

2018年度(平成30年度)版

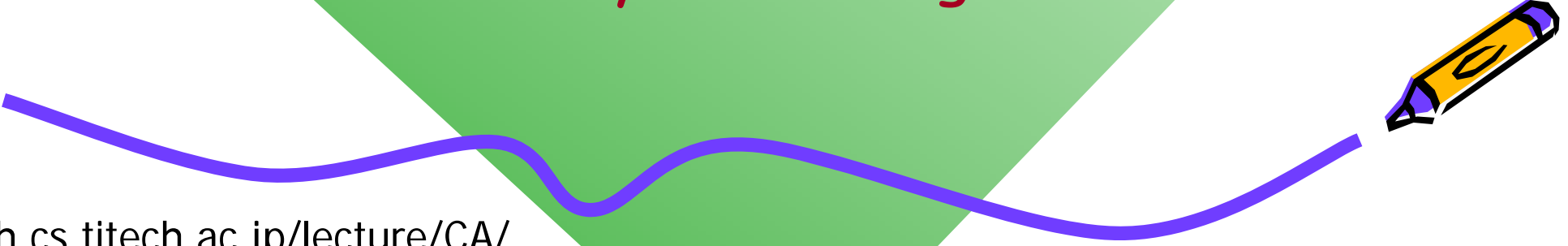
Ver. 2018-10-06a

Course number: CSC.T363



# コンピュータアーキテクチャ Computer Architecture

## 3. 半導体メモリ Memory Technologies



[www.arch.cs.titech.ac.jp/lecture/CA/](http://www.arch.cs.titech.ac.jp/lecture/CA/)  
Room No.W321  
Tue 13:20-16:20, Fri 13:20-14:50

吉瀬 謙二 情報工学系  
Kenji Kise, Department of Computer Science  
[kise\\_at\\_c.titech.ac.jp](mailto:kise_at_c.titech.ac.jp)

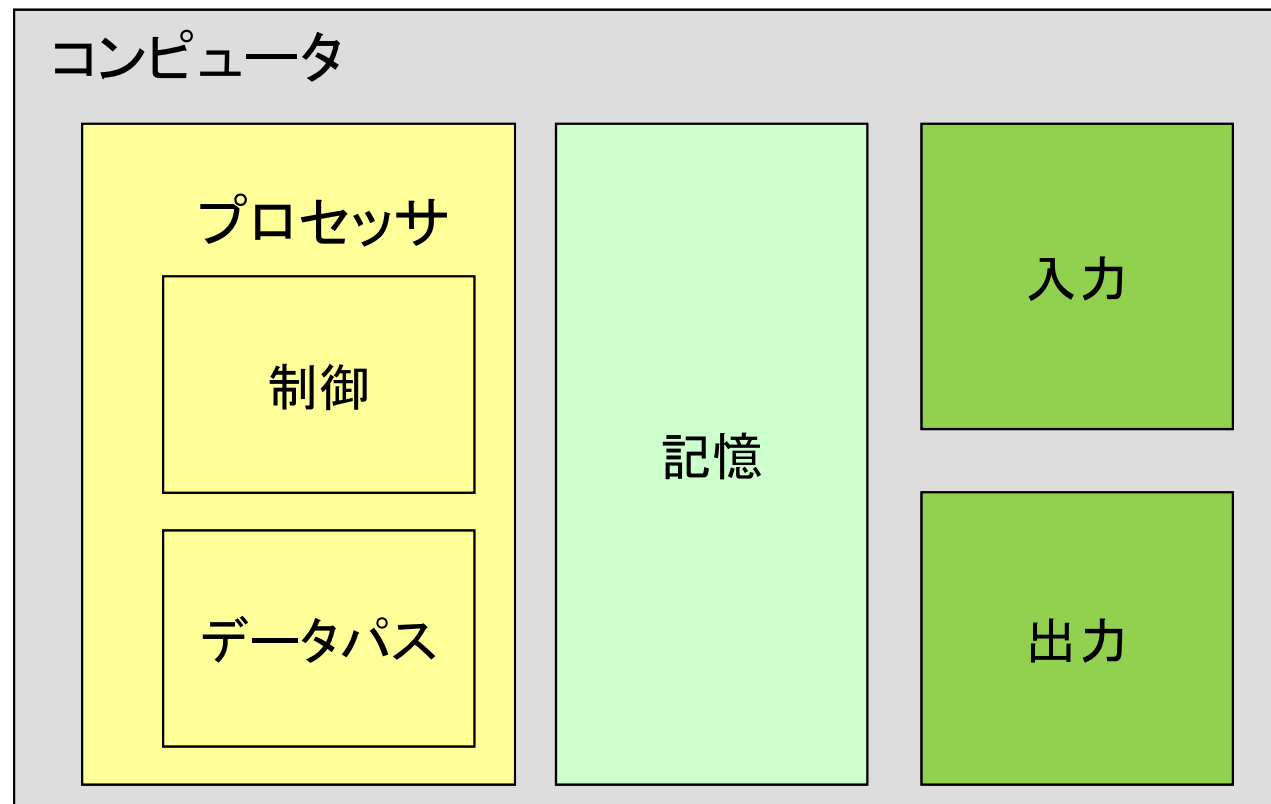
# コンピュータの古典的な要素

コンパイラ

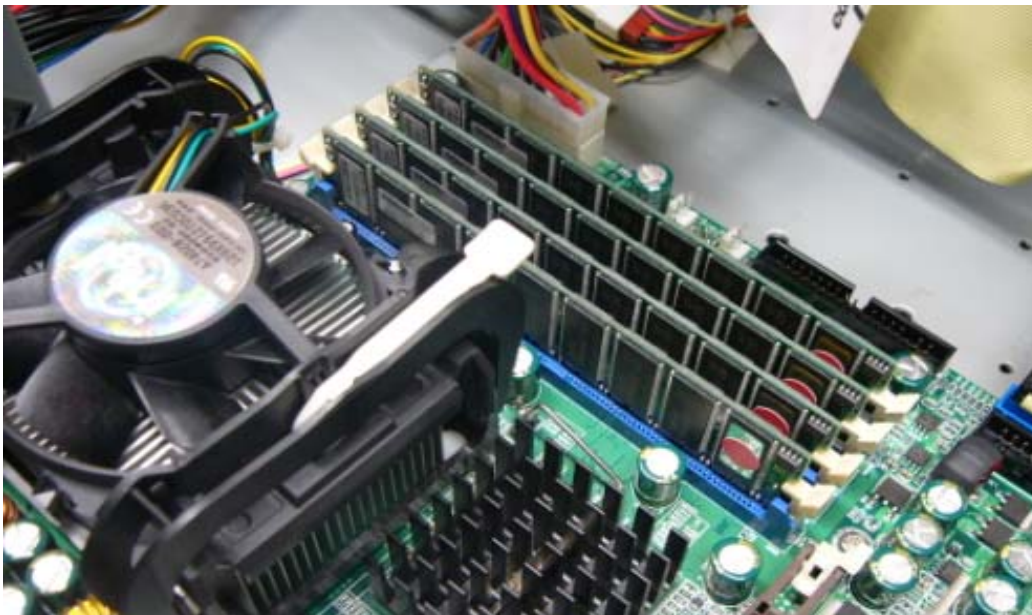
Instruction Set Architecture (ISA), 命令セットアーキテクチャ

インタフェース

性能の評価

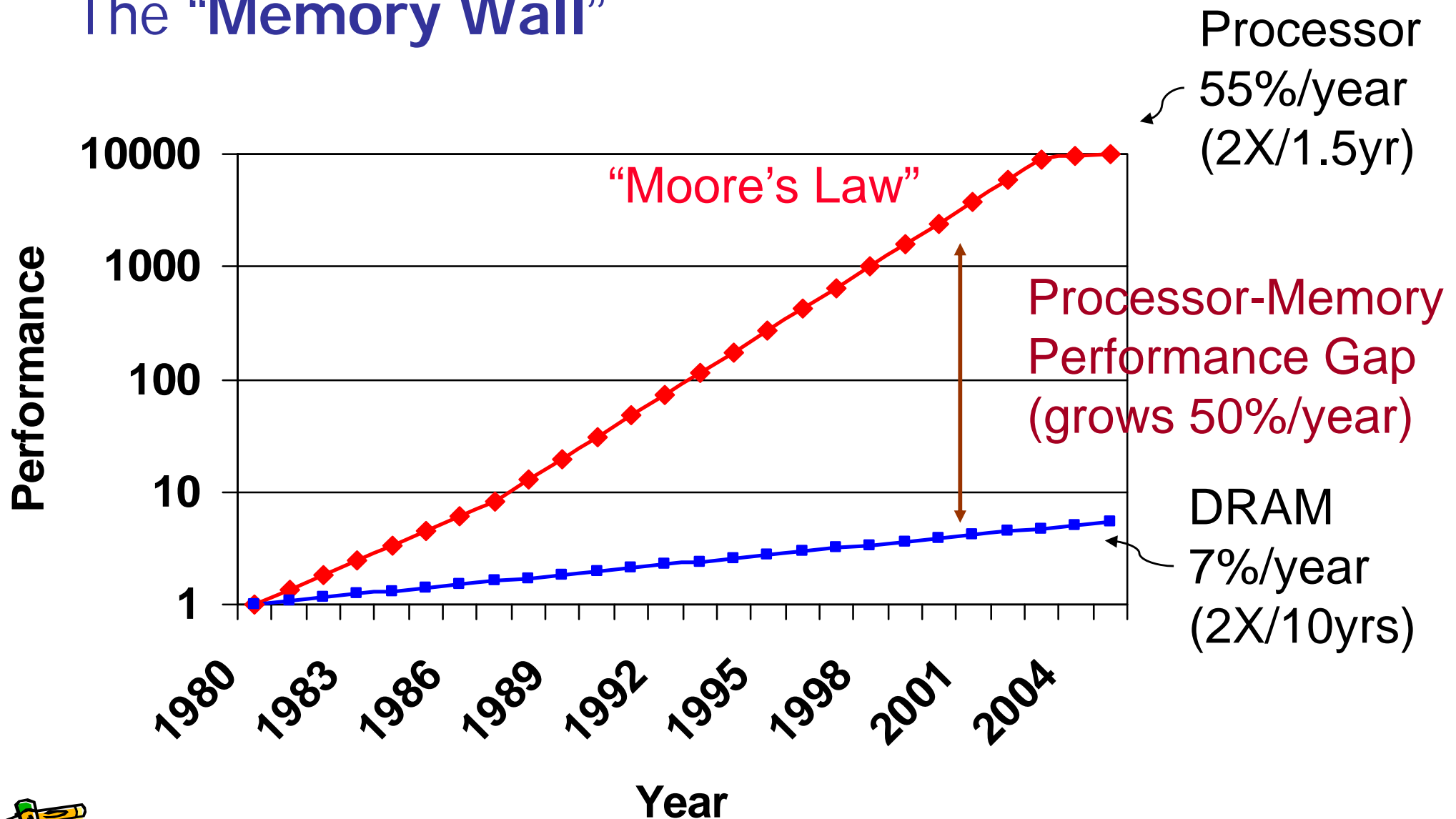


# DRAM (dynamic random access memory)



# Processor-Memory(DRAM) Performance Gap

## The “Memory Wall”



# The Memory System's Fact and 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 ?

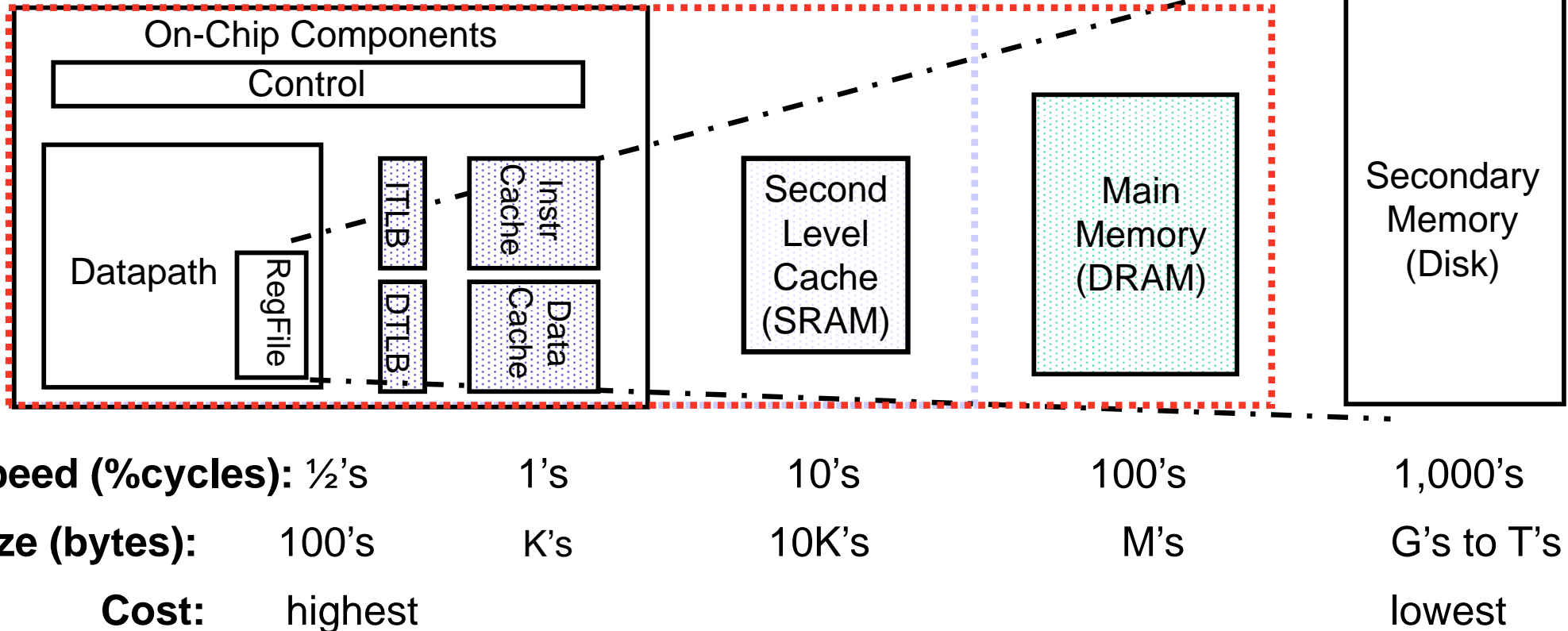
With **hierarchy** (階層)

With **parallelism** (並列性)



# A Typical Memory Hierarchy

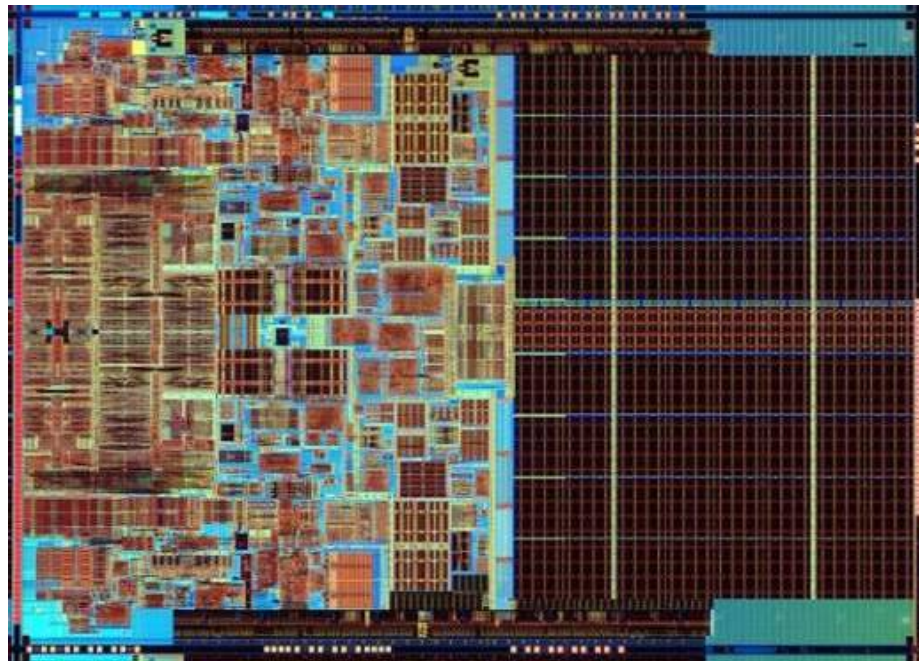
- By taking advantage of **the principle of locality** (局所性)
  - Present **much memory** in **the cheapest technology**
  - at **the speed of fastest technology**



TLB: Translation Lookaside Buffer

# Cache

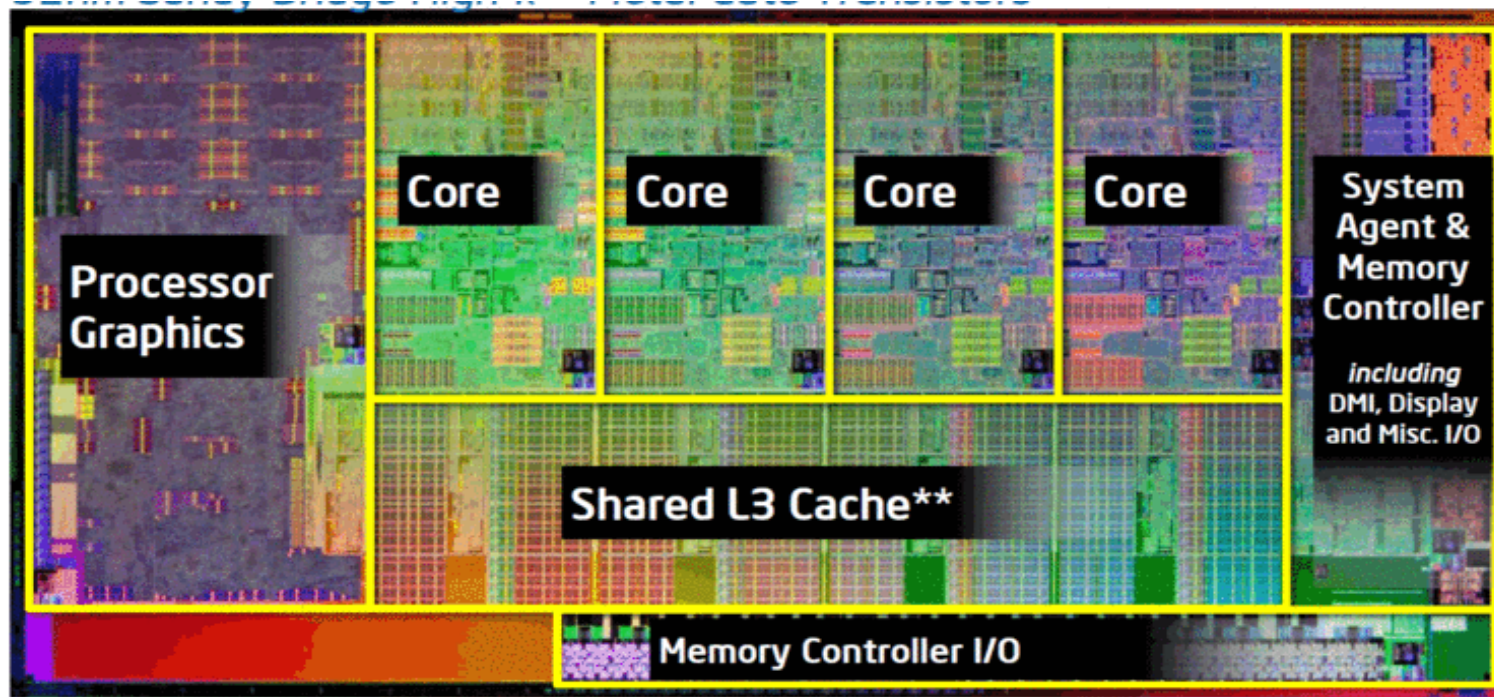
- *Cache memory* consists of a small, fast memory that acts as a buffer for the large memory.
- The nontechnical definition of *cache* is a safe place for hiding things.



Intel Core 2 Duo

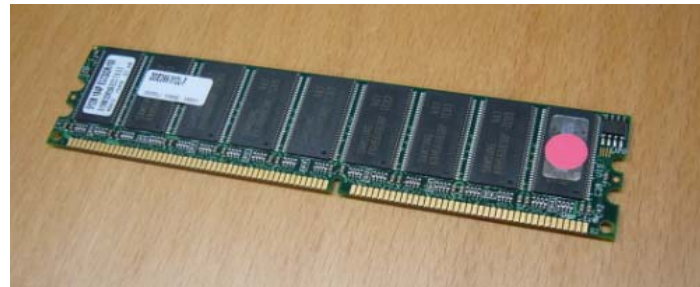


# Intel Sandy Bridge, January 2011



Processor

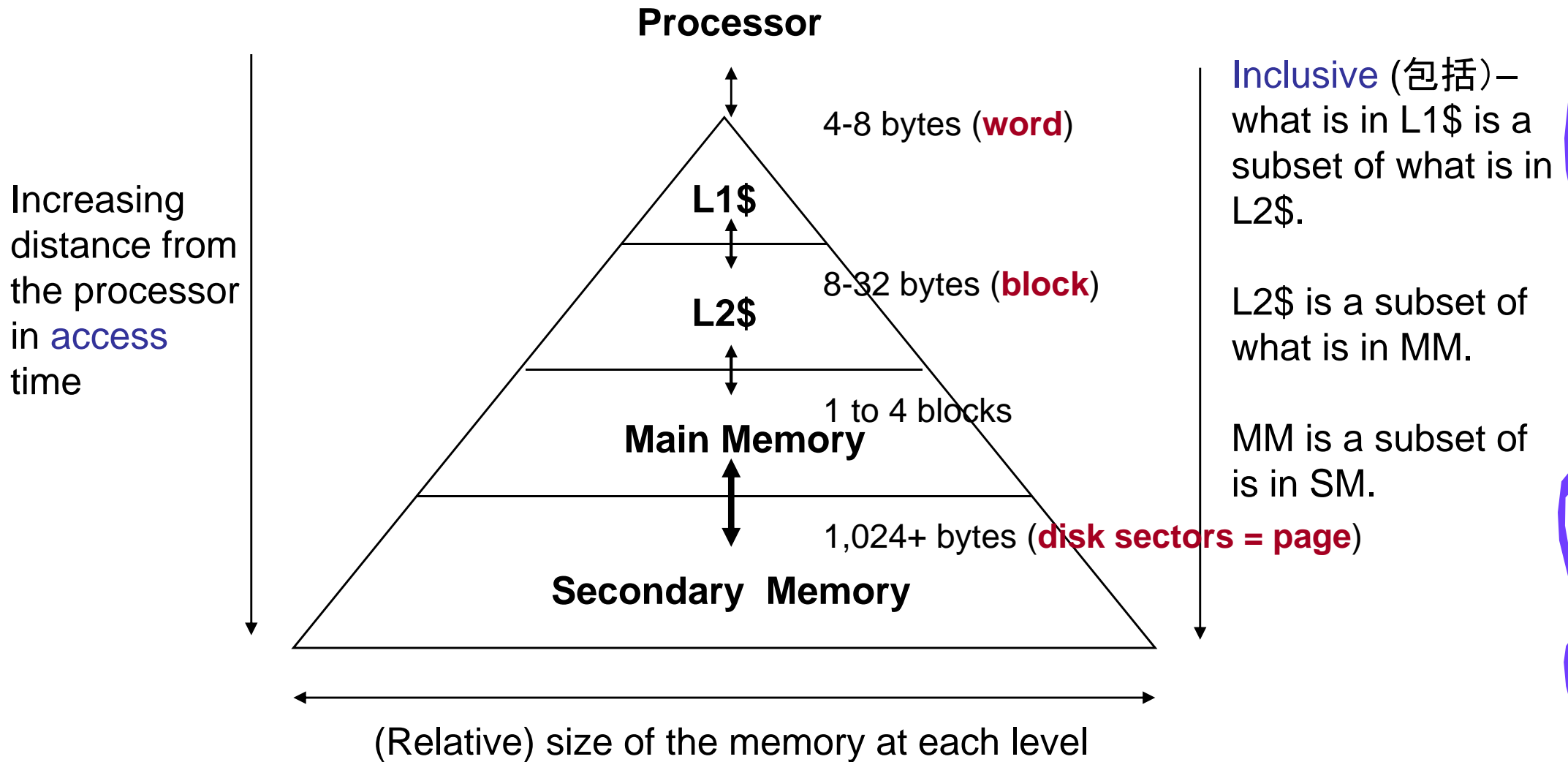
Main memory



Disk



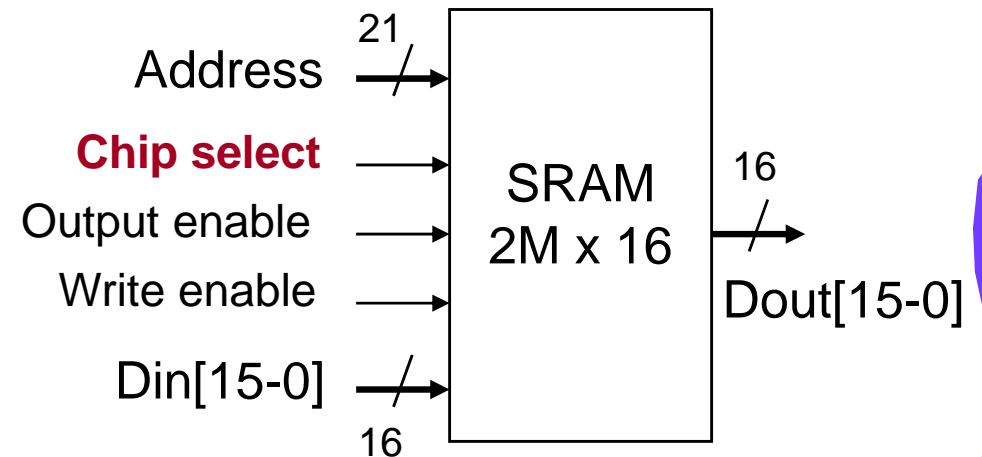
# Characteristics of the Memory Hierarchy



# Memory Hierarchy **Technologies**

- Caches use **SRAM** (static random access memory) for speed and technology compatibility

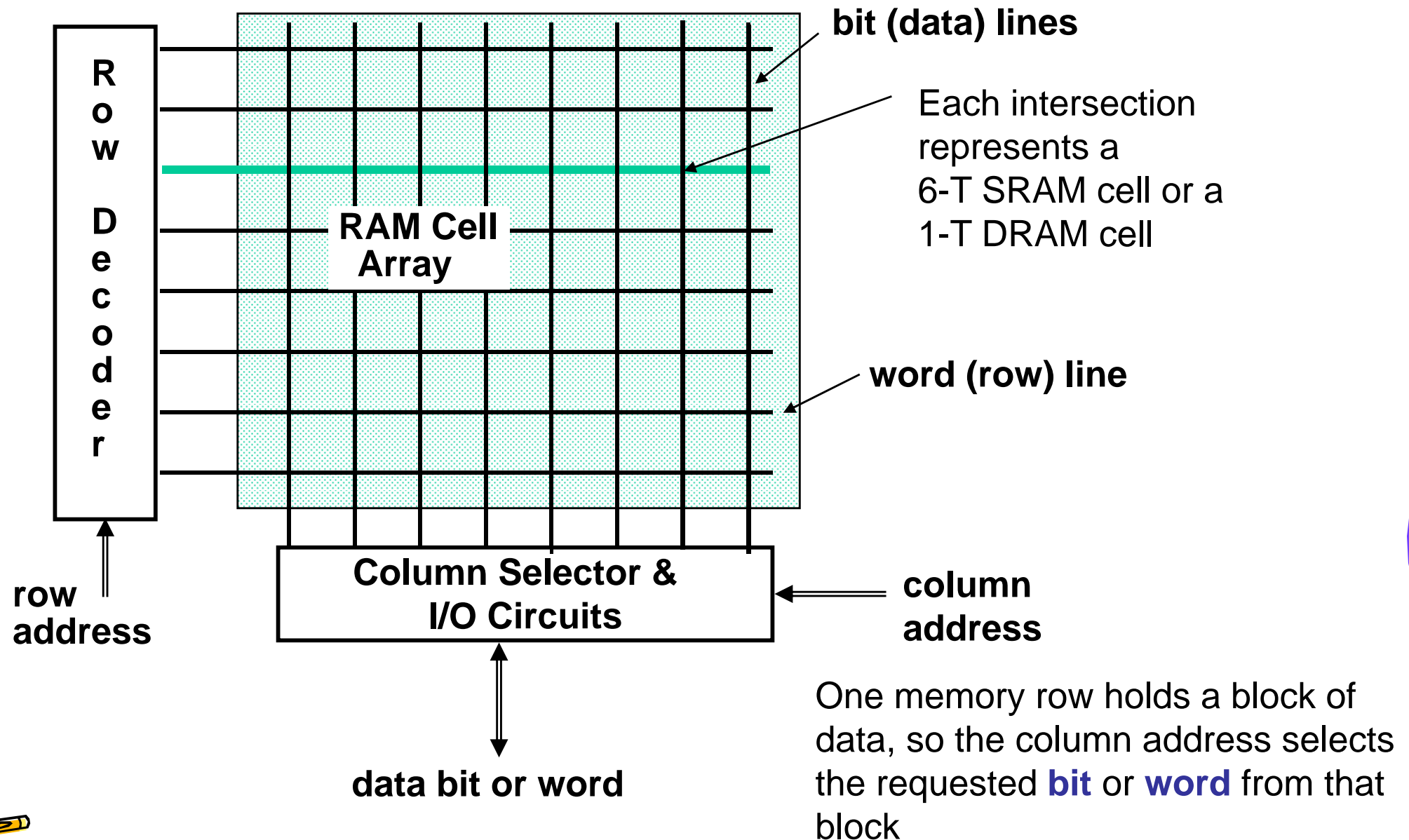
- Low density (6 transistor cells), high power, expensive, fast
- Static: content will last “forever” (until power turned off)



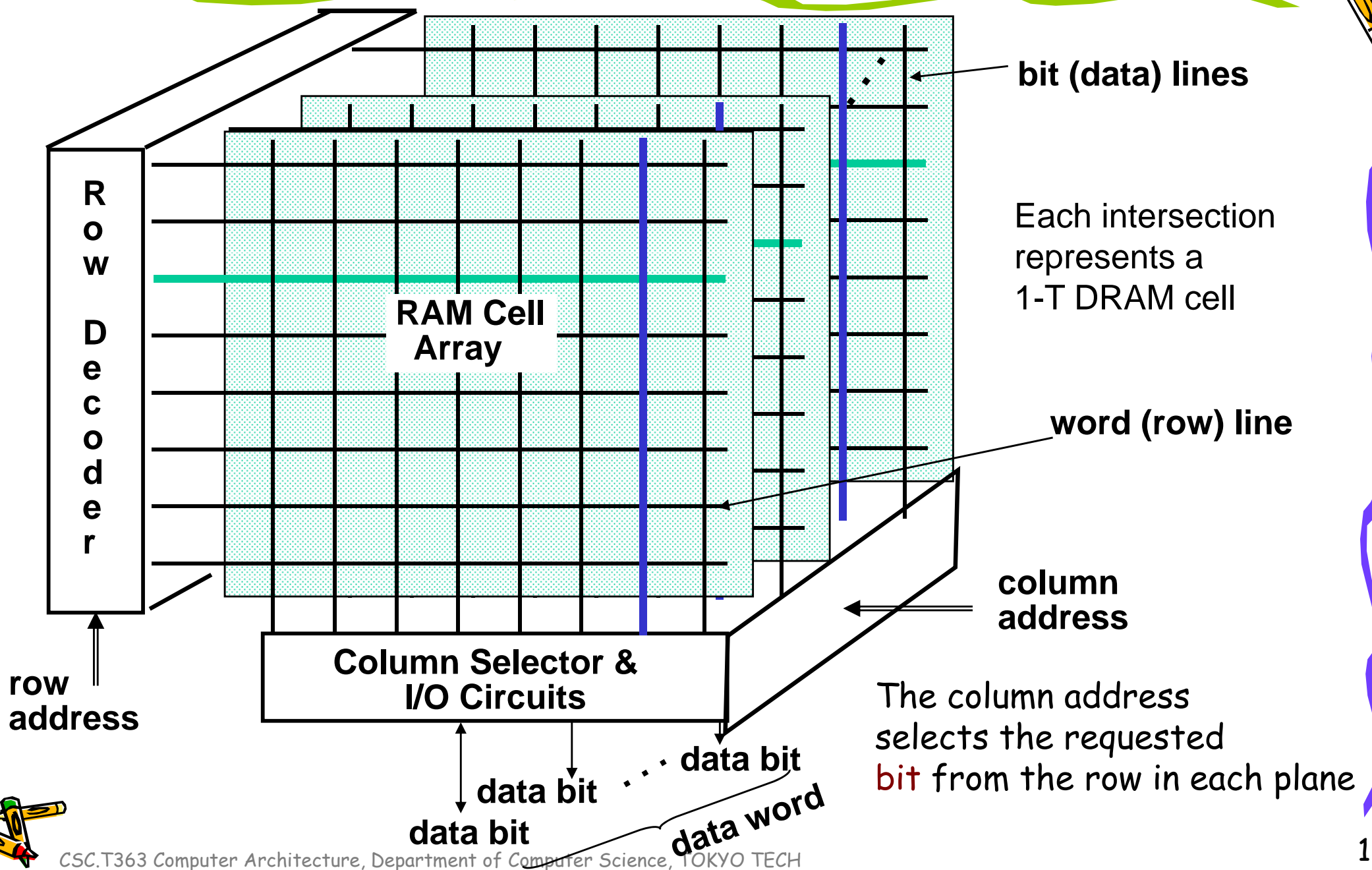
- Main Memory uses **DRAM** for size (density)

- 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 RAM Organization (~Square)



# Classical RAM Organization (~Square Planes)



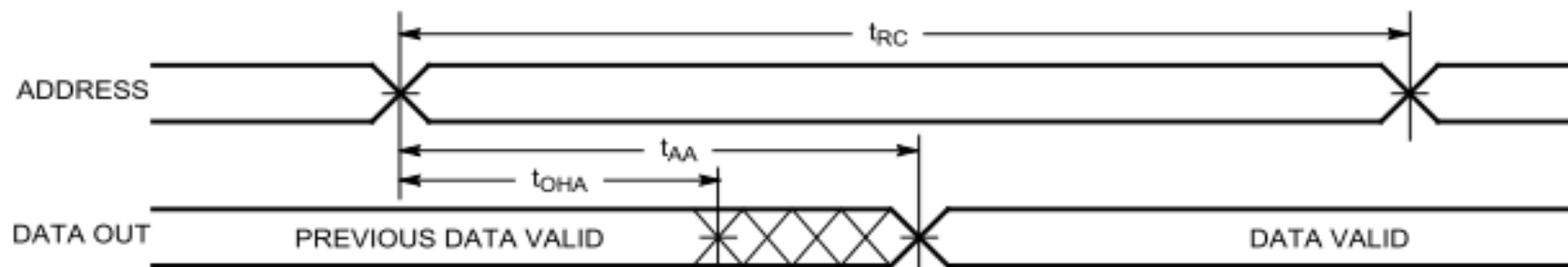
# 非同期式 SRAMメモリ



CY7C1049DV33

## Switching Waveforms

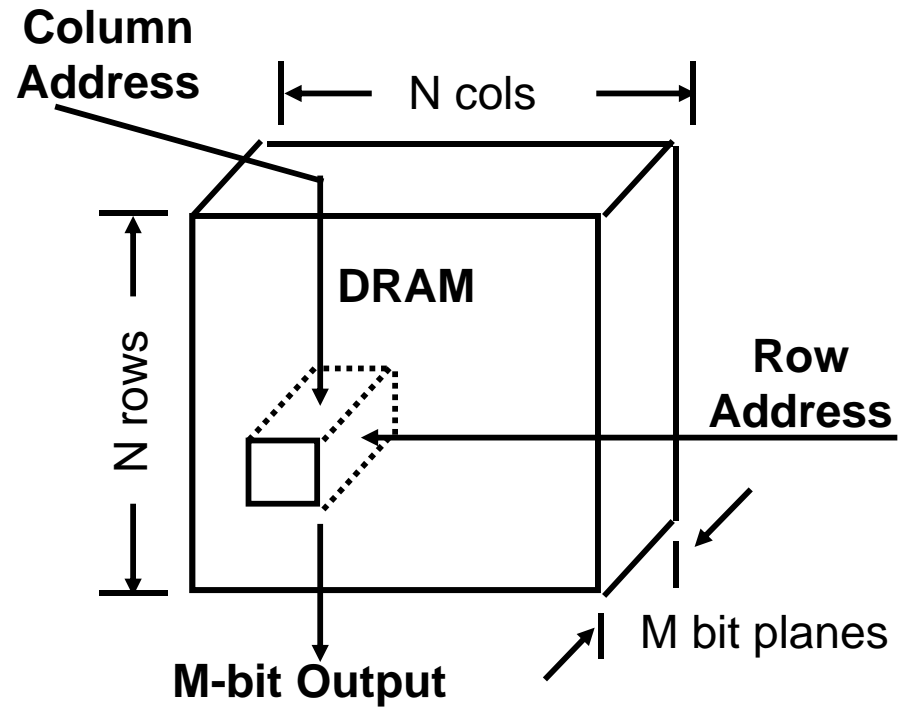
Figure 3. Read Cycle No. 1<sup>[13, 14]</sup>



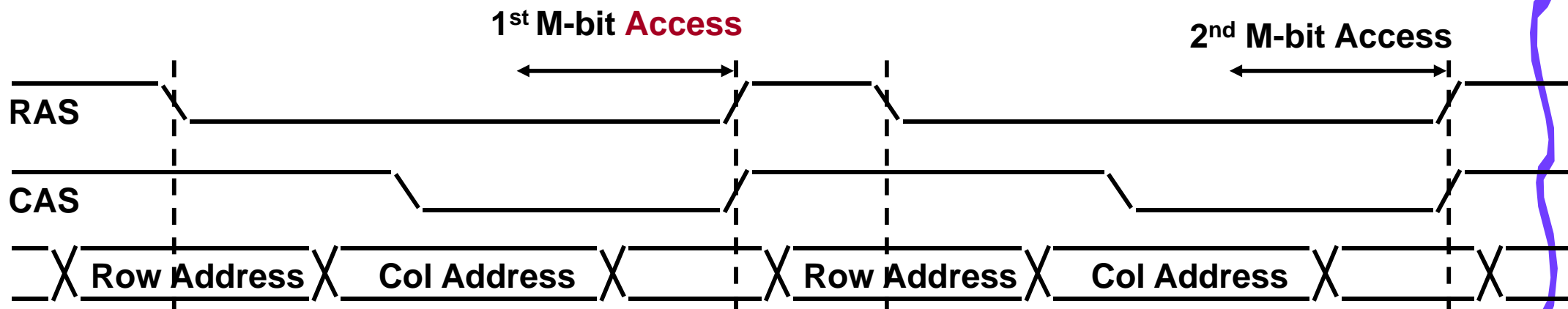
Datasheet

# Classical DRAM Operation

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

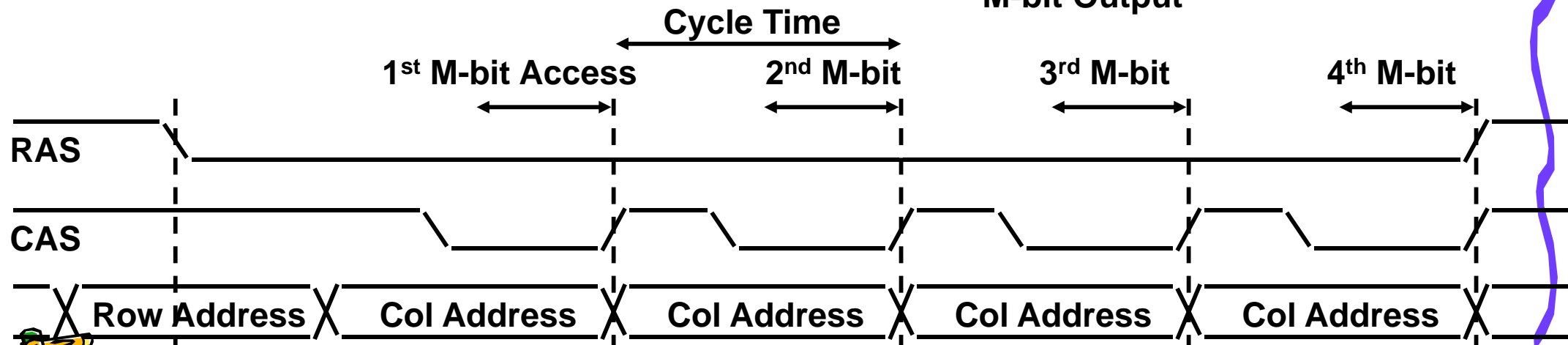
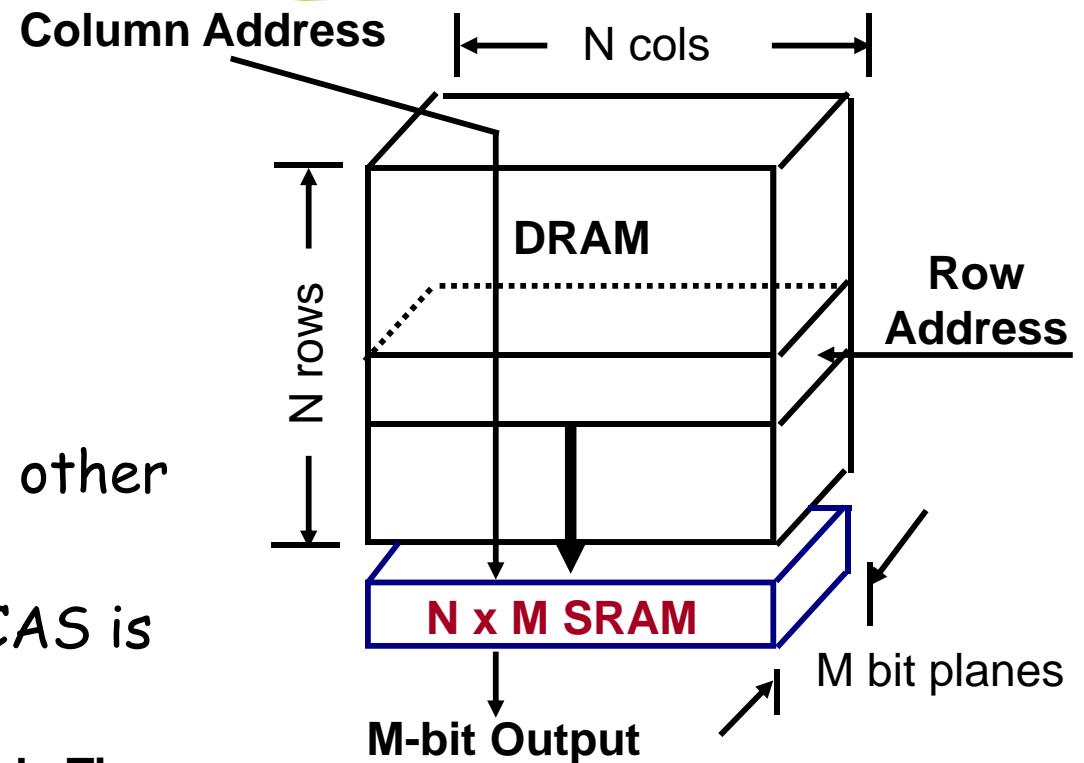


**Cycle Time**



# Page Mode DRAM Operation

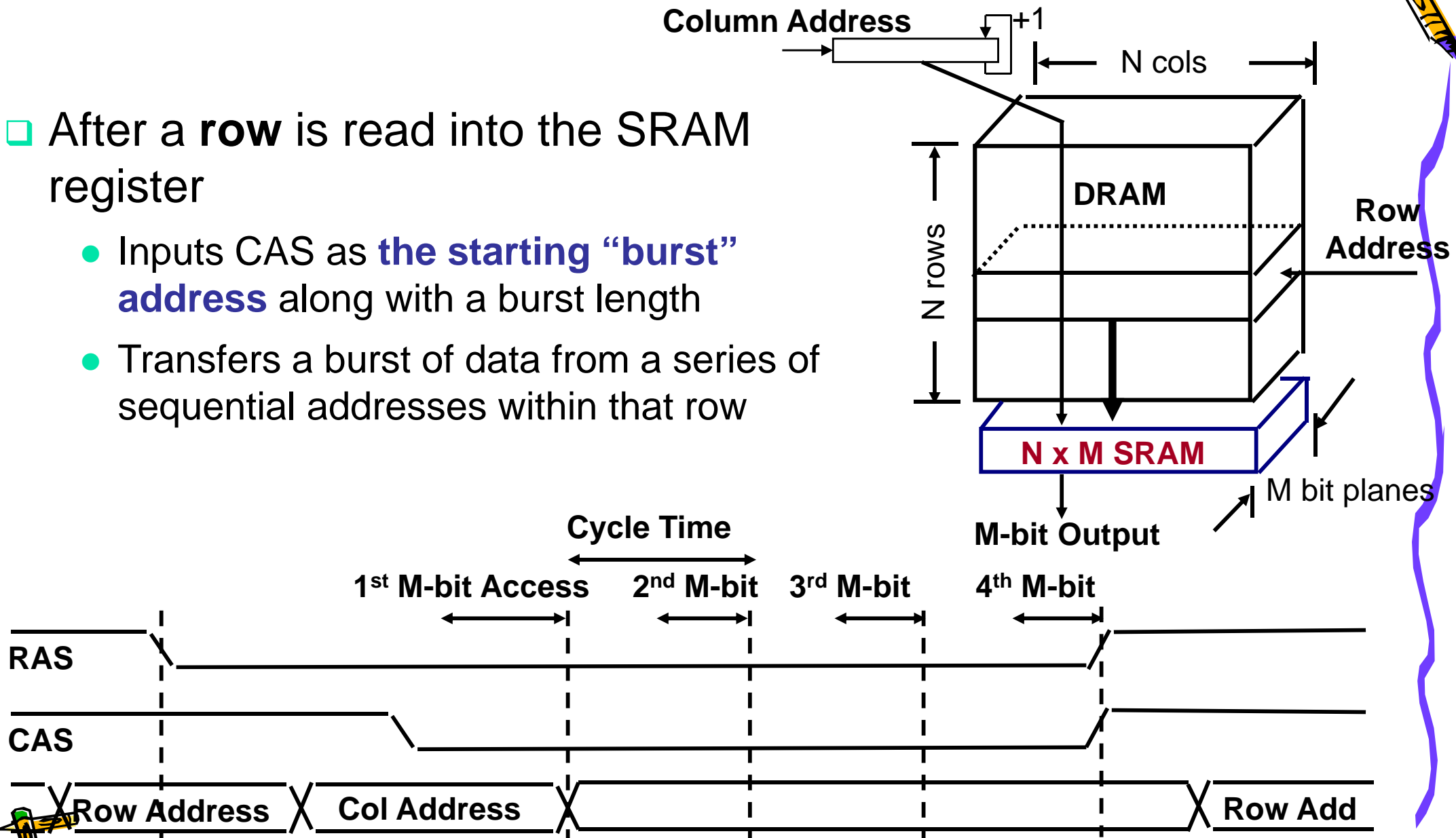
- Page Mode DRAM
  - $N \times M$  SRAM to save a row
- After a row is read into the SRAM "register"
  - Only CAS is needed to access other M-bit words on that row
  - RAS remains asserted while CAS is toggled



# 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**



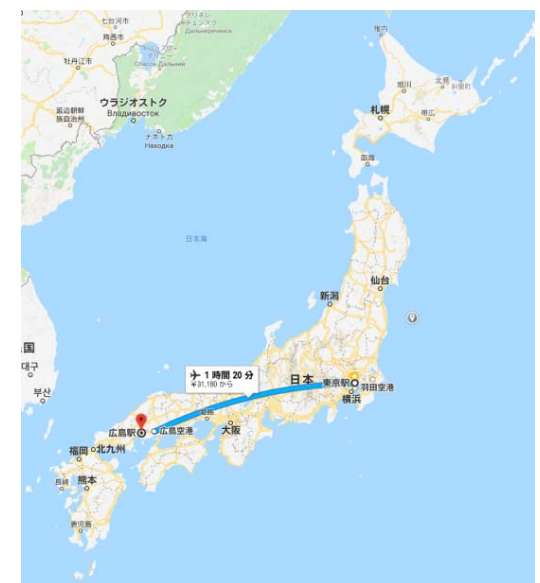
# Which is faster?

## From Tokyo to Hiroshima


	Time Cost	Max Speed	Passengers	Throughput (P x S)
Boeing 737	1:20 32,000yen	800km/h (670km)	170	85,510 (170 x 503)
Nozomi	4:00 18,000yen	270km/h (820km)	1,300	266,500 (1,300 x 205)

- Time to run the task (ExTime)
  - Execution time, response time, latency
- Tasks per day, hour, week, sec, ns ... (Performance)
  - Throughput, bandwidth


From the lecture slide of David E Culler



# 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 (mm <sup>2</sup> )	35	45	70	130	170	204
Pins/chip	16	16	18	20	54	66
<b>BWidth (MB/s)</b>	<b>13</b>	<b>40</b>	<b>160</b>	<b>267</b>	<b>640</b>	<b>1600</b>
<b>Latency (nsec)</b>	<b>225</b>	<b>170</b>	<b>125</b>	<b>75</b>	<b>62</b>	<b>52</b>



Patterson, CACM Vol 47, #10, 2004

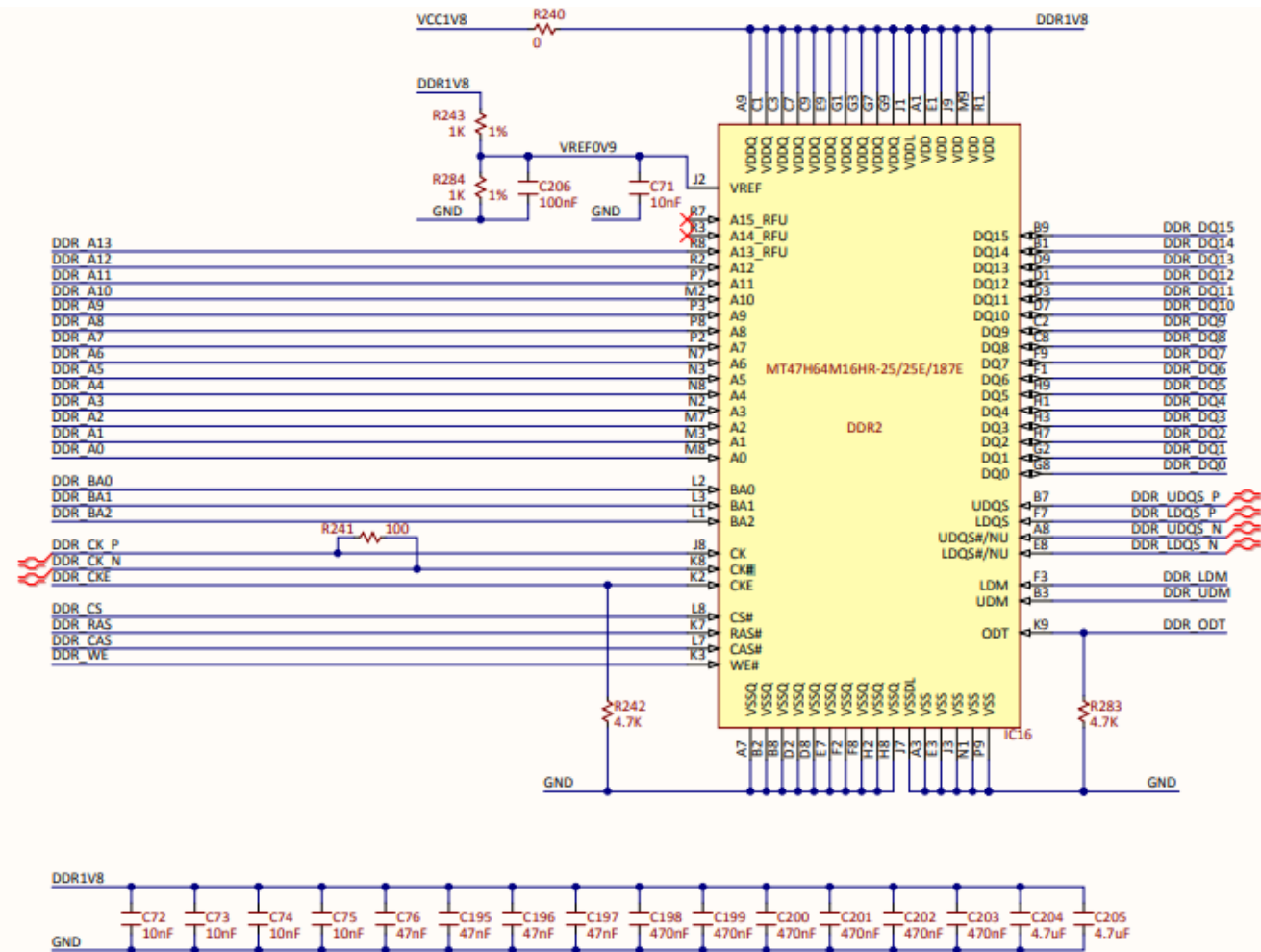
In the time that the memory to processor **bandwidth** doubles the memory **latency** improves by a factor of only 1.2 to 1.4

To deliver such high bandwidth, the internal DRAM has to be organized as **interleaved memory banks**



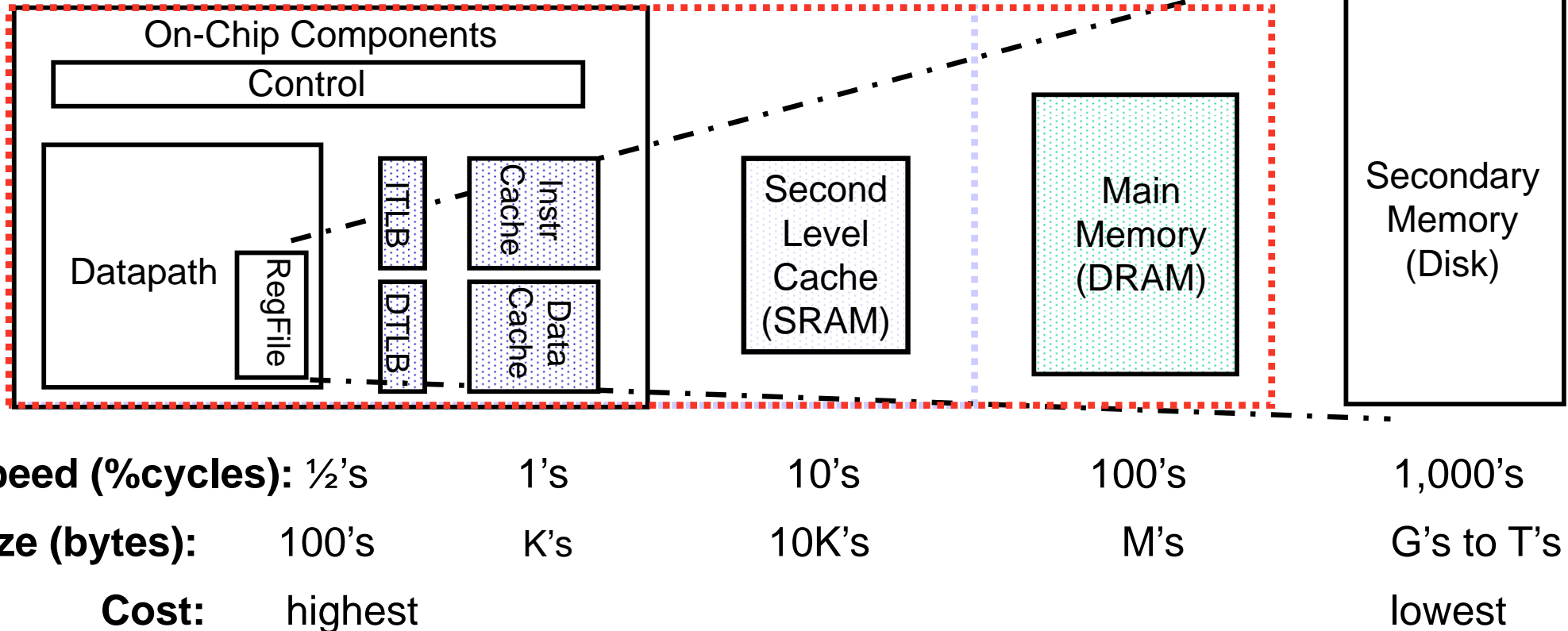
# NEXYS 4 DDR

- Micron MT47H64M16HR-25:H DDR2 memory
  - 128MiB DDR2, 16-bit wide interface



# 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**



TLB: Translation Lookaside Buffer