Lecture 13
Cache Performance

CS213 – Intro to Computer Systems

Branden Ghena – Winter 2024

Slides adapted from:
St-Amour, Hardavellas, Bustamente (Northwestern), Bryant, O’Hallaron (CMU), Garcia, Weaver (UC Berkeley)

Northwestern

Today’s Goals

- Explore impacts of cache and code design

- Calculate cache performance based on array accesses

- Understand what it means to write “cache-friendly code”

Outline

- **Memory Mountain**
- Cache Metrics
- Cache Performance for Arrays
- Improving code
  - Rearranging Matrix Math
  - Matrix Math in Blocks

Writing Cache-Friendly Code

- Caches are key to program performance
  - CPU accessing main memory = CPU twiddling its thumbs = bad
  - Want to avoid as much as possible

- Minimize cache misses in the inner loops of core functions
  - That’s usually where your program spends most of its time (“hot” code)
    - Programmers are notoriously bad at guessing these spots
    - Use a profiler to find them (e.g., **`gprof`**)

  - Repeated references to variables are good (***temporal locality***)
  - Stride-1 reference patterns are good (***spatial locality***)
    - I.e., accessing array elements in sequence, not jumping around

- Now that we know how cache memories work
  - We can quantify the effect of locality on performance

The Memory Mountain

- *Read throughput* (read bandwidth)
  - Number of bytes read from the memory subsystem per second (MB/s)
  - The higher it is, the less likely your CPU is to be waiting on memory

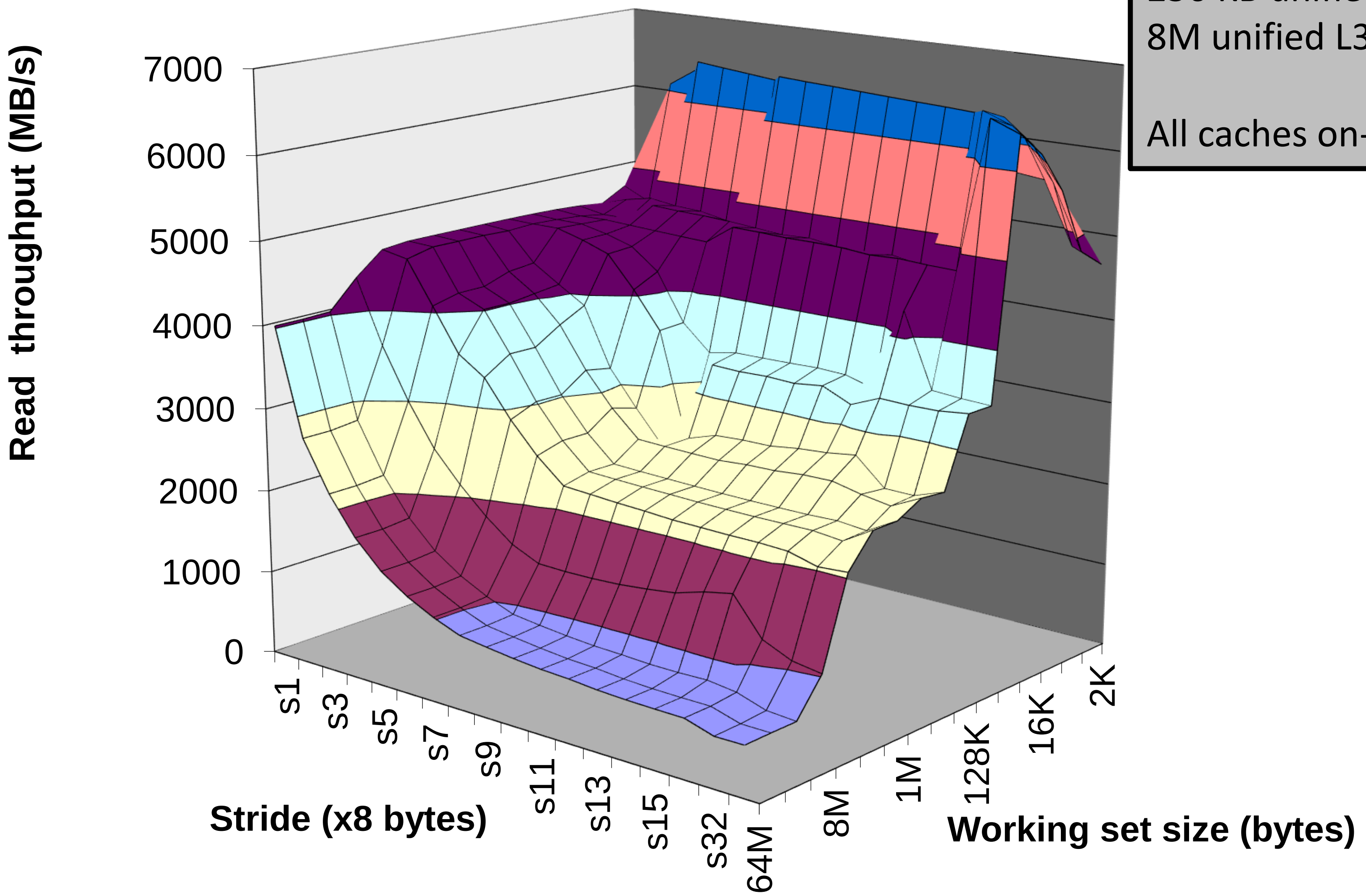
- *Memory mountain*: Measures read throughput as a function of spatial and temporal locality.
  - We run variants of the same program with different levels of spatial and temporal locality, then measure read throughput
  - Compact way to characterize memory system performance
  - Different systems (with different caches) have different mountains!

- Observation: if you decrease locality, bandwidth drops
  - As we'd expect; locality is key to having the right data in the cache
  - And if data is not in the cache, need to get it from next level down

A Memory Mountain

Intel Core i7
32 KB L1 i-cache
32 KB L1 d-cache
256 KB unified L2 cache
8M unified L3 cache

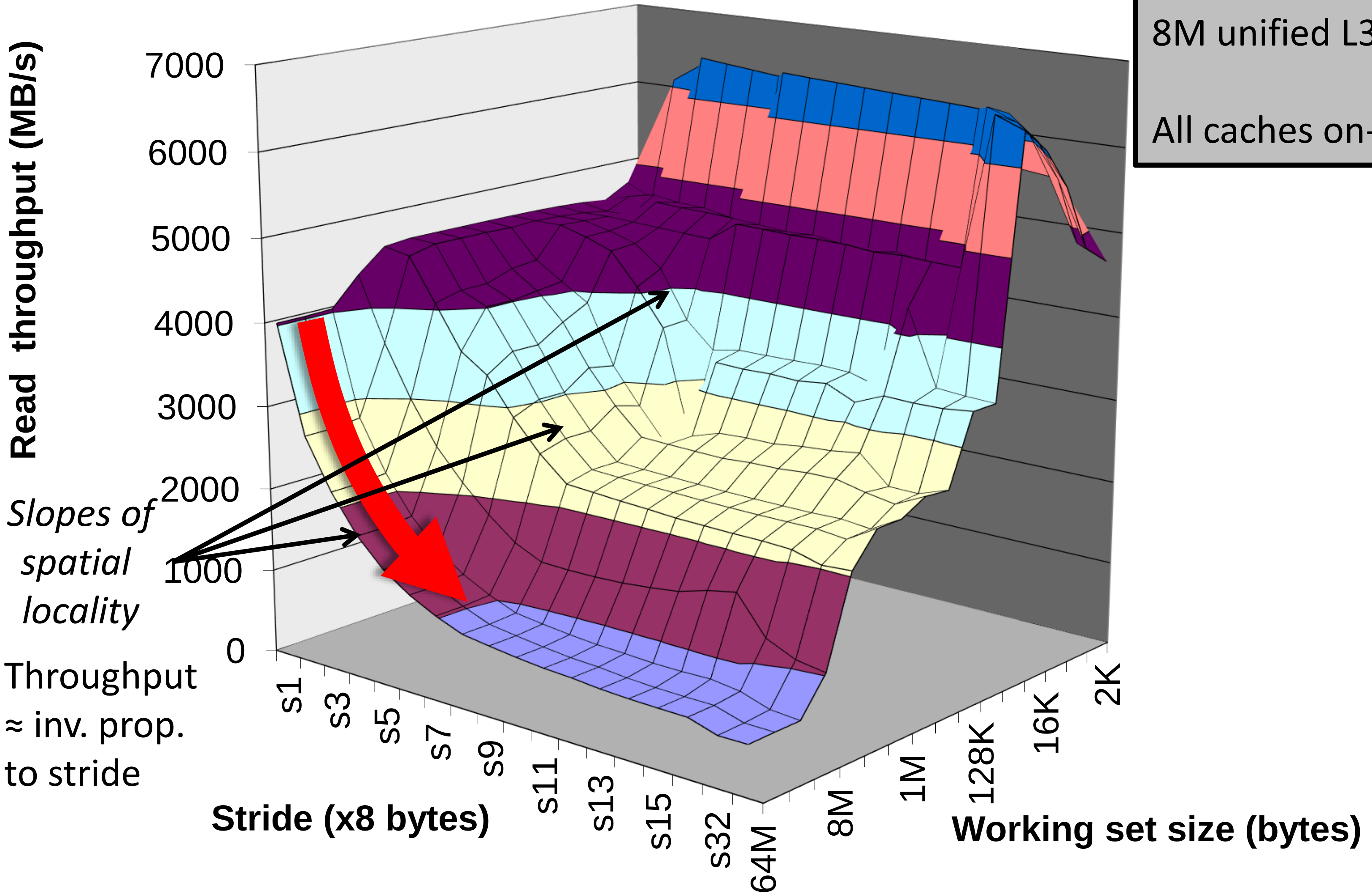
All caches on-chip



A Memory Mountain

Intel Core i7
32 KB L1 i-cache
32 KB L1 d-cache
256 KB unified L2 cache
8M unified L3 cache

All caches on-chip

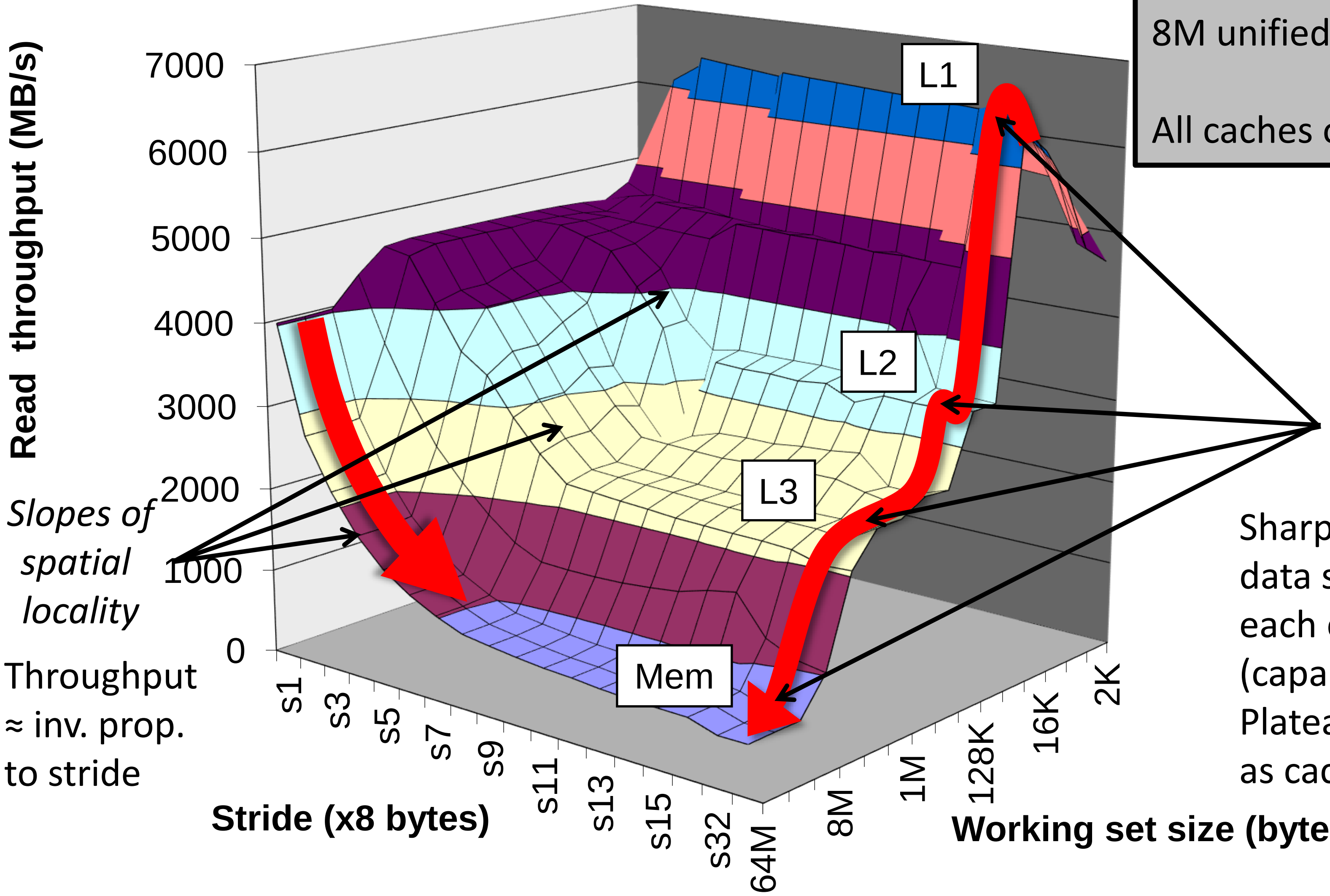


Throughput
≈ inv. prop.
to stride

A Memory Mountain

Intel Core i7
32 KB L1 i-cache
32 KB L1 d-cache
256 KB unified L2 cache
8M unified L3 cache

All caches on-chip



Ridges of Temporal locality

Sharp drops when data stops fitting in each cache level (capacity misses)

Plateaus get wider as caches get larger

Outline

- Memory Mountain
- **Cache Metrics**
- Cache Performance for Arrays
- Improving code
  - Rearranging Matrix Math
  - Matrix Math in Blocks

Cache Performance Metrics

- Miss Rate
  - Fraction of memory references not found in cache (misses / accesses) = 1 – hit rate
  - Typical numbers (in percentages):
    - 3-10% for L1
    - Can be quite small (e.g., < 1%) for L2, depending on dataset size, etc.
    - However, many applications have >30% miss rate in L2 cache

- Hit Time
  - Time to deliver a block in the cache to the processor
    - Includes time to determine whether the block is in the cache
    - Assumption: always check first cache *before* going to the next level
  - Typical numbers:
    - 1-2 clock cycles for L1
    - 5-20 clock cycles for L2

- Miss Penalty
  - Additional time required because of a miss
  - Typically 50-200 cycles for main memory
    - Not really a "penalty", just how long it takes to read from memory

Let's think about those numbers

- Huge difference between a hit and a miss
  - Could be 100x, if comparing L1 and main memory
- Would you believe a 99% hit rate is twice as good as 97%?
  - Consider:
    cache hit time of 1 cycle
    miss penalty of 100 cycles
  - Average access time:

97% hits: 100 instructions: 100*1 (L1 accesses) + 3*100 (misses)

on average: 1 cycle/instr. + 0.03 * 100 cycles/instr. = **4 cycles/instruction**

99% hits: on average: 1 cycle/instr. + 0.01 * 100 cycles/instr. = **2 cycles/instruction**

- This is why "miss rate" is used instead of "hit rate"
  - In our example, 1% miss rate vs. 3% miss rate
  - Makes the radical performance difference more obvious
  - "Computation is what happens between cache misses."

Average Memory Access Time (AMAT)

- AMAT = Hit time + Miss rate × Miss penalty
  - Generalization of previous formula

- Can extend for multiple layers of caching
  - AMAT = Hit Time L1 + Miss Rate L1 × Miss Penalty L1
    - Miss Penalty L1 = Hit Time L2 + Miss Rate L2 × Miss Penalty L2
    - Miss Penalty L2 = Hit Time Main Memory

- Generally: multi-level caching helps minimize AMAT

Example Memory Access Time Problem

- Computer specs: One layer of cache plus main memory
  - Cache Hit Time: 5 nanoseconds
  - Cache Miss Rate: 2%
    - Memory Access Time: 100 nanoseconds

- Calculate Average Memory Access Time (Hit Time + Miss Rate * Miss Penalty)
  - 5 ns + 0.02 * 100 ns
  - = 5 ns + 2 ns
  - = 7 ns

Break + Practice

- Computer specs: Two layers of cache plus main memory
  - L1 Cache Hit Time: 4 nanoseconds
  - L1 Cache Miss Rate: 10%
    - L2 Cache Hit Time: 8 nanoseconds
    - L2 Cache Miss Rate: 2%
      - Memory Access Time: 100 nanoseconds

- Calculate Average Memory Access Time (Hit Time + Miss Rate * Miss Penalty)

Break + Practice

- Computer specs: Two layers of cache plus main memory
  - L1 Cache Hit Time: 4 nanoseconds
  - L1 Cache Miss Rate: 10%
    - L2 Cache Hit Time: 8 nanoseconds
    - L2 Cache Miss Rate: 2%
      - Memory Access Time: 100 nanoseconds

- Calculate Average Memory Access Time (Hit Time + Miss Rate * Miss Penalty)
  - 4 ns + 0.10 * (8 ns + 0.02 * 100 ns)
  - = 4 ns + 0.10 * (8 ns + 2 ns)
  - = 4 ns + 0.10 * 10 ns
  - = 5 ns

Outline

- Memory Mountain
- Cache Metrics
- **Cache Performance for Arrays**
- Improving code
  - Rearranging Matrix Math
  - Matrix Math in Blocks

Contiguous Memory vs Indirection

- The rest of this lecture will focus on loops over arrays
  - I.e., operating on contiguous blocks of memory

- Not all programs are like that
  - "Pointer-chasing" is common
    - E.g., traversing a linked list, following a pointer for every node
  - (Usually) terrible for locality
    - See earlier comment about some programs having >30% L2 misses
    - A good allocator (**`malloc`**) can help some, but no miracles

- Specialized data structures can improve locality while still having a linked structure, e.g., for trees
  - E.g., ropes, B-trees, HAMTs, etc.

Understanding cache layout

- Cache parameters
  - Direct-mapped data cache
  - 256-byte total size
  - 16-byte blocks

  - Blocks per set: 1 (because direct mapped)
  - Sets: 256/16 = 16

- Assume data starts at address 0 and the cache starts empty

Set	Valid	Tag	Block
0000	0	??	
0001	0	??	
0010	0	??	
0011	0	??	
0100	0	??	
0101	0	??	
0110	0	??	
0111	0	??	
1000	0	??	
1001	0	??	
1010	0	??	
1011	0	??	
1100	0	??	
1101	0	??	
1110	0	??	
1111	0	??	

Understanding cache layout

- Cache parameters
  - Direct-mapped data cache
  - 256-byte total size
  - 16-byte blocks

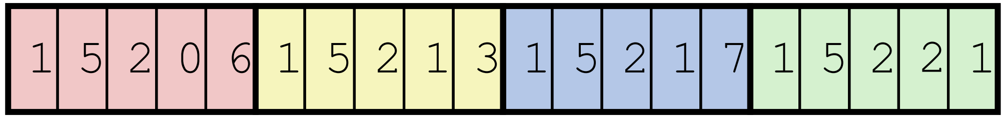
  - Blocks per set: 1 (because direct mapped)
  - Sets: 256/16 = 16

- Assume data starts at address 0 and the cache starts empty
  - Valid & Tag bits don't really matter here, so let's remove them from the diagram

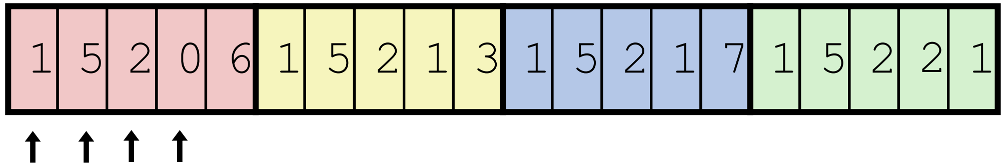
Set	**Block** **(16 byte)**
0000	
0001	
0010	
0011	
0100	
0101	
0110	
0111	
1000	
1001	
1010	
1011	
1100	
1101	
1110	
1111	

Layout of C Arrays in Memory (review)

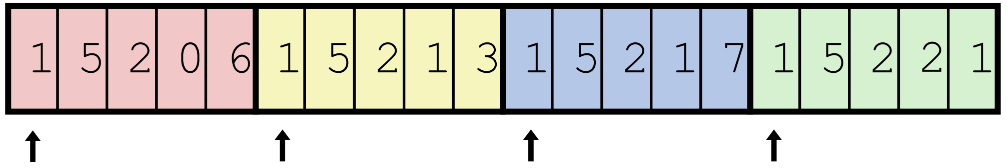
- C arrays allocated in row-major order
  - Each row in contiguous memory locations



- Stepping through columns in one row:
  - Accesses successive elements
  - Good spatial locality!
    - Miss rate ≈ 1 miss / Elements in a Block



- Stepping through rows in one column:
  - Accesses distant elements
  - Bad spatial locality!
    - Miss rate ≈ 1 (i.e. 100%) if the data is large enough



How do 1D arrays map to caches?

- How would an array of `int` map to this cache?
  - `int` -> 4 bytes. So, 4 `int` values per block
  - Example: `int array[100]`
- Where do the items go?
  - First four (0-3) go in set 0
  - Next four (4-7) go in set 1
  - Next four (8-11) go in set 2
  - etc.
- What if there are more elements in the array than there are blocks in the cache?
  - It wraps around and starts at set 0 again!
  - Indexes 60-63 go in set 15
  - Indexes 64-67 go in set 0 -> possible conflict!!

Set	**Block (16 byte)**
0000	[0-3]
0001	[4-7]
0010	[8-11]
0011	[12-15]
0100	[16-19]
0101	[20-23]
0110	[24-27]
0111	[28-31]
1000	[32-35]
1001	[36-39]
1010	[40-43]
1011	[44-47]
1100	[48-51]
1101	[52-55]
1110	[56-59]
1111	[60-63]

How do 2D arrays map to caches?

- How would a 2D array of `int` map to this cache?
  - `int` -> 4 bytes
  - So, 4 `int` values per block

- Breakdown of indexes depends on the shape of the array
  - If there are 4 values per row, entire row fits in a block
  - Example: `int array[16][4]`

Set	**Block (16 byte)**
0000	[0][0-3]
0001	[1][0-3]
0010	[2][0-3]
0011	[3][0-3]
0100	[4][0-3]
0101	[5][0-3]
0110	[6][0-3]
0111	[7][0-3]
1000	[8][0-3]
1001	[9][0-3]
1010	[10][0-3]
1011	[11][0-3]
1100	[12][0-3]
1101	[13][0-3]
1110	[14][0-3]
1111	[15][0-3]

How do 2D arrays map to caches?

- How would a 2D array of `int` map to this cache?
  - `int` -> 4 bytes
  - So, 4 `int` values per block

- Breakdown of indexes depends on the shape of the array
  - If there are 4 values per row, entire row fits in a block

  - If there are 16 values per row, ¼ of row fits in a block
  - Example: `int array[4][16]`

Set	**Block (16 byte)**
0000	[0][0-3]
0001	[0][4-7]
0010	[0][8-11]
0011	[0][12-15]
0100	[1][0-3]
0101	[1][4-7]
0110	[1][8-11]
0111	[1][12-15]
1000	[2][0-3]
1001	[2][4-7]
1010	[2][8-11]
1011	[2][12-15]
1100	[3][0-3]
1101	[3][4-7]
1110	[3][8-11]
1111	[3][12-15]

Example cache performance problem

- Cache parameters
  - Direct-mapped data cache
  - 256-byte total size
  - 16-byte blocks

  - Blocks per set: 1
  - Sets: 256/16 = 16

- Assume data starts at address 0 and cache starts empty

```
int mat[6][16];
```

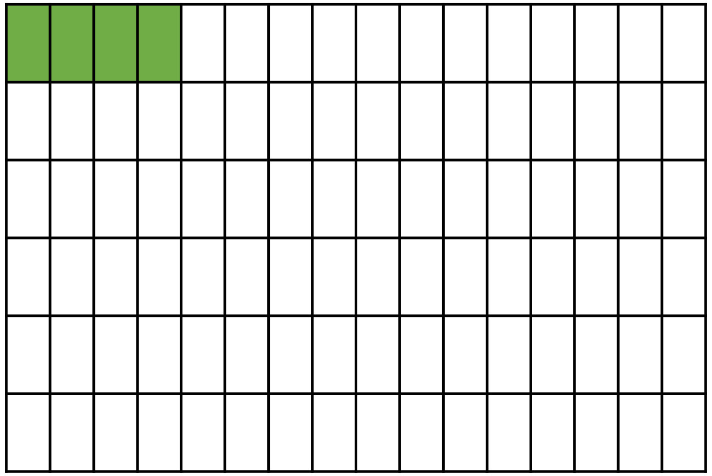
- First, think about how array maps to the cache
  - Element size: 4 bytes
  - Array size: 384 bytes (too big)

  - 4 elements per cache block
  - Array row takes up 4 cache blocks

  - First 4 rows * 16 cols fit in cache without overlap
    - Next 2 rows overlap with first 2 rows

Thinking visually about a 2D array

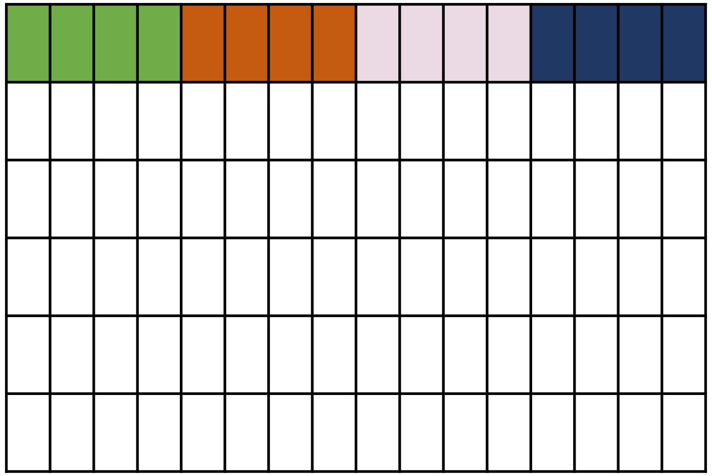
- `int mat[6][16];`

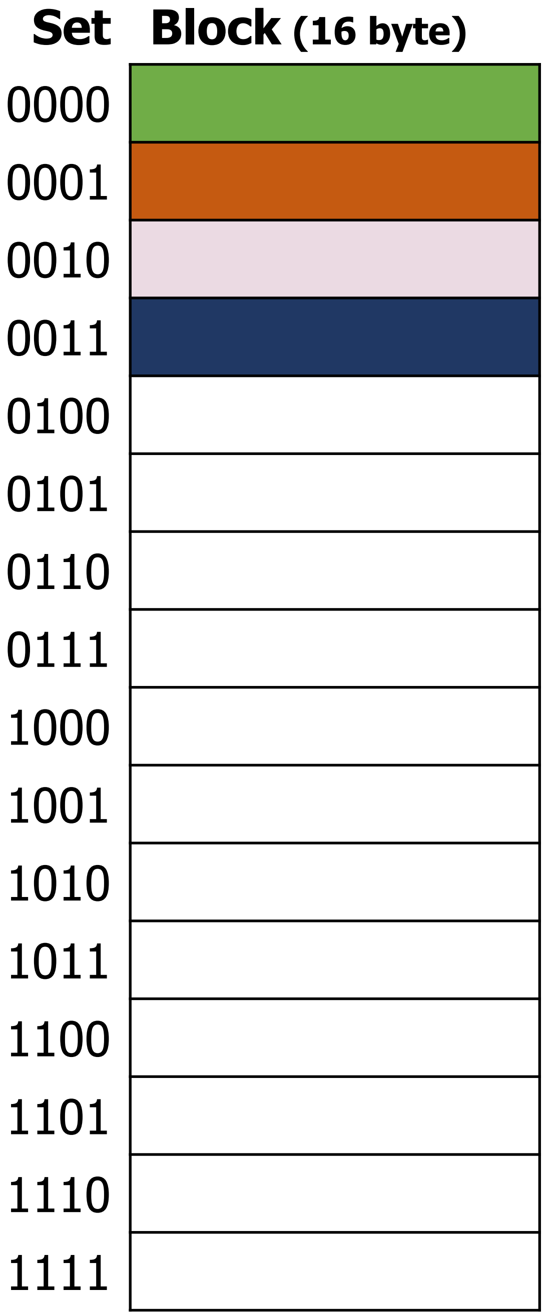


Set	Block (16 byte)
0000	
0001	
0010	
0011	
0100	
0101	
0110	
0111	
1000	
1001	
1010	
1011	
1100	
1101	
1110	
1111	

Thinking visually about a 2D array

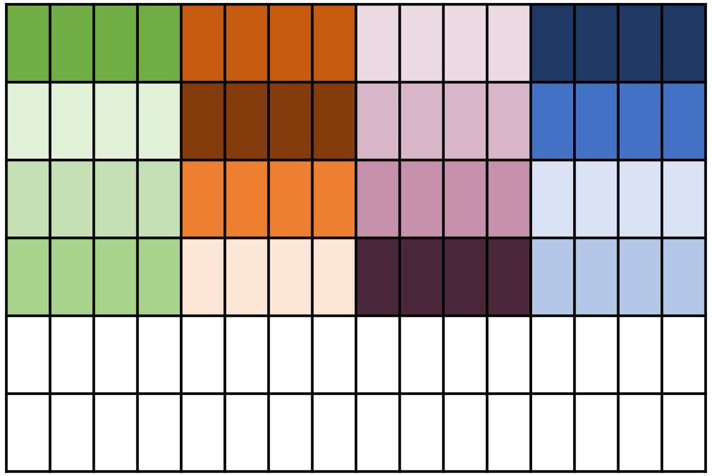
- **`int mat[6][16];`**

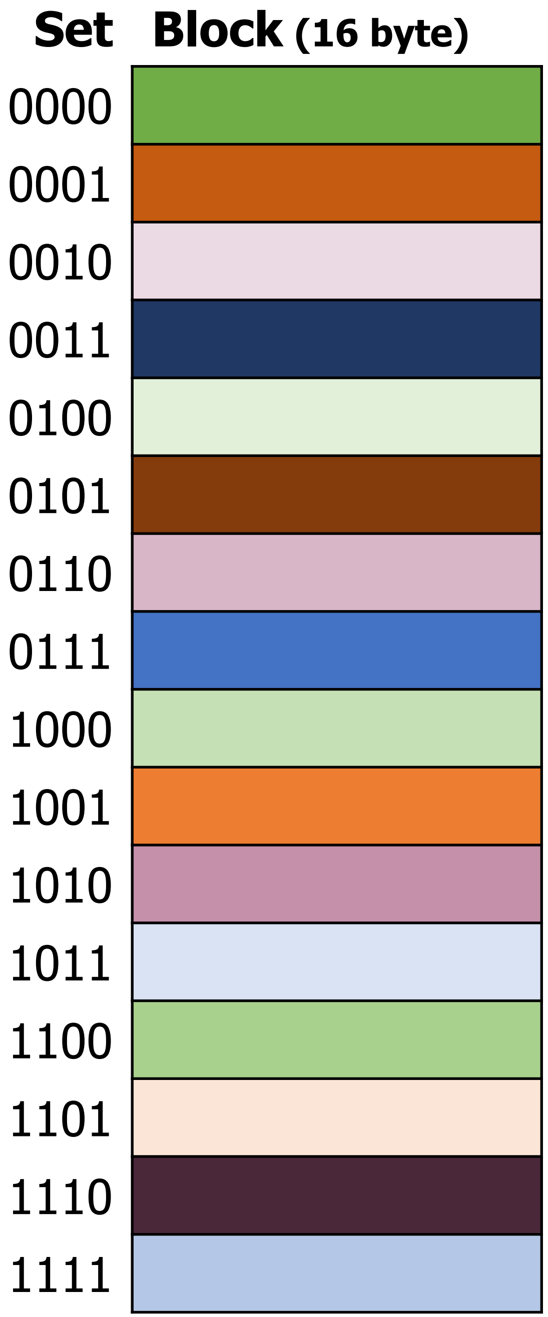




Thinking visually about a 2D array

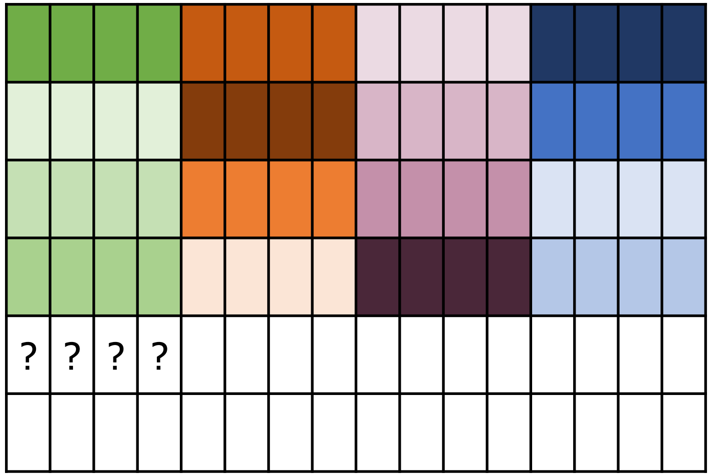
- `int mat[6][16];`





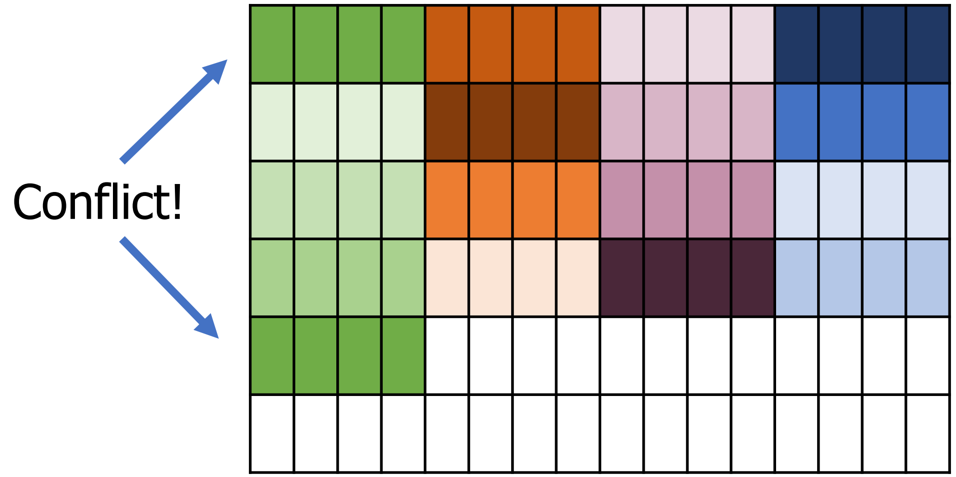
Thinking visually about a 2D array

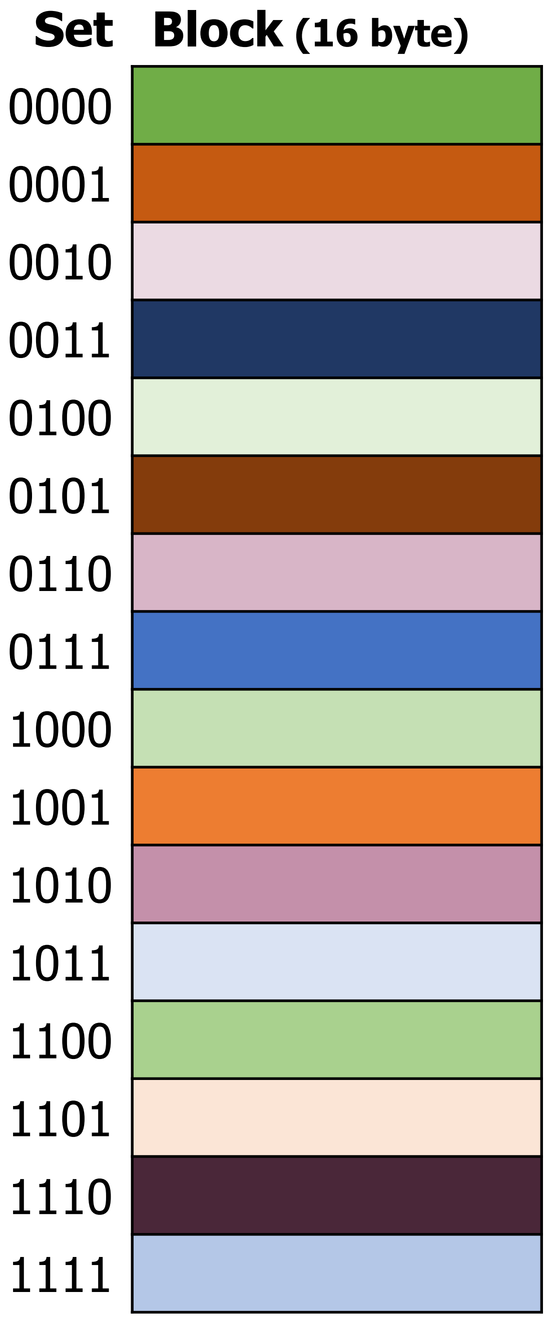
- `int mat[6][16];`



Thinking visually about a 2D array

- `int mat[6][16];`





Example: accessing elements in a row

```
int mat[6][16];
```

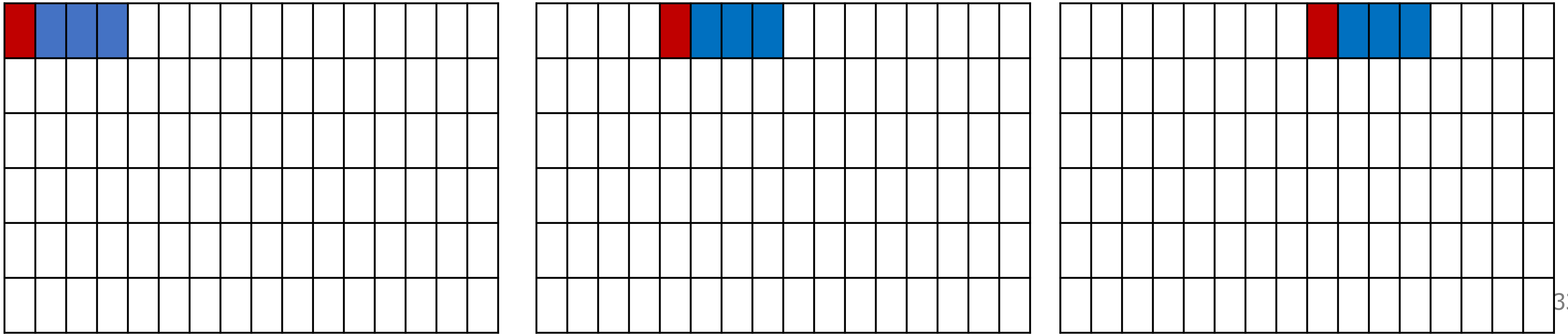
- First, think about how array maps to the cache
  - Element size: 4 bytes
  - Array size: 384 bytes (too big)

  - 4 elements per cache block
  - Array row takes up 4 cache blocks

  - First 4 rows * 16 cols fit in cache without overlap
    - Next 2 rows overlap with first 2 rows

```
for (int i = 0; i < 6; i = i+1) {
  for (int j = 0; j < 16; j = j+4) {
    mat[i][j]   = 0;
    mat[i][j+1] = 1;
    mat[i][j+2] = 2;
    mat[i][j+3] = 3;
  }
}
```

- Calculate miss rate



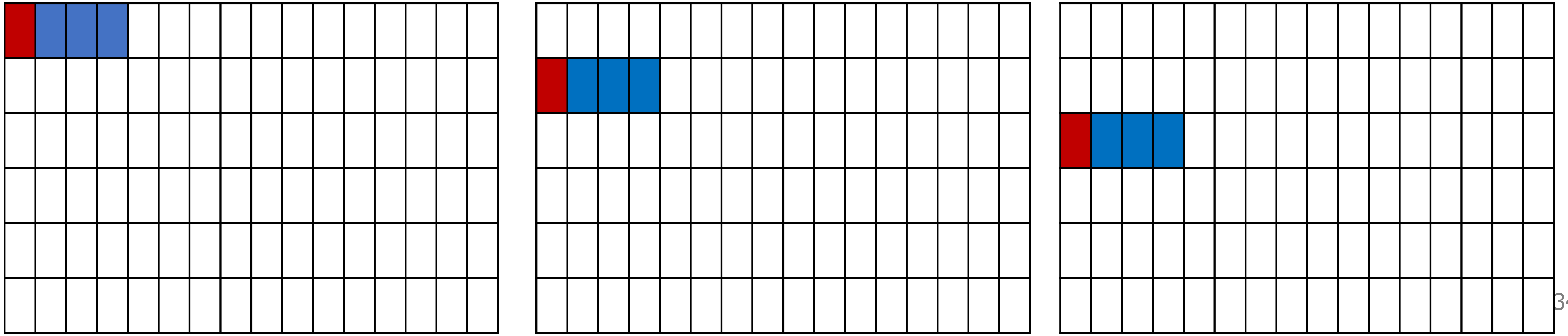
Example: accessing elements in a row

```
int mat[6][16];
```

- First, think about how array maps to the cache
  - Element size: 4 bytes
  - Array size: 384 bytes (too big)
  - 4 elements per cache block
  - Array row takes up 4 cache blocks
  - First 4 rows * 16 cols fit in cache without overlap
    - Next 2 rows overlap with first 2 rows

```
for (int i = 0; i < 6; i = i+1) {
  for (int j = 0; j < 16; j = j+4) {
    mat[i][j]   = 0;
    mat[i][j+1] = 1;
    mat[i][j+2] = 2;
    mat[i][j+3] = 3;
  }
}
```

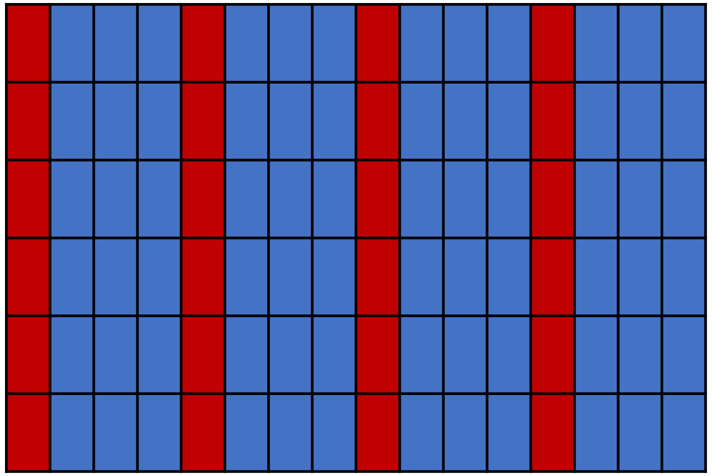
- Calculate miss rate



Example: accessing elements in a row

```
int mat[6][16];
```

- First, think about how array maps to the cache
  - Element size: 4 bytes
  - Array size: 384 bytes (too big)
  - 4 elements per cache block
  - Array row takes up 4 cache blocks
  - First 4 rows * 16 cols fit in cache without overlap
    - Next 2 rows overlap with first 2 rows



```
for (int i = 0; i < 6; i = i+1) {
  for (int j = 0; j < 16; j = j+4) {
    mat[i][j]   = 0;
    mat[i][j+1] = 1;
    mat[i][j+2] = 2;
    mat[i][j+3] = 3;
  }
}
```

- Calculate miss rate
  - All four accesses within loop fit in a cache block!
    - 1 miss, 3 hits
  - The next set of columns repeat pattern
  - The next row repeats pattern
    - Nothing already in cache from before
    - Never reference old cells again
  - **Miss rate: 25%**

Example: reordering element access

```
int mat[6][16];
```

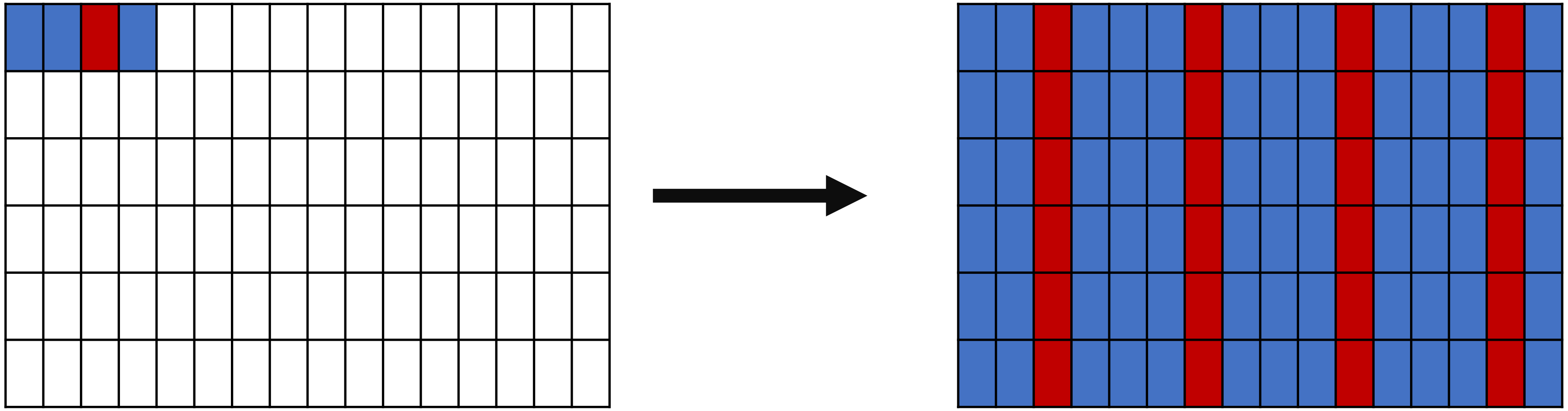
- First, think about how array maps to the cache
  - Element size: 4 bytes
  - Array size: 384 bytes (too big)

  - 4 elements per cache block
  - Array row takes up 4 cache blocks

  - First 4 rows * 16 cols fit in cache without overlap
    - Next 2 rows overlap with first 2 rows

```
for (int i = 0; i < 6; i = i+1) {
  for (int j = 0; j < 16; j = j+4) {
    mat[i][j+2] = 2;
    mat[i][j]   = 0;
    mat[i][j+3] = 3;
    mat[i][j+1] = 1;
  }
}
```

- Does this change anything?
  - No! First access brings in entire block
  - Later accesses within block are hits



Example: accessing elements by column

```
int mat[6][16];
```

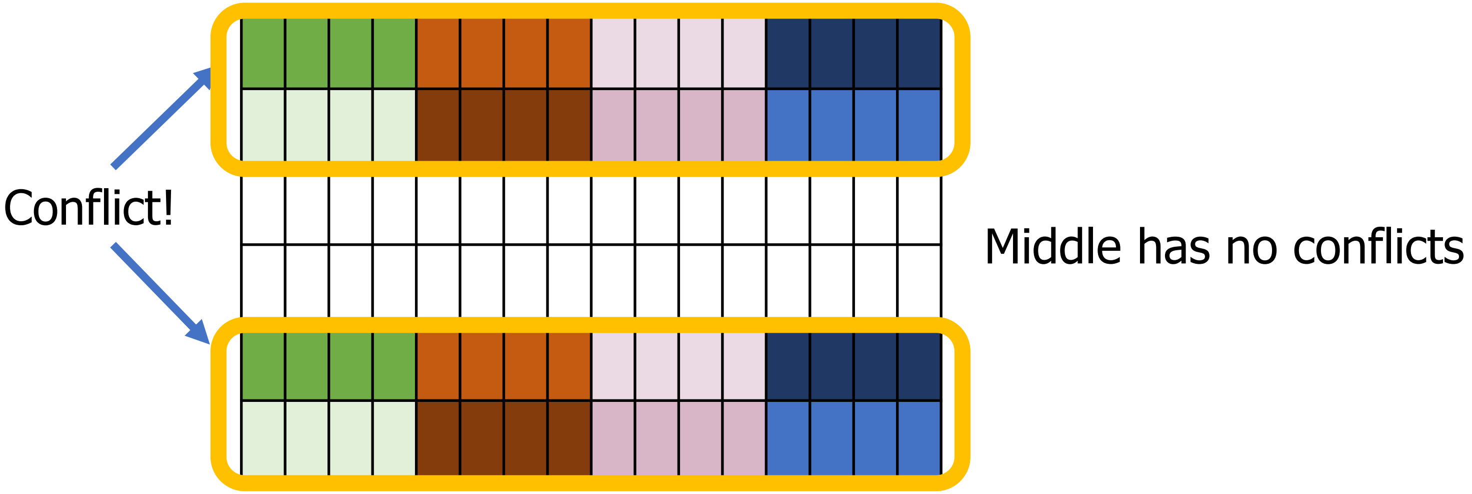
- First, think about how array maps to the cache
  - Element size: 4 bytes
  - Array size: 384 bytes (too big)
  - 4 elements per cache block
  - Array row takes up 4 cache blocks
  - First 4 row * 16 cols fit in cache without overlap
    - Next 2 rows overlap with first 2 rows

```
for (int j = 0; j < 16; j = j+1) {
  for (int i = 0; i < 6; i = i+1) {
    mat[i][j] = 7;
  }
}
```

- Calculate miss rate

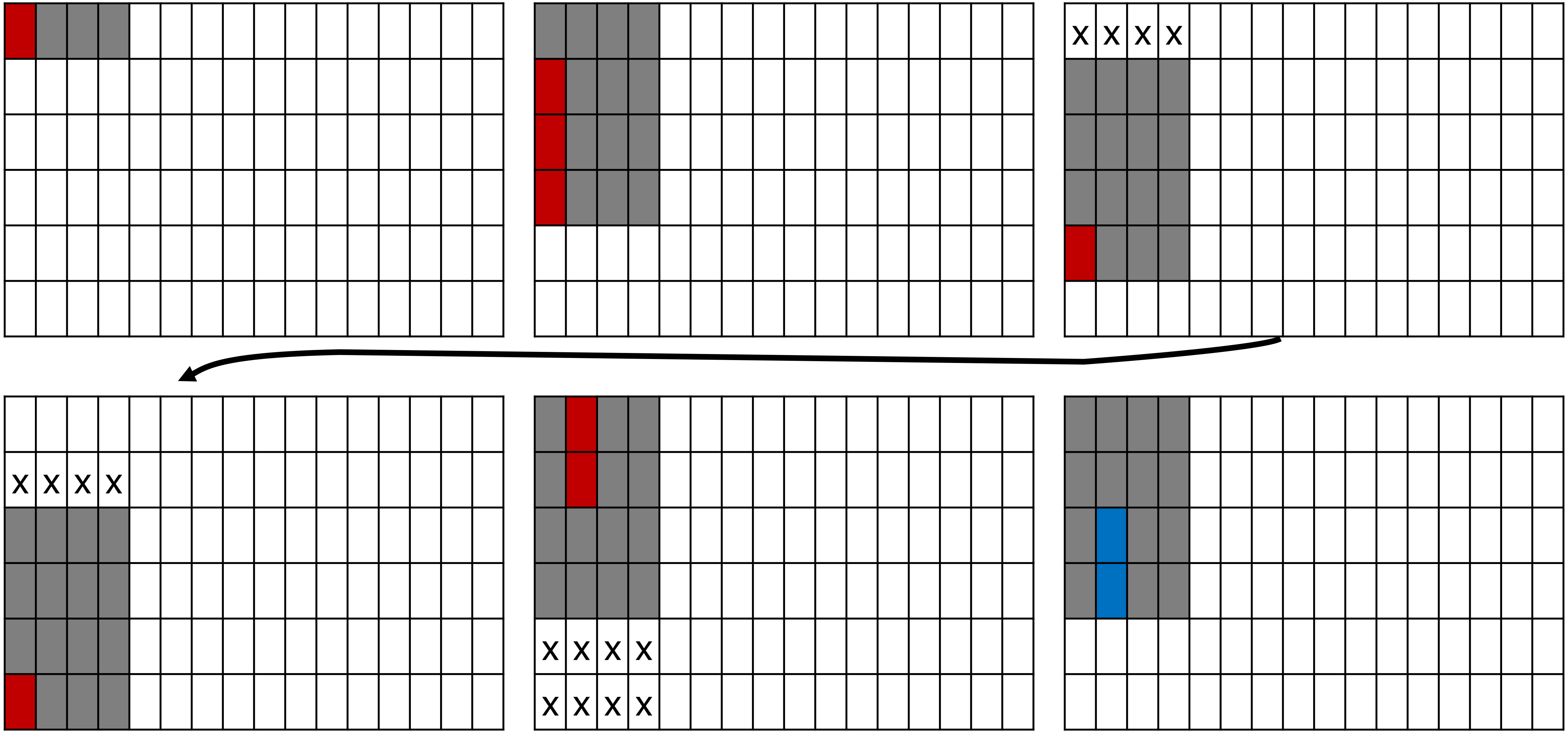
Remember, some rows are in conflict

- `int mat[6][16];`



Example: accessing elements by column (graphically)

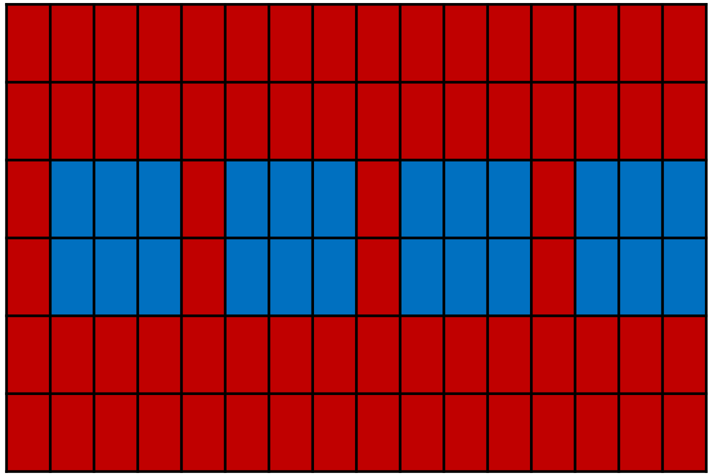
Grey blocks are loaded into the cache, but not accessed at this time



Example: accessing elements by column

```
int mat[6][16];
```

- First, think about how array maps to the cache
  - Element size: 4 bytes
  - Array size: 384 bytes (too big)
  - 4 elements per cache block
  - Array row takes up 4 cache blocks
  - First 4 row * 16 cols fit in cache without overlap
    - Next 2 rows overlap with first 2 rows



```
for (int j = 0; j < 16; j = j+1) {
  for (int i = 0; i < 6; i = i+1) {
    mat[i][j] = 7;
  }
}
```

- Calculate miss rate
  - 6 misses for 1st load of each row
  - 4 misses for 2nd column in the row (2 hits)
  - 4 misses for 3rd column in the row (2 hits)
  - 4 misses for 4th column in the row (2 hits)
  - Repeat
  - Miss rate = (6+4+4+4)/24 = 75%

Break + Question

```
int mat[4][16];
```

- Same cache from before:
  - Direct-mapped data cache
  - 256-byte total size
  - 16-byte blocks

- Change matrix to be 4 rows of 16 columns (not 6 rows)

```
for (int j = 0; j < 16; j = j+1) {
  for (int i = 0; i < 4; i = i+1) { // 4!
    mat[i][j] = 7;
  }
}
```

- Calculate miss rate

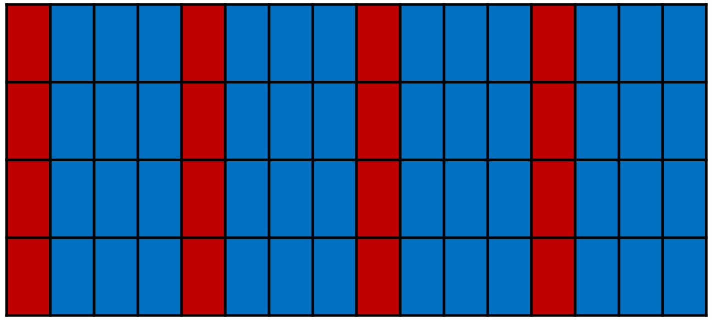
Break + Question

```
int mat[4][16];
```

- Same cache from before:
  - Direct-mapped data cache
  - 256-byte total size
  - 16-byte blocks
- Change matrix to be 4 rows of 16 columns (not 6 rows)

```
for (int j = 0; j < 16; j = j+1) {
  for (int i = 0; i < 4; i = i+1) { // 4!
    mat[i][j] = 7;
  }
}
```

- Calculate miss rate
  - Entire array fits in cache!
    - No conflicts
  - 1 miss per four accesses
  - **Miss rate = 25%**



Outline

- Memory Mountain
- Cache Metrics
- Cache Performance for Arrays
- **Improving code**
  - **Rearranging Matrix Math**
  - Matrix Math in Blocks

Our Benchmark: Matrix Multiplication

- Review from your linear algebra class

$$\begin{bmatrix} 1 & 3 \\ 2 & 4 \end{bmatrix} \times \begin{bmatrix} 5 & 6 \\ 7 & 8 \end{bmatrix} = \begin{bmatrix} 26 & 30 \\ 38 & 44 \end{bmatrix}$$

$1 \times 5 + 3 \times 7 = 26$

$1 \times 6 + 3 \times 8 = 30$

$2 \times 5 + 4 \times 7 = 38$

$2 \times 6 + 4 \times 8 = 44$

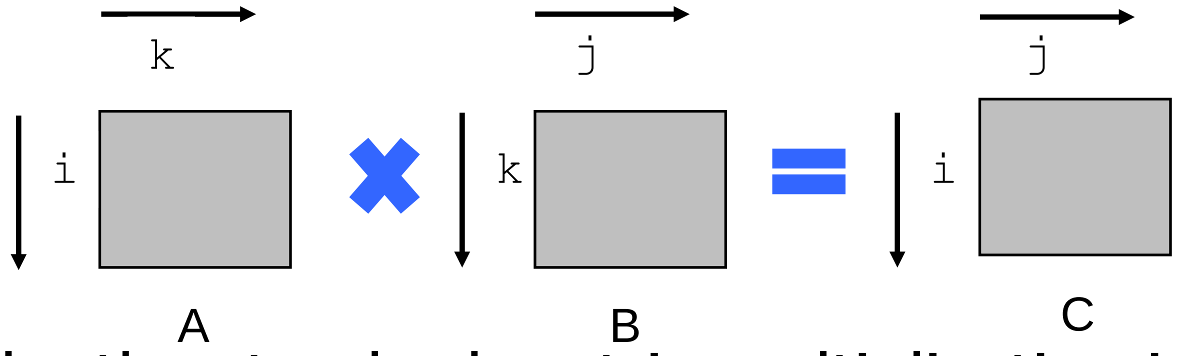
$$\begin{array}{cc} & \begin{bmatrix} 5 & 6 \\ 7 & 8 \end{bmatrix} \\ \begin{bmatrix} 1 & 3 \\ 2 & 4 \end{bmatrix} & \begin{bmatrix} 26 & 30 \\ 38 & 44 \end{bmatrix} \end{array}$$

When is matrix multiplication important?

- ML and AI algorithms!!

Miss Rate Analysis for Matrix Multiply

- Assume:
  - Block size = 32B (big enough for four 64-bit longs)
  - Matrix dimension (N) is very large
  - Cache is not big enough to hold even one row
- Analysis Method:
  - Look at access pattern of inner loop



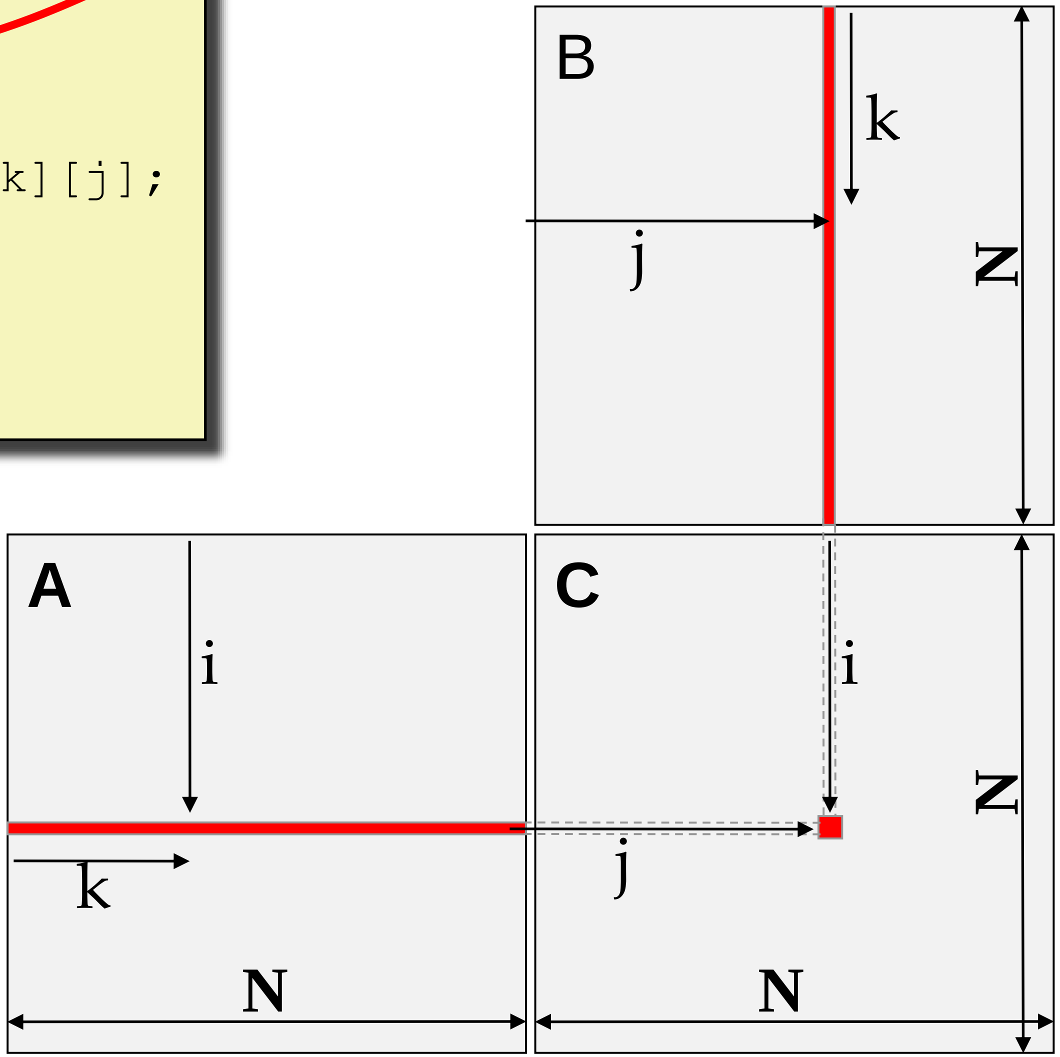
- Now we'll see why the standard matrix multiplication is bad!
  - From a performance standpoint, that is

Matrix Multiplication Example

```
/* ijk */
for (i=0; i<n; i++)  {
  for (j=0; j<n; j++) {
    sum = 0.0;
    for (k=0; k<n; k++)
      sum += a[i][k] * b[k][j];
    c[i][j] = sum;
  }
}
```

Variable `sum` held in register

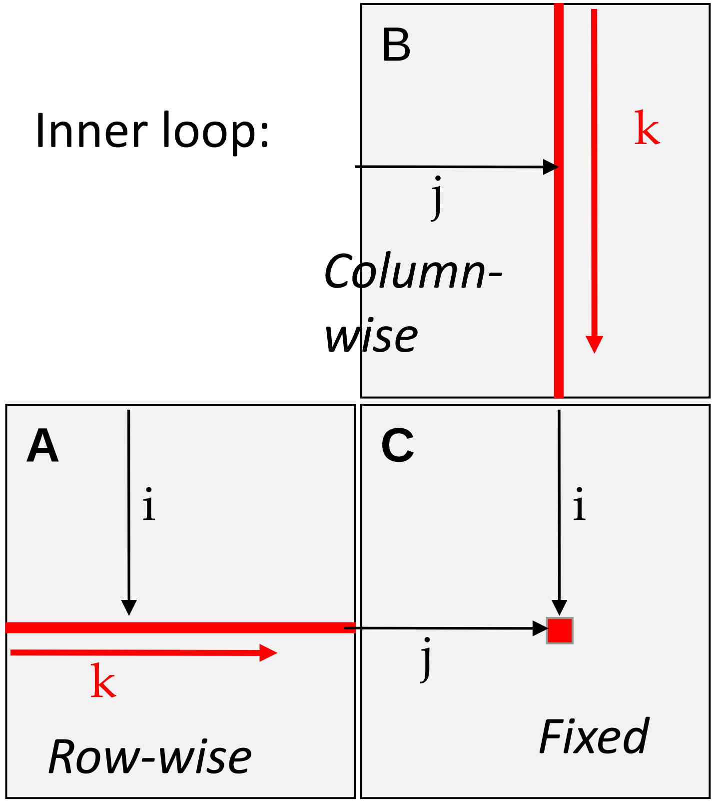
- Multiply N x N matrices
- $O(N^3)$ total operations
- Each source element read N times
- N values summed per destination



Matrix Multiplication (ijk)

```
/* ijk */
for (i=0; i<n; i++)  {
  for (j=0; j<n; j++) {
    sum = 0.0;
    for (k=0; k<n; k++)
      sum += a[i][k] * b[k][j];
    c[i][j] = sum;
  }
}
```

Inner loop:



Misses per inner loop iteration:

A	B	C
0.25	1	0

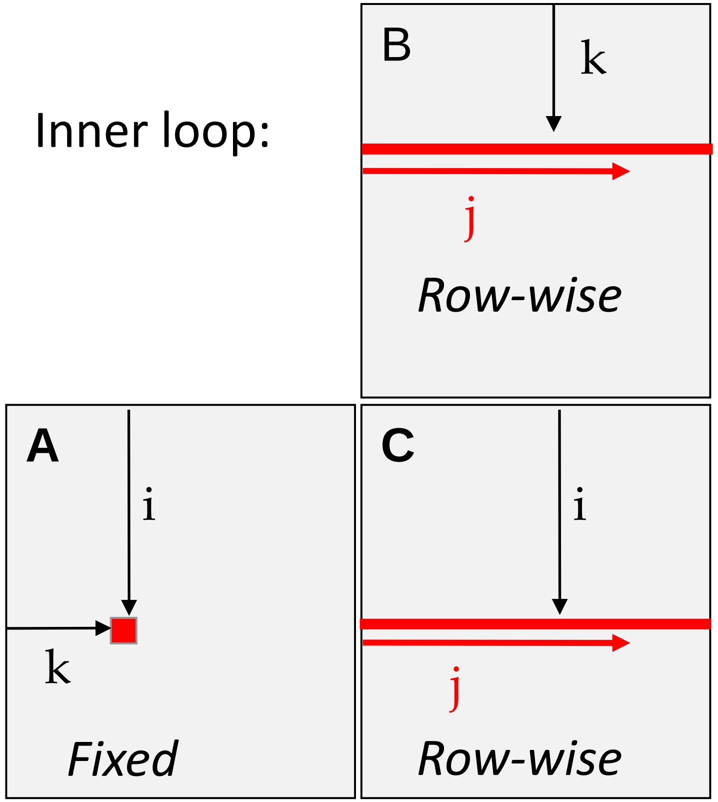
Remember: Block size = 32B
(big enough for four 64-bit longs)

Total misses/iteration: 1.25

Matrix Multiplication (kij)

```
/* kij */
for (k=0; k<n; k++) {
  for (i=0; i<n; i++) {
    r = a[i][k];
    for (j=0; j<n; j++)
      c[i][j] += r * b[k][j];
  }
}
```

Inner loop:



Misses per inner loop iteration:

A	B	C
0	0.25	0.25

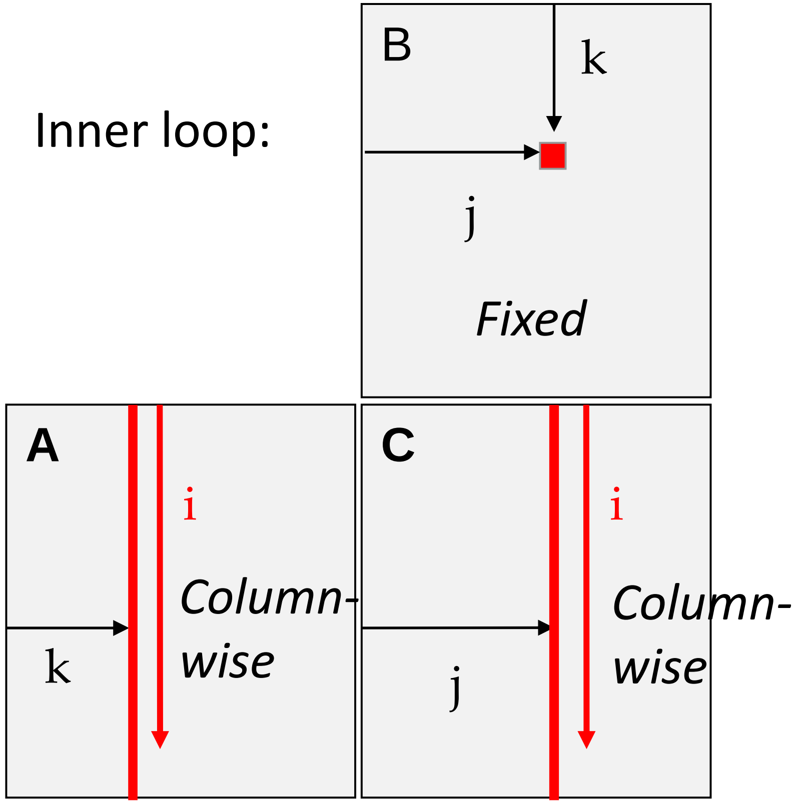
Remember: Block size = 32B
(big enough for four 64-bit longs)

Total misses/iteration: 0.5

Matrix Multiplication (jki)

```
/* jki */
for (j=0; j<n; j++) {
  for (k=0; k<n; k++) {
    r = b[k][j];
    for (i=0; i<n; i++)
      c[i][j] += a[i][k] * r;
  }
}
```

Inner loop:



Misses per inner loop iteration:

A	B	C
1	0	1

Total misses/iteration: 2

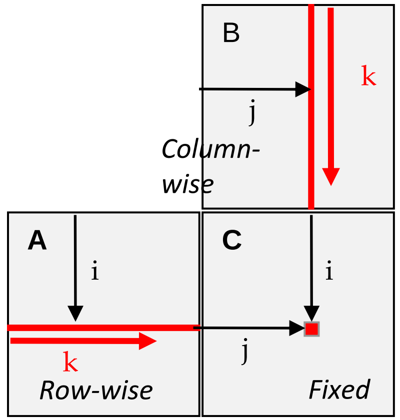
Remember: Block size = 32B
(big enough for four 64-bit longs)

Summary of Matrix Multiplication

```
for (i=0; i<n; i++) {
  for (j=0; j<n; j++) {
   sum = 0.0;
   for (k=0; k<n; k++)
     sum += a[i][k] * b[k][j];
   c[i][j] = sum;
 }
}
```

ijk (& jik):

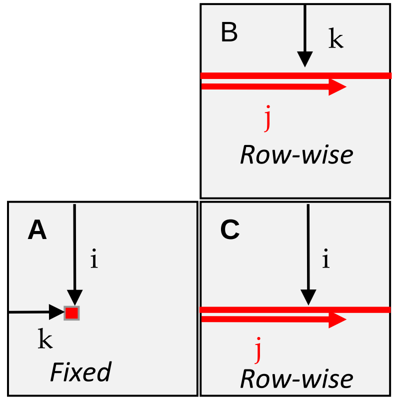
- 2 loads, 0 stores
- misses/iter = 1.25




```
for (k=0; k<n; k++) {
 for (i=0; i<n; i++) {
  r = a[i][k];
  for (j=0; j<n; j++)
   c[i][j] += r * b[k][j];
 }
}
```

kij (& ikj):

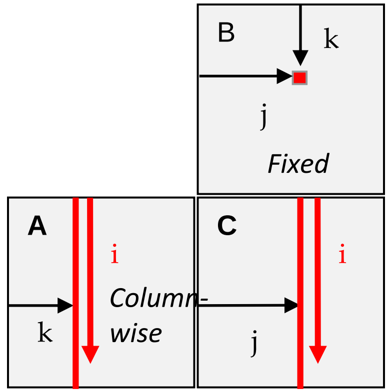
- 2 loads, 1 store
- misses/iter = 0.5




```
for (j=0; j<n; j++) {
 for (k=0; k<n; k++) {
   r = b[k][j];
   for (i=0; i<n; i++)
    c[i][j] += a[i][k] * r;
 }
}
```

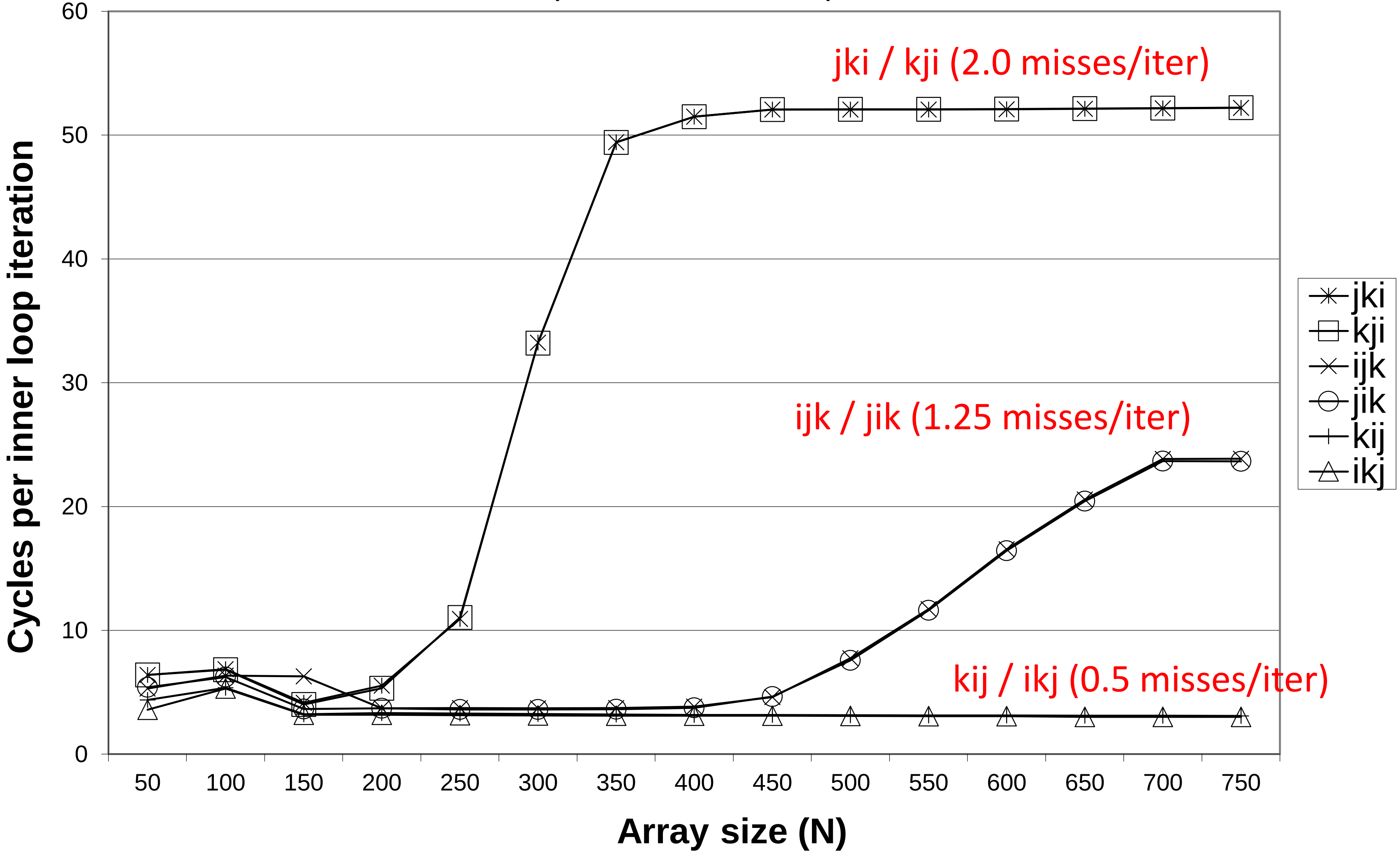
jki (& kji):

- 2 loads, 1 store
- misses/iter = 2



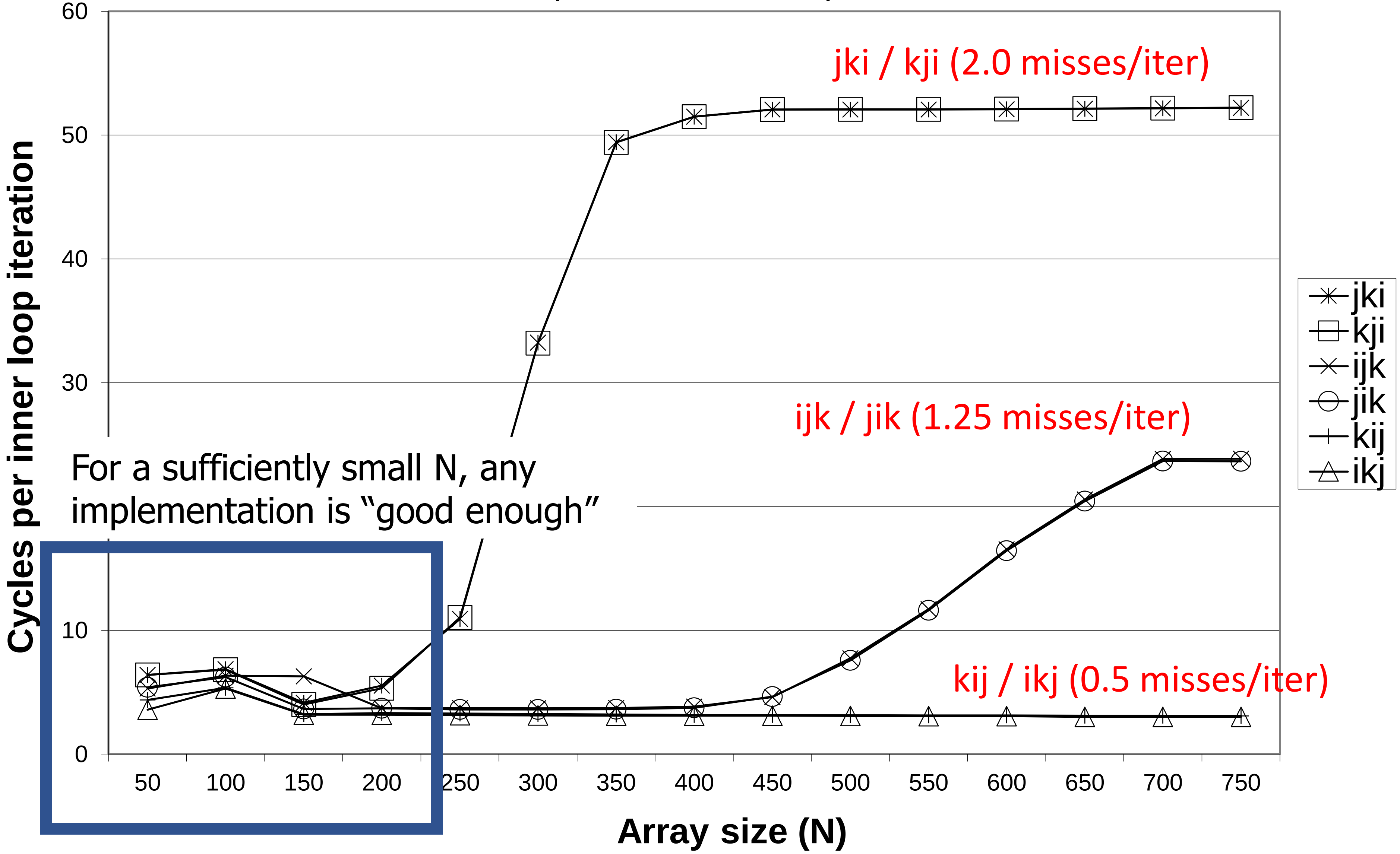
Core i7 Matrix Multiply Performance

Essentially the same algorithm, just different data access patterns!
The most natural way to write code may not be the best one!



Core i7 Matrix Multiply Performance

Essentially the same algorithm, just different data access patterns!
The most natural way to write code may not be the best one!



Break + Open Question

- What about those writes? Do they have additional costs?

Break + Open Question

- What about those writes? Do they have additional costs?
  - Assumption: write-back cache such that they don't cost more than reads until evicted

  - As long as evictions of modified (dirty) data happen once per array cell, we're equivalent to the one write outside of the `for` loop
    - This is not the case here since entire row doesn't fit in cache

  - If evictions of modified (dirty) data happen multiple times per array cell, question becomes complicated
    - How much does that hurt compared to extra cache misses?
    - Writes can happen in the background (while processor is running)
    - Likely need to measure real-world performance to understand

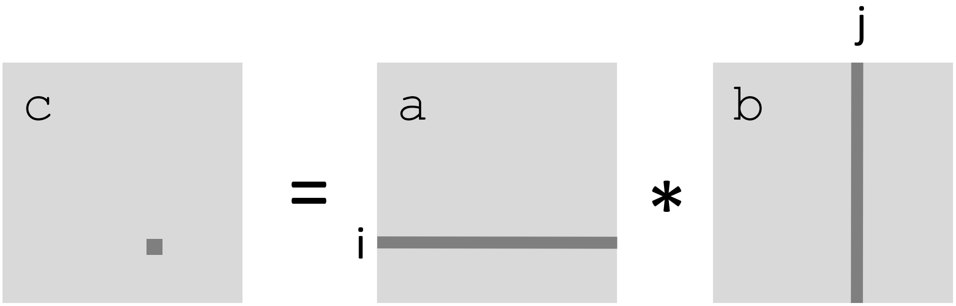
Outline

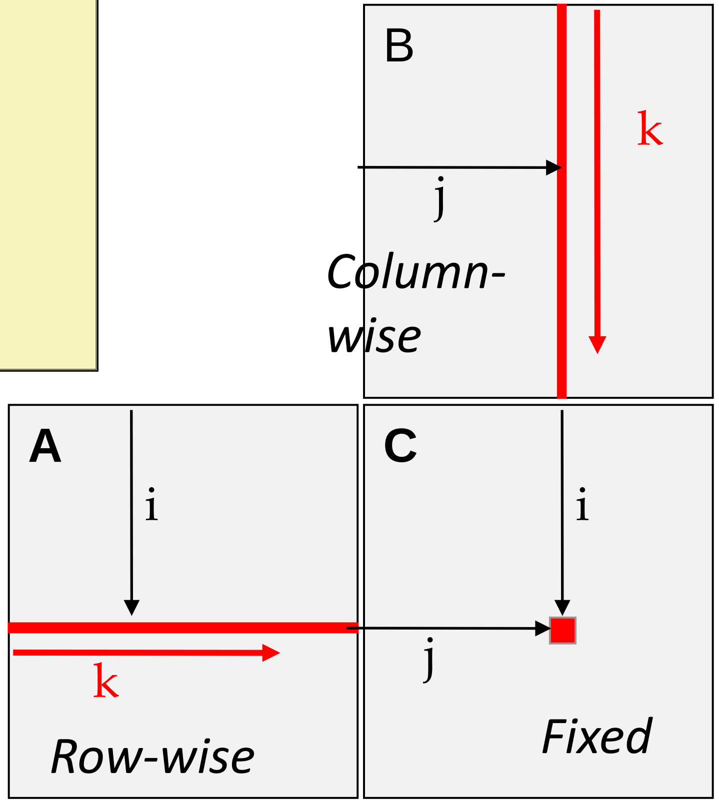
- Memory Mountain
- Cache Metrics
- Cache Performance for Arrays
- **Improving code**
  - Rearranging Matrix Math
  - **Matrix Math in Blocks**

Example: Matrix Multiplication

```
double *c = (double *) malloc(sizeof(double)*n*n);

/* Multiply n x n matrices a and b  */
void mmm(double *a, double *b, double *c, int n) {
  for (int i = 0; i < n; i++) {
    for (int j = 0; j < n; j++) {
      double sum = 0.0;
      for (int k = 0; k < n; k++) {
        sum += a[i*n + k] * b[k*n + j];
      }
      c[i*n+j] = sum;
} } }
```

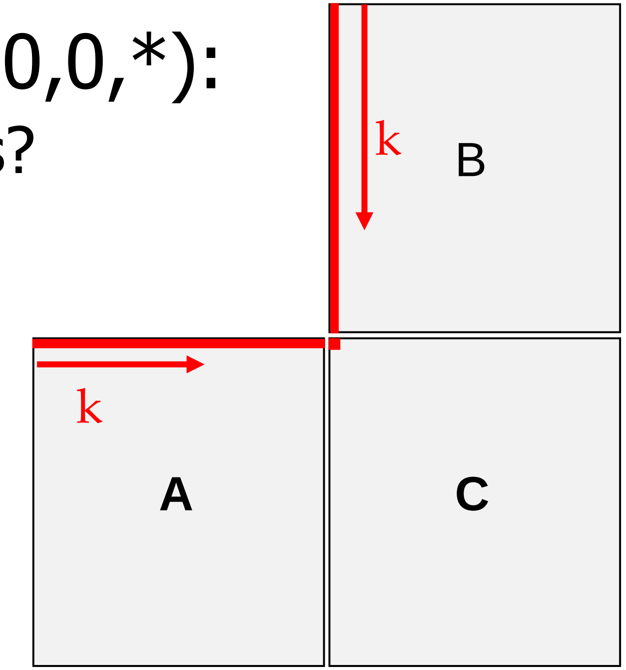




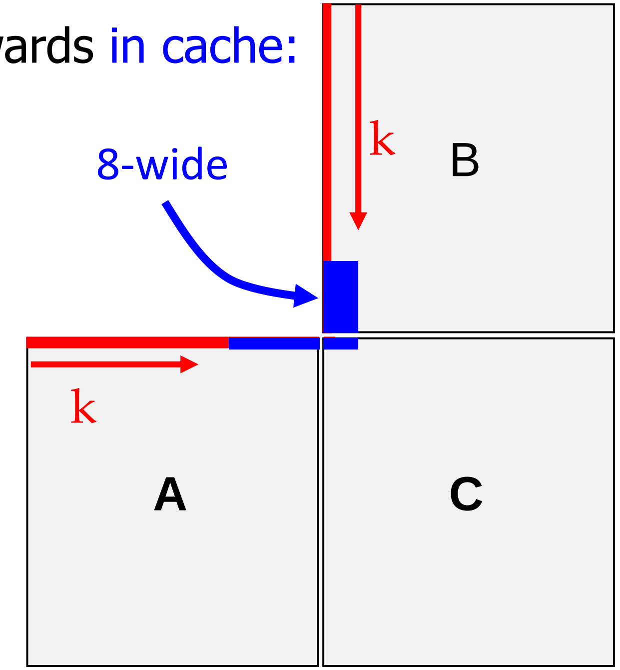
Cache Miss Analysis (approximate)

- Assume:
  - Matrix elements are doubles
  - Cache block = 8 doubles
  - Cache size C <<< n (much smaller than n)

- 1st iteration (i,j,k=0,0,*):
  - How many misses?
  - n/8 + n + 1 = 9n/8+1 misses



Afterwards in cache:

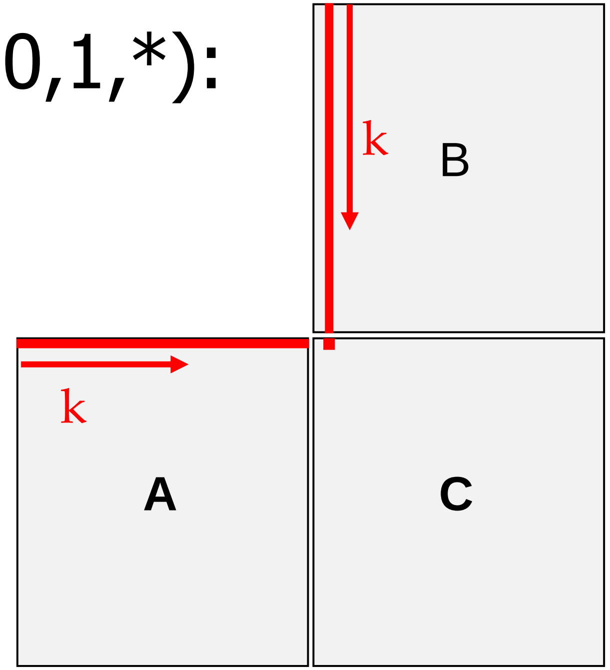


Cache Miss Analysis (approximate)

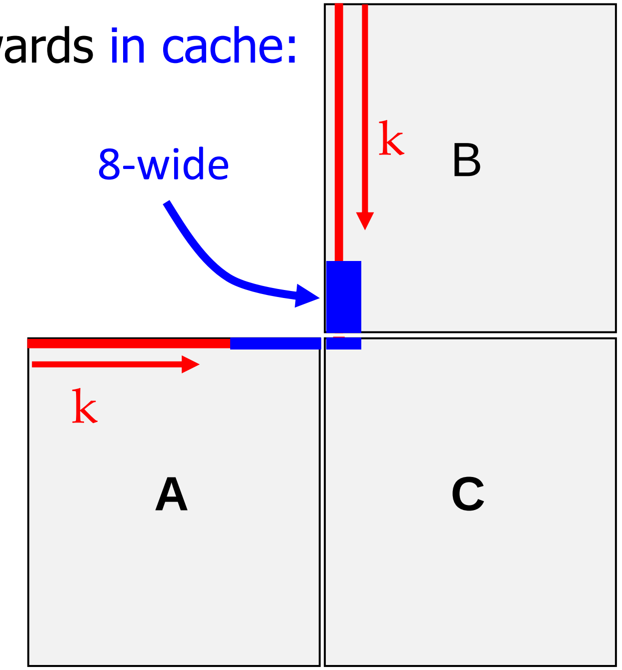
- Assume:
  - Matrix elements are doubles
  - Cache block = 8 doubles
  - Cache size C <<< n (much smaller than n)

- Total misses:
  - Every iteration: $9n/8 + 1$
  - # iterations: n^2
  - $(9n/8+1)*n^2 = (9/8)*n^3 + n^2$

- 2nd iteration (i,j,k=0,1,*):
  - Again:
    $n/8 + n + 1 =$
    $9n/8+1$ misses



Afterwards in cache:

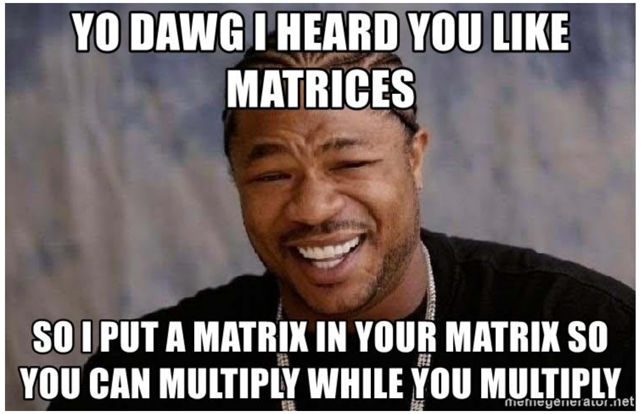


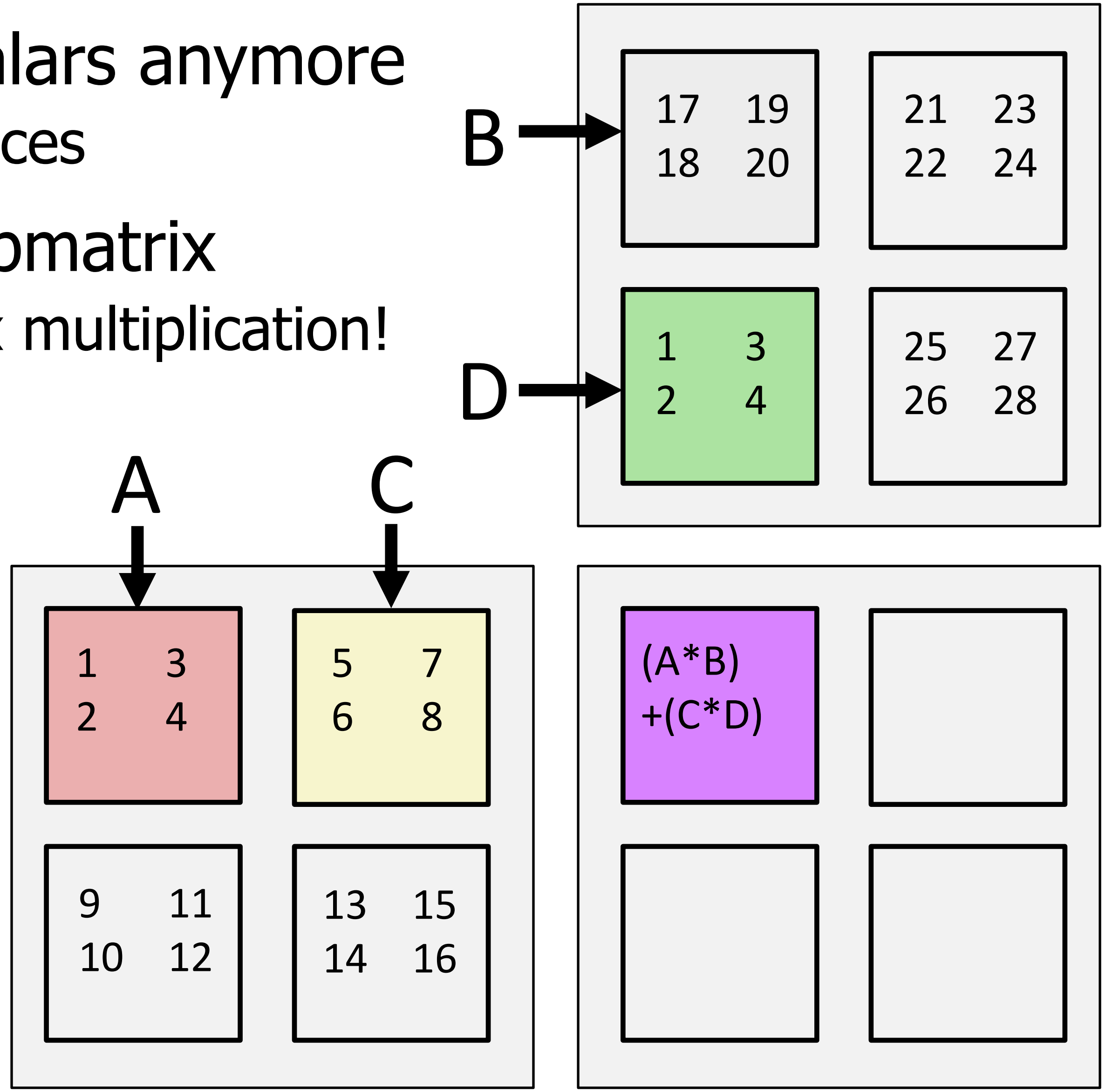
Enter Blocking Algorithms

- Special class of algorithms designed specifically to have excellent temporal and spatial locality
- Key idea: don't operate on individual elements; instead operate on *blocks* !
  - Treat the overall matrices as containing submatrices as elements
    - See next slide
- General principle: use a piece of data as much as we can
  - Then it's ok to kick it out of the cache
  - As opposed to using, kicking out, using again later, and so on
- Same result, but much nicer locality!
  - And thus can leverage the cache better (more hits, fewer misses)
  - Still same computational complexity
- May get a bit mind bending
  - I want you to understand the general principle
  - But you don't need to fully understand the details of the algorithm

Matrices as Matrices of Submatrices

- Elements of are not scalars anymore
  - But rather smaller matrices
- To compute a result submatrix
  - Just do a smaller matrix multiplication!

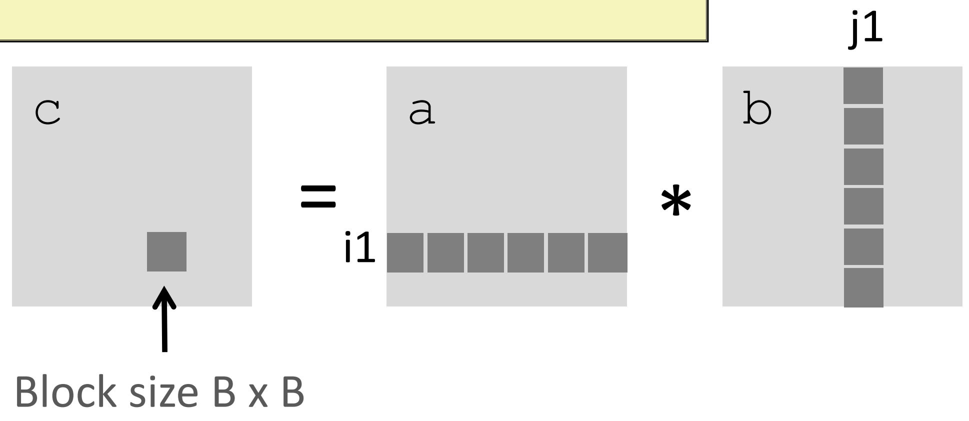




Blocked Matrix Multiplication

```
double * c = (double *) malloc(sizeof(double)*n*n);

/* Multiply n x n matrices a and b  */
void mmm(double *a, double *b, double *c, int n) {
  for (int i = 0; i < n; i+=B) {
    for (int j = 0; j < n; j+=B) {
      for (int k = 0; k < n; k+=B) {
        /* B x B mini matrix multiplications */
        for (int i1 = i; i1 < i+B; i1++) {
          for (int j1 = j; j1 < j+B; j1++) {
            double sum = 0.0;
            for (int k1 = k; k1 < k+B; k1++) {
              sum += a[i1*n + k1] * b[k1*n + j1];
            }
            c[i1*n + j1] = sum;
} } } } } }
```

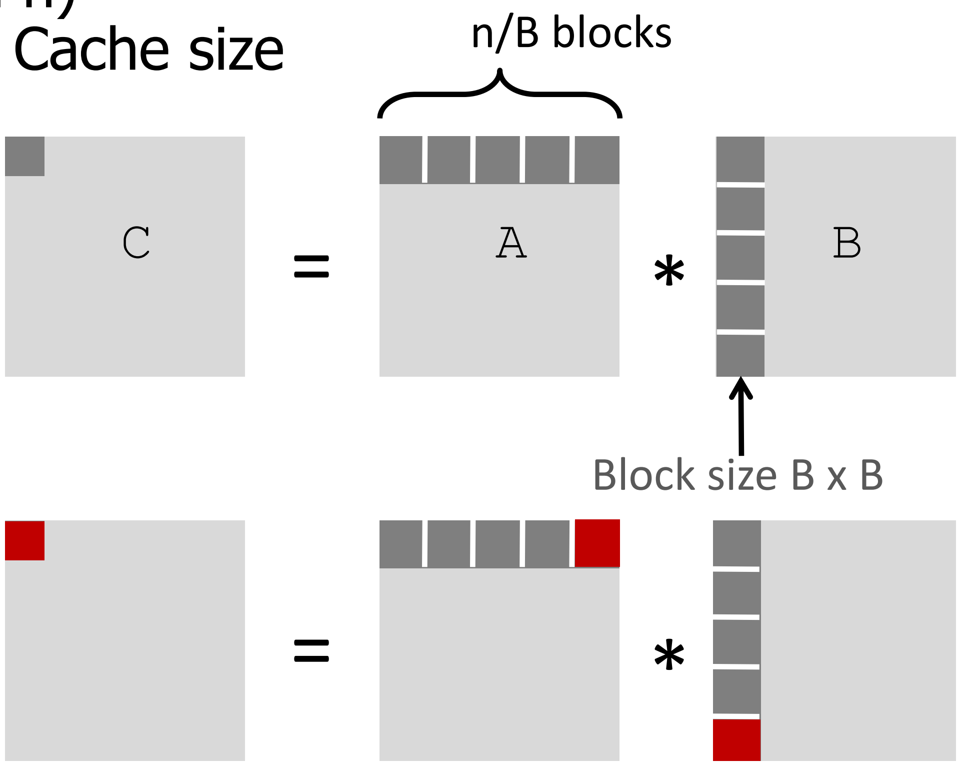


Cache Miss Analysis (approximate)

- Assume:
  - Cache block = 8 doubles
  - Cache size <<< n (much smaller than n)
  - Three blocks fit into cache: $3B^2$ < Cache size

- First (block) iteration:
  - $B^2/8$ misses for any given block
  - $2B^2/8$ misses for each BxB-block multiplication (only counting A, B misses)
  - # BxB multiplications: n/B
  - $B^2/8$ misses for C[ ] block total
  - $2B^2/8 * n/B + B^2/8 = nB/4 + B^2/8$

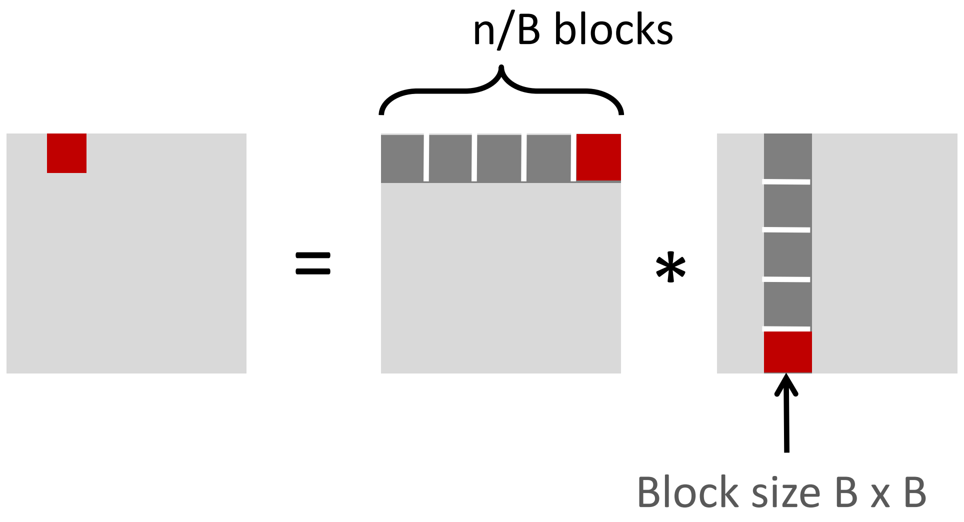
  - Afterwards in cache
    - No waste! We used all that we brought in!



Cache Miss Analysis (approximate)

- Assume:
  - Cache block = 8 doubles
  - Cache size << n (much smaller than n)
  - Three blocks fit into cache: $3B^2$ < Cache size

- Second (block) iteration:
  - Same as first iteration
  - misses = $nB/4+B^2/8$



- Total misses:
  - #block iterations: $(n/B)^2$
  - $(nB/4 + B^2/8) * (n/B)^2 = n^3/(4B) + n^2/8$

Performance Impact

- Misses without blocking: $(9/8) * n^3 + n^2$
- Misses with blocking: $1/(4B) * n^3 + 1/8 * n^2$

- Largest possible block size B, but limit $3B^2 < C \rightarrow B = \lfloor \sqrt{C/3} \rfloor$ (so it all fits in the cache)
  - e.g., Cache size = 32K = 32,768 Bytes, then pick B = 104
  - Results:

  - No blocking: $1.125*n^3 + n^2$

    468x 8x

  - Blocking: $0.0024*n^3 + 0.125*n^2$

- Reason for dramatic difference:
  - Matrix multiplication has inherent temporal locality
  - But program has to be written properly to take advantage of it

Takeaways

- Writing code to take advantage of the cache is challenging
  - It's totally possible, but high effort
- Generally: maximize spatial and temporal locality
  - Use elements close to each other (moving horizontally in 2D array)
  - Use the same element as many times as possible in a row (output)
- Well-designed math libraries will do this for you!
  - MATLAB, Mathematica, R, SciPy, etc.
  - Jack Dongarra won a Turing award for this in 2021!

Outline

- Memory Mountain
- Cache Metrics
- Cache Performance for Arrays
- Improving code
  - Rearranging Matrix Math
  - Matrix Math in Blocks