Assignment (Video- 7 to 9): Database System Impl. (COP6726)

Name: Vikas Chaubey**, UFID:** 3511 5826**, Email:** vikas.chaubey@ufl.edu

1) **Caches and Cache Lines:** Cache is the second closest memory level to the processor after register. A CPU cache is a hardware cache used by the central processing unit (CPU) of a computer to reduce the average cost (time or energy) to access data from the main memory. There are various levels of cache memories available such as: L0 cache, L1 cache, L2 cache, L3 and L4 caches. As increase the cache levels and move away from processor the cache capacity is increased but the access speed is decreased. On the other hand, a primary memory of CPU is where all the data which is being processed is kept. Data fetched from hard disk or magnetic disks is kept in primary memory for processing. Generally, the size of primary memory could be in Gigabytes and they have high access times compared to registers and cache memories. The closest memory to the processor is registers and then comes caches and finally the main memory. This is the memory hierarchy in the central processing unit of the computer machine. Closer the memory is to the processor, faster the I/O operation performed on memory by processor. when any data is loaded into main memory and it is frequently processed by processor in case of repetitive access then this data is copied in cache memory as well so that access to the main memory could be avoided and the access of data can be made faster by processor. This way processor latencies could be avoided, and program execution can be made faster. The data transfer between the main memory and cache memory is done in form of data units, these data units are called "Cache Lines". Basically, whenever the processor reads from the main memory and write it to cache memory it will read or write this data in form of a chunk of data this chunk is basically constitutes a cache line. A cache line is generally 64 bytes, the processor reads and writes an entire cache line when any location in the 64-byte region is read and written.

2) **Hard Disk and Disk Pages:** An HDD or data storage device is the main storage device in a computer with high storage capacity up to terabytes. A hard disk is the memory device with slowest access time in the memory hierarchy. whenever any data or program needs to be processed in the central processing unit, then the data is loaded from the main memory into the main memory. When data is processed by the processor. then the data is read from the hard disk in form of disk pages. A disk page is the self-designated smallest unit of transfer between main memory and disk. Hard drives have a minimum unit of transfer - sector which is 512 bytes. Most hard drives have a native 32K Page. A particular technique to speed things up is to pack more useful things within the same unit of transfer. Page size is generally dependent on the processor. A single processor may allow various sizes of pages to make the process execution faster. This process of pagination is also called swapping. The swapping size depends on various factors such as page table size, TLB usage, internal fragmentation and disk access. If page size is small then it means there would be a greater number of pages, hence in order to maintain these pages a page table is maintained, as the number of pages increase the

page table size is also increased. Every access to the memory is mapped from virtual to physical address hence reading the page table every time can be very costly, hence a very fast type of cache is used which is called translation lookaside buffer, the page size also depends on the TLB buffer size and usage. Another factor is the internal fragmentation for example there is always a possibility of the last page of data to be not filled fully , in this case if the page size is large then in this case a lot of space is wasted as more unused portions of main memory is blocked and not being used.one way to resolve this issue of fragmentation is to use smaller page size so when a memory block is associated for a page, even if it is not full with data it does not waste a lot of memory space on main memory which could be used for other tasks.

3) **Virtual Memory:** Virtual memory is a memory management technique in the modern computers, using this technique when there is a shortage of main memory to execute and process data or programs then secondary memory can be used in virtual manner as if it were a part of the main memory. virtual memory technique is a very important technique used in the operating system of modern computers. This virtualization of memory is achieved using the hardware and software of the computers, to compensate for the physical memory shortage the data is transferred from the random-access memory or main memory to the disk memory. In this whole process the disk memory is utilized as the main memory. This is done in the systems where the random-access memory is not enough or very less. Virtual memory can be used to transfer the data from random access memory to disk memory, generally this data is not used enough in recent time and hence computer can afford to transfer it temporarily to the disk memory which can free up space on main memory for storage of other important data which has to be processed immediately. Virtual memory can improve system performance significantly when the computer is running low on main memory. It can be used in case of multitasking or while running large programs and to improve flexibility. However, over reliance on the virtual memory is not good idea because virtual memory is not as fast as Main memory hence it can make a process execution feel very slow. This is called thrashing.

4) **How Virtual Memory Works:** Virtual memory uses both computer hardware and software to work. when a program or data is being processed then it is loaded in physical address location in the RAM main memory. Specifically, virtual memory will map that address to RAM using a memory management unit (MMU). the Operating system manages the memory mappings by using page tables and other data structures are also used. The memory management unit of OS acts as a address translator hardware and it automatically translates the address. If at any point, there is shortage of main memory then least used data already present on the main memory is transferred to the virtual memory. The memory management system of OS is responsible to keep track of shifts or swaps between the physical and virtual memory. If data is needed again a context switch could be used to process that data again. While copying virtual memory into physical memory, the OS divides memory into page files or swap files with a fixed number of addresses. Each page is stored on a disk, and

when the page is needed, the OS copies it from the disk to main memory and translates the virtual addresses into real addresses. However, the memory swap between physical and main memory is slow. This means using virtual memory can cause significant reduction in performance of the system.