Caching Issues in Multicore Performance

Problem: The Path Between a CPU Chip and Off-chip Memory is Slow

This path is relatively slow, forcing the CPU to wait for up to 200 clock cycles just to do a store to, or a load from, memory.

Depending on your CPU’s ability to process instructions out-of-order, it might go idle during this time.

This is a huge performance hit!

Solution: Hierarchical Memory Systems, or “Cache”

The solution is to add intermediate memory systems. The one closest to the ALU (L1) is small and fast. The memory systems get slower and larger as they get farther away from the ALU.

L3 cache also exists on some high-end CPU chips.

Cache and Memory are Named by “Distance Level” from the ALU

L3 cache has been added to many CPU chips as well.

Storage Level Characteristics

<table>
<thead>
<tr>
<th>Type of Storage</th>
<th>L1</th>
<th>L2</th>
<th>L3</th>
<th>Memory</th>
<th>Disk</th>
</tr>
</thead>
<tbody>
<tr>
<td>Typical Size</td>
<td>102 KB</td>
<td>8 MB</td>
<td>32 MB</td>
<td>32 GB</td>
<td>Many GB</td>
</tr>
<tr>
<td>Typical Access Time (ns)</td>
<td>25</td>
<td>50</td>
<td>10.8</td>
<td>50</td>
<td>5,000,000</td>
</tr>
<tr>
<td>Access Time</td>
<td>1 second</td>
<td>2 seconds</td>
<td>43 seconds</td>
<td>3.3 minutes</td>
<td>231 days</td>
</tr>
<tr>
<td>Managed by</td>
<td>Hardware</td>
<td>Hardware</td>
<td>Hardware</td>
<td>OS</td>
<td>OS</td>
</tr>
</tbody>
</table>

Usually there are two L1 caches – one for Instructions and one for Data. You will often see this referred to in data sheets as: “L1 cache: 32KB + 32KB” or “I and D cache”.

Cache Hits and Misses

When the CPU asks for a value from memory, and that value is already in the cache, it is a cache hit.

This is called a cache hit.

When the CPU asks for a value from memory, and that value is not already in the cache, it will have to go off the chip to get it.

This is called a cache miss.

While cache might be multiple kilo- or megabytes, the bytes are transferred in much smaller quantities, each called a cache line. The size of a cache line is typically just 64 bytes.

Performance programming should strive to avoid as many cache misses as possible. That’s why it is very helpful to know the cache structure of your CPU.
Computer Graphics

Spatial and Temporal Coherence

Successful use of the cache depends on Spatial Coherence:

"If you need one memory address’s contents now, then you will probably also need the contents of some of the memory locations around it soon."

Successful use of the cache depends on Temporal Coherence:

"If you need one memory address’s contents now, then you will probably also need its contents again soon."

If these assumptions are true, then you will generate a lot of cache hits.

If these assumptions are not true, then you will generate a lot of cache misses, and you end up re-loading the cache a lot.

How Bad Is It? -- Demonstrating the Cache-Miss Problem

C and C++ store 2D arrays a row-at-a-time, like this, \( A[i][j] \):

\[ \begin{array}{cccccc}
0 & 1 & 2 & 3 & 4 \\
5 & 6 & 7 & 8 & 9 \\
10 & 11 & 12 & 13 & 14 \\
15 & 16 & 17 & 18 & 19 \\
20 & 21 & 22 & 23 & 24 \\
\end{array} \]

For large arrays, would it be better to add the elements by row, or by column? Which will avoid the most cache misses?

Sequential memory order

Jump-around-in-memory order

Demonstrating the Cache-Miss Problem – Across Rows

\[ \text{float sum = 0.;} \]
\[ \text{for(int i = 0; i < NUM; i++)} \]
\[ \text{for(int j = 0; j < NUM; j++)} \]
\[ \text{sum += Array[i][j]; // access across a row} \]
\[ \text{double row_finish = MyTimer();} \]
\[ \text{double row_secs = finish - start;} \]

Demonstrating the Cache-Miss Problem – Down Columns

\[ \text{float sum = 0.;} \]
\[ \text{double start = MyTimer();} \]
\[ \text{for(int i = 0; i < NUM; i++)} \]
\[ \text{for(int j = 0; j < NUM; j++)} \]
\[ \text{sum += Array[j][i]; // access down a column} \]
\[ \text{double col_finish = MyTimer();} \]
\[ \text{double col_secs = col_finish - start;} \]

Demonstrating the Cache-Miss Problem

Time, in seconds, to compute the array sums, based on by-row versus by-column order:

<table>
<thead>
<tr>
<th>Dimension (NUM)</th>
<th>(Total array size = NUM*NUM)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1000</td>
<td>0.05</td>
</tr>
<tr>
<td>10000</td>
<td>0.5</td>
</tr>
<tr>
<td>100000</td>
<td>5.0</td>
</tr>
</tbody>
</table>

Good Object-Oriented Programming Style can sometimes be Inconsistent with Good Cache Use:

```
class xyz
{
public:
    float x, y, z;
    xyz( );
    static xyz *Head = NULL;
};
xyz::xyz( )
{
    xyz * n = new xyz;
    n->next = Head;
    Head = n;
}
```

This is good OO style – it encapsulates and isolates the data for this class. Once you have created a linked list whose elements are all over memory, is it the best use of the cache?
But, Here Is a Compromise:

It might be better to create a large array of xyz structures and then have the constructor method pull new ones from that list. That would keep many of the elements close together while preserving the flexibility of the linked list.

When you need more, allocate another large array and link to it.

Remember: in this scheme, you cannot delete an individual node because it was allocated as part of an array. The best you can do is track which nodes can be deleted and then when all of an array’s nodes are flagged, delete the whole array.

Why Can We Get This Kind of Performance Decrease as Data Sets Get Larger?

We are violating Temporal Coherence

An Example of Where Cache Coherence Really Matters: Matrix Multiply

The usual approach is multiplying the entire A row * entire B column. This is equivalent to computing a single dot product.

```
for (k = 0; k < SIZE; k++)
    for (j = 0; j < SIZE; j++)
        for (i = 0; i < SIZE; i++)
```

Problem: Column j of B matrix is not doing a unit stride.

An Example of Where Cache Coherence Really Matters: Matrix Multiply

Scalable Universal Matrix Multiply Algorithm (SUMMA):

Entire A row * one element of B row. Equivalent to computing one item in many separate dot products.

```
for (j = 0; j < SIZE; j++)
    for (k = 0; k < SIZE; k++)
        for (i = 0; i < SIZE; i++)
```

We will cover this in further detail when we discuss SIMD.

We Can Help the Temporal Problem with Pre-Fetching

![Graph showing speed vs array size](image)
**Cache Architectures**

N-way Set Associative – a cache line from a particular block of memory can appear in a limited number of places in cache. Each “limited place” is called a set of cache lines. A set contains N cache lines.

The memory block can appear in any cache line in its set.

Most Caches today are N-way Set Associative

N is typically 4 for L1 and 8 or 16 for L2

**A Specific Example with Numbers**

Memory address = 1234 bytes

Memory address in bytes

Cache Line Block in Memory = 1234 / 64 = 19

Because there are 64 bytes in a cache line

Cache Set # = 19 % 4 = 3

Because there are 4 sets to rotate through

Offset in the Cache Line = 1234 – 19*64 = 18

Because there are 18 bytes left after filling 19 complete cache lines

**How Different Cores’ Cache Lines Keep Track of Each Other**

Each core has its own separate L2 cache, but a write by one can impact the state of the others.

For example, if one core writes a value into one of its own cache lines, any other core using a copy of that same cache line can no longer count on its values being up-to-date. In order to regain that confidence, the core that wrote must flush that cache line back to memory and the other core must then reload its copy of that cache line.

To maintain this organization, each core’s L2 cache has 4 states (MESI):

1. Modified
2. Exclusive
3. Shared
4. Invalid
### A Simplified View of How MESI Works

1. **Core A reads a value.** Those values are brought into its cache. That cache line is now tagged **Exclusive**.

2. **Core B reads a value from the same area of memory.** Those values are brought into its cache, and now both cache lines are re-tagged **Shared**.

3. **If Core B writes into that value.** Its cache line is re-tagged **Modified** and Core A’s cache line is re-tagged **Invalid**.

4. **Core A tries to read a value from that same part of memory.** But its cache line is tagged **Invalid**. So, Core B’s cache line is flushed back to memory and then Core A’s cache line is re-loaded from memory. Both cache lines are now tagged **Shared**.

Note: False Sharing doesn’t create incorrect results – it just creates a performance hit. If anything, False Sharing prevents getting incorrect results.

### False Sharing – Fix #1

Adding some padding:

```c
#include <stdlib.h>

struct s { float value; } Array[4];

omp_set_num_threads( 4 );

#pragma omp parallel for
for( int i = 0; i < 4; i++ )
{
#pragma omp for
do( int j = 0; j < SomeBigNumber; ++j )
{
    Array[i].value = Array[i].value + (float)rand();
}
}
```

This works because successive Array elements are forced onto different cache lines, so less (or no) cache line conflicts exist.

### False Sharing – Fix #2

Adding more padding:

```c
const int SomeBigNumber = 100000000; // keep less than 2B

#pragma omp parallel for
for( int i = 0; i < 4; i++ )
{
#pragma omp for
do( int j = 0; j < SomeBigNumber; ++j )
{
    Array[i].value = Array[i].value + (float)rand();
}
}
```

This is a huge performance hit, and is referred to as False Sharing.

### False Sharing – the Effect of Spreading Your Data to Multiple Cache Lines

![Graph showing the effect of spreading data to multiple cache lines](image)

**Why do these curves look this way?**
False Sharing – the Effect of Spreading Your Data to Multiple Cache Lines

Numpad = 1

False Sharing – the Effect of Spreading Your Data to Multiple Cache Lines

Numpad = 2

False Sharing – the Effect of Spreading Your Data to Multiple Cache Lines

Numpad = 3

False Sharing – the Effect of Spreading Your Data to Multiple Cache Lines

Numpad = 4

False Sharing – the Effect of Spreading Your Data to Multiple Cache Lines

Numpad = 5

False Sharing – Fix #1

Graph showing the effect of spreading data to multiple cache lines and the fix for this issue.
False Sharing – Fix #1

NUMPAD = 11

False Sharing – the Effect of Spreading Your Data to Multiple Cache Lines

NUMPAD = 12

False Sharing – the Effect of Spreading Your Data to Multiple Cache Lines

NUMPAD = 13

False Sharing – the Effect of Spreading Your Data to Multiple Cache Lines

NUMPAD = 14

False Sharing – the Effect of Spreading Your Data to Multiple Cache Lines

NUMPAD = 15
False Sharing – Fix #2

```c
#include <stdlib.h>
struct s {
    float value;
} Array[4];
omp_set_num_threads(4);
const int SomeBigNumber = 100000000;
#pragma omp parallel for
for (int i = 0; i < 4; i++) {
    float tmp = Array[i].value;
    for (int j = 0; j < SomeBigNumber; j++)
        tmp = tmp + (float)rand();
    Array[i].value = tmp;
}
```

This works because a localized temporary variable is created in each core’s stack area, so little or no cache line conflict exists.

False Sharing – Fix #2 vs. Fix #1

Note that Fix #2 with {1, 2, 4} threads gives the same performance as NUMPAD= {0,7,15}

malloc’ing on a cache line

What if you are malloc’ing, and want to be sure your data structure starts on a cache line boundary?

Knowing that cache lines start on fixed 64-byte boundaries lets you do this. Consider a memory address. The top 9-6 bits tell you what cache line number this address is a part of. The bottom 6 bits tell you what offset that address has within that cache line. So, for example, on a 32-bit memory system:

<table>
<thead>
<tr>
<th>Cache line number</th>
<th>Offset in that cache line</th>
</tr>
</thead>
<tbody>
<tr>
<td>32 - 6 = 26 bits</td>
<td>6 bits: 0-63</td>
</tr>
</tbody>
</table>

For example 101010₂ = 42

So, if you see a memory address whose bottom 6 bits are 000000, then you know that memory location begins on a cache line boundary.