
  - With **parallelism**（並列性）

A Typical Memory Hierarchy

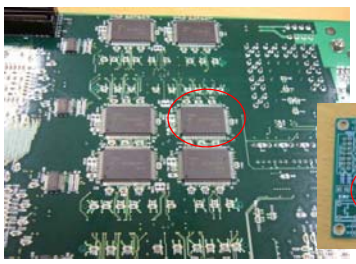
- By taking advantage of **the principle of locality**（局所性）
  - Present **much memory** in **the cheapest technology**
  - at **the speed of fastest technology**



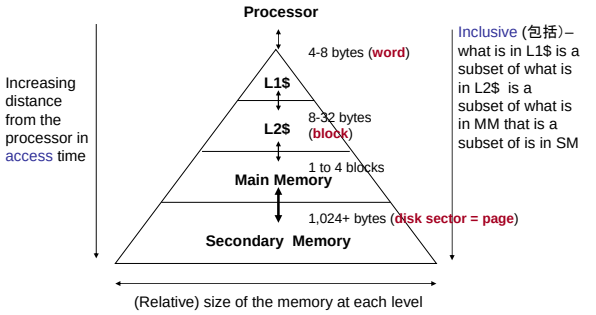
Speed (%cycles):	½'s	1's	10's	100's	1,000's
Size (bytes):	100's	K's	10K's	M's	G's to T's
Cost:	highest				lowest

DRAM (dynamic random access memory)

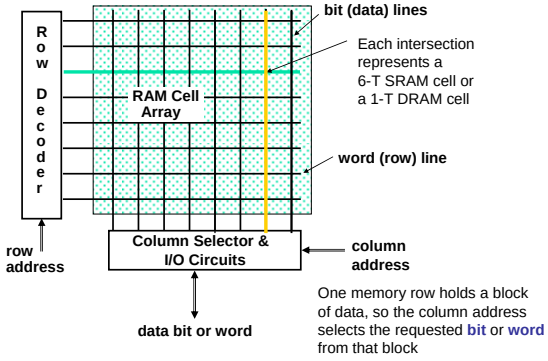
SRAM (static random access memory)



Characteristics of the Memory Hierarchy

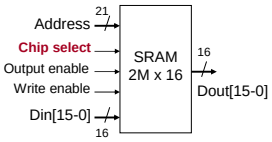


Classical RAM Organization (~Square)



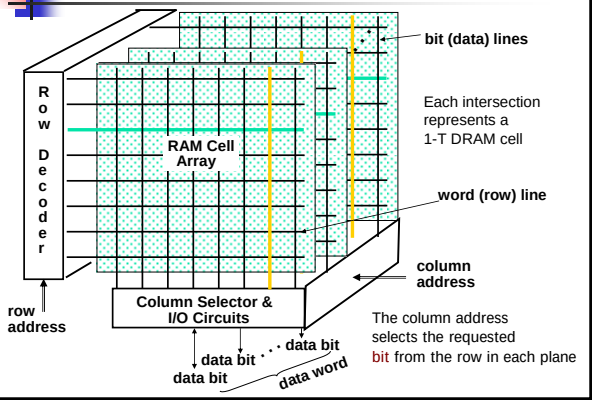
Memory Hierarchy Technologies

- Caches use SRAM for speed and technology compatibility
  - **Low density** (**6 transistor cells**), high power, expensive, fast
  - **Static**: content will last "forever" (until power turned off)



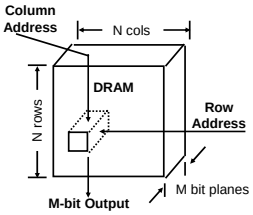
- Main Memory uses DRAM for size (density)
  - **High density** (**1 transistor cells**), low power, cheap, slow
  - **Dynamic**: needs to be "refreshed" regularly (~ every 8 ms)
    - 1% to 2% of the active cycles of the DRAM
  - Addresses divided into 2 halves (row and column)
    - **RAS** or Row Access Strobe triggering row decoder
    - **CAS** or Column Access Strobe triggering column selector

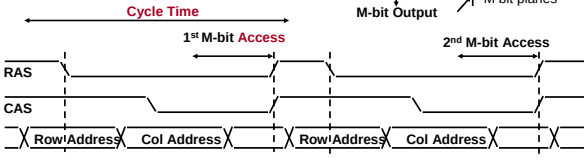
Classical DRAM Organization (~Square Planes)



Classical DRAM Operation

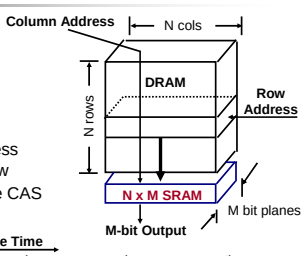
- DRAM Organization:
  - N rows x N column x M-bit
  - Read or Write M-bit at a time
  - Each M-bit access requires a RAS / CAS cycle

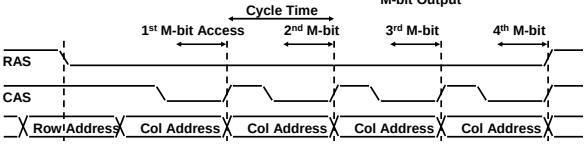




Page Mode DRAM Operation

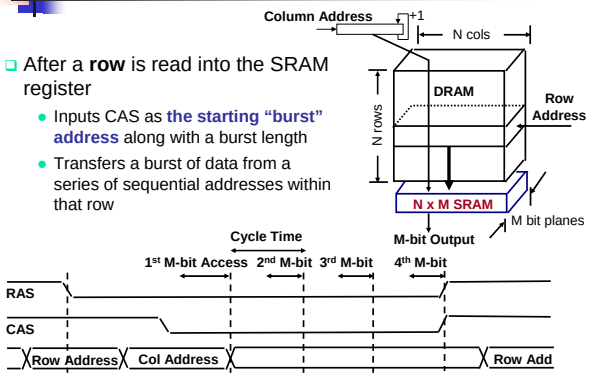
- Page Mode DRAM
  - N x M SRAM to save a row
- After a row is read into the SRAM "register"
  - Only CAS is needed to access other M-bit words on that row
  - RAS remains asserted while CAS is toggled





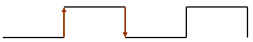
Synchronous DRAM (SDRAM) Operation

- After a **row** is read into the SRAM register
  - Inputs CAS as **the starting "burst" address** along with a burst length
  - Transfers a burst of data from a series of sequential addresses within that row



Other DRAM Architectures

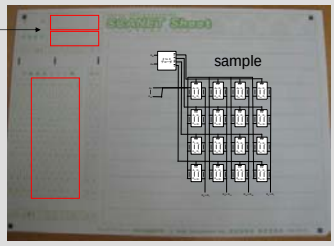
- Double Data Rate SDRAMs – **DDR-SDRAMs** (and DDR-SRAMs)
  - Double data rate because they transfer data on both the rising and falling edge of the clock
  - Are the most widely used form of SDRAMs
- **DDR2-SDRAMs**
- **DDR3-SDRAMs**



演習

- 512K x 8ビット（512KB）のSRAMを用いて、32ビットデータ幅の4MB のメモリを実現したい.
- 8個のメモリチップ，チップ選択信号CS， データ信号，アドレス信号の接続を示せ.

氏名，学籍番号，
学籍番号マーク欄(**右詰で**)



DRAM Memory Latency & Bandwidth Milestones

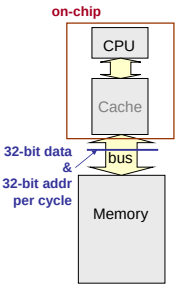
	DRAM	Page DRAM	FastPage DRAM	FastPage DRAM	Synch DRAM	DDR SDRAM
Module Width	16b	16b	32b	64b	64b	64b
Year	1980	1983	1986	1993	1997	2000
Mb/chip	0.06	0.25	1	16	64	256
Die size (mm²)	35	45	70	130	170	204
Pins/chip	16	16	18	20	54	66
BWidth (MB/s)	**13**	**40**	**160**	**267**	**640**	**1600**
Latency (nsec)	**225**	**170**	**125**	**75**	**62**	**52**

Patterson, CACM Vol 47, #10, 2004

- In the time that the memory to processor bandwidth doubles the memory latency improves by a factor of only 1.2 to 1.4
- To deliver such high bandwidth, the internal DRAM has to be organized as **interleaved memory banks**

Memory Systems that Support Caches

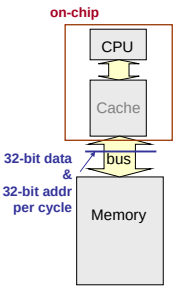
- The off-chip interconnect and memory architecture can affect overall system performance **in dramatic ways**



One word wide organization (one word wide bus and one word wide memory)

- **Assume（前提）**
  1. **1** clock cycle to send the address
  2. **25** clock cycles for DRAM **cycle time**, **8** clock cycles **access time**
  3. **1** clock cycle to return a word of data
- Memory-Bus to Cache **bandwidth**
  - **number of bytes transferred from memory to cache per clock cycle**

One Word Wide Memory Organization

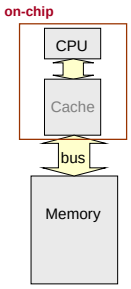


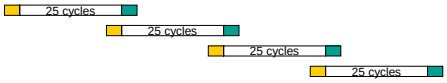
- The pipeline stalls the number of cycles for **one word** (32bit) from memory
  - **1** cycle to send address
  - **25** cycles to read DRAM
  - **1** cycle to return data
  - **27** **total clock cycles** miss penalty

- Number of bytes transferred per clock cycle (**bandwidth**) for a single miss
  - **4 / 27 = 0.148** bytes per clock

One Word Wide Memory Organization, con't

- What if the block size is **four words**?
  - **1** cycle to send 1st address
  - **4 * 25 = 100** cycles to read DRAM
  - **1** cycle to return last data word
  - **102 total clock cycles** miss penalty

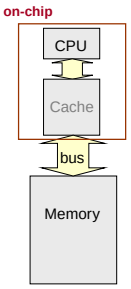




- Number of bytes transferred per clock cycle (**bandwidth**) for a single miss
  - **(4 x 4) / 102 = 0.157** bytes per clock

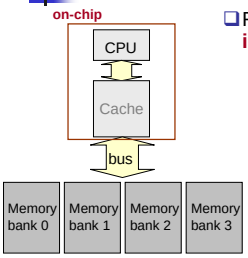
One Word Wide Memory Organization, con't

- What if the block size is **four words** and if a **fast page mode DRAM** is used?
  - **1** cycle to send 1st address
  - **25 + (3 * 8) = 49** cycles to read DRAM
  - **1** cycle to return last data word
  - **51 total clock cycles** miss penalty

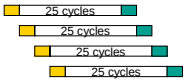


- Number of bytes transferred per clock cycle (**bandwidth**) for a single miss
  - **(4 x 4) / 51 = 0.314** bytes per clock

Interleaved(インターリーブ) Memory Organization



- For a block size of **four words with interleaved memory (4 banks)**
  - **1** cycle to send 1st address
  - **25 + 3 = 28** cycles to read DRAM
  - **1** cycle to return last data word
  - **30 total clock cycles** miss penalty



- Number of bytes transferred per clock cycle (**bandwidth**) for a single miss
  - **(4 x 4) / 30 = 0.533** bytes per clock

アナウンス

- 講義スライドおよびスケジュール
  - www.arch.cs.titech.ac.jp
  - 講義日程が変更になることがあるので頻繁に確認すること.

22 / 42