Computer architecture: Quick reference

vishvAs vAsuki

December 20, 2011

1 Architecture of a computer

1.1 Gross view

Visualize a processor connected to some working memory and to an input output device. This is aka von-Neumann architecture.

1.2 Motherboard

A mother-board provides suitable interface and circuits for communication and power-supply among/ for the processor, working memory, I/O, persistent storage etc..
The motherboard contains the BIOS memory where the boot-program is located.

1.2.1 System Bus

The system bus is a common communication channel on the motherboard. It has many different dedicated branches for communication between the different components and certain controller hubs.
A computer bus operating with *double data rate* transfers data on both the rising and falling edges of the clock signal.

1.2.2 Quality

Important factors determining quality are: the bus clock frequency, the choice of I/O protocols supported.

1.3 Processor

1.3.1 Components

Every processor consists of an arithmetic and logic unit/ circuit (ALU), a few memory registers with a few layers of cache, memory and graphics controller circuits in order to be able to perform I/O operations and using extended working memory.

1.3.2 Tight knit Parallelism

1.3.2.1 Multiple cores

Usually multiple cores share memory.

1.3.2.2 Pipelines

Aka instruction level parallelism.
Processing often has pipeline structure; eg: fetch instruction, fetch data, execute instruction, store data. There are usually separate functional units within a single processor which take care of these separate tasks in each clock cycle.
'Superscalar processors' try to parallelize these tasks. Branch prediction problem: how to pipeline if the instruction to be executed is a conditional: multiple choices of what to prefetch to the pipeline.

Simultaneous multithreading Parallelism in use of functional units within processors to execute two instruction threads at once, rather than one instruction thread faster is a feature provided in 'SMT' processors. This is called hyperthreading by Intel.

1.3.3 Memory cache

There is a tradeoff between latency and miss-rate: the larger the cache, the lower the possibility of necessary data not being cached, but latency is higher because of greater addressing/ accounting needs. Hence, there are multiple cache-levels, denoted L1 .. Lk in increasing order of size.

1.3.4 Quality

The frequency indicates the speed at which a processor can be operated. Processors also differ in the ability to undertake 16 vs 32 vs 64 bit arithmetic. They also differ in the number of cores to accommodate parallelism.
They differ in the number and sizes of memory-cache levels they offer, and the speed with which they can be accessed.
They differ in the power they consume - for mobile devices, the ability to adjust processing speed (stepping) according to computing needs is important.

1.3.4.1 Best and economical

2011: 4-core 64 bit processors running at 2.7 GHz with 4MB L3 cache size by Intel and AMD are available for $\approx$ 110$.

1.3.4.2 Trends

Graphene replacing silicon can lead to continuation of moore's law - enabling smaller and smaller transistors on the Integrated circuit/ IC chip. But smaller communication channels implies greater power costs for pushing electrons across the channels within the chip.
Then, it will be possible to have a thousand-core processor.

1.4 Working Memory

1.4.1 DDR SDRAM

This is Double data rate synchronous dynamic random access memory. The terms are explained below.
Working memory is capable of random, rather than serial, access; though it may still be efficient to read big blocks of memory together into the CPU cache for processing. Also, it is dynamic - it retains data only when supplied electricity to periodically refresh its memory. It is commonly Synchronous - synchronized with the system bus.
Current bottleneck!

1.4.1.1 Quality

They mainly vary in the clock/ data transfer rate (Eg: DDR-333 > DDR-200) and in size. PC2-xxxx denotes theoretical bandwidth (with the last two digits truncated).

Best and economical As of 2011: 2*2 = 4GB RAM modules are economical - sometimes insufficient for programming and surfing the web simultaneously - some website/browser combos consume much memory! Macbook pro comes with 8GB RAM.

1.4.1.2 Interface

DDR SDRAM modules come in different interface sizes; the number of pins can be different: laptop units are smaller. Vostro 1000 accepts DDR2 PC2-5300 RAM, even 2GB DDR2 800MHz/PC2-6400 200-pin SODIMM.
Some motherboards support dual channel memory, which theoretically doubles the throughput if both SDRAM-slots in the contain identical memory modules.

1.4.2 Memory hierarchy

Memory access could be 100 times slower than flops: this is an important consideration when optimizing algorithms. Hence, there is a hierarchy of memory: Registers in the processor > On chip Cache: many layers > main memory > secondary memory located on the hard-disk for example.

1.5 Graphics processing (GPU)

GPU's, implementing common graphics tasks on specialized circuits, speed up display tasks involved in personal computing.

1.5.1 High data parallelism

GPU's follow a dataflow architecture: Highly pipelined, parallel with many small cores - much more than CPU's. Usually, these parallel cores are divided profitably and easily: Like one core for a bunch of pixels.
Earlier these processing elements were specialized for graphics, now these shaders are more programmable.

1.5.2 For general computing

With GPGPU, one would Disguise program as geometry computation. Now, can do such computation directly: eg: NVidia CUDA.

2 Storage devices

2.1 Magnetic storage

Unlike optical storage, magnetic storage is re-writable.

2.1.1 Bus interface

While SCSI has a processor integrated into the controller, SATA makes greater use of the system processor to serve that function; the former is costlier.

2.1.2 Quality

Hard disk drives differ mainly in speed, storage capacity, size, redundancy (RAID), error correction, shock resistance.

2.1.2.1 Best and economical

Areal density doubling every two to four years since their invention. So, even 1TB is economical.

2.2 Optical storage

They are compact, but have the disadvantage of being over-writable a limited number of times.

2.2.1 Quality

2.2.1.1 Best and economical

Blu-Ray - where blue laser is used - is the rage. It is even more burdened than the DVD with copy-protection mechanisms like region code and DRM.

2.3 Flash memory

These use floating gate MOSFET transistors. They are fast and compact.

2.3.1 SD cards

2.3.1.1 Size

They come in 3 sizes: micro, mini and regular. A smaller card can be used in place of a larger card using an adapter.

2.3.1.2 Speed

Depending on transfer speed (in MBps), they are classified into various classes, with class 4 commonly used. Class 10 is recommended for watching movies etc..

3 Display

3.1 Display screens

Liquid Crystal displays tend to be more ergonomically friendly than CRT tube monitors.

3.1.1 Quality

3.1.1.1 Image quality

Display screens vary in image quality parameters: screen size, maximum resolution: SVGA (800 x 600), XGA (1024 x 768), SXGA (1280 x 1024), or UXGA (1600 x 1200), contrast ratio.

They also vary in refresh rate, in the case of images generated from projection, as in CRT monitors or projectors: low refresh rate results in flicker and eye strain - especially from typing distance. Higher resolutions require correspondingly higher refresh rates.

3.1.1.2 LCD screens

In addition LCD screens vary in adjustibility of height, orientation etc. and whether they also function as a touch sensitive input device. Screen size of around 14' has been quite adequate in my laptop computer.

3.1.1.3 Best and economical

23 inch LCD without touch sensitivity was available for around 110$ in a sale.

3.2 Projectors

Using projectors for display allows display size to be set dynamically; allowing a variety of postures and ability to view from a distance, it may be ergonomically better. But, it is not suitable for close viewing (and there fore use as a second monitor), because of the flickering in the display.

3.2.1 Quality

Besides image quality parameters described in the Display screens section, Projectors vary in portability, the noise they make while operating due to cooling needs, the brightness of the image, lamp life, energy consumption.

3.2.2 Best and economical

Image brightness of atleast 1400 is good (2300 was tested successfully). Lamp life should be atleast 2000 lumens.
For 340$, good projectors are available.

4 Input

4.1 Gaming input

4.1.1 Motion detection

Microsoft Kinect detects hand motions at a certain optimum distance using a camera - no external devices are required.

4.2 Containment/ Size

4.2.1 Computer case

The case/ tower provides power-supply and cooling. Good cooling systems try to be noiseless - they rely on airflow or ensure that the fan speed is usually low.

4.2.2 Size

Computer's processing and storage components are sometimes fit into the monitor or under keyboard in case of a notebook computer, or in a small case.

4.2.2.1 Best and economical

Demand and competition ensure that fairly powerful laptops are now available at a price comparable to equally powerful desktop machines.

5 Interfaces and networking

5.1 Device interfaces

Common standards are often used to connect a wide variety of device pairs.

5.1.1 Universal serial bus

USB is currently most popular. USB 3 enables superior data-transfer speeds. USB allows for a tiered star topology: so one can connect many devices to a particular device with a single usb wire branching out into many others. Power dissipation limits the extant to which this can be done.

5.1.2 USB Plugs

USB wires have 4 (sometimes 5) terminals arranged in various shapes.
Type A is most common.
Micro B is a small trapezoid - often used with pocket computers.
Mini A and B are also sometimes used.
Type B is used with monitors.

5.2 Modems

Modems provide a link to a wider network - usually the Internet. They differ based on the means of communication used in providing this link, and the links it can provide to devices seeking connection to the wider network.

5.2.1 Mobile Wifi (MiFi)

These connect to the wider network using 3G or 4G compatible protocols; and allow devices to connect to it using WiFi protocol - for which reason, they are called 'hotspots'.

6 Special computers

6.1 Pocket computers

6.1.1 Components

These devices are very compact, and often combined with specialized equipment like ability to connect to certain mobile phone networks, GPS, gyroscopes, compasses, cameras, flash led's, accelerometers. Such components often make single function devices much less attractive.

6.1.2 Customizability

The hardware customizability is often very limited; but communication interfaces (eg: bluetooth, card slots) enable expansion.

6.1.3 OS

The operating systems used are often not easily switched; but may be upgradeable or hacked in minor ways. Popular OSes are Apple iOS, Windows, HP's OS, Blackberry OS, Google's Android (often allied with HTC sense UI). Applications/ programs for these operating systems can be bought/ distributed on the internet.

6.1.4 Quality

A major feature of the device is its interface. Touch screen devices which recognize gestures were pioneered by Apple's iphones.

6.1.4.1 Best and economical (2011)

Small size of hard-disks enable provision of large internal storage space.
Popular features: 1GHz processors with 512MB RAM, 5 to 8MP cameras, atleast 8GB storage, 800*480 screens, video cards capable of 20 frames per second.
Current devices weigh around 4.5 oz.

7 Procurement and maintenance

7.1 Maintenance

7.1.1 Fan vent cleaning

In case of a laptop, the fan-vent needs to be cleaned periodically. CPU temperature can be monitored in the GUI using common applications.

7.1.2 Battery

If laptop is always connected, best to disconnect battery [**Check**].

7.1.3 Vostro 1000 disassembly procedure

Remove panel above keyboard. Disconnect keyboard, wires to monitor, trackpad. Remove screws in the front and in the back. Remove the panel from above the keyboard.

7.2 Upgrades

7.2.1 Understand needs

In order to understand what needs to be change, it is good to understand resource limitations: is working memory low, or is the CPU too slow. Hence, run a system monitor application.

8 Parallel computers

8.1 Architecture

Possible 4 uses of multiple processing units were named by Flynn as single/ multiple instructions single/ multiple data-streams.

8.2 Communication

Shared memory vs message passing architectures. Can use shared memory abstraction over message passing, vice-versa.

9 Number storage formats

9.1 Design factors

9.1.1 Words and bits

Suppose b bits are available to store a certain type of number: this is usally expressed in terms of multiples of words (collection of w bits). Because binary logic is used for addressing and processing, w is generally a power of 2. In old days, it used to be $w = 2^2$, now 64 bit words are common.

9.1.2 Simplicity of computation logic

Some representation formats require simpler and more efficient circuits for performing basic arithmetic operations than others. This is an important factor in choosing the representation format.

9.1.3 Special numbers

ONe may want to reserve space in representation set for storing special numbers like +Inf, -Inf, NaN (for storing results of illegal operations).

9.2 Integers

Suppose b bits are available to store a number.

9.2.1 Unsigned

Any $x \in \{0\} \cup N$ can be stored in the natural binary representation. So, in b bits, 2^b unsigned numbers $x \in (0 : 2^b - 1)$ can be stored.

9.2.2 Signed

While storing negative numbers along with positive numbers, one has to distinguish it from positive numbers, one requires a sign bit.

9.2.2.1 Use sign bit + absolute value

A straightforward way to store $x \in Z^-$ is to set the sign bit to 1 and store $-x \in N$ in the remaining $b-1$ bits. In this case, $x \in (-2^{b-1}+1 : 2^{b-1}-1)$ can be stored. Note that there is redundancy in possible representations of 0: the sign bit may be 1 or 0 (one of which can then be taken to mean NaN). Thus, a total of $2^b - 1$ numbers can be stored.

9.2.2.2 Offset/ biased storage

One can take a large natural number called k the bias. Then, one can store 2^b (as against $2^b - 1$ in another representation) numbers $x \in -k : 2^b - k - 1$ by simply storing $x + k \in \{0\} \cup N$.

9.2.2.3 1's complement storage

Suppose we use the bias $2^b - 1$ to store $x \in Z^-$. This amounts to inverting bits in the binary representation of $-x \in N$, is called (one's) complement representation of $x \in Z^-$.

9.2.2.4 2's complement bias

Here again, we used biased representation only to store -ve numbers, we store $2^b + x = 2^b - |x|$: this is the 2's complement of x - we discuss this below.
If b=3, the numbers $-4 : -1$ have representations $100 : 111$.
Note that this representation of x effectively constitutes the use of the most significant (b-th) bit as a sign bit - distinguishing $x \in Z^-$ from $x \in \{0\} \cup N$. Thus, we can store $x \in -2^{b-1} : 2^{b-1} - 1$: a total of 2^b numbers.
Addition of -ve and +ve numbers (ie subtraction) becomes slightly easier: the circuit used to add two unsigned numbers will work fine.

9.3 IEEE floating point

9.3.1 Division of bits

The bits provided for storage are divided into the following components: 1 sign ($\pm$) bit, M bits to store part of the mantissa, E bits to store a function of the exponent. These components are described below.

9.3.2 Number stored

This imitates scientific notation of numbers: $1.2332 * 10^{-9}$. Stores $\pm(1 + f/2^M)2^{e+2^{E-1}-1}$. $\pm(1 + f/2^M)$:= mantissa, f:= significand or precision; $e :=$ exponent stored in the biased representation; $k = 2^{E-1} - 1 :=$ exponent bias. Note that rather than use a sign bit in the exponent, the biased representation is used.
As scientific notation is used, 1 in (1+f) assumed, so the number of bits needed is effectively reduced by 1 bit!

9.3.3 IEEE standards

Single precision: M= 23 bits, E = 8 bits.
Double precision: M= 52 bits, E = 11 bits.

9.3.4 Reserved numbers

0 := is stored as $\pm 1\ 2^{-k}$, which amounts setting all non-sign bits to 0: note that -0 and +0 are distinct (to indicate different underflow conditions while performing arithmetic).
$\pm\infty = \pm 1.0\ 2^{k+2^{E-1}}$, which amounts to setting f=0, exponent bits being 1111...
NaN = $\pm 1.f 2^{k+2^{E-1}}$: identical to $\pm\infty$, except $f \neq 0$.

9.3.5 Range

Allowing for the reserved numbers and considering the range of M and E bits, we can observe the range.
Smallest non 0: $\pm 1.0..01\ 2^{-k}$. Largest num: $\pm 1.11...\ 2^{k+2^{E-1}-1} = \pm 2^{2^{E-1}}$.
So, underflow or overflow rare.

9.3.6 Increasing gaps in different ranges

In [1,2]: 2^{-M}; In [2, 4]: 2^{-M+1}; In $[2^j, 2^{j+1}]$: 2^{-M+j}; relative gap only 2^{-M}. So, 'Floating point'. Matlab eps: (num next to 1) - 1: $2^{-M} = 2^{-52} \approx 2.2204\ 10^{-16}$.

9.3.7 Representation Accuracy

Let $fl : R \to Q$ be a function which maps any number to its floating point representation.

9.3.7.1 Machine epsilon

In case of floating point representations, we can guarantee that $\forall x \in \mathbb{R}$, given that x in range of floating point number system: $\frac{fl(x)-x}{x} \leq \epsilon$.
In case of a floating point number system with M mantissa bits, $\epsilon_M = 2^{-M-1} = 2^{-53}$. This is because, in storing the number $1.f * 2^t$: 1] We assume that

sufficient bits are available to store the exponent, 2]M bits are available to store f.
Yet, note that $fl(\epsilon_M) = \epsilon_M$.

9.3.7.2 Error guarantee view!

Then, roundoff error $[fl(a \odot \epsilon) = a]$ guaranteed. So, $fl(1 + 10^{-16}) = 1$.

9.3.8 Accuracy of arithmetic operations

IEEE ensures: $fl(x \odot y) = (x \odot y)(1 + \epsilon), \epsilon \leq \epsilon_M$ if $\odot = + - / \times$. See this by finding $x(1+\epsilon) \odot y(1+\epsilon)$. $fl(x \odot y)$ is written as $\oplus \ominus \otimes$.
For complex numbers $\otimes$ and div, use $\epsilon_M = 2^{-M-2}$.