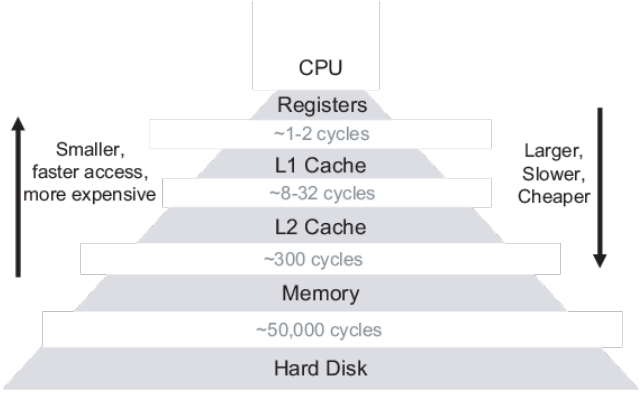
Lab 01: GEMM

Bei Yu
Department of Computer Science & Engineering
Chinese University of Hong Kong
byu@cse.cuhk.edu.hk

July 8, 2025

Memory Hierarchy



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

Memory Hierarchy

- 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

Memory Hierarchy

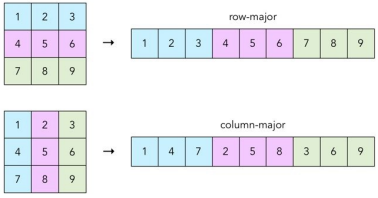
- However ...
  - Cache memory has limited capacity
  - It is very expensive
- Primary Memory (Main Memory):
  - Usually volatile memory
  - Working memory of the computer
  - Faster than secondary memories
  - A computer cannot run without the primary memory

Cache Performance

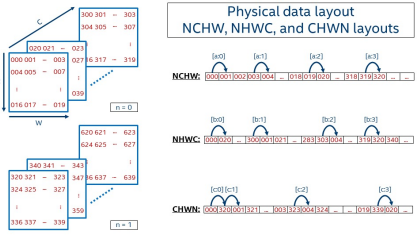
- If the processor finds that the memory location is in the cache, a **cache hit** has occurred and data is read from cache
- If the processor **does not** find the memory location in the cache, a **cache miss** has occurred. For a cache miss, the cache allocates a new entry and copies in data from main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

Data Layout

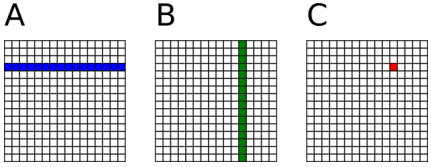
- Matrix:



- Tensor:



Matrix Multiplication




```
// ********************* 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]
      }
    }
  }
}
```

Matrix Multiplication

- 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?

Matrix Multiplication

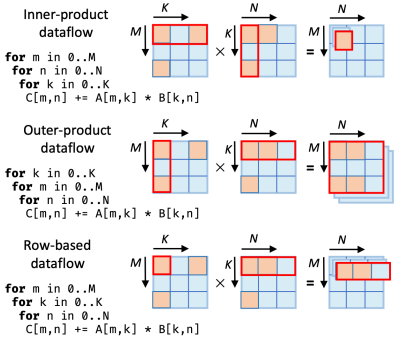
- Meanwhile, loop order (dataflow) may also affect the performance.
- Think about using additional optimization approaches to improve the hit ratio.

Matrix Multiplication

- Meanwhile, loop order (dataflow) may also affect the performance.
- Think about using additional optimization approaches to improve the hit ratio.

```
void matmul_tiling() {
  memset(C, 0, sizeof(C));
  int iTile = 64, jTile = 64, kTile = 32;
  for (int i_outer = 0; i_outer < n / iTile; i_outer++) {
    // Write your code here
  }
}
```

Matrix Multiplication



For each element's value on output matrix C:

- Inner: read A, B for K times, write C for once
- Outer: read A, B only for once, write C for K times (by frame)
- Row-based: read A for once, read B for K times, write C for K times (by row)

THANK YOU!