Cache Memories, Cache Complexity

Marc Moreno Maza

University of Western Ontario, London, Ontario (Canada)

CS3101 and CS4402-9535
Plan

Hierarchical memories and their impact on our programs

Cache Analysis in Practice

The Ideal-Cache Model

Cache Complexity of some Basic Operations

Matrix Transposition

A Cache-Oblivious Matrix Multiplication Algorithm
Plan

Hierarchical memories and their impact on our programs

Cache Analysis in Practice

The Ideal-Cache Model

Cache Complexity of some Basic Operations

Matrix Transposition

A Cache-Oblivious Matrix Multiplication Algorithm
## Capacity

<table>
<thead>
<tr>
<th>Access Time</th>
<th>Cost</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Staging Xfer Unit
- CPU Registers
  - 100s Bytes
  - 300 – 500 ps (0.3-0.5 ns)
- L1 and L2 Cache
  - 10s-100s K Bytes
  - ~1 ns - ~10 ns
  - $1000s/ GByte
- Main Memory
  - G Bytes
  - 80ns- 200ns
  - ~ $100/ GByte
- Disk
  - 10s T Bytes, 10 ms (10,000,000 ns)
  - ~ $1 / GByte
- Tape
  - infinite sec-min
  - ~$1 / GByte

### Upper Level
- Registers
- Instr. Operands
- Blocks
- Memory
- Blocks
- Pages
- Blocks
- Disk
- Files
- Blocks
- Lower Level
- Larger
- sec-min
- ~$1 / GByte

### Lower Level
- Staging
- Xfer Unit
- Cache
- Blocks
- prog./compiler
  - 1-8 bytes
- cache cntl
  - 32-64 bytes
- cache cntl
  - 64-128 bytes
- OS
  - 4K-8K bytes
- user/operator
  - Mbytes
- ~$1 / GByte
A CPU cache is an auxiliary memory which is smaller, faster memory than the main memory and which stores copies of the main memory locations that are expectedly frequently used.

Most modern desktop and server CPUs have at least three independent caches: the data cache, the instruction cache and the translation look-aside buffer.
Each location in each memory (main or cache) has
- a datum (cache line) which ranges between 8 and 512 bytes in size, while a datum requested by a CPU instruction ranges between 1 and 16.
- a unique index (called address in the case of the main memory)
- In the cache, each location has also a tag (storing the address of the corresponding cached datum).
CPU Cache (3/3)

When the CPU needs to read or write a location, it checks the cache:

- if it finds it there, we have a cache hit
- if not, we have a cache miss and (in most cases) the processor needs to create a new entry in the cache.

Making room for a new entry requires a replacement policy: the Least Recently Used (LRU) discards the least recently used items first; this requires to use age bits.
CPU Cache (4/7)

- **Read latency** (time to read a datum from the main memory) requires to keep the CPU busy with something else:
  - **out-of-order execution**: attempt to execute independent instructions arising after the instruction that is waiting due to the cache miss
  - **hyper-threading (HT)**: allows an alternate thread to use the CPU
Modifying data in the cache requires a **write policy** for updating the main memory

- **write-through cache**: writes are immediately mirrored to main memory
- **write-back cache**: the main memory is mirrored when that data is evicted from the cache

The cache copy may become out-of-date or stale, if other processors modify the original entry in the main memory.
CPU Cache (6/7)

- The replacement policy decides where in the cache a copy of a particular entry of main memory will go:
  - fully associative: any entry in the cache can hold it
  - direct mapped: only one possible entry in the cache can hold it
  - $N$-way set associative: $N$ possible entries can hold it
Cache Performance for SPEC CPU2000 by J.F. Cantin and M.D. Hill.

The SPEC CPU2000 suite is a collection of 26 compute-intensive, non-trivial programs used to evaluate the performance of a computer’s CPU, memory system, and compilers (http://www.spec.org/osg/cpu2000).
Cache issues

- **Cold miss:** The first time the data is available. Cure: Prefetching may be able to reduce this type of cost.

- **Capacity miss:** The previous access has been evicted because too much data touched in between, since the *working data set* is too large. Cure: Reorganize the data access such that *reuse* occurs before eviction.

- **Conflict miss:** Multiple data items mapped to the same location with eviction before cache is full. Cure: Rearrange data and/or pad arrays.

- **True sharing miss:** Occurs when a thread in another processor wants the same data. Cure: Minimize sharing.

- **False sharing miss:** Occurs when another processor uses different data in the same cache line. Cure: Pad data.
A typical matrix multiplication C code

```c
#define IND(A, x, y, d) A[(x)*(d)+(y)]
uint64_t testMM(const int x, const int y, const int z)
{
    double *A; double *B; double *C;
    long started, ended;
    float timeTaken;
    int i, j, k;
    srand(getSeed());
    A = (double *)malloc(sizeof(double)*x*y);
    B = (double *)malloc(sizeof(double)*x*z);
    C = (double *)malloc(sizeof(double)*y*z);
    for (i = 0; i < x*z; i++) B[i] = (double) rand() ;
    for (i = 0; i < y*z; i++) C[i] = (double) rand() ;
    for (i = 0; i < x*y; i++) A[i] = 0 ;
    started = example_get_time();
    for (i = 0; i < x; i++)
        for (j = 0; j < y; j++)
            for (k = 0; k < z; k++)
                // A[i][j] += B[i][k] + C[k][j];
                IND(A,i,j,y) += IND(B,i,k,z) * IND(C,k,j,z);
    ended = example_get_time();
    timeTaken = (ended - started)/1.f;
    return timeTaken;
}
```
Issues with matrix representation

- Contiguous accesses are better:
  - Data fetch as cache line (Core 2 Duo 64 byte per cache line)
  - With contiguous data, a single cache fetch supports 8 reads of doubles.
  - Transposing the matrix C should reduce L1 cache misses!
Transposing for optimizing spatial locality

```c
float testMM(const int x, const int y, const int z)
{
    double *A; double *B; double *C; double *Cx;
    long started, ended; float timeTaken; int i, j, k;
    A = (double *)malloc(sizeof(double)*x*y);
    B = (double *)malloc(sizeof(double)*x*z);
    C = (double *)malloc(sizeof(double)*y*z);
    Cx = (double *)malloc(sizeof(double)*y*z);
    srand(getSeed());
    for (i = 0; i < x*z; i++) B[i] = (double) rand() ;
    for (i = 0; i < y*z; i++) C[i] = (double) rand() ;
    for (i = 0; i < x*y; i++) A[i] = 0 ;
    started = example_get_time();
    for(j =0; j < y; j++)
        for(k=0; k < z; k++)
            IND(Cx,j,k,z) = IND(C,k,j,y);
    for (i = 0; i < x; i++)
        for (j = 0; j < y; j++)
            for (k = 0; k < z; k++)
                IND(A, i, j, y) += IND(B, i, k, z) * IND(Cx, j, k, z);
    ended = example_get_time();
    timeTaken = (ended - started)/1.f;
    return timeTaken;
}
```
Issues with data reuse

▶ Naive calculation of a row of A, so computing 1024 coefficients: 1024 accesses in A, 384 in B and $1024 \times 384 = 393,216$ in C. Total = 394,524.

▶ Computing a $32 \times 32$-block of A, so computing again 1024 coefficients: 1024 accesses in A, $384 \times 32$ in B and $32 \times 384$ in C. Total = 25,600.

▶ The iteration space is traversed so as to reduce memory accesses.
float testMM(const int x, const int y, const int z)
{
    double *A; double *B; double *C;
    long started, ended; float timeTaken; int i, j, k, i0, j0, k0;
    A = (double *)malloc(sizeof(double)*x*y);
    B = (double *)malloc(sizeof(double)*x*z);
    C = (double *)malloc(sizeof(double)*y*z);
    srand(getSeed());
    for (i = 0; i < x*z; i++) B[i] = (double) rand();
    for (i = 0; i < y*z; i++) C[i] = (double) rand();
    for (i = 0; i < x*y; i++) A[i] = 0;
    started = example_get_time();
    for (i = 0; i < x; i += BLOCK_X)
        for (j = 0; j < y; j += BLOCK_Y)
            for (k = 0; k < z; k += BLOCK_Z)
                for (i0 = i; i0 < min(i + BLOCK_X, x); i0++)
                    for (j0 = j; j0 < min(j + BLOCK_Y, y); j0++)
                        for (k0 = k; k0 < min(k + BLOCK_Z, z); k0++)
                            IND(A,i0,j0,y) += IND(B,i0,k0,z) * IND(C,k0,j0,y);
    ended = example_get_time();
    timeTaken = (ended - started)/1.f;
    return timeTaken;
}
float testMM(const int x, const int y, const int z)
{
    double *A; double *B; double *C;
    long started, ended; float timeTaken; int i, j, k, i0, j0, k0;
    A = (double *)malloc(sizeof(double)*x*y);
    B = (double *)malloc(sizeof(double)*x*z);
    C = (double *)malloc(sizeof(double)*y*z);
    srand(getSeed());
    for (i = 0; i < x*z; i++) B[i] = (double) rand();
    for (i = 0; i < y*z; i++) C[i] = (double) rand();
    for (i = 0; i < x*y; i++) A[i] = 0;
    started = example_get_time();
    for (i = 0; i < x; i += BLOCK_X)
        for (j = 0; j < y; j += BLOCK_Y)
            for (k = 0; k < z; k += BLOCK_Z)
                for (i0 = i; i0 < min(i + BLOCK_X, x); i0++)
                    for (j0 = j; j0 < min(j + BLOCK_Y, y); j0++)
                        for (k0 = k; k0 < min(k + BLOCK_Z, z); k0++)
                            IND(A,i0,j0,y) += IND(B,i0,k0,z) * IND(C,j0,k0,z);
    ended = example_get_time();
    timeTaken = (ended - started)/1.f;
    return timeTaken;
}
Experimental results

Computing the product of two $n \times n$ matrices on my laptop (Core2 Duo CPU P8600 @ 2.40GHz, L1 cache of 3072 KB, 4 GBytes of RAM)

<table>
<thead>
<tr>
<th>$n$</th>
<th>naive</th>
<th>transposed</th>
<th>speedup</th>
<th>64 × 64-tiled</th>
<th>speedup</th>
<th>t. &amp; t.</th>
<th>speedup</th>
</tr>
</thead>
<tbody>
<tr>
<td>128</td>
<td>7</td>
<td>3</td>
<td></td>
<td>7</td>
<td></td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>256</td>
<td>26</td>
<td>43</td>
<td>6.81</td>
<td>1928</td>
<td>0.936</td>
<td>187</td>
<td>9.65</td>
</tr>
<tr>
<td>512</td>
<td>1805</td>
<td>265</td>
<td>6.62</td>
<td>14020</td>
<td>1.76</td>
<td>1490</td>
<td>16.59</td>
</tr>
<tr>
<td>1024</td>
<td>24723</td>
<td>3730</td>
<td>9.11</td>
<td>112298</td>
<td>2.41</td>
<td>11960</td>
<td>22.69</td>
</tr>
<tr>
<td>2048</td>
<td>271446</td>
<td>29767</td>
<td>9.83</td>
<td>1009445</td>
<td>2.32</td>
<td>101264</td>
<td>23.15</td>
</tr>
<tr>
<td>4096</td>
<td>2344594</td>
<td>238453</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Timings are in milliseconds.

The cache-oblivious multiplication (more on this later) runs within 12978 and 106758 for $n = 2048$ and $n = 4096$ respectively.
Other performance counters

Hardware count events

- **CPI** *Clock cycles Per Instruction*: the number of clock cycles that happen when an instruction is being executed. With pipelining we can improve the CPI by exploiting instruction level parallelism.

- **L1 and L2 Cache Miss Rate**.

- **Instructions Retired**: In the event of a misprediction, instructions that were scheduled to execute along the mispredicted path must be canceled.

<table>
<thead>
<tr>
<th></th>
<th>CPI</th>
<th>L1 Miss Rate</th>
<th>L2 Miss Rate</th>
<th>Percent SSE Instructions</th>
<th>Instructions Retired</th>
</tr>
</thead>
<tbody>
<tr>
<td>In C</td>
<td>4.78</td>
<td>0.24</td>
<td>0.02</td>
<td>43%</td>
<td>13,137,280,000</td>
</tr>
<tr>
<td>Transposed</td>
<td>1.13</td>
<td>0.15</td>
<td>0.02</td>
<td>50%</td>
<td>13,001,486,336</td>
</tr>
<tr>
<td>Tiled</td>
<td>0.49</td>
<td>0.02</td>
<td>0</td>
<td>39%</td>
<td>18,044,811,264</td>
</tr>
</tbody>
</table>
Analyzing cache misses in the naive and transposed multiplication

Let $A$, $B$ and $C$ have format $(m, n)$, $(m, p)$ and $(p, n)$ respectively.

- $A$ is scanned one, so $mn/L$ cache misses if $L$ is the number of coefficients per cache line.
- $B$ is scanned $n$ times, so $mnp/L$ cache misses if the cache cannot hold a row.
- $C$ is accessed “nearly randomly” (for $m$ large enough) leading to $mnp$ cache misses.
- Since $2mnp$ arithmetic operations are performed, this means roughly one cache miss per flop!
Analyzing cache misses in the tiled multiplication

Let $A$, $B$ and $C$ have format $(m, n)$, $(m, p)$ and $(p, n)$ respectively.

Assume all tiles are square of order $B$ and three fit in cache.

If $C$ is transposed, then loading three blocks in cache cost $3B^2/L$.

This process happens $n^3/B^3$ times, leading to $3n^3/(BL)$ cache misses.

Three blocks fit in cache for $3B^2 < Z$, if $Z$ is the cache size.

So $O(n^3/\sqrt{ZL})$ cache misses, if $B$ is well chosen, which is optimal.
Counting sort: the algorithm

- **Counting sort** takes as input a collection of n items, each of which known by a key in the range $0 \cdots k$.
- The algorithm computes a *histogram* of the number of times each key occurs.
- Then performs a *prefix sum* to compute positions in the output.

```plaintext
allocate an array Count[0..k]; initialize each array cell to zero
for each input item x:
    Count[key(x)] = Count[key(x)] + 1
total = 0
for i = 0, 1, ... k:
    c = Count[i]
    Count[i] = total
    total = total + c
allocate an output array Output[0..n-1]
for each input item x:
    store x in Output[Count[key(x)]]
    Count[key(x)] = Count[key(x)] + 1
return Output
```
Counting sort: cache complexity analysis

allocate an array Count[0..k]; initialize each array cell to zero
for each input item x:
    Count[key(x)] = Count[key(x)] + 1
total = 0
for i = 0, 1, ... k:
    c = Count[i]
    Count[i] = total
    total = total + c
allocate an output array Output[0..n-1]
for each input item x:
    store x in Output[Count[key(x)]]
    Count[key(x)] = Count[key(x)] + 1
return Output

1. \( n/L \) to compute \( k \).
2. \( k/L \) cache misses to initialize Count.
3. \( n/L + n \) cache misses for the histogram (worst case).
4. \( k/L \) cache misses for the prefix sum.
5. \( n/L + n + n \) cache misses for building Output (worst case).

Total: \( 3n + 3n/L + 2k/L \) cache misses (worst case).
Counting sort: cache complexity analysis: explanations

1. \( n/L \) to compute \( k \): this can be done by traversing the items linearly.
2. \( k/L \) cache misses to initialize \( \text{Count} \): this can be done by traversing the \( \text{Count} \) linearly.
3. \( n/L + n \) cache misses for the histogram (worst case): items accesses are linear but \( \text{Count} \) accesses are potentially random.
4. \( k/L \) cache misses for the prefix sum: \( \text{Count} \) accesses are linear.
5. \( n/L + n + n \) cache misses for building \( \text{Output} \) (worst case): items accesses are linear but \( \text{Output} \) and \( \text{Count} \) accesses are potentially random.

**Total:** \( 3n + 3n/L + 2k/L \) cache misses (worst case).
How to fix the poor data locality of counting sort?

allocate an array Count[0..k]; initialize each array cell to zero
for each input item x:
    Count[key(x)] = Count[key(x)] + 1
total = 0
for i = 0, 1, ... k:
    c = Count[i]
    Count[i] = total
    total = total + c
allocate an output array Output[0..n-1]
for each input item x:
    store x in Output[Count[key(x)]]
    Count[key(x)] = Count[key(x)] + 1
return Output

- Recall that our worst case is $3n + 3n/L + 2k/L$ cache misses.
- The troubles come from the irregular accesses which experience capacity misses and conflict misses.
- To solve this problem, we preprocess the input so that counting sort is applied in succession to several smaller input item sets with smaller value ranges.
- To put it simply, so that $k$ and $n$ are small enough for Output and Count to incur cold misses only.
Counting sort: bucketing the input

allocate an array bucketsize[0..m-1]; initialize each array cell to zero
for each input item x:
    bucketsize[floor(key(x) m/(k+1))] := bucketsize[floor(key(x) m/(k+1))] + 1
         total = 0
for i = 0, 1, ... m-1:
    c = bucketsize[i]
    bucketsize[i] = total
    total = total + c
allocate an array bucketedinput[0..n-1];
for each input item x:
    q := floor(key(x) m/(k+1))
    bucketedinput[bucketsize[q]] := key(x)
    bucketsize[q] := bucketsize[q] + 1
return bucketedinput

▶ Intention: after preprocessing, the arrays Count and Output incur cold misses only.
▶ To this end we choose a parameter $m$ (more on this later) such that
  1. any key in the range $[ih, (i + 1)h − 1]$ is always $\leq$ any key in the range $[(i + 1)h, (i + 2)h − 1]$, for $i = 0 \cdots m − 2$, with $h = k/m$,
  2. bucketsize and $m$ cache-lines from bucketedinput all fit in cache. That is, counting cache-lines, $mL + m \leq Z$. 
Counting sort: cache complexity with bucketing

allocate an array bucketsize[0..m-1]; initialize each array cell to zero
for each input item x:
    bucketsize[floor(key(x) m/(k+1))] := bucketsize[floor(key(x) m/(k+1))] + 1
total = 0
for i = 0, 1, ... m-1:
    c = bucketsize[i]
    bucketsize[i] = total
    total = total + c
allocate an array bucketedinput[0..n-1];
for each input item x:
    q := floor(key(x) m/(k+1))
    bucketedinput[bucketsize[q] ] := key(x)
    bucketsize[q] := bucketsize[q] + 1
return bucketedinput

1. $3m/L + n/L$ caches misses to compute bucketsize
2. **Key observation:** bucketedinput is traversed regularly by segment.
3. Hence, $2n/L + m + m/L$ caches misses to compute bucketedinput

Preprocessing: $3n/L + 3m/L + m$ cache misses.
Counting sort: cache complexity with bucketing: explanations

1. $3m/L + n/L$ caches misses to compute bucketsize:
   ▶ $m/L$ to set each cell of bucketsize to zero,
   ▶ $m/L + n/L$ for the first for loop,
   ▶ $m/L$ for the second for loop.

2. **Key observation**: bucketedinput is traversed regularly by segment:
   ▶ So writing bucketedinput means writing (in a linear traversal) $m$ consecutive arrays, of possibly different sizes, but with total size $n$.
   ▶ Thus, because of possible misalignments between those arrays and their cache-lines, this writing procedure can yield $n/L + m$ cache misses (and not just $n/L$).

3. Hence, $2n/L + m + m/L$ caches misses to compute bucketedinput:
   ▶ $n/L$ to read the items,
   ▶ $n/L + m$ to write bucketedinput,
   ▶ $m/L$ to load bucketsize.
Cache friendly counting sort: complete cache complexity analysis

- **Assumption:** the preprocessing creates buckets of average size $n/m$.
- After preprocessing, counting sort is applied to each bucket whose values are in a range $[ih, (i + 1)h − 1]$, for $i = 0 \cdots m − 1$.
- To be cache-friendly, this requires, for $i = 0 \cdots m − 1$, $h + |\{\text{key} \in [ih, (i + 1)h − 1]\}| < Z$ and $m < Z/(1 + L)$. These two are very realistic assumption considering today’s cache size.
- And the total complexity becomes;

$$Q_{\text{total}} = Q_{\text{preprocessing}} + Q_{\text{sorting}}$$

$$= Q_{\text{preprocessing}} + mQ_{\text{sorting of one bucket}}$$

$$= Q_{\text{preprocessing}} + m(3\frac{n}{mL} + 3\frac{n}{mL} + 2\frac{k}{mL})$$

$$= Q_{\text{preprocessing}} + 6n/L + 2k/L$$

$$= 3n/L + 3m/L + m + 6n/L + 2k/L$$

$$= 9n/L + 3m/L + m + 2k/L$$

Instead of $3n + 3n/L + 2k/L$ for the naive counting sort.
Cache friendly counting sort: experimental results

- Experimentation on an Intel(R) Core(TM) i7 CPU @ 2.93GHz. It has L2 cache of 8MB.
- CPU times in seconds for both classical and cache-friendly counting sort algorithm.
- The keys are random machine integers in the range $[0, n]$.

<table>
<thead>
<tr>
<th>n</th>
<th>classical counting sort</th>
<th>cache-oblivious counting sort (preprocessing + sorting)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1000000000</td>
<td>13.74</td>
<td>4.66 (3.04 + 1.62 )</td>
</tr>
<tr>
<td>2000000000</td>
<td>30.20</td>
<td>9.93 (6.16 + 3.77)</td>
</tr>
<tr>
<td>3000000000</td>
<td>50.19</td>
<td>16.02 (9.32 + 6.70)</td>
</tr>
<tr>
<td>4000000000</td>
<td>71.55</td>
<td>22.13 (12.50 + 9.63)</td>
</tr>
<tr>
<td>5000000000</td>
<td>94.32</td>
<td>28.37 (15.71 + 12.66)</td>
</tr>
<tr>
<td>6000000000</td>
<td>116.74</td>
<td>34.61 (18.95 + 15.66)</td>
</tr>
</tbody>
</table>
Plan

Hierarchical memories and their impact on our programs

Cache Analysis in Practice

The Ideal-Cache Model

Cache Complexity of some Basic Operations

Matrix Transposition

A Cache-Oblivious Matrix Multiplication Algorithm
Basic idea of a cache memory (review)

- A cache is a smaller memory, faster to access.
- Using smaller memory to cache contents of larger memory provides the illusion of fast larger memory.
- Key reasons why this works: **temporal locality** and **spatial locality**.
A simple cache example

- Byte addressable memory
- Cache of 32Kbyte with direct mapping and 64 byte lines (thus 512 lines) so the cache can fit $2^9 \times 2^4 = 2^{13}$ int.
- “Therefore” successive 32Kbyte memory blocks line up in cache
- A cache access costs 1 cycle while a memory access costs 100 = $99 + 1$ cycles.
- How addresses map into cache
  - Bottom 6 bits are used as offset in a cache line
Exercise 1 (1/2)

// sizeof(int) = 4 and Array laid out sequentially in memory
#define S ((1<<20)*sizeof(int))
int A[S];
// Thus size of A is 2^(20) x 4
for (i = 0; i < S; i++) {
    read A[i];
}

Total access time? What kind of locality? What kind of misses?
Exercise 1 (2/2)

```c
#define S ((1<<20)*sizeof(int))
int A[S];
for (i = 0; i < S; i++) {
    read A[i];
}
```

- S reads to A.
- 16 elements of A per cache line
- 15 of every 16 hit in cache.
- Total access time: $15(S/16) + 100(S/16)$.
- spatial locality, cold misses.
Exercise 2 (1/2)

```c
#define S ((1<<20)*sizeof(int))
int A[S];
for (i = 0; i < S; i++) {
    read A[0];
}
```

Total access time? What kind of locality? What kind of misses?
#define S ((1<<20)*sizeof(int))
int A[S];
for (i = 0; i < S; i++) {
    read A[0];
}

- S reads to A
- All except the first one hit in cache.
- Total access time: $100 + (S - 1)$.
- Temporal locality
- Cold misses.
Exercise 3 (1/2)

// Assume 4 <= N <= 13
#define S ((1<<20)*sizeof(int))
int A[S];
for (i = 0; i < S; i++) {
    read A[i % (1<<N)];
}

Total access time? What kind of locality? What kind of misses?
Exercise 3 (2/2)

```c
// Assume 4 <= N <= 13
#define S ((1<<20)*sizeof(int))
int A[S];
for (i = 0; i < S; i++) {
    read A[i % (1<<N)];
}
```

- S reads to A
- One miss for each accessed line, rest hit in cache.
- Number of accessed lines: $2^{N-4}$.
- Total access time: $2^{N-4}100 + (S - 2^{N-4})$.
- Temporal and spatial locality
- Cold misses.
Exercise 4 (1/2)

// Assume 14 <= N
#define S ((1<<20)*sizeof(int))
int A[S];
for (i = 0; i < S; i++) {
    read A[i % (1<<N)];
}

Total access time? What kind of locality? What kind of misses?
Exercise 4 (2/2)

// Assume 14 <= N
#define S ((1<<20)*sizeof(int))
int A[S];
for (i = 0; i < S; i++) {
    read A[i % (1<<N)];
}

▶ S reads to A.
▶ First access to each line misses
▶ Rest accesses to that line hit.
▶ Total access time: $15(S/16) + 100(S/16)$.
▶ Spatial locality
▶ Cold and capacity misses.
Exercise 5 (1/2)

// Assume 14 <= N
#define S ((1<<20)*sizeof(int))
int A[S];
for (i = 0; i < S; i++) {
    read A[(i*16) % (1<<N)];
}

Total access time? What kind of locality? What kind of misses?
Exercise 5 (2/2)

// Assume 14 <= N
#define S ((1<<20)*sizeof(int))
int A[S];
for (i = 0; i < S; i++) {
    read A[(i*16) % (1<<N)];
}

- S reads to A.
- First access to each line misses
- One access per line.
- Total access time: 100S.
- No locality!
- Cold and conflict misses.
Exercise 6 (1/2)

```c
#define S ((1<<20)*sizeof(int))
int A[S];
for (i = 0; i < S; i++) {
    read A[random()%S];
}
```

Total access time? What kind of locality? What kind of misses?
#define S ((1<<20)*sizeof(int))

int A[S];

for (i = 0; i < S; i++) {
    read A[random()%S];
}

- S reads to A.
- After N iterations, for some N, the cache is full.
- Then the chance of hitting in cache is $32Kb/16b = 1/512$
- Estimated total access time: $S(511/512)100 + S(1/512)$.
- Almost no locality!
- Cold, capacity conflict misses.
Exercise 7 (1/2)

```c
#define S ((1<<19)*sizeof(int))
int A[S];
int B[S];
for (i = 0; i < S; i++) {
    read A[i], B[i];
}
```

Total access time? What kind of locality? What kind of misses?
Exercise 7 (2/2)

```
#define S ((1<<19)*sizeof(int))
int A[S];
int B[S];
for (i = 0; i < S; i++) {
    read A[i], B[i];
}
```

- S reads to A and B.
- A and B interfere in cache: indeed two cache lines whose addresses differ by a multiple of $2^9$ have the same way to cache.
- Total access time: $200S$.
- Spatial locality but the cache cannot exploit it.
- Cold and conflict misses.
Exercise 8 (1/2)

```c
#define S ((1<<19+4)*sizeof(int))
int A[S];
int B[S];
for (i = 0; i < S; i++) {
    read A[i], B[i];
}
```

Total access time? What kind of locality? What kind of misses?
#define S ((1<<19+4)*sizeof(int))
int A[S];
int B[S];
for (i = 0; i < S; i++) {
read A[i], B[i];
}

- S reads to A and B.
- A and B almost do not interfere in cache.
- Total access time: 2(15S/16 + 100S/16).
- Spatial locality.
- Cold misses.
Set associative caches have sets with multiple lines per set.

Each line in a set is called a way.

Each memory line maps to a specific set and can be put into any cache line in its set.

In our example, we assume a 32 Kbyte cache, with 64 byte lines, 2-way associative. Hence we have:

- 256 sets
- Bottom six bits determine offset in cache line
- Next 8 bits determine the set.
Exercise 9 (1/2)

```c
#define S ((1<<19)*sizeof(int))
int A[S];
int B[S];
for (i = 0; i < S; i++) {
    read A[i], B[i];
}
```

Total access time? What kind of locality? What kind of misses?
Exercise 9 (2/2)

```c
#define S ((1<<19)*sizeof(int))
int A[S];
int B[S];
for (i = 0; i < S; i++) {
    read A[i], B[i];
}
```

- S reads to A and B.
- A and B lines hit same set, but enough lines in a set.
- Total access time: $2\left(\frac{15S}{16} + \frac{100S}{16}\right)$.
- Spatial locality.
- Cold misses.
Tuned cache-oblivious square matrix transposition

```c
void DC_matrix_transpose(int *A, int lda, int i0, int i1,
    int j0, int dj0, int j1 /*, int dj1 = 0 */){
    const int THRESHOLD = 16; // tuned for the target machine
    tail:
    int di = i1 - i0, dj = j1 - j0;
    if (dj >= 2 * di && dj > THRESHOLD) {
        int dj2 = dj / 2;
        cilk_spawn DC_matrix_transpose(A, lda, i0, i1, j0, dj0, j0 + dj2);
        j0 += dj2; dj0 = 0; goto tail;
    } else if (di > THRESHOLD) {
        int di2 = di / 2;
        cilk_spawn DC_matrix_transpose(A, lda, i0, i0 + di2, j0, dj0, j1);
        i0 += di2; j0 += dj0 * di2; goto tail;
    } else {
        for (int i = i0; i < i1; ++i) {
            for (int j = j0; j < j1; ++j) {
                int x = A[j * lda + i];
                A[j * lda + i] = A[i * lda + j];
                A[i * lda + j] = x;
            }
            j0 += dj0;
        }
    }
}
```
Tuned cache-oblivious matrix transposition benchmarks

<table>
<thead>
<tr>
<th>size</th>
<th>Naive</th>
<th>Cache-oblivious</th>
<th>ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>5000x5000</td>
<td>126</td>
<td>79</td>
<td>1.59</td>
</tr>
<tr>
<td>10000x10000</td>
<td>627</td>
<td>311</td>
<td>2.02</td>
</tr>
<tr>
<td>20000x20000</td>
<td>4373</td>
<td>1244</td>
<td>3.52</td>
</tr>
<tr>
<td>30000x30000</td>
<td>23603</td>
<td>2734</td>
<td>8.63</td>
</tr>
<tr>
<td>40000x40000</td>
<td>62432</td>
<td>4963</td>
<td>12.58</td>
</tr>
</tbody>
</table>

- Intel(R) Xeon(R) CPU E7340 @ 2.40GHz
- L1 data 32 KB, L2 4096 KB, cache line size 64 bytes
- Both codes run on 1 core
- The ratio comes simply from an optimal memory access pattern.
Tuned cache-oblivious matrix multiplication
#define S ((1<<19)*sizeof(int))
int A[S];
int B[S];
int C[S];
for (i = 0; i < S; i++) {
    C[i] := A[i] + B[i];
}

For the above 2-way associative cache (of size 32 Kbyte cache, and with 64 byte lines): Total access time? What kind of locality? What kind of misses?
Extra Exercise B

Let $A$ be a $n \times n$ invertible lower triangular matrix. A simple divide-and-conquer strategy to invert $A$ is described below. Let $A$ be partitioned into $(n/2) \times (n/2)$ blocks as follows:

\[
A = \begin{bmatrix}
A_1 & 0 \\
A_2 & A_3
\end{bmatrix},
\]

(1)

where $n$ is assumed to be a power of 2. Clearly $A_1$ and $A_3$ are invertible lower triangular matrices. $A^{-1}$ is given by

\[
A^{-1} = \begin{bmatrix}
A_1^{-1} & 0 \\
-A_3^{-1}A_2A_1^{-1} & A_3^{-1}
\end{bmatrix}
\]

(2)

Write a Cilk-like program computing $A^{-1}$. Analyze the work and critical path of your parallel program. We can assume that we have a sub-routine computing the product (resp. sum) of 2 square matrices of order $n$ in work $\Theta(n^3)$ (resp. $\Theta(n^2)$) and span $\Theta(\log^2(n))$ (resp. $\Theta(\log(n))$).

Same question using a space-efficient (i.e. in place) multiplication.
Plan

Hierarchical memories and their impact on our programs

Cache Analysis in Practice

The Ideal-Cache Model

Cache Complexity of some Basic Operations

Matrix Transposition

A Cache-Oblivious Matrix Multiplication Algorithm
The \((Z, L)\) ideal cache model (1/4)

**Figure 1:** The ideal-cache model
The \((Z, L)\) ideal cache model (2/4)

- Computer with a **two-level memory hierarchy**:
  - an ideal (data) cache of \(Z\) words partitioned into \(Z/L\) cache lines, where \(L\) is the number of words per cache line.
  - an arbitrarily large main memory.
- Data moved between cache and main memory are always cache lines.
- The cache is **tall**, that is, \(Z\) is much larger than \(L\), say \(Z \in \Omega(L^2)\).
The \((Z, L)\) ideal cache model (3/4)

- The processor can only reference words that reside in the cache.
- If the referenced word belongs to a line already in cache, a **cache hit** occurs, and the word is delivered to the processor.
- Otherwise, a **cache miss** occurs, and the line is fetched into the cache.

**Figure 1:** The ideal-cache model
The $(Z, L)$ ideal cache model (4/4)

- The ideal cache is **fully associative**: cache lines can be stored anywhere in the cache.
- The ideal cache uses the **optimal off-line strategy of replacing** the cache line whose next access is furthest in the future, and thus it exploits temporal locality perfectly.

**Figure 1**: The ideal-cache model
Cache complexity

- For an algorithm with an input of size $n$, the ideal-cache model uses two complexity measures:
  - the **work complexity** $W(n)$, which is its conventional running time in a RAM model.
  - the **cache complexity** $Q(n; Z, L)$, the number of cache misses it incurs (as a function of the size $Z$ and line length $L$ of the ideal cache).
  - When $Z$ and $L$ are clear from context, we simply write $Q(n)$ instead of $Q(n; Z, L)$.

- An algorithm is said to be **cache aware** if its behavior (and thus performances) can be tuned (and thus depend on) on the particular cache size and line length of the targeted machine.
- Otherwise the algorithm is **cache oblivious**.
Cache complexity of the naive matrix multiplication

```c
// A is stored in ROW-major and B in COLUMN-major
for(i=0; i < n; i++)
    for(j=0; j < n; j++)
        for(k=0; k < n; k++)
            C[i][j] += A[i][k] * B[j][k];
```

- Assuming $Z \geq 3L$, computing each $C[i][j]$ incurs $O(1 + n/L)$ caches misses.
- If $Z$ large enough, say $Z \in \Omega(n)$ then the row $i$ of $A$ will be remembered for its entire involvement in computing row $i$ of $C$.
- For column $j$ of $B$ to be remembered when necessary, one needs $Z \in \Omega(n^2)$ in which case the whole computation fits in cache. Therefore, we have

$$Q(n, Z, L) = \begin{cases} 
O(n + n^3/L) & \text{if } 3L \leq Z < n^2 \\
O(1 + n^2/L) & \text{if } 3n^2 \leq Z.
\end{cases}$$
A cache-aware matrix multiplication algorithm (1/2)

// A, B and C are in row-major storage
for(i =0; i < n/s; i++)
  for(j =0; j < n/s; j++)
    for(k=0; k < n/s; k++)
      blockMult(A,B,C,i,j,k,s);

- Each matrix $M \in \{A, B, C\}$ consists of $(n/s) \times (n/s)$ submatrices $M_{ij}$ (the blocks), each of which has size $s \times s$, where $s$ is a tuning parameter.
- Assume $s$ divides $n$ to keep the analysis simple.
- $\text{blockMult}(A,B,C,i,j,k,s)$ computes $C_{ij} = A_{ik} \times B_{kj}$ using the naive algorithm
A cache-aware matrix multiplication algorithm (2/2)

```c
// A, B and C are in row-major storage
for(i =0; i < n/s; i++)
  for(j =0; j < n/s; j++)
    for(k=0; k < n/s; k++)
      blockMult(A,B,C,i,j,k,s);
```

- Choose $s$ to be the largest value such that three $s \times s$ submatrices simultaneously fit in cache, that is, $Z \in \Theta(s^2)$, that is, $s \in \Theta(\sqrt{Z})$.
- An $s \times s$ submatrix is stored on $\Theta(s + s^2/L)$ cache lines.
- Thus `blockMult(A,B,C,i,j,k,s)` runs within $\Theta(s + s^2/L)$ cache misses.
- Initializing the $n^2$ elements of $C$ amounts to $\Theta(1 + n^2/L)$ caches misses. Therefore we have

$$Q(n, Z, L) \in \Theta(1 + n^2/L + (n/\sqrt{Z})^3(\sqrt{Z} + Z/L))$$

$$\in \Theta(1 + n^2/L + n^3/Z + n^3/(L\sqrt{Z})).$$
Plan

Hierarchical memories and their impact on our programs

Cache Analysis in Practice

The Ideal-Cache Model

Cache Complexity of some Basic Operations

Matrix Transposition

A Cache-Oblivious Matrix Multiplication Algorithm
Scanning

Scanning $n$ elements stored in a contiguous segment (= cache lines) of memory costs at most $\lceil n/L \rceil + 1$ cache misses.

- In the above, $N = n$ and $B = L$. The main issue here is alignment.
- Let $(q, r)$ be the quotient and remainder in the integer division of $n$ by $L$. Let $u$ (resp. $w$) be \# words in a fully (not fully) used cache line.
- If $w = 0$ then $r = 0$ and the conclusion is clear.
- If $w < L$ then $r = w$ and the conclusion is clear again.
- If $L \leq w < 2L$ then $qL = u + 1$ and the conclusion follows.

Figure 2. Scanning an array of $N$ elements arbitrarily aligned with blocks may cost one more memory transfer than $\lceil N/B \rceil$.
Array reversal

Reversing an array of $n$ elements stored in a contiguous segment (= cache lines) of memory costs at most $\lceil n/L \rceil + 1$ cache misses, provided that $Z \geq 2L$ holds. Exercise!

Figure 3. Bentley’s reversal of an array.
A selection algorithm is an algorithm for finding the \( k \)-th smallest number in a list. This includes the cases of finding the minimum, maximum, and median elements.

A worst-case linear algorithm for the general case of selecting the \( k \)-th largest element was published by Blum, Floyd, Pratt, Rivest, and Tarjan in their 1973 paper *Time bounds for selection*, sometimes called BFPRT.

The principle is the following:
- Find a pivot that allows splitting the list into two parts of nearly equal size such that
- the search can continue in one of them.
Median and selection (2/8)

select(L,k)
{
if (L has 10 or fewer elements)
{
    sort L
    return the element in the kth position
}

partition L into subsets S[i] of five elements each
   (there will be n/5 subsets total).

for (i = 1 to n/5) do
    x[i] = select(S[i],3)

M = select({x[i]}, n/10)

partition L into L1<M, L2=M, L3>M
if (k <= length(L1))
    return select(L1,k)
else if (k > length(L1)+length(L2))
    return select(L3,k-length(L1)-length(L2))
else return M
Median and selection (3/8)

For an input list of \( n \) elements, the number \( T(n) \) of comparisons satisfies

\[
T(n) \leq 12n/5 + T(n/5) + T(7n/10).
\]

- We always throw away either \( L_3 \) (the values greater than \( M \)) or \( L_1 \) (the values less than \( M \)). Suppose we throw away \( L_3 \).
- Among the \( n/5 \) values \( x[i] \), \( n/10 \) are larger than \( M \), since \( M \) was defined to be the median of these values.
- For each \( i \) such that \( x[i] \) is larger than \( M \), two other values in \( S[i] \) are also larger than \( x[i] \).
- So \( L_3 \) has at least \( 3n/10 \) elements. By a symmetric argument, \( L_1 \) has at least \( 3n/10 \) elements.
- Therefore the final recursive call is on a list of at most \( 7n/10 \) elements and takes time at most \( T(7n/10) \).
Median and selection (4/8)

How to solve

$$T(n) \leq 12n/5 + T(n/5) + T(7n/10)?$$

- We “try” $T(n) \leq cn$ by induction. The substitution gives
  $$T(n) \leq n(12/5 + 9c/10).$$

  From $n(12/5 + 9c/10) \leq cn$ we derive $c \leq 24$.

- The tree-based method also brings $T(n) \leq 24n$.

- The same tree-expansion method then shows that, more generally, if $T(n) \leq cn + T(an) + T(bn)$, where $a + b < 1$, the total time is $c(1/(1 - a - b))n$.

- With a lot of work one can reduce the number of comparisons to $2.95n$ [D. Dor and U. Zwick, Selecting the Median, 6th SODA, 1995].
In order to analyze its cache complexity, let us review the algorithm and consider an array instead of a list.

**Step 1:** Conceptually partition the array into \( \frac{n}{5} \) quintuplets of five adjacent elements each.

**Step 2:** Compute the median of each quintuplet using \( O(1) \) comparisons.

**Step 3:** Recursively compute the median of these medians (which is not necessarily the median of the original array).

**Step 4:** Partition the elements of the array into three groups, according to whether they equal, or strictly less or strictly greater than this median of medians.

**Step 5:** Count the number of elements in each group, and recurse into the group that contains the element of the desired rank.
Median and selection (6/8)

To make this algorithm cache-oblivious, we specify how each step works in terms of memory layout and scanning. We assume that $Z \geq 3L$.

**Step 1:** Just conceptual; no work needs to be done.

**Step 2:** requires two parallel scans, one reading the 5 element arrays at a time, and the other writing a new array of computed medians, incurring $\Theta(1 + n/L)$.

**Step 3:** Just a recursive call on size $n/5$.

**Step 4:** Can be done with three parallel scans, one reading the array, and two others writing the partitioned arrays, incurring again $\Theta(1 + n/L)$.

**Step 5:** Just a recursive call on size $7n/10$.

This leads to

$$Q(n) \leq Q(n/5) + Q(7n/10) + \Theta(1 + n/L).$$
Median and selection (7/8)

How to solve

\[ Q(n) \leq Q(n/5) + Q(7n/10) + \Theta(1 + n/L)? \]

The unknown is what is the base-case?

- Suppose the base case is \( Q(0(1)) \in O(1) \).
- Following Master Theorem proof the number of leaves \( L(n) = n^c \) satisfies in \( N(n) = N(n/5) + N(7n/10), N(1) = 1 \), which brings

\[ \left( \frac{1}{5} \right)^c + \left( \frac{7}{10} \right)^c = 1 \]

leading to \( c \approx 0.8397803 \).
- Since each leaf incurs a constant number of cache misses we have \( Q(n) \in \Omega(n^c) \), which could be larger or smaller than \( \Theta(1 + n/L) \) . . .
Median and selection (8/8)

How to solve

\[ Q(n) \leq Q(n/5) + Q(7n/10) + \Theta(1 + n/L) \]

- Fortunately, we have a better base-case: \( Q(0(L)) \in O(1) \).
- Indeed, once the problem fits into \( O(1) \) cache-lines, all five steps incur only a constant number of cache misses.
- Thus we have only \( (n/L)^c \) leaves in the recursion tree.
- In total, these leaves incur \( O((n/L)^c) = o(n/L) \) cache misses.
- In fact, the cost per level decreases geometrically from the root, so the total cost is the cost of the root. Finally we have

\[ Q(n) \in \Theta(1 + n/L) \]
Plan

Hierarchical memories and their impact on our programs

Cache Analysis in Practice

The Ideal-Cache Model

Cache Complexity of some Basic Operations

Matrix Transposition

A Cache-Oblivious Matrix Multiplication Algorithm
Matrix transposition: various algorithms

- **Matrix transposition problem:** Given an $m \times n$ matrix $A$ stored in a row-major layout, compute and store $A^T$ into an $n \times m$ matrix $B$ also stored in a row-major layout.

- We shall describe a recursive cache-oblivious algorithm which uses $\Theta(mn)$ work and incurs $\Theta(1 + mn/L)$ cache misses, which is optimal.

- The straightforward algorithm employing doubly nested loops incurs $\Theta(mn)$ cache misses on one of the matrices when $m \gg Z/L$ and $n \gg Z/L$.

- We shall start with an apparently good algorithm and use complexity analysis to show that it is even worse than the straightforward algorithm.
Matrix transposition: a first divide-and-conquer (1/4)

- For simplicity, assume that our input matrix $A$ is square of order $n$ and that $n$ is a power of 2, say $n = 2^k$.
- We divide $A$ into four square quadrants of order $n/2$ and we have

\[
\begin{pmatrix} A_{1,1} & A_{1,2} \\ A_{2,1} & A_{2,2} \end{pmatrix} \Rightarrow \begin{pmatrix} ^tA_{1,1} & ^tA_{2,1} \\ ^tA_{1,2} & ^tA_{2,2} \end{pmatrix}.
\]

- This observation yields an “in-place” algorithm:
  1. If $n = 1$ then return $A$.
  2. If $n > 1$ then
    2.1 recursively compute $^tA_{1,1}, ^tA_{1,2}, ^tA_{2,1}, ^tA_{2,2}$ in place as

\[
\begin{pmatrix} ^tA_{1,1} & ^tA_{1,2} \\ ^tA_{2,1} & ^tA_{2,2} \end{pmatrix}
\]

    2.2 exchange $^tA_{1,2}$ and $^tA_{2,1}$.

- What is the number $M(n)$ of memory accesses to $A$, performed by this algorithm on an input matrix $A$ of order $n$?
Matrix transposition: a first divide-and-conquer (2/4)

- $M(n)$ satisfies the following recurrence relation

$$M(n) = \begin{cases} 
0 & \text{if } n = 1 \\
4M(n/2) + 2(n/2)^2 & \text{if } n > 1.
\end{cases}$$

- Unfolding the tree of recursive calls or using the Master’s Theorem, one obtains:

$$M(n) = 2(n/2)^2 \log_2(n).$$

- This is worse than the straightforward algorithm (which employs doubly nested loops). Indeed, for this latter, we have $M(n) = n^2 - n$. Explain why!

- Despite of this negative result, we shall analyze the cache complexity of this first divide-and-conquer algorithm. Indeed, it provides us with an easy training exercise

- We shall study later a second and efficiency-optimal divide-and-conquer algorithm, whose cache complexity analysis is more involved.
Matrix transposition: a first divide-and-conquer (3/4)

- We shall determine $Q(n)$ the number of cache misses incurred by our first divide-and-conquer algorithm on a $(Z, L)$-ideal cache machine.
- For $n$ small enough, the entire input matrix or the entire block (input of some recursive call) fits in cache and incurs only the cost of a scanning. Because of possible misalignment, that is, $n(\lceil n/L \rceil + 1)$.
- **Important:** For simplicity, some authors write $n/L$ instead of $\lceil n/L \rceil$. This can be dangerous.
- **However:** these simplifications are fine for asymptotic estimates, keeping in mind that $n/L$ is a rational number satisfying

$$n/L - 1 \leq \lfloor n/L \rfloor \leq n/L \leq \lceil n/L \rceil \leq n/L + 1.$$ 

Thus, for a fixed $L$, the functions $\lfloor n/L \rfloor$, $n/L$ and $\lceil n/L \rceil$ are asymptotically of the same order of magnitude.

- We need to translate “for $n$ small enough” into a formula. We claim that there exists a real constant $\alpha > 0$ s.t. for all $n$ and $Z$ we have

$$n^2 < \alpha Z \implies Q(n) \leq n^2/L + n.$$
Matrix transposition: a first divide-and-conquer (4/4)

- **Q(n)** satisfies the following recurrence relation

\[
Q(n) = \begin{cases} 
  \frac{n^2}{L} + n & \text{if } n^2 < \alpha Z \quad \text{(base case)} \\
  4Q(n/2) + \frac{n^2}{2L} + n & \text{if } n^2 \geq \alpha Z \quad \text{(recurrence)}
\end{cases}
\]

- Indeed, **exchanging 2 blocks** amount to \(2((n/2)^2/L + n/2)\) accesses.

- Unfolding the recurrence relation \(k\) times (more details in class) yields

\[
Q(n) = 4^k Q\left(\frac{n}{2^k}\right) + k \frac{n^2}{2L} + (2^k - 1)n.
\]

- The minimum \(k\) for reaching the base case satisfies \(\frac{n^2}{4^k} = \alpha Z\), that is, \(4^k = \frac{n^2}{\alpha Z}\), that is, \(k = \log_4\left(\frac{n^2}{\alpha Z}\right)\). This implies \(2^k = \frac{n}{\sqrt{\alpha Z}}\) and thus

\[
Q(n) \leq \frac{n^2}{\alpha Z} \left(\alpha Z/L + \sqrt{\alpha Z}\right) + \log_4\left(\frac{n^2}{\alpha Z}\right) \frac{n^2}{2L} + \frac{n}{\sqrt{\alpha Z}} n \\
\leq n^2/L + 2 \frac{n^2}{\sqrt{\alpha Z}} + \log_4\left(\frac{n^2}{\alpha Z}\right) \frac{n^2}{2L}.
\]
A matrix transposition cache-oblivious algorithm (1/2)

- If \( n \geq m \), the \texttt{Rec-Transpose} algorithm partitions

\[
A = (A_1 \ A_2) , \quad B = \begin{pmatrix} B_1 \\ B_2 \end{pmatrix}
\]

and recursively executes \texttt{Rec-Transpose}(\(A_1, B_1\)) and \texttt{Rec-Transpose}(\(A_2, B_2\)).

- If \( m > n \), the \texttt{Rec-Transpose} algorithm partitions

\[
A = \begin{pmatrix} A_1 \\ A_2 \end{pmatrix} , \quad B = (B_1 \ B_2)
\]

and recursively executes \texttt{Rec-Transpose}(\(A_1, B_1\)) and \texttt{Rec-Transpose}(\(A_2, B_2\)).
Recall that the matrices are stored in row-major layout.

Let $\alpha$ be a constant sufficiently small such that the following two conditions hold:

(i) two sub-matrices of size $m \times n$ and $n \times m$, where $\max\{m, n\} \leq \alpha L$, fit in cache

(ii) even if each row starts at a different cache line.

We distinguish three cases for the input matrix $A$:

- Case I: $\max\{m, n\} \leq \alpha L$.
- Case II: $m \leq \alpha L < n$ or $n \leq \alpha L < m$.
- Case III: $m, n > \alpha L$. 
Case I: \( \max \{m, n\} \leq \alpha L \).

- Both matrices fit in \( O(1) + 2mn/L \) lines.

- From the choice of \( \alpha \), the number of lines required for the entire computation is at most \( Z/L \).

- Thus, no cache lines need to be evicted during the computation. Hence, it feels like we are simply scanning \( A \) and \( B \).

- Therefore \( Q(m, n) \in O(1 + mn/L) \).
Case II: $m \leq \alpha L < n$ or $n \leq \alpha L < m$.

- Consider $n \leq \alpha L < m$. The $\text{Rec-Transpose}$ algorithm divides the greater dimension $m$ by 2 and recurses.
- At some point in the recursion, we have $\alpha L/2 \leq m \leq \alpha L$ and the whole computation fits in cache. At this point:
  - the input array resides in contiguous locations, requiring at most $\Theta(1 + nm/L)$ cache misses
  - the output array consists of $nm$ elements in $n$ rows, where in the worst case every row starts at a different cache line, leading to at most $\Theta(n + nm/L)$ cache misses.
- Since $m/L \in [\alpha/2, \alpha]$, the total cache complexity for this base case is $\Theta(1 + n)$, yielding the recurrence (where the resulting $Q(m, n)$ is a worst case estimate)

$$Q(m, n) = \begin{cases} 
\Theta(1 + n) & \text{if } m \in [\alpha L/2, \alpha L], \\
2Q(m/2, n) + O(1) & \text{otherwise} 
\end{cases}$$

whose solution satisfies $Q(m, n) = \Theta(1 + mn/L)$. 
Case III: $m, n > \alpha L$.

- As in Case II, at some point in the recursion both $n$ and $m$ fall into the range $[\alpha L/2, \alpha L]$.
- The whole problem fits into cache and can be solved with at most $\Theta(m + n + mn/L)$ cache misses.
- The worst case cache miss estimate satisfies the recurrence
  \[
  Q(m, n) = \begin{cases} 
  \Theta(m + n + mn/L) & \text{if } m, n \in [\alpha L/2, \alpha L], \\
  2Q(m/2, n) + O(1) & \text{if } m \geq n, \\
  2Q(m, n/2) + O(1) & \text{otherwise};
  \end{cases}
  \]
  whose solution is $Q(m, n) = \Theta(1 + mn/L)$.
- Therefore, the Rec-Transpose algorithm has optimal cache complexity.
- Indeed, for an $m \times n$ matrix, the algorithm must write to $mn$ distinct elements, which occupy at least $\lceil mn/L \rceil$ cache lines.
Plan

Hierarchical memories and their impact on our programs

Cache Analysis in Practice

The Ideal-Cache Model

Cache Complexity of some Basic Operations

Matrix Transposition

A Cache-Oblivious Matrix Multiplication Algorithm
A cache-oblivious matrix multiplication algorithm (1/3)

- We describe and analyze a cache-oblivious algorithm for multiplying an $m \times n$ matrix by an $n \times p$ matrix cache-obliviously using
  - $\Theta(mnp)$ work and incurring
    - $\Theta(m + n + p + (mn + np + mp)/L + mnp/(L\sqrt{Z}))$ cache misses.

- This straightforward divide-and-conquer algorithm contains no voodoo parameters (tuning parameters) and it uses cache optimally.

- Intuitively, this algorithm uses the cache effectively, because once a subproblem fits into the cache, its smaller subproblems can be solved in cache with no further cache misses.

- These results require the tall-cache assumption for matrices stored in row-major layout format,

- This assumption can be relaxed for certain other layouts, see (Frigo et al. 1999).

- The case of Strassen’s algorithm is also treated in (Frigo et al. 1999).
To multiply an $m \times n$ matrix $A$ and an $n \times p$ matrix $B$, the \texttt{ReC-Mult} algorithm halves the largest of the three dimensions and recurs according to one of the following three cases:

\[(A_1) \quad B = (A_1 B) \quad , \quad (A_2 B) , \quad (3)\]

\[(A_1 \quad A_2) \quad (B_1) = A_1 B_1 + A_2 B_2 , \quad (4)\]

\[A \quad (B_1 \quad B_2) = (A B_1 \quad A B_2) . \quad (5)\]

- In case (3), we have $m \geq \max \{n, p\}$. Matrix $A$ is split horizontally, and both halves are multiplied by matrix $B$.
- In case (4), we have $n \geq \max \{m, p\}$. Both matrices are split, and the two halves are multiplied.
- In case (5), we have $p \geq \max \{m, n\}$. Matrix $B$ is split vertically, and each half is multiplied by $A$.
- The base case occurs when $m = n = p = 1$. 

let \( \alpha > 0 \) be the largest constant sufficiently small that three submatrices of sizes \( m' \times n', n' \times p', \) and \( m' \times p' \) all fit completely in the cache, whenever \( \max \{ m', n', p' \} \leq \alpha \sqrt{Z} \) holds.

We distinguish four cases depending on the initial size of the matrices.

- **Case I:** \( m, n, p > \alpha \sqrt{Z} \).
- **Case II:** \( (m \leq \alpha \sqrt{Z} \) and \( n, p > \alpha \sqrt{Z} \) or \( (n \leq \alpha \sqrt{Z} \) and \( m, p > \alpha \sqrt{Z} \)) or \( (p \leq \alpha \sqrt{Z} \) and \( m, n > \alpha \sqrt{Z} \))
- **Case III:** \( (n, p \leq \alpha \sqrt{Z} \) and \( m > \alpha \sqrt{Z} \) or \( (m, p \leq \alpha \sqrt{Z} \) and \( n > \alpha \sqrt{Z} \)) or \( (m, n \leq \alpha \sqrt{Z} \) and \( p > \alpha \sqrt{Z} \))
- **Case IV:** \( m, n, p \leq \alpha \sqrt{Z} \).

Similarly to matrix transposition, \( Q(m, n, p) \) is a **worst case cache miss estimate**.
Case I: $m, n, p > \alpha \sqrt{Z}$. (1/2)

$$Q(m, n, p) = \begin{cases} 
\Theta((mn + np + mp)/L) & \text{if } m, n, p \in [\alpha \sqrt{Z}/2, \alpha \sqrt{Z}], \\
2Q(m/2, n, p) + O(1) & \text{ow. if } m \geq n \text{ and } m \geq p, \\
2Q(m, n/2, p) + O(1) & \text{ow. if } n > m \text{ and } n \geq p, \\
2Q(m, n, p/2) + O(1) & \text{otherwise}. 
\end{cases}$$

- The base case arises as soon as all three submatrices fit in cache:
  - The total number of cache lines used by the three submatrices is $\Theta((mn + np + mp)/L)$.
  - The only cache misses that occur during the remainder of the recursion are the $\Theta((mn + np + mp)/L)$ cache misses required to bring the matrices into cache.
Case I: $m, n, p > \alpha \sqrt{Z}$. (2/2)

$$Q(m, n, p) =$$

\begin{align*}
\Theta((mn + np + mp)/L) & \quad \text{if } m, n, p \in \left[\alpha \sqrt{Z}/2, \alpha \sqrt{Z}\right], \\
2Q(m/2, n, p) + O(1) & \quad \text{ow. if } m \geq n \text{ and } m \geq p, \\
2Q(m, n/2, p) + O(1) & \quad \text{ow. if } n > m \text{ and } n \geq p, \\
2Q(m, n, p/2) + O(1) & \quad \text{otherwise}.
\end{align*}

- In the recursive cases, when the matrices do not fit in cache, we pay for the cache misses of the recursive calls, plus $O(1)$ cache misses for the overhead of manipulating submatrices.
- The solution to this recurrence is
  $$Q(m, n, p) = \Theta(mnp/(L\sqrt{Z})).$$

- Indeed, for the base-case $m, m, p \in \Theta(\alpha \sqrt{Z})$. 
Case II: \((m \leq \alpha \sqrt{Z})\) and \((n, p > \alpha \sqrt{Z})\).

- Here, we shall present the case where \(m \leq \alpha \sqrt{Z}\) and \(n, p > \alpha \sqrt{Z}\).
- The \textsc{Rec-Mult} algorithm always divides \(n\) or \(p\) by 2 according to cases (4) and (5).
- At some point in the recursion, both \(n\) and \(p\) are small enough that the whole problem fits into cache.
- The number of cache misses can be described by the recurrence

\[
Q(m, n, p) =
\begin{cases}
\Theta(1 + n + m + np/L) & \text{if } n, p \in [\alpha \sqrt{Z}/2, \alpha \sqrt{Z}], \\
2Q(m, n/2, p) + O(1) & \text{otherwise if } n \geq p, \\
2Q(m, n, p/2) + O(1) & \text{otherwise};
\end{cases}
\]

whose solution is \(Q(m, n, p) = \Theta(np/L + mnp/(L\sqrt{Z}))\).
- Indeed, in the base case: \(mnp/(L\sqrt{Z}) \leq \alpha np/L\).
- The term \(\Theta(1 + n + m)\) appears because of the row-major layout.
Case III: \((n, p \leq \alpha \sqrt{Z} \text{ and } m > \alpha \sqrt{Z})\)

- In each of these cases, one of the matrices fits into cache, and the others do not.
- Here, we shall present the case where \(n, p \leq \alpha \sqrt{Z}\) and \(m > \alpha \sqrt{Z}\).
- The \texttt{Rec-Mult} algorithm always divides \(m\) by 2 according to case (3).
- At some point in the recursion, \(m\) falls into the range \(\alpha \sqrt{Z}/2 \leq m \leq \alpha \sqrt{Z}\), and the whole problem fits in cache.
- The number cache misses can be described by the recurrence

\[
Q(m, n, p) = \begin{cases} 
\Theta(1 + m) & \text{if } m \in [\alpha \sqrt{Z}/2, \alpha \sqrt{Z}], \\
2Q(m/2, n, p) + O(1) & \text{otherwise};
\end{cases}
\]

whose solution is \(Q(m, n, p) = \Theta(m + mnp/(L\sqrt{Z}))\).
- Indeed, in the base case: \(mnp/(L\sqrt{Z}) \leq \alpha \sqrt{Z}m/L\); moreover \(Z \in \Omega(L^2)\) (tall cache assumption).
Case IV: $m, n, p \leq \alpha \sqrt{Z}$.

- From the choice of $\alpha$, all three matrices fit into cache.
- The matrices are stored on $\Theta(1 + mn/L + np/L + mp/L)$ cache lines.
- Therefore, we have $Q(m, n, p) = \Theta(1 + (mn + np + mp)/L)$. 
Figure 2: Layout of a 16 × 16 matrix in (a) row major, (b) column major, (c) 4 × 4-blocked, and (d) bit-interleaved layouts.
Acknowledgements and references

Acknowledgements.

▶ Charles E. Leiserson (MIT) and Matteo Frigo (Intel) for providing me with the sources of their article *Cache-Oblivious Algorithms*.

▶ Charles E. Leiserson (MIT) and Saman P. Amarasinghe (MIT) for sharing with me the sources of their course notes and other documents.

References.


▶ *Cache-Oblivious Algorithms and Data Structures* by Erik D. Demaine.