M-way Set-Associative and Sector Mapped Cache in Verilog- Design and Testing

EE 275 Advanced Computer Architecture

Final Project Submission

By:

Kamlesh Kumar Student ID- 017047212 December 06, 2024

Abstract

This project focuses on developing an advanced cache simulator to analyze the performance of both m-way set-associative caches and sector-mapped caches across various configurations. These configurations range from simple direct-mapped designs to fully associative and sector-mapped approaches. The simulator evaluates performance metrics such as hit ratio and miss ratio using Verilog, employing the Least Recently Used (LRU) replacement algorithm for set-associative caches and sector-specific strategies for sector-mapped caches. By experimenting with different cache sizes, line sizes, associativity, and sector-mapping techniques, we aim to understand their impact on cache performance and identify key factors that enhance efficiency.

Through detailed simulations and analyses, this study optimizes cache design by balancing hit rate improvement with minimal storage costs. For m-way caches, line size and associativity significantly influence performance, while sector mapping introduces a new dimension of block segmentation and mapping efficiency. Ultimately, the objective is to provide insights into cache operation and configuration, equipping system designers with the knowledge to make informed decisions and enhance overall system performance.

1. Introduction

The primary goal of this project is to develop a highly accurate cache simulator that evaluates the performance of m-way set-associative caches and sector-mapped caches. Cache memory, a critical component of modern computer architecture, significantly reduces memory access time compared to traditional hard drives. By enabling faster access to frequently used data, cache memory plays a vital role in enhancing system performance.

In this project, we present a Verilog-based cache simulator capable of analyzing two prominent cache memory designs:

- (1) **M-way Set-Associative Cache**: Configurable from direct-mapped (m = 1) to fully associative (m = cache size/line size), supporting associativity values such as 2, 4, 8, 16, and 32.
- (2) **Sector-Mapped Cache**: A design that divides cache lines into smaller sectors, optimizing storage efficiency and access speed for specific workloads.

The simulator supports cache sizes up to 128 KB and incorporates the Least Recently Used (LRU) algorithm for block replacement in set-associative caches. For sector-mapped caches, a sector-based management strategy ensures efficient handling of data blocks. To evaluate cache performance, the simulator uses an address trace file where each entry logs a byte address, and differences between consecutive addresses are recorded to optimize storage.

This dual-design simulator focuses on key parameters such as block (or line) size, cache size, associativity, and sector configuration. By analyzing the sequence of memory load instructions executed by programs, the simulator provides comprehensive insights into how these parameters influence performance.

The accompanying report presents a detailed analysis of simulation results, exploring the technical intricacies of m-way and sector-mapped cache designs. This documentation aims to aid developers and researchers in optimizing cache memory systems, advancing the understanding of cache behaviour and fostering innovation in computer architecture.

2. M-way Set Associative Cache

2.1.Design Methodology

Understanding cache design and operation is fundamental to mastering efficient memory management techniques. Cache memory is a specialized, high-speed memory that bridges the performance gap between the CPU and slower memory components, such as RAM and disk storage. While less costly than CPU registers, cache memory is more expensive than main memory but offers significantly faster access times. Positioned as a buffer between the CPU and RAM, it stores frequently accessed data and instructions, enabling the CPU to retrieve information quickly and thereby reducing the average time required to access main memory.

Cache memory operates by maintaining copies of frequently accessed data from main memory in a smaller, faster memory space. Modern CPUs typically include separate instruction and data caches to streamline processing. When the CPU requires data, it first checks the cache. If the requested data is found—a cache hit—the data is retrieved directly from the cache. Conversely, a cache miss occurs when the data is absent, prompting the cache to fetch it from main memory, create a new cache entry, and subsequently serve the request.

For instance, consider a system with a main memory comprising 128 words, where each block contains four words. This setup results in 32 memory blocks, each holding four words. If the cache size is limited to 16 words, the cache accommodates four lines (16/4 = 4), with each line corresponding to a block size of four words. The mapping of main memory blocks to cache lines is determined by the modulo operation (k mod n), where k is the block number, and n is the number of cache lines. In a direct-mapped cache, block 4 maps exclusively to line 0, as $4 \mod 4 = 0$. Similarly, block 13 maps to line 1 $(13 \mod 4 = 1)$, irrespective of the occupancy of other lines.

This mapping principle, illustrated in Figure 1, demonstrates how main memory blocks are allocated to cache lines in a direct-mapped configuration. By efficiently managing these mappings and leveraging cache memory, systems achieve enhanced performance through reduced memory access times.

2.2.Design Description

Blocks of different sizes are used to segment caches. The number of blocks in a cache is typically a power of two. The simplest method is the Direct-Mapped Cache. In Direct Mapped Cache, every memory address correlates to a single cache block. One method for figuring out which cache block belongs at a given memory location is the mod (remainder) operation. The data at the memory address would be delivered to the cache block index if the cache contained 2^k blocks, giving I mod 2^k . At any one time, the CPU only needs a specific quantity of data. The cache memory duplicates the contents of the memory locations ahead of time to identify the range of memory locations that the processor will need shortly. The CPU loads data from the main memory into the cache first when it needs it, and if it cannot find it there, it must wait until the main memory is filled with the necessary data. At this point, data from nearby sites that match the requested address are also transferred to the cache. The basic idea behind cache operation is the locality of reference, sometimes referred to as the concept of locality. According to this principle, every program or application running on a processor ought to only use a tiny portion of the total address space at any one time. Reference locations come in two varieties.

- (1) Temporal locality, or time locality: There is a good chance that a certain data item will be needed again shortly if it is utilized at one point in time.
- (2) Spatial locality: There is a high probability that data items from nearby addresses will be needed shortly if a data item from a particular memory location is used at a given moment.

A cache controller receives the address when the CPU tries to read from memory. A block in the cache will be indexed using the lowest k bits of the address. If the block is valid and the tag matches the top (m-k) bits of the m-bit address, the data will be sent to the CPU. What we've been expecting to be memory delays are cache hits. If the CPU implementations had gone straight to the main memory, the cycle times would have been much longer. However, a much slower main memory access is necessary upon a cache miss. A total of 2n words and n-bit address lines make up the main memory. It is arranged into several blocks, with k words in each block. Consequently, its total number of main memory blocks is $\frac{2n}{k}MB$. Similarly, a few lines, each containing k words, make up the cache memory.

There are far fewer cache lines (CL) than there are main memory blocks (MB) in total. Because of this, only a small number of main memory blocks are mapped to the cache at any given moment. The full block containing the requested data item is mapped to one of the cache lines when the processor starts a read request for data that is not in the cache. A cache line cannot be unique to a single main memory block due to the vast number of blocks compared to cache lines. Consequently, a tag designating the physical location linked to every cache line is assigned. How soon it returns the requested data is what defines the cache's performance. It is computed using the access time and hit/miss ratio. A cache hit occurs when the requested data is found in the cache; a cache miss occurs when it is not. The number of memory accesses that are discovered in the cache relative to the total number of memory accesses requested is known as the hit ratio. One may describe the miss ratio as (1 - hit ratio). Hit time is the amount of time it takes the cache to provide the requested data in the case of a hit. The necessary data is acquired from the main memory and mapped to the cache in the event of a cache miss. The time it takes to get and map the data is called a miss penalty. Conventional methods of enhancing cache performance (by raising the hit or lowering the miss ratio) are generally categorized as

- (1) Increasing block and cache size,
- (2) Increasing associativity,
- (3) Cache probing,
- (4) Supplementing the regular cache with victim cache,
- (5) Hardware data prefetching, and
- (6) Including additional cache hierarchy.

In the direct mapped cache, or one-way-set associative cache, every block from the main memory is allocated to a particular cache line. There are two benefits to the direct mapped cache: a straightforward design and extremely easy hardware implementation. But this architecture's mapping of every block from main memory to a single cache line is its lone flaw. The Direct Mapped Cache architecture is shown in Figure 1.

In set-associative cache design, the cache memory is divided into several tiny, direct-mapped modules, each of which is referred to as a set. A cache composed of M sets is referred to as M-way set associative. The performance of the set-associative cache lies in the center of these two cache architectures as it is essentially a compromise between completely associative and direct-mapped

caches. Two-way and four-way set-associative caches offer the best performance in terms of hit ratio and access time for embedded systems. A 4-way set associative cache's design is shown in Figure 2.

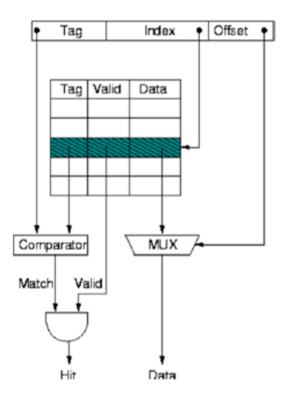


Figure 1: Architecture of Direct Mapped Cache.

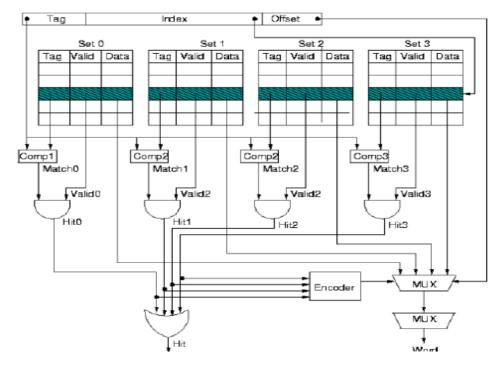


Figure 2: Architecture of 4-way set-associative cache.

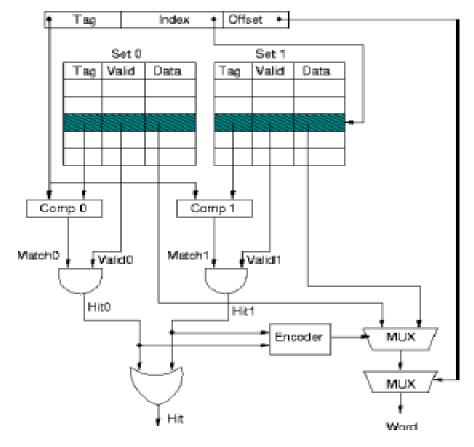


Figure 3: Architecture of fully associative cache.

In a completely associative cache, every block from the main memory can be allocated to any of the cache lines, giving the best possible hit ratio. It also involves the expense, intricacy, hardware, and access time related to looking up the requested location in the cache. Every cache block is compared simultaneously to determine if a requested memory location is present in the cache. The building of a completely associative cache is shown in Figure 3.

2.3. Simulations

First, we compare the miss ratio to the variable cache size. In the first example, with the fixed block size 32 and fixed associativity 4, a line will be formed with the miss rate as the Y-axis and the cache size as the X-axis for the graph.

Data collection and visualization of results for cache sizes of 16KB, 32KB, 64KB, and 128KB:

S.No.	Cache Size (in KB)	Block Size	Set Associativity	Total Access Count	Cache Hits	Hit Ratio	Miss Ratio
1	16	32	4	1500000	1488766	99.25	0.75
2	32	32	4	1500000	1493426	99.56	0.44
3	64	32	4	1500000	1495663	99.71	0.29
1	120	22	1	1500000	1407210	00.91	0.10

Table 1: Miss rate variation with 32 bytes fixed block size, 4-way set associativity, and varying cache size.

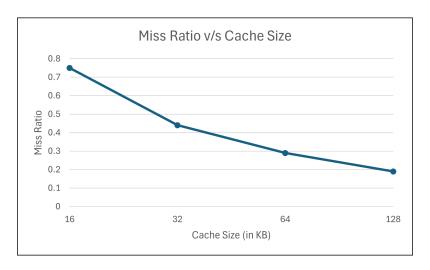


Figure 6: Miss Ratio Vs Cache Size Chart.

Furthermore, it is important to consider the changeable line size of the Miss Ratio. In the second example, given the fixed cache size of 8192 and fixed associativity of 4, a line will be formed with the block size as the X- axis and the miss rate as the Y- axis for the graph. The results are obtained and a graph based on the results for block sizes 16, 32, 64, and 128 is produced, as shown in Fig. 7.

S.No.	Cache Size (in KB)	Block size	Set Associativity	Total Access Count	Cache Hits	Hit Ratio	Miss Ratio
1	8	16	4	1500000	1474170	98.28	1.72
2	8	32	4	1500000	1471126	98.08	1.92
3	8	64	4	1500000	1459207	97.28	2.78
4	8	128	4	1500000	1430504	95 37	4 63

Table 2: Miss rate variation with varying block size, 4-way set associativity, and 8KB cache size.

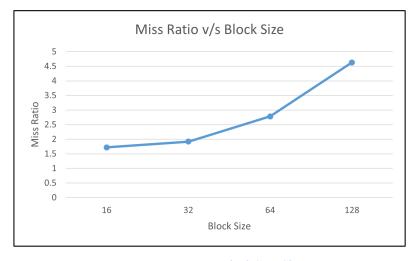


Figure 7: Miss Ratio Vs Block Size Chart.

Third, contrast the Miss Ratio with Variable Associativity. In the third example, given the fixed cache size of 8 KB and the fixed block size of 32 bytes, a line will be drawn with the associativity

acting as the X-axis and the miss rate acting as the Y-axis for the graphs. The results for associativity of 1, 2, 4, 8, 16, and completely associative are generated and displayed, as seen in Fig. 8.

S.No.	Cache Size (in KB)	Block Size	Set Associativity	Total Access Count	Cache Hits	Hit Ratio	Miss Ratio
1	8	32	1	1500000	1417382	94.49	5.51
2	8	32	2	1500000	1456971	97.13	2.87
3	8	32	4	1500000	1471126	98.08	1.92
4	8	32	8	1500000	1476941	98.48	1.52
5	8	32	16	1500000	1479609	98.64	1.36
6	8	32	32	1500000	1480144	98.68	1.32

Table 3: Miss rate variation with 32 bytes fixed block size, 8KB fixed cache size, and varying associativity.

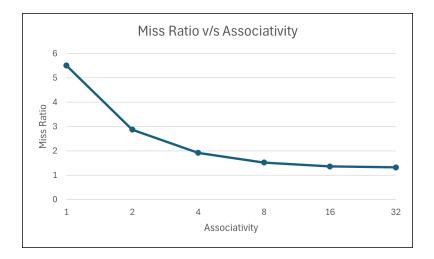


Figure 8: Miss Ratio Vs Associativity Chart.

3. Sector-Mapped Cache

3.1. Design Methodology

Sector-mapped cache design introduces an innovative approach to cache memory management by segmenting cache lines into smaller, independently addressable units called sectors. This technique optimizes storage utilization and reduces the overhead associated with caching large blocks, particularly for workloads where only portions of a block are frequently accessed. Understanding the sector-mapped cache architecture is essential to exploring its benefits and implications for system performance.

In a sector-mapped cache, each cache line is divided into multiple sectors, and each sector can hold a smaller subset of data from a memory block. This segmentation enables fine-grained data storage, allowing only the relevant parts of a memory block to occupy cache space. The key components of sector-mapped cache design include the **sector tag**, which identifies the memory address mapped to the sector, and a **valid bit**, which indicates whether the sector contains valid data.

Cache operations in a sector-mapped cache follow a unique flow:

(1) When the CPU requests data, the cache controller extracts the memory block address and checks the

sector tags within the corresponding cache line.

- (2) If the requested data is found in a valid sector (sector hit), it is retrieved directly from the cache.
- (3) If the requested data is not found (sector miss), the cache fetches the specific sector from the main memory rather than the entire block, minimizing memory traffic.
- (4) Replacement policies, such as Least Recently Used (LRU) or variations adapted for sectors, manage sector replacement within cache lines.

To illustrate, consider a main memory with 128 words divided into 32 blocks, each containing four words. A sector-mapped cache with a line size of four sectors (each holding one word) and a total size of 16 words would have four cache lines. Each line accommodates a block of four sectors, with each sector independently mapped and validated. In this architecture, the block number determines the cache line, while the sector index specifies the target sector within the line.

3.2. Design Description

The sector-mapped cache architecture offers several advantages over traditional set-associative caches:

- (1) **Fine-Grained Storage**: By caching only, the accessed sectors of a block, it reduces the overhead of storing unused data.
- (2) **Efficient Miss Handling**: Fetching individual sectors instead of entire blocks decreases memory bandwidth usage and improves system responsiveness.
- (3) **Enhanced Flexibility**: The sector-based approach adapts well to workloads with irregular access patterns, where only small portions of a block are frequently used.

Sector-mapped caches operate under the principle of **locality of reference**:

- (1) **Temporal Locality**: Recently accessed sectors are likely to be accessed again soon.
- (2) **Spatial Locality**: Sectors close to the currently accessed sector have a high probability of being accessed soon.

The cache controller plays a crucial role in managing sector mapping and replacement. Upon receiving a memory address, the controller decodes it into a block number and a sector offset. The block number determines the cache line, while the offset identifies the target sector. If the sector is valid and its tag matches the memory address, the controller serves the data directly to the CPU, achieving a sector hit.

In the event of a sector miss, the controller fetches the specific sector from main memory. This operation involves updating the sector tag and valid bit, ensuring efficient cache utilization. Replacement strategies such as sector-LRU manage the eviction of sectors when cache space is fully utilized.

3.3. Simulations

As first experiment, we perform performance evaluation of sector mapped cache by comparing miss rate variation with respect to variation in cache memory size. Table 5 presents the results from this experiment and Figure 9 illustrates how miss ratio varies with increment in cache size.

Table 4: Performance of Sector-mapped cache with variation in Cache Size keeping Block Size, Associativity, and Sector Size, constant.

S.No.	Cache Size (in KB)	Block Size	Set Associativity	Sector Size	Total Access Count	Cache Hits	Hit Ratio	Miss Ratio
1	8	32	4	4	1500000	1477927	98.53	1.47
2	16	32	4	4	1500000	1490650	99.38	0.62
3	32	32	4	4	1500000	1495824	99.72	0.28
4	64	32	4	4	1500000	1497872	99.86	0.14
5	128	32	4	4	1500000	1497880	99.86	0.14

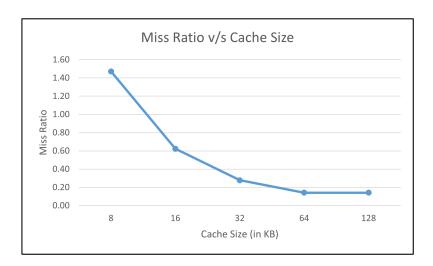


Figure 9: Miss Ratio variation with respect to Cache Size keeping other parameters constant.

Further, we perform performance evaluation of sector mapped cache by comparing miss rate variation with respect to variations in Block Size and Associativity as well, for which results are noted in Table 5 and 6, and Figure 10 and 11. These results demonstrate that miss ratio increases only with increasing Block Size and shows downward trend in all other cases.

Table 5: Cache Performance variation with variation in Block Size.

S.No.	Cache Size (in KB)	Block Size	Set Associativity	Sector Size	Total Access Count	Cache Hits	Hit Ratio	Miss Ratio
1	8	8	4	4	1500000	1476092	98.41	1.59
2	8	16	4	4	1500000	1479373	98.62	1.38
3	8	32	4	4	1500000	1477927	98.53	1.47
4	8	64	4	4	1500000	1497872	97.88	2.12
5	8	128	4	4	1500000	1497880	96.24	3.76

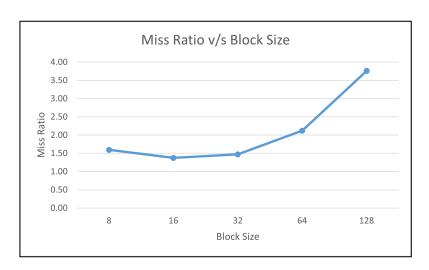


Figure 10: Cache Miss Ratio visualization with increment in Block Size, keeping all other parameters constant.

T 11 (O 1	D C				(C 1 , 22)
Table 6: Cache	Pertormance c	amparisan with	variation in	i associativity i	trom 1 to 3/1
Tubic o. Cuche	1 cijoi manee e	omparison wiin	variation in	associativity (ji 0111 1 to 52).

S.No.	Cache Size (in KB)	Block Size	Set Associativity	Sector Size	Total Access Count	Cache Hits	Hit Ratio	Miss Ratio
1	8	32	1	4	1500000	1419474	94.63	5.37
2	8	32	2	4	1500000	1458418	97.23	2.77
3	8	32	4	4	1500000	1477927	98.53	1.47
4	8	32	8	4	1500000	1488655	99.24	0.76
5	8	32	16	4	1500000	1490635	99.38	0.62
6	8	32	32	4	1500000	1492758	99.52	0.48

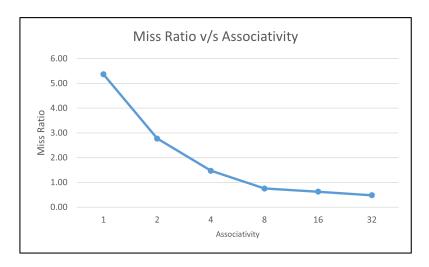


Figure 11: Miss Ratio v/s Associativity curve keeping other conditions constant.

We also performed experiments with variable sector size (2, 4, 8, 16, 32, and 64) to observe the performance of sector-mapped cache memory with respect to change in sector size. Table 7 and Figure 12, illustrate comparison between Miss Ratio and Sector Size, keeping Block Size (32 bytes), Cache Size (8 KB), and Associativity (4-way) as constants. One may observe that with increment in Sector Size, Miss Ratio is showing downward trend which

implies that increasing Sector Size, may improve cache memory performance by increasing the Cache Hits.

S. No.	Cache Size (in KB)	Block Size	Set Associativity	Sector Size	Total Access Count	Cache Hits	Hit Ratio	Miss Ratio
1	8	32	4	2	1500000	1475487	98.37	1.63
2	8	32	4	4	1500000	1477927	98.53	1.47
3	8	32	4	8	1500000	1480129	98.68	1.32
4	8	32	4	16	1500000	1481955	98.80	1.20
5	8	32	4	32	1500000	1483879	98.93	1.07
6	8	32	4	64	1500000	1496211	99.75	0.25

Table 7: Cache performance with variation in Sector size.

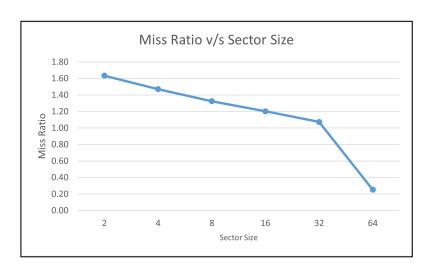


Figure 12: Miss Ratio variation with change in Sector size, keeping other parameters constant.

Conclusions

The research focused on developing a data cache simulator capable of evaluating m-way set-associative caches and sector-mapped caches across various configurations. Using Verilog and the Least Recently Used (LRU) algorithm, the simulator was implemented to handle complex cache architectures, requiring 2 GB of RAM for the given address space. The study explored three primary cache parameters—associativity, block (line) size, and cache size—and analyzed their influence on hit and miss rates for both m-way set-associative and sector-mapped caches.

M-Way Set-Associative Cache:

The results for m-way caches demonstrated that **cache size** significantly impacts performance. Increasing the cache size reduced capacity misses, thereby enhancing the hit ratio. However, larger cache sizes increased storage costs, emphasizing the trade-off between performance improvement and implementation expense.

• Block size played a critical role by affecting spatial locality. Larger blocks improved the hit ratio by fetching more adjacent data, but excessively large blocks reduced efficiency, as

unused data displaced frequently accessed blocks. This finding highlights the importance of selecting an optimal block size for balancing locality and cache utilization.

• Associativity also influenced performance, with higher associativity reducing conflict misses and improving the hit ratio. Fully associative caches achieved the lowest miss rates but involved higher hardware complexity and cost. These results underscore the trade-offs in designing efficient, cost-effective cache systems.

Sector-Mapped Cache:

The sector-mapped cache experiments added new dimensions to the findings by introducing finer granularity through sector division. Key observations from the results include:

- Cache size: Increasing cache size from 8 KB to 128 KB with a fixed block size of 32 bytes and associativity of 4-way yielded a consistent improvement in the hit ratio, reaching as high as 99.86%. This trend mirrors the behaviour of m-way caches but demonstrates sector-mapped caches' efficiency in reducing memory traffic due to fine-grained sector fetching.
- **Block size:** Smaller blocks generally resulted in higher hit ratios due to better utilization of spatial locality within sectors. However, extremely small blocks increased sector management overhead, indicating the need for balanced block size selection.
- Associativity: Higher associativity reduced conflict misses, as seen in configurations ranging from direct-mapped (94.63% hit ratio) to fully associative (99.52% hit ratio) caches. Sector-mapped caches further benefit from associativity by leveraging their finer granularity to minimize unnecessary block evictions.
- Sector size: Smaller sector sizes provided better cache utilization by reducing the storage of unused data. For example, reducing the sector size to 4 bytes increased the hit ratio to 98.53%, while larger sector sizes, such as 64 bytes, achieved 99.75%, balancing spatial locality with fewer cache misses.

General Insights:

The experimental results revealed how cache size, associativity, block size, and sector size interact to influence hit and miss rates. Larger cache sizes and higher associativity improved performance by reducing misses. Sector-mapped caches added efficiency by enabling fine-grained data handling, which was particularly beneficial for workloads with irregular access patterns. Optimal block and sector sizes further enhanced the hit ratio by balancing locality and utilization.

By integrating both m-way and sector-mapped cache designs, this research provides comprehensive insights into cache memory optimization. These findings equip system designers with a deeper understanding of the trade-offs and opportunities in selecting cache configurations for modern computing applications, contributing to the development of high-performance memory systems.

REFERENCES:

- 1. Lecture Notes EE275-Fall 2024: Dr. Harish Hiriyannaiah.
- 2. Computer Organization and Design: The Hardware/Software Interface, Fourth Edition by John L. Hennessy & David A. Patterson
- 3. Computer system architecture textbook by M Morris Mano ISBN: 81-317-0070-4
- 4. <u>Cache Memory in Computer Organization GeeksforGeeks</u>

Supplementary Material

A. M-Way Associative Cache Simulation

A.1. Design File Verilog

```
define cache size (1024*8)
define line size 128
define Associativity 4
define Index bit ('Associativity==0)? 0: $clog2('cache size/('line size*'Associativity))
define Offset bit $clog2(`line size)
define Tag bit 31-('Offset bit+'Index bit)
module test(adder 41,clk 41, rst 41,misses 41,hits 41);
 input clk 41, rst 41;
 input [30:0] adder 41;
 output [30:0] misses 41,hits 41;
                                                                                                                    [30:0]
Num Blocks 41, Num Sets 41, misses 41, hits 41, cache block 41, cache set 41, set index 41, Curr Block 41, Curr Cou
nt 41;
 reg data present 41;
 integer i,j;
 integer k;
 parameter CS = 'cache size;
 parameter LS = `line size;
 parameter Assoc = ('Associativity==0)? (CS/LS): 'Associativity;
 parameter Tbit = `Tag bit;
 parameter Ob = 'Offset bit;
 parameter Ib = `Index bit;
 reg [(LS*8)-1:0] cache [0:(CS/LS) - 1];
 reg [Tbit-1:0] tag array [0:(CS/LS) - 1]; // for all tags in cache
 reg valid array [0:(CS/LS) - 1]; //0 - there is no data 1 - there is data
 reg [Tbit-1:0] tag; // for current tag
 reg [Assoc-1:0] counter [0:(CS/LS) - 1];
```

```
hits 41 = 0;
misses 41 = 0;
Num Blocks 41 = CS/LS;
Num Sets 41 = Num Blocks 41/Assoc;
for (k = 0; k < Num Blocks 41; k = k + 1)
    valid array[k] = 0;
    tag array[k] = 0;
 end
for (j = 0; j < Num Sets 41; j = j + 1)
 for (k = 0; k < Assoc; k = k + 1)
  counter[(j*Assoc)+k] = k;
always@(posedge clk 41) begin
 cache block 41 = (adder 41/LS)% Num Blocks 41;
 cache set 41 = ((adder 41/LS)\% (Num Blocks 41/Assoc));
 set index 41 = cache set 41*Assoc; //pointing at first 41 block of current set
 data present 41 = 0;
 tag = adder 41[30:(Ob+Ib)];
 for(i=0; i<Assoc; i=i+1) begin
  if ((valid array [set index 41+i] == 1) && (tag == tag array[set index 41+i])) begin
  hits 41 = \text{hits } 41 + 1;
  data present 41 = 1;
  Curr Block 41 = set index 41+i;
  Curr Count 41 = counter[Curr Block 41];
 // $display("hits 41=%d",hits 41);
 if (data present 41 == 0) begin
  misses 41 = misses 41+1;
 // $display("misses 41=%d",misses 41);
  for(i=0; i<Assoc; i=i+1) begin
   if (counter[set index 41+i]==0)begin
   tag array[set index 41+i] = tag;
    valid array [set index 41+i] = 1;
    Curr Block 41 = set index 41+i;
    Curr Count 41 = 0;
```

```
end
end
end
for(i=0; i<Assoc; i=i+1) begin // counter

if (counter[set_index_41+i]>Curr_Count_41) begin
    counter[set_index_41+i] = counter[set_index_41+i] - 1;
end

counter[Curr_Block_41] = Assoc - 1;
end
end
end
end
```

A.2. Test-bench Verilog

```
module tb();
 reg clk 41, rst 41;
 reg [30:0] adder_41;
 reg [8:0]LS 41;
 reg [16:0] CS_41;
 reg [5:0] Assoc 41;
 wire [30:0] misses 41,hits 41;
 real hit_ratio_41,per = 100.00;
 real m,h;
 'define Length 1500000
 integer p,r,c;
  integer file, stat, out, i, j;
 reg signed [30:0] face[0:`Length-1];
 reg [30:0] actual[0:`Length-1];
 test t1(adder 41,clk 41, rst 41,misses 41,hits 41);
 initial begin
 file=$fopen("addr trace.txt","r");
            i=0;
       while (! $feof(file))
```

```
stat=$fscanf(file,"%d\n",face[i]);
                 i=i+1;
       $fclose(file);
  for(i=0;i<`Length;i=i+1)
          if (i==0) actual [0] = face [0];
           else actual[i]=actual[i-1]+face[i];
 always #5 clk 41 = \sim \text{clk } 41;
  clk 41 = 0;
  for (j = 0; j \le Length; j = j+1) begin
   @(posedge clk 41) adder 41 = actual[j];
  @(posedge clk 41)
  m=misses 41;
  h=hits 41;
 hit_ratio_41 = (h/(m+h))*per;
       $display("Hit Ratio=%7.2f Percentage,
                                                     Cache
                                                              Hits = %d, Total
                                                                                         Simulation Addresses
%d",hit ratio 41,hits 41,(misses 41 + hits 41));
  $finish;
endmodule
```

A.3. Simulation Results

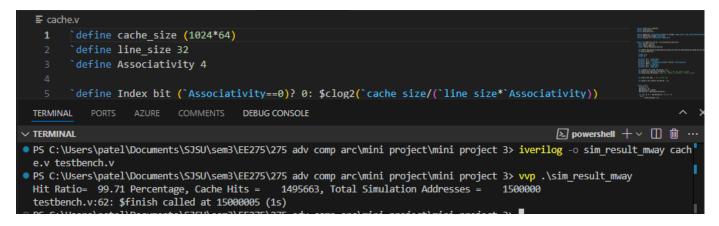
First case: Varying Cache size (Fixed block size= 32 Bytes, Fixed Associativity=4)

Cache size = 16KB:

Cache Size = 32KB:



Cache size = 64KB:



Cache size = 128KB:



Second case: Varying Block size (fixed cache size = 8192 KB, fixed associativity =4)

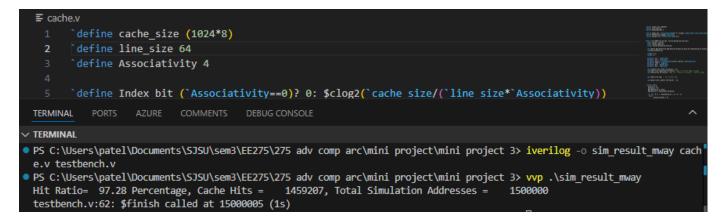
Block size = 16:



Block size = 32:



Block size = 64:



Block size = 128:



Third case: Varying Associativity (fixed cache size = 8192 KB, block size = 32KB)

1-way Set Associativity (Direct mapped cache):

2-way set associativity:

4-way set associativity:

```
cache.v

define cache_size (1024*8)

define line_size 32

define Associativity 4

define Index bit (`Associativitv==0)? 0: $clog2(`cache size/(`line size*`Associativitv))

TERMINAL PORTS AZURE COMMENTS DEBUG CONSOLE

TERMINAL

PS C:\Users\patel\Documents\SJSU\sem3\EE275\275 adv comp arc\mini project\mini project 3> iverilog -o sim_result_mway cache.v testbench.v

PS C:\Users\patel\Documents\SJSU\sem3\EE275\275 adv comp arc\mini project\mini project 3> vvp .\sim_result_mway
Hit Ratio= 98.08 Percentage, Cache Hits = 1471126, Total Simulation Addresses = 1500000
testbench.v:62: $finish called at 15000005 (1s)
```

8-way set associativity:

16-way set associativity:

32-way set associativity (Fully Associative):

B. Sector Mapped Cache Simulation

B.1. Design File Verilog

```
define cache size (1024*8)
 define line size 32
 define sector size 64 // new definition for sector size
 define Associativity 4
 define Index_bit (`Associativity==0)? 0: $clog2(`cache_size/(`line_size*`Associativity))
 define Offset_bit $clog2(`line_size)
 define Sector Offset Bit $clog2(`sector size)
 define Tag bit 31-(`Offset bit+`Index bit+`Sector Offset Bit)
module test(adder 41,clk 41, rst 41,misses 41,hits 41);
  input clk_41, rst_41;
  input [30:0] adder 41;
  output [30:0] misses_41,hits_41;
                                                                                     [30:0]
                                         reg
Num_Blocks_41,Num_Sets_41,misses_41,hits_41,cache_block_41,cache_set_41,set_index_41,Curr
 Block 41, Curr Count 41;
  reg data_present_41;
  integer i,j;
  integer k;
  parameter CS = `cache size;
  parameter LS = `line size;
  parameter SS = `sector size; // sector size
  parameter Assoc = (`Associativity==0)? (CS/LS):`Associativity;
  parameter Tbit = `Tag_bit;
  parameter Ob = `Offset bit;
```

```
parameter Ib = `Index_bit;
parameter SOB = `Sector Offset Bit; // sector offset bit
reg [(LS*8)-1:0] cache [0:(CS/LS) - 1];
reg [Tbit-1:0] tag_array [0:(CS/LS) - 1]; // for all tags in cache
reg valid_array [0:(CS/LS) - 1]; //0 - there is no data 1 - there is data
// sector information
reg [SOB-1:0] sector_offset [0:(CS/LS) - 1];
reg [Tbit-1:0] tag; // for current tag
reg [Assoc-1:0] counter [0:(CS/LS) - 1];
initial begin
 hits_41 = 0;
 misses 41 = 0;
 Num_Blocks_41 = CS/LS;
 Num_Sets_41 = Num_Blocks_41/Assoc;
 for (k = 0; k < Num Blocks 41; k = k + 1)
  begin
        valid_array[k] = 0;
        tag_array[k] = 0;
        sector_offset[k] = 0; // initialize sector offset
 for (j = 0; j < Num_Sets_41; j = j + 1)
   for (k = 0; k < Assoc; k = k + 1)
     counter[(j*Assoc)+k] = k;
always@(posedge clk_41)begin
  cache block 41 = (adder 41/LS)% Num Blocks 41;
  cache_set_41 = ((adder_41/LS)% (Num_Blocks_41/Assoc));
  set_index_41 = cache_set_41*Assoc; //pointing at first block of current set
  data_present_41 = 0;
  tag = adder_41[30:(0b+Ib+SOB)];
  for(i=0; i<Assoc; i=i+1) begin</pre>
```

```
if ((valid_array [set_index_41+i] == 1) && (tag == tag_array[set_index_41+i]))
begin
      hits_41 = hits_41+1;
      data_present_41 = 1;
      Curr_Block_41 = set_index_41+i;
      Curr_Count_41 = counter[Curr_Block_41];
     // $display("hits_41=%d",hits_41);
    if (data_present_41 == 0) begin
      misses_41 = misses_41+1;
      for(i=0; i<Assoc; i=i+1) begin</pre>
        if (counter[set_index_41+i]==0)begin
          tag_array[set_index_41+i] = tag;
          valid_array [set_index_41+i] = 1;
              sector_offset[set_index_41+i] = adder_41[(0b+Ib+S0B-1):(0b+Ib)]; // store
sector offset
          Curr_Block_41 = set_index_41+i;
          Curr Count 41 = 0;
      for(i=0; i<Assoc; i=i+1) begin // counter</pre>
        if (counter[set_index_41+i]>Curr_Count_41)begin
          counter[set_index_41+i] = counter[set_index_41+i] - 1;
        counter[Curr_Block_41] = Assoc - 1;
      end
endmodule
```

B.2. Test-bench Verilog

```
module tb();
  reg clk_41, rst_41;
  reg [30:0] adder_41;
  reg [8:0]LS_41;
  reg [16:0] CS_41;
  reg [5:0] Assoc_41;
  wire [30:0] misses 41, hits 41;
  real hit_ratio_41,per = 100.00;
  real m,h;
  `define Length 1500000
  integer p,r,c;
    integer file,stat,out,i,j;
  reg signed [30:0] face[0:`Length-1];
  reg [30:0] actual[0:`Length-1];
  test t1(adder_41,clk_41, rst_41,misses_41,hits_41);
  initial begin
   file=$fopen("addr_trace.txt","r");
                   i=0;
            while (! $feof(file))
                begin
                    stat=$fscanf(file,"%d\n",face[i]);
                             i=i+1;
                end
            $fclose(file);
    for(i=0;i<`Length;i=i+1)</pre>
                begin
                  if (i==0) actual[0] = face[0];
                  else actual[i]=actual[i-1]+face[i];
  always #5 clk_41 = \sim clk_41;
```

```
initial begin

clk_41 = 0;
  rst_41 = 1;
  @(posedge clk_41) rst_41 = 0;

for (j = 0 ; j<`Length ; j=j+1) begin

    @(posedge clk_41) adder_41 = actual[j];
    //$display("Address=%d",adder_41);
  end

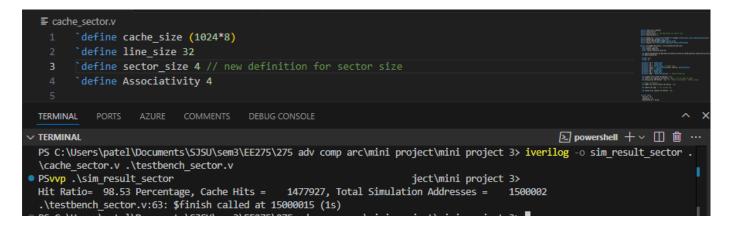
    @(posedge clk_41)
    m=misses_41;
    h=hits_41;
    hit_ratio_41 = (h/(m+h))*per;
    $display("Hit Ratio=%7.2f Percentage, Cache Hits = %d, Total Simulation Addresses = %d",hit_ratio_41,hits_41,(misses_41 + hits_41));
    $finish;
  end

endmodule</pre>
```

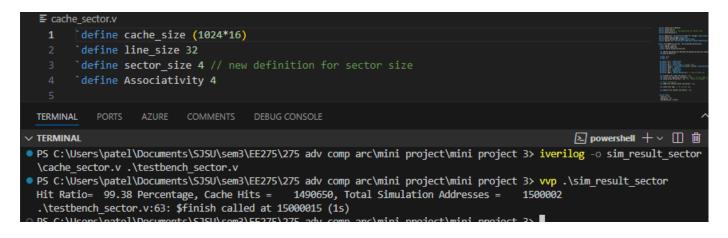
B.3. Simulation Results

First case: Varying Cache size (Fixed block size= 32 Bytes, Fixed Associativity=4, Fixed sector size=4)

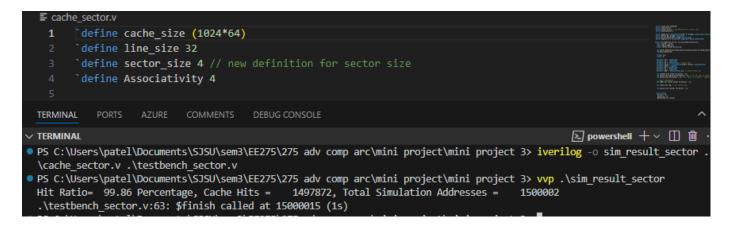
Cache size = 8KB:



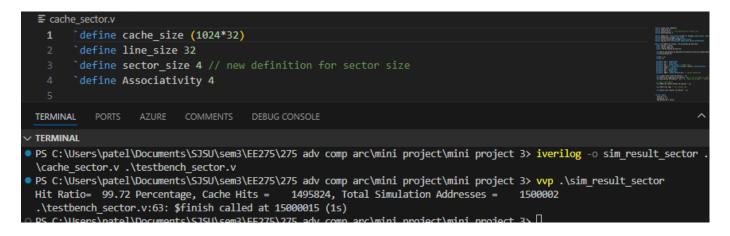
Cache size = 16KB:



Cache size = 32KB:



Cache size = 64KB:



Cache size = 128KB:

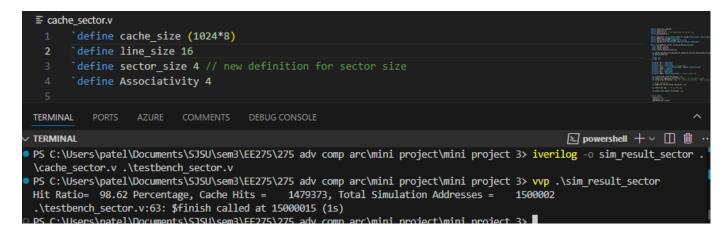


Second case: Varying Block Size (Fixed cache size= 8KB, Fixed Associativity=4, Fixed sector size=4)

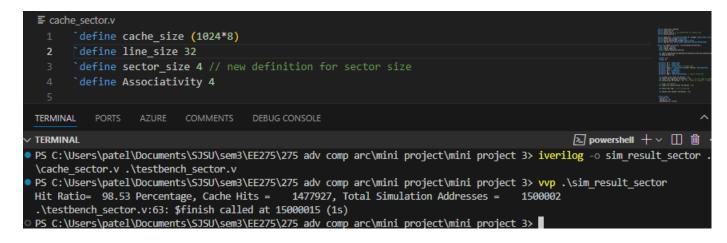
Block size = 8 bytes:



Block size = 16 bytes:



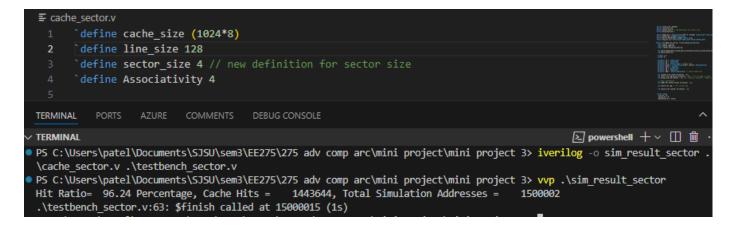
Block size = 32 bytes:



Block size = 64 bytes:



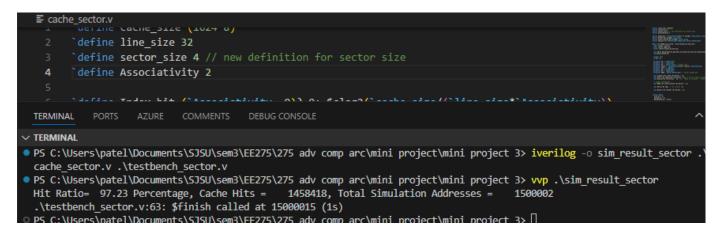
Block size = 128 bytes:



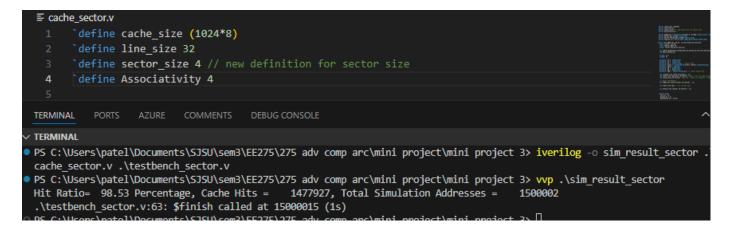
Third case: Varying Associativity (Fixed cache size= 8KB, Fixed Block Size=32 bytes, Fixed sector size=4)

Associativity=1 (Direct Mapped):

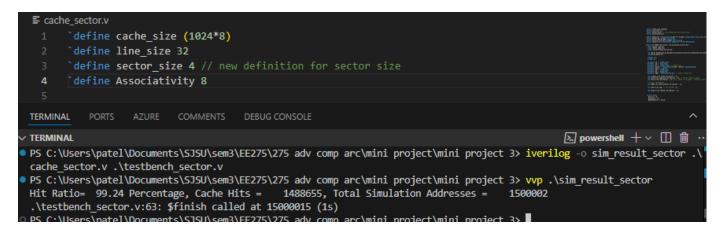
Associativity = 2:



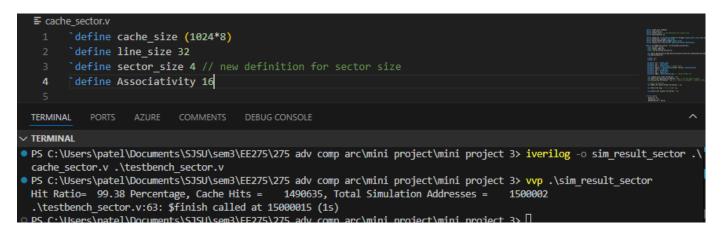
Associativity = 4:



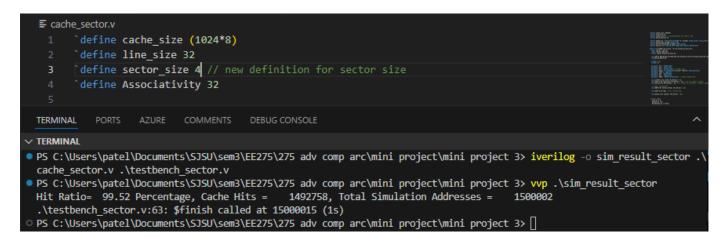
Associativity = 8:



Associativity = 16:

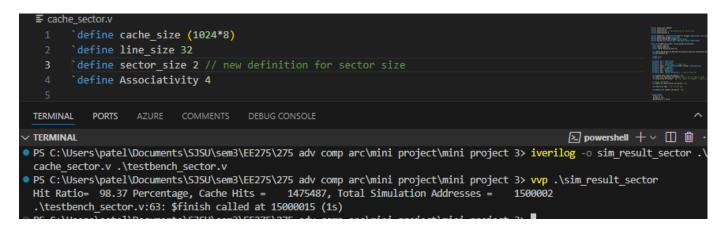


Associativity = 32 (Fully mapped):

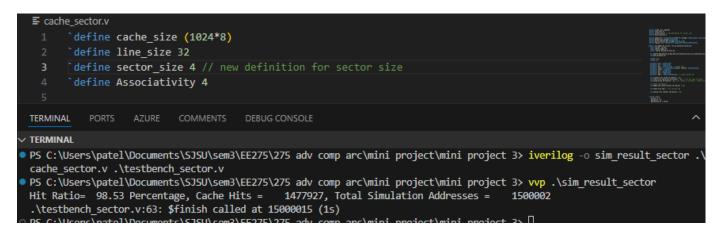


Fourth case: Varying Sector Size (Fixed cache size= 8KB, Fixed Block Size=32 bytes, Fixed associativity = 4)

Sector Size=2:



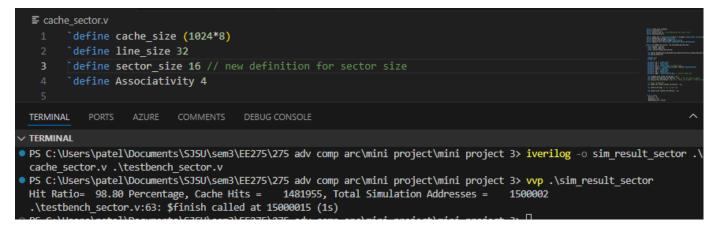
Sector Size=4:



Sector Size=8:



Sector Size=16:



Sector Size=32:



Sector Size=64:

