Lesson 1　CPU

Text

CPU is the abbreviation of Central Processing Unit，and pronounced as separate letters. Sometimes it is referred to simply as the processor or central processor. The CPU is the brains of the computer，it is responsible for handling all instructions and calculations it receives from other hardware components in the computer and software programs running on the computer.[1] For example，the CPU runs the operating system，the software programs installed on your computer，and device peripherals such as printers and flatbed scanners. In terms of computing power，the CPU is the most important element of a computer system.

All the CPU does is run programs by fetching instructions from RAM，evaluating them，and executing them in sequence. The instructions are numbers of the binary system，in a special format that is unique for each machine. The CPU breaks an instruction into parts to see if it has to do something. For instance，a “1” in a certain position in an instruction could mean that the CPU would have to load data in from RAM，or that it would have to add two numbers.[2] After the CPU determines what an instruction is supposed to do，it tells its component parts what to do to complete the instruction.

A CPU has three typical components，that is Arithmetic and Logic Unit(ALU)，Control Unit(CU)and Registers. The ALU is made up of devices called gates that receive one or more inputs and，based upon what function they are designed to perform，output a result.[3] It performs simple arithmetic and logical operations，such as NOT，Left Shift，Right Shift，Add，Subtract，AND，and OR. CU extracts instructions from memory，decodes and executes them，calling on the ALU when necessary. Registers are temporary memory units that store words. The registers are located in the processor，instead of in RAM，so data can be accessed and stored faster.

On large machines，CPUs require one or more printed circuit blocks. On personal computers(PC)and small workstations，the CPU is housed in a single chip called a microprocessor.[4]

Key Words & Terms

access	存取
arithmetic logic unit	算术逻辑单元
base on	基于……

binary system	二进制
chip	芯片
control unit	控制单元
execute	执行
extract	取出，提取
load	载入，加载
microprocessor	微处理器
peripheral	外围部件
register	寄存器
word	字，命令
workstation	工作站

Abbreviations

ALU(Arithmetic and Logic Unit)	算术逻辑部件，运算器
BIOS(Basic Input/Output System)	基本输入/输出系统
CPU(Central Processing Unit)	中央处理器
PC(Personal Computer)	个人计算机
PCB(Printed Circuit Block)	印刷电路板

Notes

[1] **The CPU is the brains of the computer, it is responsible for handling all instructions and calculations it receives from other hardware components in the computer and software programs running on the computer.**

句中的 it receives from other hardware components in the computer and software programs running on the computer 是个从句，作定语修饰 instructions and calculation; be responsible for 意为“为……负责”，后面跟动名词形式。

译文：CPU 犹如计算机的大脑。它负责处理所有的指令和进行计算，这些指令或者计算是该计算机的其他硬件部件或者运行在计算机上的软件发出的。

[2] **For instance, a “1” in a certain position in an instruction could mean that the CPU would have to load data in from RAM, or that it would have to add two numbers.**

句中的 or 连接的两个并列宾语从句作谓语 mean 的宾语；load 意为“载入”。

译文：例如，在一条指令中，某一个位置上的“1”可能表示 CPU 要从 RAM 中加载数据，也可能表示要进行两个数的加法。

[3] **The ALU is made up of devices called gates that receive one or more inputs and, based upon what function they are designed to perform, output a result.**

句中的 based upon what function they are designed to perform 作状语修饰 output; be made up of 意为“由……组成”,指某个整体是由各个部分组成的。

译文:算术逻辑运算单元由被称为门限的设备组成,它接受一个或多个输入并且根据设计所要执行的功能输出结果。

［4］**On personal computers(PC)and small workstations, the CPU is housed in a single chip called a microprocessor.**

句中的 microprocessor 是一个合成词,前缀 micro 表示“微”、“较小的”。

译文:在个人计算机和小型工作站上,CPU 被集成在一个被称为微处理器的单独芯片里。

Exercises

Ⅰ. Translate the following phrases into English.

1. 操作系统

2. 取指—译码—执行

3. 前端总线

4. 双处理器

5. 基本输入/输出系统

Ⅱ. Translate the following sentences into Chinese.

1. The instructions are numbers of the binary system, in a special format that is unique for each machine.

2. CPU is the abbreviation of central processing unit, and pronounced as separate letters.

3. The central processor is where most calculations take place.

4. Dual-core refers to a CPU that includes two complete execution cores per physical processor.

5. Processor: Short for microprocessor or CPU.

6. Integrated circuit: Another name for a chip, an integrated circuit (IC) is a small electronic device made out of a semiconductor material.

Ⅲ. Identify the following to be true or false according to the text.

1. If you want to add two numbers, the operation will be performed by CU.

2. The instructions that are stored in registers will be lost when the computer is shut off.

3. One number of binary system may represent different means in different computers.

Ⅳ. Filling in each of the following blanks according to the text.

1. A CPU has three typical components, that is ________, ________, and ________.

2. CPU extracts instructions from ________.

3. The registers are located in ________.

4. In terms of computing power, ________ is the most important element of a computer system.

Reading Material

Try to find out:

◇ How many processors are on a dual-core computer?

◇ What's the advantage of dual-core?

◇ What's the difference between multi-thread technology and HT technology?

Dual-core Processor

Dual core refers to integrated circuits (silicon chips, 硅片) that contain two complete physical computer processors (cores) in the same IC package. Typically, this means that two identical processors are manufactured so they reside side-by-side on the same die (并行驻留在同一内核上). It is also possible to (vertically) stack two separate processor die and place them in the same IC package. Each of the physical processor cores has its own resources (architectural state, registers, execution units, etc.). The multiple cores on-die may or may not share several layers of the on-die cache.

Dual-processor, Dual-core, and Multi-core: Keeping It Straight

Dual-processor (DP) systems are those that contains two separate physical computer processors in the same chassis (底盘). In dual-processor systems, the two processors can either be located on the same motherboard or on separate boards. In a dual-core configuration, an integrated circuit (IC) contains two complete computer processors.

Usually, the two identical processors are manufactured so they reside side-by-side on the same die, each with its own path to the system front-side bus. Multi-core is somewhat of an expansion to dual-core technology and allows for more than two separate processors.

Taking Advantage of Dual-core Technology

A dual-core processor has many advantages especially for those looking to boost their system's multitasking computing power. Dual-core processors provide two complete execution cores instead of one, each with an independent interface to the frontside bus. Since each core has its own cache, the operating system has sufficient resources to handle intensive tasks in parallel(并行), which provides a noticeable improvement to multitasking.

Complete optimization for the dual-core processor requires both the operating system and applications running on the computer to support a technology called thread-level parallelism, or TLP. Thread-level parallelism is the part of the OS or application that runs multiple threads simultaneously, where threads refer to the part of a program that can execute independently of other parts.

Even without a multithread-enabled application, you will still see benefits of dual-core processors if you are running an OS that supports TLP. For example, if you have Microsoft Windows XP(which supports multithreading), you could have your Internet browser open along with a virus scanner running in the background, while using Windows Media Player to stream your favorite radio station and the dual-core processor will handle the multiple threads of these programs running simultaneously with an increase in performance and efficiency.

Today Windows XP and hundreds of applications already support multithread technology, especially applications that are used for editing and creating music files, videos and graphics because types of programs need to perform operations in parallel. As dual-core technology becomes more common in homes and the workplace, you can expect to see more applications support thread-level parallelism.

Intel & AMD Dual-core Desktop Processors

The Intel Pentium Processor Extreme Edition 840 running at 3.2 GHz and Intel 955X Express Chipsets(芯片组)are being built into computers that are now entering the market. This is Intel's first desktop dual-core product supporting Hyper Threading Technology(HT Technology). Processor features include the following:

Hyper-Threading Technology(超线程技术): With HT technology, two threads can execute on the same single processor core simultaneously in parallel rather than context switching between the threads.

Intel Extended Memory（扩展内存）64 Technology：Provides flexibility for future applications that support both 32-bit and 64-bit computing.

Dual-Core（双核）：Two physical cores in one processor support better system responsiveness and multi-tasking capability than a comparable single core processor.

AMD also announced its line of desktop dual-core processors，the AMD Athlon 64 X2 processor family. The initial model numbers in the new family include the 4200＋，4400＋，4600＋and 4800＋(2.2GHz to 2.4GHz). The processors are based on AMD64 technology and are compatible with the existing base of x86 software，whether single-threaded or multithreaded. Software applications will be able to support AMD64 dual-core processors with a simple BIOS upgrade and no substantial code changes.

Do you know

If we assume that the number of transistors per processor core remains relatively fixed，it is reasonable to assume that the number of processor cores follows Moore's Law，which states that the number of transistors per a certain area on the chip will double approximately every 18 months.

Supplementary Reading

About AMD

"With AMD's tremendous set of technology assets，diverse industry partnerships，and rich talent base，we are uniquely capable of creating value for our customers as we change the dynamics of the industry."

—Hector Ruiz，Chairman of the Board and CEO，AMD

Igniting a New Level of Innovation

At AMD，innovation that truly works for people inspires us. For almost 40 years，the technology needs of customers have shaped our strategic vision，leading to breakthroughs in energy efficiency，digital entertainment，and affordable Internet access worldwide. Now we're pushing further. With the acquisition of ATI and by closely collaborating with world-class technology companies，AMD is helping to reinvent the industry through a strategy that delivers relevant innovations，compelling experiences，and meaningful differences throughout the world.

The new AMD is a processing solutions powerhouse. We're working together with computing，digital TV，and handheld companies as well as industry partners to create smarter choices for customers — from high-performance computing to wireless handsets to HD digital video. AMD is inspiring the technology products of the future.

Providing the Building Blocks of Innovation

We believe that fair and open competition drives innovation. And when new technologies flourish, businesses can reach higher levels of productivity. People can explore a world of unlimited creativity. And everyone benefits from more choice in the global marketplace.

To help our technology partners compete more effectively, we provide the building blocks of innovation, integrating indrustry-leading graphics and computer processing technologies with advanced development resources and world-class engineering expertise. Through AMD's ecosystem, we enable industry partners to build on existing innovations, add their unique value, and create differentiated products to help them gain a competitive advantage.

Delivering Compelling Experiences

As demand grows for more intense, more immersive digital experiences, we're up to the challenge. The new AMD combines superior computing and graphics processing to create powerful visual experiences on everything from PCs to digital televisions to handheld devices.

For this next generation of computing, AMD processor-based systems and graphics are designed to help customers realize the rich visual sophistication and true power of Windows Vista. With graphically advanced processing and video technologies like ATI Avivo. and platforms like the AMD LIVE! PC, we're changing the nature of competition to deliver the total interactive entertainment experience, not just in computing, but in gaming and digital TV.

However, this is only possible through AMD's commitment to being a good collaborator. Our continuous drive to simplify our business process and provide the right talent, tools, and technologies helps our industry partners anticipate future needs and respond faster to new market opportunities.

Making a Difference in the World

At AMD, we believe that companies can do well while doing good. And we're dedicated to making a difference in the success of businesses and the lives of people throughout the world.

We're pushing the edge of performance and efficiency. AMD is committed to designing and building platforms that deliver more computing power using less energy and less space. AMD is one of 11 founding members of the Green Grid, a global consortium dedicated to developing and promoting energy efficiency for data centers and information services.

Through programs such as 50X15, AMD is driven to make the advantages of digital technology ubiquitous. 50X15 is a global initiative founded by AMD that aims to enable affordable Internet access and computing capabilities to 50 per cent of the world's population by the year 2015. To do this, we're leveraging our global ecosystem of strategic relationships to develop usable technologies for the people who need them the most.

AMD at-a-glance

Founded in 1969 and based in Sunnyvale, California, AMD is a leading global provider of innovative processing solutions in the computing, graphics, and consumer electronics markets. AMD is dedicated to delivering standards-based, customer-focused solutions for technology users, ranging from enterprises and governments to individual consumers.

As a company, we are more than 16,000 strong with operations around the globe.

Deriving more than 75 percent of our revenues from international markets in 2006, we truly are a company of the world.

Servers and Workstations

AMD is leading the way in crucial computing technologies like performance-per-watt, virtualization, Direct Connect Architecture, and 64-bit and multi-core computing. AMD Opteron. processors with DDR2 memory offer a seamless upgrade path to quad-core computing to help lower total cost of ownership — scaling business applications and increasing capacity without incurring additional data center infrastructure costs.

Desktop and Notebook PCs

All AMD64 processors feature advanced technology like Direct Connect Architecture, simultaneous 32-bit and 64-bit processing capabilities, and multi-core architecture. AMD Athlon. 64 processors offer powerful performance for a rich digital experience. AMD Turion. 64 Mobile Technology provides the ultimate in mobility, taking advantage of the latest innovations in mobile computing. And the AMD Sempron. processor redefines affordable performance.

Digital TV, Handhelds, and Game Consoles

AMD's digital television business unit is a world leader in the design and manufacturing of silicon and software solutions for integrated digital televisions. ATI Avivo technology delivers PVR, HDTV, and next-generation HD disc capabilities with vibrant, sharp images, and smooth video playback. The AMD Imageon. product family of media processors provides the high-quality, feature-rich multimedia experience demanded by mobile users.

AMD technology can be found in video game consoles, like Nintendo's revolutionary new Wii and Microsoft's next generation Xbox 360; in more than 50 handheld devices on the market, including Motorola multimedia cell phones; and in digital televisions from Samsung, Sony, and others.

Multi-media and PC Gaming

Powered by AMD Athlon 64 X2 dual-core processors and AMD Turion 64 mobile technology, and AMD's award-winning ATI Radeon graphics processors, AMD LIVE! desktop and notebook PCs let users consolidate all of their photos, videos, music, movies and more, and then access them through their own PC or a range of digital devices. Featuring AMD's most advanced processor technology, the AMD Athlon 64 FX processor is designed specifically for PC enthusiasts. Powering the AMD Quad FX platform with Dual Socket Direct Connect Architecture, AMD Athlon 64 FX processors enable users to immerse themselves in their digital world.

Lesson 2 Computer Memory

Text

Memory is something that stores, preserves and recalls data when needed. Your brain has this capability and memory inside computers is an electronic incarnation of this concept.[1]

Computer memory is used to store data that needs to be accessed by the Central Processing Unit (CPU). It is the CPU that performs the laborious tasks, the memory acts as storage for uncompleted tasks and other relevant information needed to accomplish those tasks.

All the information in memory is encoded in fixed size cells called bytes.[2] A byte can hold a small amount of information, such as a single character or a numeric value between 0 and 255. The CPU will perform its operations on groups of one, two, four, or eight bytes, depending on the interpretation being placed on the data, and the operations required.

Forms of Memory

Although the term "computer memory" is commonly used to refer to RAM, there are various other forms of memory inside a computer — such as the hard disk drive.

The illustration depicted in Figure 2.1 below outlines the common memory architecture adopted within most modern computers.

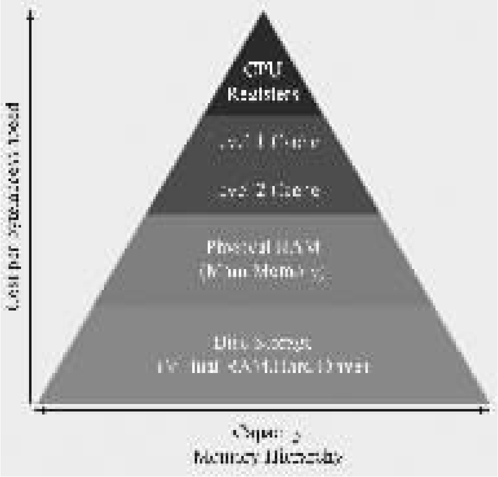

Figure 2.1

Any of the four major categories of memory in the diagram above can feed information directly to the CPU. Each form of memory feeds the CPU at different speeds and efficiency due to their different technological make up.[3]

Registers and cache will transfer data to the CPU at greater speeds than RAM and hard disk drives. As the forms of memory in the top of the pyramid are costly to make, their size are limited to make computers affordable. The size of memory forms towards the bottom of the pyramid is made larger to hold all the other data that other forms cannot handle.[4]

It should also be noted here that physical RAM and cache are volatile in nature — meaning they store, preserve and recall data so long as there is electrical power flowing through the system. Once a computer system is shut off, the physical RAM and cache are cleared. Disc storage by way of floppy disk, hard disk, CD-ROM and DVD-ROM drives holds information those are required to be non-volatile in nature.

There is a trade-off between speed and cost, resulting in the development of such a pyramid-like architecture. Information is prioritized in terms of importance and stability to determine which form of memory would hold the data. [5]

Registers

The bottle-neck in a memory and CPU architecture is the slow transfer speeds between the two. The fastest, and sadly the most expensive, form of memory resolves this problem by having the memory within the CPU itself. Data within registers are instantly fed to the Arithmetic and Logic Unit (ALU) portion of the CPU making the relevant data instantly available.

Registers are typically small in size and is controlled by the CPU's compiler.

Cache

This form of memory can be considered as an intermediary between the main physical RAM and the CPU. The cache makes any data frequently used by CPU instantly available. If the required information is not located in the cache, a fetch is made from the main memory.

There are two levels of cache: level 1 cache (primary cache) and level 2 cache (secondary cache).

Level 1 cache is built directly on the CPU, just like the registers. It is small in size, ranging anywhere between 2 kilobytes (KB) and 128KB. As this cache is closer to the CPU than level 2 cache, its transfer speeds are faster as a result.

Level 2 cache is usually situated in close proximity to, but off, the CPU chip.[6] However, there are certain systems where the cache is built onto the CPU as like the level 1 cache. The size of level 2 cache ranges from 256KB to 2 megabytes (MB).

Both levels of cache use Static Random Access Memory (SRAM) to hold the data.

Main Memory

This is where most of the information that a CPU requires resides. "Main Memory" commonly refers to Physical Memory, although a computer uses an operating systemimposed Virtual Memory in addition to physical memory. The amount of main memory on a computer is crucial because it determines how many programs can be executed at one time and how much data can be readily available to a program.

Physical memory uses Dynamic Random Access Memory (DRAM) to store the data, and is considerably slower than the SRAM used by the cache. Information is exchanged between the main memory and the cache to ensure that the more commonly accessed information is placed in the cache to allow faster access speed. Operating system's memory management will automatically remove data held on both the physical and virtual memory.

The physical memory acts as an Input/Output (I/O) channel for data exchanged between the computer memory and other forms of electronic storages.

Virtual Memory

Most operating systems have a form of memory management that caters for memory needs beyond a computer system's physical memory through the use of a swap file.[7] There is a need for such memory management as operating systems themselves occupy a significant portion of physical memory.

A swap file is a file located on a computer's hard disk drive (HDD) that acts as an extension to physical memory. However, the HDD has much slower access times than any of the forms of memory discussed above. Hence, information is swapped between the main memory and the swap file to ensure that the more frequently used information is located in the main memory for faster access speeds.

Key Words & Terms

anywhere between	数目在……之间
architecture	体系结构
bottle-neck	瓶颈
cache	缓冲存储器
compiler	编译器
considerably	相当地
disc storage	磁盘存储器
encode	编码
feed to	供应给……
hierarchy	层级
incarnation	化身
intermediary	调节者,中介
laborious	费力的,艰苦的
level 1 cache	一级缓存
level 2 cache	二级缓存
numeric	数值的
physical memory	物理存储器

prioritize	把……区分优先顺序
proximity	接近
swap file	交换文件
trade-off	平衡，协定
transfer speed	传输速度
virtual memory	虚拟存储器
volatile	易失性的

Abbreviations

DRAM(Dynamic Random Access Memory)	动态随机存储器
HDD(Hard Disk Drive)	硬磁盘驱动器
I/O(Input/Output)	输入/输出(设备，数据)
SRAM(Static Random Access Memory)	静态随机存储器

Notes

[1] **Your brain has this capability and memory inside computers is an electronic incarnation of this concept.**

本句是由 and 连接的两个并列句，其中的 this concept 指代上文大脑能存储和回忆的现象。

译文：你的大脑有这种能力，而计算机内部的存储器正是这一概念的电子化身。

[2] **All the information in memory is encoded in fixed size cells called bytes.**

本句中的 called bytes 修饰 cells。

译文：存储器上的所有信息都以固定大小的单元为单位进行编码，这样的单元被称为字节。

[3] **Each form of memory feeds the CPU at differing speeds and efficiency due to their different technological make up.**

句中的 at ... speed 意为“以……速度”；due to 译为“由于，因为”；make up 本身是动词词组，在这里当名词来用，意为“制作”。

译文：每种存储器向 CPU 提供信息的速度和效率都不一样，这是由存储器不同的制作技术决定的。

[4] **The size of memory forms towards the bottom of the pyramid is made larger to hold all the other data that other forms cannot handle.**

句中的 to hold all the other data that other forms cannot handle 作目的状语，其中，that other forms cannot handle 又是一个定语从句，修饰 data。

译文：越往金字塔底部去，存储器尺寸做得越大，目的是保存其他存储形式不能处理的数据。

[5] Information is prioritized in terms of importance and stability to determine which form of memory would hold the data.

句中的 prioritized 意为“把……区分优先顺序”；in terms of 意为“按照……，根据……”；data 意为“数据”，是复数名词形式，其单数为 datum。

译文：信息按照重要性和稳定性被划分为不同的优先级，这样做的目的是决定用哪种存储器保存这些数据。

[6] Level 2 cache is usually situated in close proximity to, but off, the CPU chip.

句中的 proximity 意为“接近”；句中 but off 是对 proximity to 的补充，翻译为“但有一定距离”。

译文：二级缓存通常位于非常靠近 CPU 芯片的位置，但在 CPU 芯片的外部。

[7] Most operating systems have a form of memory management that caters for memory needs beyond a computer system’s physical memory through the use of a swap file.

句中的 swap file 意为“交换文件”；cater for 意为“迎合”；句中 that caters for memory needs beyond a computer system’s physical memory through the use of a swap file 是一个定语从句，作 a form of memory management 的定语；through the use of a swap file 在从句中作状语，修饰 caters for。

译文：大部分操作系统都有一种内存管理的形式，即通过交换文件的方式满足超过计算机系统物理存储器容量的访问需求。

Exercises

Ⅰ. Translate the following phrases into English.

1. 静态随机存储器

2. 动态随机存储器

3. 虚拟存储器

4. 物理存储器

5. 一级缓存

6. 二级缓存

7. 硬盘驱动器访问速度

Ⅱ. Translate the following sentences into Chinese.

1. DRAM gets the “dynamic” in its name because it is refreshed thousands of times per second.

2. RAM：A temporary storage memory area in computer where the operating system，application programs，and data in current use are kept.

3. ROM is sustained by a small long-life battery in your computer.

4. RAM cache is a portion of memory made of high-speed static RAM.

Ⅲ. Identify the following to be true or false according to the text.

1. The speed of RAM is higher than the speed of cache.

2. Level 2 cache is built in the CPU.

3. The CPU can perform operations on group of ten bytes.

4. The capacity of Disc storage is usually larger than that of cache.

Ⅳ. Filling in each of the following blanks according to the text.

1. Disc storage holds information those are required to be ________ in nature.

2. Registers are typically small in size and is controlled by the CPU’s ________.

3. RAM and cache are ________ in nature—meaning they store，preserve and recall data so long as there is electrical power flowing through the system.

4. Physical memory uses ________ to store the data，and is considerably slower than the SRAM used by the cache.

Reading Material

Try to find out：

◇ What components is a DRAM cell made up of?

◇ What are the two functions performed by transistor?

◇ Why is DRAM slower than SRAM?

◇ Which one is more expensive，SRAM or DRAM?

Types of RAM：Static and Dynamic

Dynamic Random Access Memory (DRAM)

DRAM is the more common type of the two. A memory cell of DRAM consists of a transistor and a capacitor that holds a charge to represent one bit of binary information (0 or 1).

The transistor performs two functions：it reads the charge of the capacitor and it changes the state of the capacitor (from 1 to 0 and vice versa).

When the capacitor is charged (充电), usually more than 50%, it represents the binary "1" in the memory cell. When the capacitor is discharged, it represents "0".

DRAM gets the "dynamic" in its name because it is refreshed thousands of times per second. The reason for this is that the capacitors within DRAM do not hold their charge (and thus the value of "1") very well. It must be constantly recharged by a memory controller before the capacitor drops to below 50% of its charge (which will represent "0").

As DRAM must always refresh itself, it is slower (but cheaper!) than static RAM where such a refresh operation is not needed. DRAM supports access times of 60 nanoseconds.

Static Random Access Memory (SRAM)

SRAM uses a different architecture than DRAM to hold data and is more reliable and faster as a result. SRAM supports access times of just 10 nanoseconds.

The memory cell of SRAM consists of four or six transistors to place the cell in a state of true or false (akin to the binary "1" and "0"). As a SRAM memory cell consists of more components than DRAM, this result in a higher price tag and more space consumption.

As there are no capacitors within a memory cell, power is required to always flow to the cell to maintain the data.

SRAM and DRAM are both volatile forms of memory as they require constant electrical power to maintain the data. Once your computer is switched off, the RAM is cleared of data.

Due to price discrepancy between SRAM and DRAM, the majority of the space in a computer's RAM is made of DRAM while SRAM is used as a memory cache.

Do you know

Data stored within RAM can be randomly accessed if you know the physical address of the memory cell — hence the "R" in RAM. Contrast this notion with Serial Access Memory (SAM), where each memory cell is accessed sequentially until the relevant information is found.

As RAM only accesses one address while SAM accesses multiple addresses before reaching the desired data, random access is on average, faster than the serial access method.

Supplementary Reading

Virtual Memory

A cache stores a subset of the address space of RAM. An address space is the set of valid addresses. Thus, for each address in cache, there is a corresponding address in RAM. This subset of addresses (and corresponding copy of data) changes over time, based on the behavior of your program.

Cache is used to keep the most commonly used sections of RAM in the cache, where it can be accessed quickly. This is necessary because CPU'speed increases much faster than the speed of memory access. If we could access RAM at 3 GHz, there wouldn't be any need for cache, but RAM can't keep up, we use cache.

What if we wanted more RAM than we had available. For example, we might have 1M of RAM, what if we want 10M? How could we manage?

One way to extend the amount of memory accessible by a program is to use disk. Thus, we can use 10 Megs of disk space. At any time, only 1 Meg resides in RAM.

In effect, RAM acts like cache for disk. This idea of extending memory is called virtual memory. It's called "virtual", which doesn't mean that it's fake, but only because it's not RAM.

The real problem with disk is that it's really, really slow to access. If registers can be accessed in 1 nanosecond and cache in 5 ns and RAM in about 100 ns, then disk is accessed in fractions of seconds. It can be a million times slower to access disk than a register.

The advantage of disk is it's easy to get lots of disk space for a small cost.

Still, because disk is so slow to access, we want to avoid accessing disk unnecessarily.

Uses of Virtual Memory

Virtual memory is an old concept. Before computers had cache, they had virtual memory. For a long time, virtual memory only appeared on mainframes. Personal computers in the 1980s did not use virtual memory. In fact, many good ideas that were in common use in the UNIX operating systems didn't appear until the mid 1990s in personal computer operating systems (pre-emptive multitasking and virtual memory).

Initially, virtual memory meant the idea of using disk to extend RAM. Programs wouldn't have to care whether the memory was "real" memory (i. e. , RAM) or disk. The operating system and hardware would figure that out.

Later on, virtual memory was used as a means of memory protection. Every program uses a range of addressed called the address space.

The assumption of operating systems developers is that any user program can not be trusted. User programs will try to destroy themselves, other user programs, and the operating system itself. That seems like such a negative view, however, it's how operating systems are designed. It's not necessary that programs have to be deliberately malicious. Programs can be accidentally malicious (modify the data of a pointer pointing to garbage memory).

Virtual memory can help there too. It can help prevent programs from interfering with other programs. Occasionally, you want programs to cooperate, and share memory. Virtual memory can also help in that respect.

How Virtual Memory Works

When a computer is running, many programs are simultaneously sharing the CPU. Each running program, plus the data structures needed to manage it, is called a process.

Each process is allocated an address space. This is a set of valid addresses that can be used. This address space can be changed dynamically. For example, the program might request additional memory (from dynamic memory allocation) from the operating system.

If a process tries to access an address that is not part of its address space, an error occurs, and the operating system takes over, usually killing the process (core dumps, etc).

How does virtual memory play a role? As you run a program, it generates addresses. Addresses are generated (for RISC machines) in one of three ways:

- A load instruction
- A store instruction
- Fetching an instruction

Load/store instructions create data addresses, while fetching an instruction creates instruction addresses. Of course, RAM doesn't distinguish between the two kinds of addresses. It just sees it as an address.

Each address generated by a program is considered virtual. It must be translated to a real physical address. Thus, address translation is occurring all the time. As you might imagine, this must be handled in hardware, if it's to be done efficiently.

You might think translating each address from virtual to physical is a crazy idea, because of how slow it is. However, you get memory protection from address translation, so it's worth the hardware needed to get memory protection.