Performance Improvement of Average Access Time by Optimization of Cache Memory Access Time

Mirza Moazzem Hossain[#1]
Core Network Engineer, *Apple GlobalTel Communications Limited (IGW).*
Dhaka, Bangladesh.
E-mail- mirza.moazzem.hossain@gmail.com

Mohammad Shamsul Alam[#2]
Trainee Engineer, *ADN Telecom Limited.*
Dhaka, Bangladesh.
E-mail- shiblyshamsul@yahoo.com

Jamal Uddin Ahmed[#3]
Expert Faculty,
Department of EEE, Ahsanullah University of Science and Technology.
Dhaka, Bangladesh.
E-mail- jamaluddinahmedjua@gmail.com

Mohammad Khairul Bashar[#4]
Assistant Engineer,
AKIJ Group (DTI-LPF).
Dhaka, Bangladesh.
E-mail- basharaust@yahoo.com

Abstract **-** The cache is a small amount of high speed memory. When a program loop is executed, the CPU repeatedly refers to the set of instructions in memory that constitute the loop. If an active portion of the program and data are placed in a fast small memory, the average memory access time can be reduced, thus reducing the total execution time of the program. Such a fast small memory is referred to as a cache memory. It is placed between the CPU and main memory. In this view, we consider in this paper to optimize the cache memory access time which eventually optimizing the average access time.

Index Terms **-** Cache memory access time, Main memory access time, Hit ratio, Miss Ratio, Average access time

I. INTRODUCTION

To alleviate the effect of the rising gap between CPU speed andmain memory performance, today's computer architectures implement hierarchical memory structures. The main aim is to hide both the low main memory bandwidth and the latency of main memory accesses which is slow in contrast to the floating point performance of the CPUs. Generally, a small and expensive high speed memory sitting on top of the hierarchy which is usually integrated within the processor chip to offer data with low latency and high bandwidth, i.e.: the CPU registers. Moving further away from the CPU, the layers of memory consecutively become larger and slower. The memory components which are located between the processor core and main memory are called cache memories or caches. They are intended to contain copies of main memory blocks to speed up accesses to frequently needed data [1, 2].The following lower level of the memory hierarchy is the main memory which is large but also moderately slow.

II. TYPES OF MEMORY

A.CACHE MEMORY

The cache memory access time is less than the access time of main memory by a factor of 5 to 10.The basic characteristic of cache memory is its fast access time. Therefore very little or no time must be wasted when searching for words in the cache. The cache is the fastest component in the memory hierarchy and approaches the speed of CPU components. The fundamental idea of cache organization is that by keeping the most frequently accessed instructions and data in the fast cache memory, the average memory access time will approach the access time of the cache. Although the cache is only a small fraction of the size of main memory, a large fraction of memory requests will be found in the fast cache memory because of the locality of reference property of programs. Analysis of a large number of typical programs has shown that the references to memory at any given interval of time tend to be confined within a few localized areas in memory. This phenomenon is known as the property of locality of reference.

There are three levels of cache memory level 1(L1), level 2(L2), and level 3(L3). Level 1 (L1) is very close to the processor. Data and instructions are used most frequently. Level 2 (L2) is the second closest to the CPU. It is more common to be on the motherboard. This is used for the most frequently used data and instructions. Level 3 (L3) is the most advanced cache and will speed up the memory even further. This is used for the most frequently used data and instructions.

B. MAIN MEMORY

Main Memory is the physical memory that is internal to the computer. The word main is used to distinguish it from external mass storage devices such as disk drives. The main memory in a computer is called Random Access Memory. It is also known as RAM.This is the part of the computer that stores operating system software, software applications and other information for the central processing unit (CPU) to have fast and direct access when needed to perform tasks. It is called "random access" because the CPU can go directly to any section of main memory, and does not have go about the process in a sequential order.

The computer can manipulate only data that is in main memory. Therefore, every program we execute and every file we access must be copied from a storage device into main 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.Because computers often have too little main memory to hold all the data they need, computer engineers invented a technique called swapping, in which portions of data are copied into main memory as they are needed. Swapping occurs when there is no room in memory for needed data. When one portion of data is copied into memory, an equal-sized portion is copied (swapped) out to make room.

Now, we can usually increase the amount of memory by inserting extra memory in the form of chips.

RAM

RAM is the best known form of memory which our computer uses. Every file or application opened is placed in RAM. Any information the computer needs or uses becomes part of a continuous cycle where the CPU requests data from RAM, processes it and then writes new data back to RAM. This can happen millions of times a second. However, this is usually just for temporary file storage, so unless the data is saved somewhere, it is deleted when the files or applications are closed.

RAM is one of the faster types of memory, and has the capacity to allow data to be read and written. When the computer is shut down, all of the content held in RAM is purged. Main memory is available in two types: Dynamic Random Access Memory (DRAM) and Static Random Access Memory (SRAM).Dynamic random access memory (DRAM) is the most common kind of main memory in a computer. It is a prevalent memory source in PCs, as well as workstations. Dynamic random access memory is constantly restoring whatever information is being held in memory. It refreshes the data by sending millions of pulses per second to the memory storage cell.Static Random Access Memory (SRAM) is the second type of main memory in a computer. It is commonly used as a source of memory in embedded devices. Data held in SRAM does not have to be continually refreshed; information in this main memory remains as a "static image" until it is overwritten or is deleted when the power is switched off. Since SRAM is less dense and more power-efficient when it is not in use; therefore, it is a better choice than DRAM for certain uses like memory caches located in CPUs. Conversely, DRAM's density makes it a better choice for main memory.

III.PROCEDURE OF CACHE MEMORY

When the CPU needs to access memory, the cache is examined. If the word is found in the cache, it is read from the fast memory. If the word addressed by the CPU is not found in the cache, the main memory accessed to read the word. A block of word containing the one just accessed is then transferred from main memory to cache memory. The block size may vary from one to about sixteen (1 to 16) word adjacent to the one just accessed. In this manner, some data are transferred to cache so that future references to memory find the required words in the fast cache memory.

The performance of cache memory is frequently measured in terms of a quantity called “Hit ratio”. When the CPU refers to memory and finds the words in cache it is said to produce a “Hit”. If the word is not found in cache, it is in main memory and it counts as a “Miss”. The ratio of the number of hits divided by the total CPU reference to memory (hit + miss) is the “Hit ratio”. Hit ratio of 0.9 and higher have been reported.

The average memory access time of a computer system can be improved considerably by use of a cache. If the hit ratio is high enough so that most of the time the CPU accesses the cache instead of main memory, the average access time is closer to the access time of the fast cache memory.

IV. THEORETICAL ANALYSIS

First consider the following two equations.

HIT RATIO+MISS RATIO=1

Where, Hit ratio of 0.9 and higher have been reported.

Average access time =(Cache access time X Hit ratio) +(Main memory access time X Miss ratio)

We have assumed that the Main memory access time is 1000ns and keeping it fixed and changed the Cache access time from 20ns to 100ns by varying hit ratio from 0.91 to 0.99.

The system described above is simulated using Matlab. The simulation parameters are given in Table I.

Table I: System Parameters used for computation

CACHE ACCESS TIME(ns)	MAIN MEMORY ACCESS TIME(ns)	HIT RATIO	MISS RATIO (1- HIT RATIO)	AVERAGE ACCESS TIME(ns)
100	1000	0.91	0.09	181
90		0.92	0.08	162.8
80		0.93	0.07	144.4
70		0.94	0.06	125.8
60		0.95	0.05	107
50		0.96	0.04	88
40		0.97	0.03	68.8
30		0.98	0.02	49.4
20		0.99	0.01	29.8

Figure 1 shows the plot of Cache access time with respect to Average access time. From the plot it is visible that when the cache access time is higher then the average access time is also high and vice versa. And for cache access time 100ns the average access time is 181ns,for cache access time 20ns the average access time is 29.8ns.

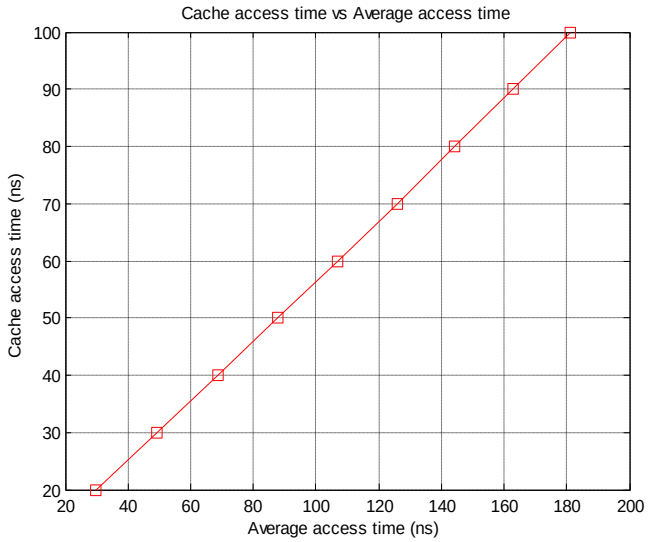

Fig.1: Plot of Cache access time vs Average access time

Figure 2 showsthe plot of hit ratio vsAverage access time. Where for highest hit the average access time becomes lower and vice versa.

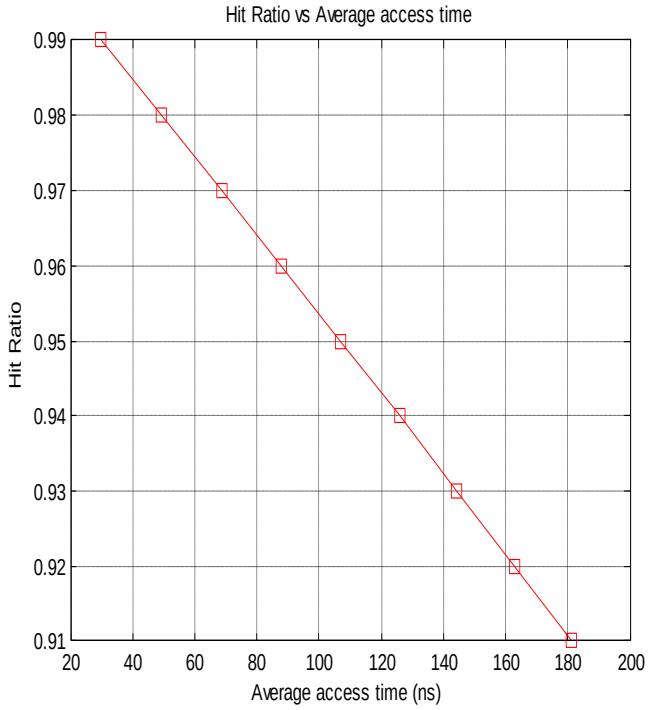

Fig.2: Plot of Hit Ratio vsAverage access time

Figure 3 shows the plot of Miss Ratio vs Average access time, The figure exactly matches to the figure 1 which is Cache access time vs Average access time.

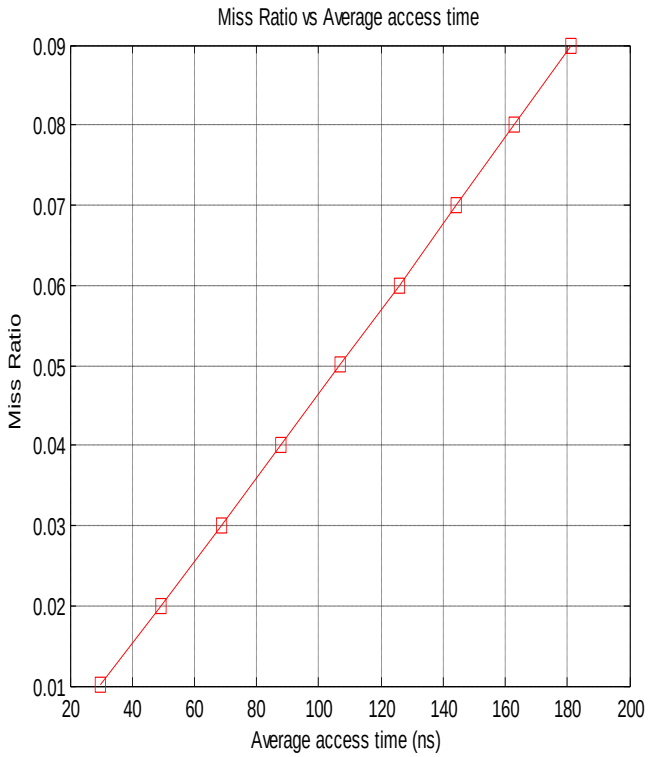

Fig.3: Plot of Miss Ratio vsAverage access time

The last figure depicts comparison of the three value i.e: Cache access time, Hit ratio, Miss Ratio with respect to Average access time. From the figure it is conspicuous that the cache access time crosses the Hit ratio in one point which is the optimum result. The value which we found is 60ns for Cache access time and 0.95 is the Hit ratio and 0.05 is the miss ratio and the most importantly the average access time is 107ns.

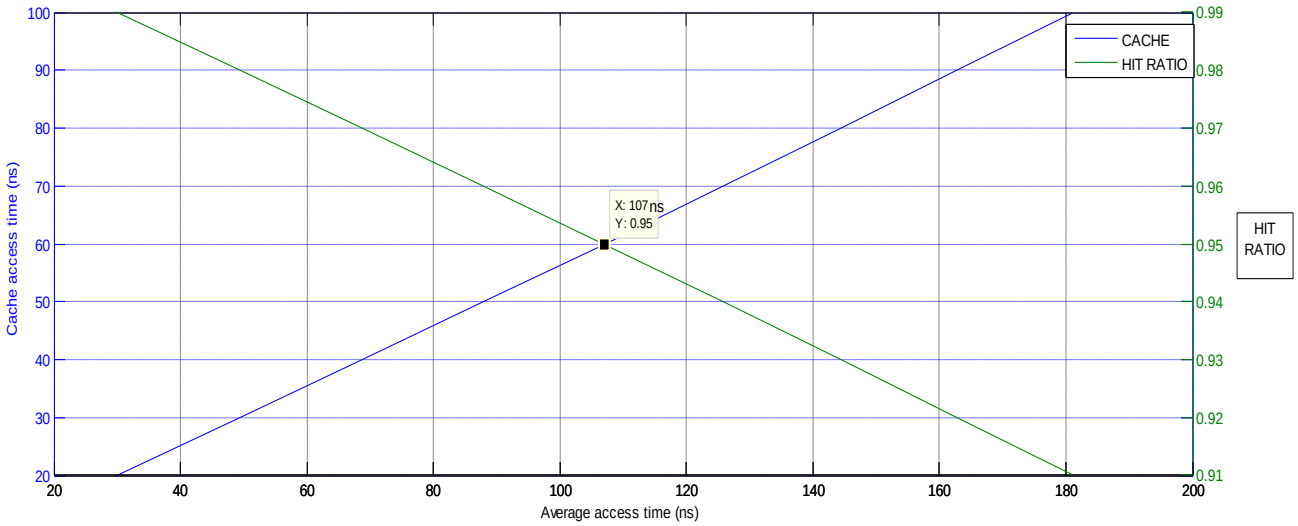

Fig.4 : Plot of Cache access time VS Average access time&Hit ratiovsAverage access time

V. CONCLUSIONS

We have analyzed the three main parts of memory systems, which are Cache access time, Hit ratio, Miss Ratio. By varying Cache access time from20ns to 100ns and hit ratio from 0.91 to 0.99 we have found the most optimum value for average access time which is 107ns corresponds to hit ratio of 0.95 and cache access time 60ns.Finally we can say that if we can reduce the Cache access time to significant level then the average access time will reduce. An important research can be done with the help frompaperas if the cache memory size doubled then the cache memory access time should also double butcache memory access time will probably increase 10% from the previous value.

REFERENCES

1.J. Handy. The Cache Memory Book.Academic Press, second edition, 1998.

2. J.L. Hennessy and D.A. Patterson. Computer Architecture: A Quantitative Approach. Morgan Kaufmann Publisher, Inc., San Francisco, California, USA, second edition, 1996.