



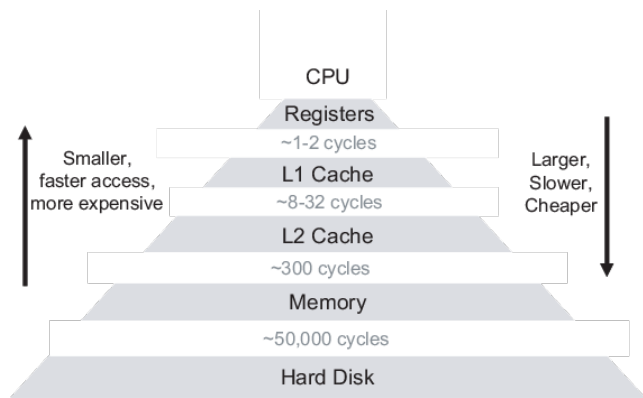
香港中文大學
The Chinese University of Hong Kong

CMSC5743 Lab 02 Additional Materials

Wenqian Zhao
Department of Computer Science & Engineering
Chinese University of Hong Kong
wqzhaoa@cse.cuhk.edu.hk

September 17, 2021





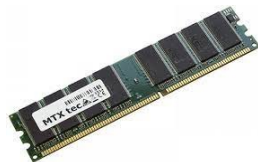
- Memory is primarily of three types :
 - Cache Memory
 - Primary Memory/Main Memory
 - Secondary Memory

- Cache Memory
 - Cache memory is faster than main memory
 - Less access time as compared to main memory
 - Stores the program that can be executed within a short period of time
 - Stores data for temporary use





- However ...
 - Cache memory has limited capacity
 - It is very expensive
- Primary Memory (Main Memory):
 - Usually volatile memory
 - Working memory of the computer
 - Faster than secondary memories
 - A computer cannot run without the primary memory



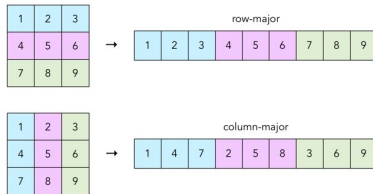


- If the processor finds that the memory location is in the cache, a **cache hit** has occurred and data is read from cache
- If the processor **does not** find the memory location in the cache, a **cache miss** has occurred. For a cache miss, the cache allocates a new entry and copies in data from main memory, then the request is fulfilled from the contents of the cache
- Hit ratio = $\text{hit} / (\text{hit} + \text{miss}) = \text{no. of hits} / \text{total accesses}$

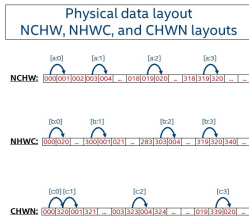
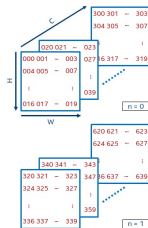


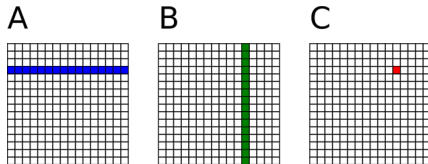
Data Layout

- Matrix:



- Tensor:





```
// ***** C = A x B ***** //  
void matmul() {  
    memset(C, 0, sizeof(C));  
    for (int i = 0; i < n; i++) {  
        for (int j = 0; j < n; j++) {  
            for (int k = 0; k < n; k++) {  
                C[i][j] += A[i][k] * B[k][j]  
            }  
        }  
    }  
}
```



- What if we use the transpose to change the visit order of the matrix?

```
// ***** C = A^T x B ***** //
void matmul_AT() {
    memset(C, 0, sizeof(C));
    for (int i = 0; i < n; i++) {
        for (int j = 0; j < n; j++) {
            AT[i][j] = A[j][i];
        }
    }
    for (int i = 0; i < n; i++) {
        for (int j = 0; j < n; j++) {
            for (int k = 0; k < n; k++) {
                C[i][j] += AT[k][i] * B[k][j];
            }
        }
    }
}
```

- What is the difference on hit ratio?

THANK YOU!