**Assignment 3**

**Part 1: Understanding Memory Hierarchy**

**Memory Hierarchy Design in Achieving High-Performance Computing Systems**

**Introduction**

In a computerized system, there are different types of storage or memory devices. These mentioned storage devices are there based on their costs, size, access speeds and their roles in operating an application process. Thus, a memory hierarchy has been introduced so that all of the available memory devices are arranged together and minimize the access time for ease of use and speed up application execution. Considering the memory hierarchy in a computerized system, different storage devices like CPU registers, main memory, cache memory, optical disks and magnetic-type disks are usually available. The arrangements of these memory devices are done based on their access as well as response speeds (Ayers, 2018).

**Memory Technologies**

Memory technologies are having a huge impact on the performance of application processes which are placed in different levels of the memory hierarchy in a computer system. Memory technologies are spreading from fast but expensive (Static Random Access Memory or SRAM) to slow but cost-cutting technologies (Dynamic Random-Access Memory or DRAM). Considering SRAM which is known for its low latency and speed is used to build cache memory. This sort of memory is costly thus it is not suitable for large-scale storage but its features make it ideal for accessing information rapidly as well as frequently. On the other hand, DRAM is known for its slower access time and cost efficiency. Thus, the main memory of a computer system is using these sorts of technologies as it requires large-scale storage. Apart from that the usage of Non-Volatile Memory or NVM technologies are also used in memory hierarchy designs. Mostly NVM technology is used in secondary storage options like SSDs (Stevens, 2013).

**Advanced Cache Optimization**

The cache acts like a bridge between the CPU and the main memory in terms of speed. In memory hierarchy, Advanced Cache Optimization techniques are utilized to mitigate the chance of cache misses so that the computer system’s throughput can be improved. In Advanced Cache Optimization techniques, prefetching, victim cache, cache partitioning etc. are used to reduce wait times, access recently used data and improve efficiency. Moreover, with the utilization of the Advanced Cache Optimization technique, the effectiveness and efficiency of the cache are enhanced and thus less time is required for accessing the required data and processing becomes comparatively faster.

**Virtual Memory and Virtual Machines**

Virtual memory is a concept that allows users to access more memory than the physically available memory in a computer system. Mostly, these techniques are created with the help of page tables, address translation, page replacement algorithms etc. mechanisms. These mechanisms allow systems to enable efficient memory management by mapping the virtual memory addresses to physical addresses. The address translation mechanism helps in the conversion of virtual addresses to physical addresses where the Memory Management Unit (MMU) plays an effective role. Least Recently Used (LRU) and First-In-First-Out (FIFO) algorithms are helping systems determine which pages need to be omitted whenever a new page is needed for operational purposes. On the other hand, the Virtual Memory (VM) supports the isolation of the memories required for multiple process execution and provides security and stability within the system. With the help of this VM, users are allowed to use multiple operating systems on a single physical machine and each of the operating systems has its own virtual memory hierarchy (Basu, 2013).

**Cross-Cutting Issues**

There are several factors that are included in designing a memory hierarchy. In this scope, several factors like cost, complexity, power consumption, workload etc. need to be balanced for a system. For example, high-performance memory technologies like SRAM are costly though they are very efficient. So, design must be there considering the budget as well as the performance of a system. Nowadays, machine learning-based memory management and usage of NVM-like memory technologies create a revolution for the future of computer systems which promises greater performance as well as efficiency in all aspects.

**Conclusion**

Memory hierarchy design is a concept where high performance within a computer system can be achieved. Various mechanisms, as well as techniques, are hereby incorporated to make the entire memory within a budget without compromising the efficiency of the system. Also, these techniques of memory hierarchy can create less complex and faster execution of processes in a computerized system.

**References**

Ayers, G., Ahn, J. H., Kozyrakis, C., & Ranganathan, P. (2018, February). Memory hierarchy for web search. In *2018 IEEE International Symposium on High Performance Computer Architecture (HPCA)* (pp. 643-656). IEEE.

Basu, A., Gandhi, J., Chang, J., Hill, M. D., & Swift, M. M. (2013). Efficient virtual memory for big memory servers. *ACM SIGARCH Computer Architecture News*, *41*(3), 237-248.

Stevens, J., Tschirhart, P., Chang, M. T., Bhati, I., Enns, P., Greensky, J., ... & Jacob, B. (2013). An integrated simulation infrastructure for the entire memory hierarchy: Cache, dram, nonvolatile memory, and disk. *Intel Technology Journal*, *17*(1), 184-200.