COMP253 — Lecture 14: **Performance Factors**

If all computers came from the same manufacturer, it would be simple to compare two computers and select the "best" one. But not all computers come from one source (except for Apple, of course).

Further, what does it really mean for one system to be "better" than another? The answer is in the factors being compared and the purpose the system will satisfy (along with another important factor: $$$$).

The topics below are general, and in no way complete; the goal is provide a foundation for understanding how the aspects of a system affect the overall performance. You are encouraged to consult the course text for a detailed discussion of each.

1. Megahertz (MHz) and other Speed Ratings

- describes the number of "ticks" of the clock crystal on the motherboard, or the number of individual "power cycles" within a component,
  - one (1) data set can move from one device to another in <u>one mainboard "clock tick"</u>; in which case the "megahertz" rating refers to the mainboard (internal, system bus) speed
  - [essentially] one (1) instruction or data exchange can occur in <u>one CPU "clock tick"</u>, but CPU speeds are multiplications (hence, based) on the crystal speed setting (pg **187** (4th), **119** (3rd))

- specifically, in describing the internal speed of the CPU,
  - one "step" in the execution cycle is performed in <u>one CPU clock tick</u>.
  - the CPU must slow down to the motherboard (internal bus) speed to access other devices and RAM
  - some CPUs are capable of performing more than one instruction per "tick"

- for the system bus (ISA, PCI, etc.) (pg **207** (4th), **142** (3rd))
  - "front-side" bus (inside the CPU); "back-side" bus (external to CPU)—the *system bus*
  - controlled by the chipset (bus controllers) that are divided into,
    - "North Bridge" bus—between the CPU, RAM (and AGP) and the system
    - "South Bridge" bus—between the PCI, ISA expansion slots, and USB and IDE busses
  - expansion slots run at different (slower) speeds than the motherboard because of <u>legacy designs</u>
    - ISA: 5/8 MHz; EISA: 8 MHz; VL-Bus: 33 MHz; PCI: 33/66 and 100/133 MHz
    - mainboard speeds also vary, with different speeds of the North Bridge and even the RAM slots

- different CPU-performance rating systems: **MIPS** (millions of instructions per second) and **FLOPS** (floating-point operations per second).
  - the best rating: "number of <u>actual instruction</u> executions per second, *regardless of speed rating*"
  - popular in determining the effectiveness of a complete CPU+system bus+memory system
  - although closer indicator to "true performance," it is not used widely in CPU comparisons, since each CPU may execute different instruction sequences differently; for example, CISC vs. RISC processors, or primarily sequential instruction execution vs. mainly parallel instruction execution

- other ratings (seen in popular diagnostic utilities, such as **HwInfo** or **Norton System/Speed Index**)
  - Dhrystones (integer calculations per second) and Whetstones (floating-point calculations per second)
    - all such metrics are *very specific* to the diagnostic performing the calculations; diagnostic utilities from different companies, both measuring the same aspect, may result in different results on the same system—the approach should be: *use the same diagnostic on different systems and compare the results*
  - "binary exchange" – # of bits (Kbps, Mbps, Gbps) or bytes (KBps, MBps, GBps) exchanged per second
    - used in describing the speed of communication (network, telephone modem, wireless)
  - "data throughput" – number of bytes (or Kbytes, Mbytes) transferred per second
    - used when describing video, peripheral bus communication, or storage device read/write speeds
  - for hard disks and other "disk" storage systems
    - "average seek time" - average time it takes for the head to locate a particular track/sector (in ms)
    - "average access time" - the total time to position head at a track/sector and read the data (in ms)
      - ms - milliseconds, thousandth of a second; average hard disk access time between 9 and 12 ms
    - Cache memories and the OS's swap/virtual memory techniques may distort these ratings; for example, making the system appear better in Windows than it does in DOS

2. Memory Speed

- along with the CPU and system buses, memory is the one of the most important components in the system
- essentially, the faster the available RAM, the more efficiently the CPU can perform its functions (which is the whole purpose of Cache: to make the RAM *seem faster*; see below)
- memory is measured in nanoseconds (ns): one billionth of a second
  - measures of how quickly a particular memory location can be accessed for read or write — the smaller the number, the better the RAM performance
- each memory type follows its own standard for speed and capacity,
  - most SIMMS work at 80, 70, 60, or 50 ns (lower number => faster access)
  - DIMM memory comes in many forms with speeds based on multiples of the system bus speed

- "dynamic RAM" (DRAM) must be "refreshed" (statically recharged), which means that a "clock cycle" is required on the system bus during which time no memory access is possible; this is called a wait state (data can not be transferred during a wait state),
  - most 486 and Pentium-class systems (and PowerPC) can refresh memory in parallel with a *non-RAM* data transfer (such moving data from the CPU to the video system), or during a *RAM access for write*
  - Static RAM (SRAM) requires no refresh, since it maintains values until modified; a great example of the use of SRAM is Cache RAM (L1, L2, and L3); but SRAM still requires power and is not permanent
- Cache increases overall performance by offering an intermediate memory that does not require refresh (it is SRAM) along with working at a much faster speed (20 ns or lower)
  - cache is also available on the physical hard disk, but controlled by the disk logic not the mainboard controllers, and so is not given an *Lx* number
- although new memory forms (such as, SD-RAM, DDR-RAM, RDRAM) are rated based on system bus speeds (or some other "external" speed), the chip on the memory modules still have *nanosecond ratings*

3. Storage Devices

- with the increased need of virtual memory, hard disks with fast spins and access times are a must for quick memory swaps,
  - along with fast access, virtual memory techniques also demand ample disk space
  - in Windows (9x, NT/XP), the swap space is just a file that changes size depending in memory demands, a technique that is not as efficient (or fast) as the "dedicated swap partition" concept used in UNIX/Linux

- a disk (hard disk, CD, DVD) is rated in terms of seek time and access time (see above)
- to increase the access time, hard disks also incorporate a small amount of cache onboard the logic circuitry to reduce the need for continuous physical seeks, reads, and writes (the cache stock piles numerous accesses)
- for microcomputer hard disks, the average measure of access time is between 9 and 12 ms
- **RAID** ("redundant arrays of inexpensive/independent disks") incorporates two (2) or more disks working in tandem (together) to create a "logical disk" that is more efficient, and secure, than a single disk (pg **368** (4th))
  - there are 8 single RAID levels (0-7), with each organising data to be written & read across multiple disks
  - multiple (nested) RAID levels combine two single levels to form an even more robust RAID config.
  - to learn about RAID, start at: *http://techrepublic.com.com/5100-6255-1043757.html*

4. Video Output

- video is probably the most used (and vital) output device, so a reliable and fast display unit is a must
- a "video bottleneck" occurs when the rest of the computer must wait while the video system is producing an image on the display screen; to reduce such wait times,
  - offload more functions to the video processor, freeing the bus and CPU for other tasks
    - this is done by using a video card with a video processor that can be directly accessed by the video API (application program interface) of the OS; examples are cards that support Direct-X or OpenGL
  - increase the video memory (VRAM) to allow for multiple display pages to be prepared during the time one is being displayed, (clearly, this requires more memory)
    - most modern video cards have 128MB, 256MB, up to 1 GB or VRAM, where only a part is used for display and the rest is used for image processing
  - the screen resolution and bit-colour depth demand a specific amount of memory for display
    - the screen is viewed in 3 dimensions: horizontal pixels (X), vertical pixels (Y), colour-depth (Z)
    - at any resolution (X, Y) and colour-depth (Z), a specific amount of memory is needed to retain the complete screen display descriptions (and the higher the resolution, the longer it takes to draw),
    - 800 x 600 @ 16-bit [standard] colour: 800 * 600 * 16 / 8 = 960,000 bytes ~ 1 MB VRAM
      800 x 600 @ 24-bit ["true"] colour: 800 * 600 * 24 / 8 = 1,440,000 bytes ~ 1.5 MB VRAM
      800 x 600 @ 32-bit ["full"] colour: 800 * 600 * 32 / 8 = 1,920,000 bytes ~ 2 MB VRAM
    - 1024 x 768 @ 16-bit [standard] colour: 1024 * 768 * 16 / 8 = 1,572,864 bytes ~ 1.5 MB VRAM
      1024 x 768 @ 24-bit ["true"] colour: 1024 * 768 * 24 / 8 = 2,359,296 bytes ~ 2.3 MB VRAM
      1024 x 768 @ 32-bit ["full"] colour: 1024 * 768 * 32 / 8 = 3,145,728 bytes ~ 3 MB VRAM

- modern video cards segment memory into multiple "screen pages" (at least 2) for quick refresh display, along with memory required by the video processor for calculations and graphic details,
  - to calculate the minimum required display memory required by a video card, consider the following example:
    - multiply a single page memory requirement by (*min. pages*1.5)*:
      - 2 display pages: 1024 x 768 @ 24-bit with 2 pages: 2.3 * (2*1.5) ~ 7 MB VRAM
      - 4 display pages: 1024 x 768 @ 24-bit with 4 pages: 2.3 * (4*1.5) ~ 14 MB VRAM
    - add to this the memory for 3D image calculation, texture mapping, etc. (see below)
  - this is the reason why graphic-intensive games require a minimum of 64 MB of VRAM

- rating of video performance,
  - video throughput, measured in bytes, or Kbytes, per second, describing the amount of data accepted by the video system over the internal bus; which can either describe *text throughput* or *graphic throughput*
  - frames per second, measuring the number of individual screen frames (and essentially, graphic memory pages) displayed in one second—a high rating indicates smooth animation and video playback

- 3D, high-performance video cards require an extra consideration in terms of polygon construction, transformations, texture mapping, lighting effects – "how fast to draw and reshape a 3D image"
  - 3D graphics requires an even faster video processor and larger amount of VRAM than regular use
  - as video processors advance, less work is required by the CPU for decomposing the screen image; new video systems from ATI and Nvidia require only the most basic instructions from the CPU (*details for how graphic processors have advanced is left for students interested in this area*)

5. Bus Width

- defines the number of bits transferred across the bus, in one direction, during one mainboard "clock tick"
- buses on motherboard,
  - control bus - indicates whether the address (in address bus) and data (in data bus) are for read or write from the CPU
  - data bus – carries the data being transferred in 16, 32, or 64 bit-widths, but this width can be doubled (or quadrupled) by producing a series of multiple data passes
    - a bottleneck on data transfer usually occurs when passing through an expansion slot or port, since the width is reduced from the full data bus: (ISA-8/16 bit, EISA/VL-Bus-32 bit)
  - address bus - contains the memory address (RAM location) being accessed for read or write
    - width of the address bus defines the total addressable memory by the CPU (and hence, the system),
      - PC/XT—20 bits; AT—24 bits; 386/486/Pentium (I-IV)—32 bits
      - formula: $2^{(\textit{bus width})}$ = *total addressable memory (in bytes)*
        [to calculate memory in KB, MB, or GB, divide above result by 2^{10} *(KB),* 2^{20} *(MB), or* 2^{30} *(GB)]*
        - ex: PC/XT: $2^{20}/2^{10}$ = 1024 KB or $2^{20}/2^{20}$ = 1 MB of total addressable RAM
          ex: AT: $2^{24}/2^{10}$ = 16,384 KB or $2^{24}/2^{20}$ = 16 MB of total addressable RAM
    - the total addressable RAM does not imply that the memory can be physically attached,
      - the *address bus* defines the maximum logical RAM that could be accessed (theoretical)
      - the bus architecture and BIOS define the maximum physical RAM that can be added

- new CPU and chipsets combinations exist with address buses higher than 32 bits (such as the Intel Itanium and AMD Hammer series)
  - advanced concepts from Intel, IBM, and Motorola are suggesting *virtual addressing schemes*, that allow for addressing beyond the limit placed on the actual address bus
    - a possibility is using a scheme similar to how the data bus creates a series of data passes (to have a 32-bit bus act like a 64-bit bus), providing two addresses passes (one for lower-order bits, and one for high-order bits); so 64-bit address bus would act like a 128-bit address bus

6. Advances

- although currently beyond the scope of the course (but probably not for long),
  - common parallel CPUs architectures, in which multiple (2 or more) CPUs work together in the same system, or within a group of systems
  - distributed processing, in which multiple systems are connected together to share resources: CPU, RAM, disk, and network access; examples, are <u>clustering</u> and <u>grid technologies</u>
  - secondary storage running of serial ATA (or serial SCSI) interfaces, with advances leading to directly shared secondary storage that does not require a regular network connection
  - complete "all-in-one" mainboard systems that incorporate all processing, communication (network), video, multimedia, removable Flash memory, on a single mainboard, eliminating the need for expansion card slots and large cooling systems
  - removable memory systems (such as USB drives, and USB card readers) that eliminate the need for floppies, CD-ROMs, or other disk devices
  - wireless technology, beyond just networking, to allow for all system components connected together using radio or infra-red communication technologies—*the user would never have to plug anything in!!*
  - finally, acquiring the ability to *convincing management that a system with a 2.5 GHz Xeon CPU, 8 GB RAM, 2 GB video, and a 1000 Mbps network is <u>not required</u> to run a word-processor or just surf the Internet…although it would probably play a mean game of Quake Arena or Unreal Tournament!*