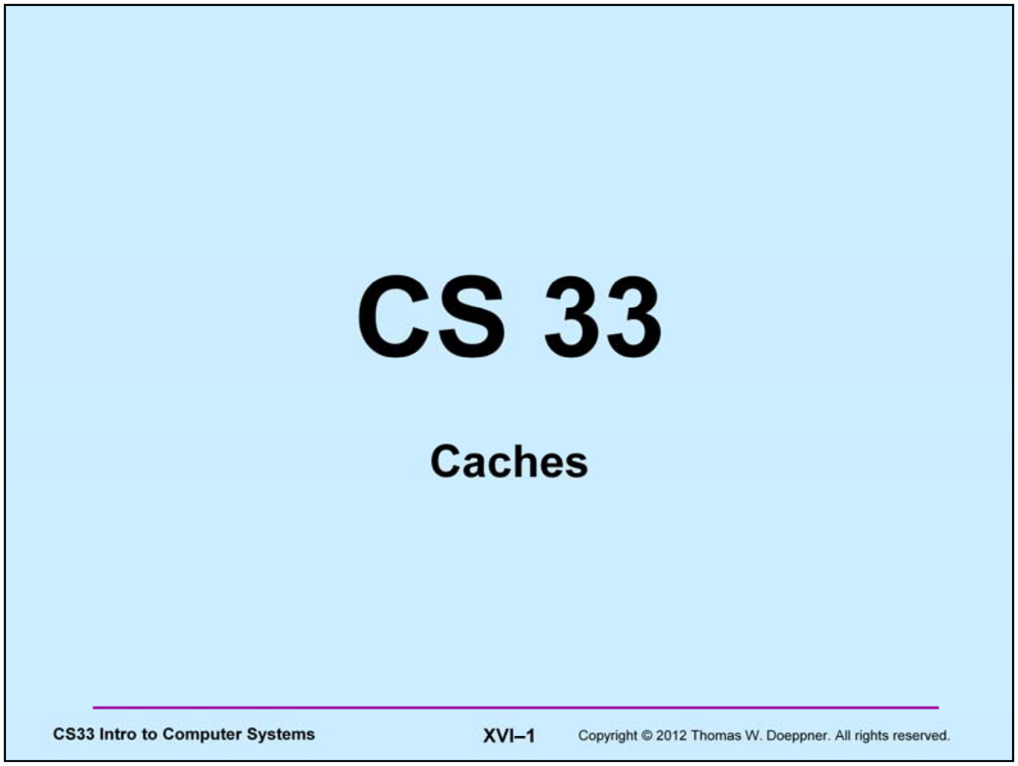


Most of the slides in this lecture are either from or adapted from slides provided by the authors of the textbook "Computer Systems: A Programmer's Perspective," 2nd Edition and are provided from the website of Carnegie-Mellon University, course 15-213, taught by Randy Bryant and David O'Hallaron in Fall 2010. These slides are indicated "Supplied by CMU" in the notes section of the slides.

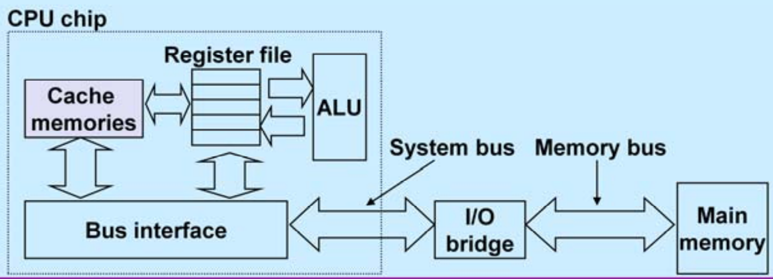
Today

- **Cache memory organization and operation**
- Performance impact of caches
  - The memory mountain
  - Rearranging loops to improve spatial locality
  - Using blocking to improve temporal locality

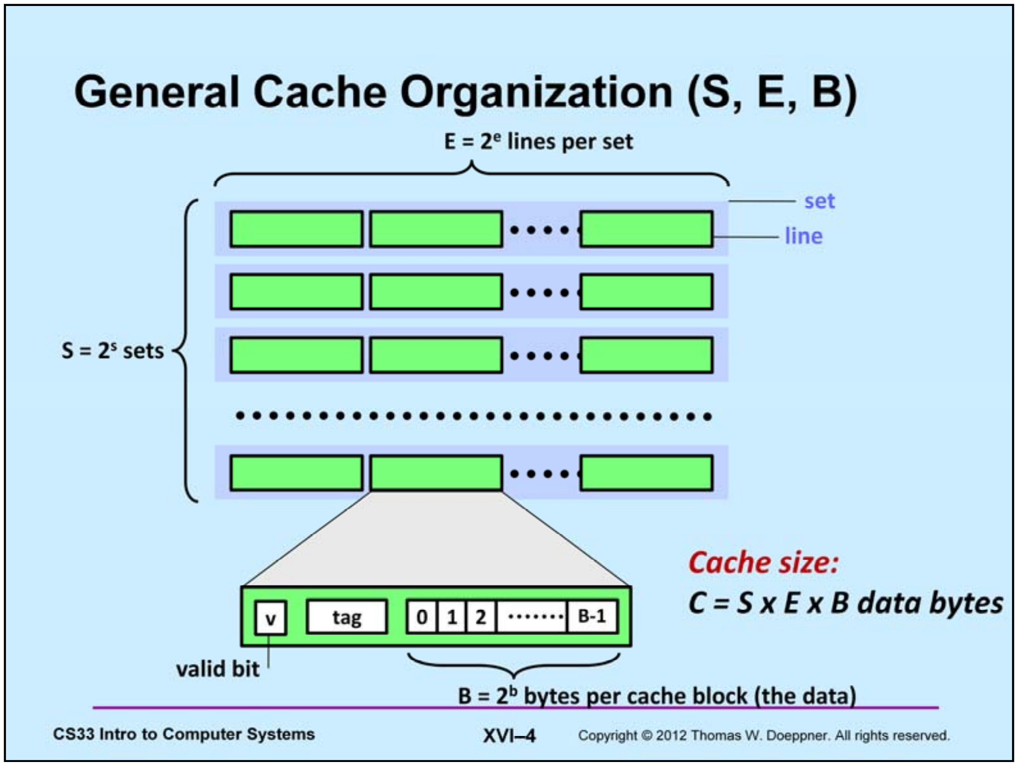
Supplied by CMU.

Cache Memories

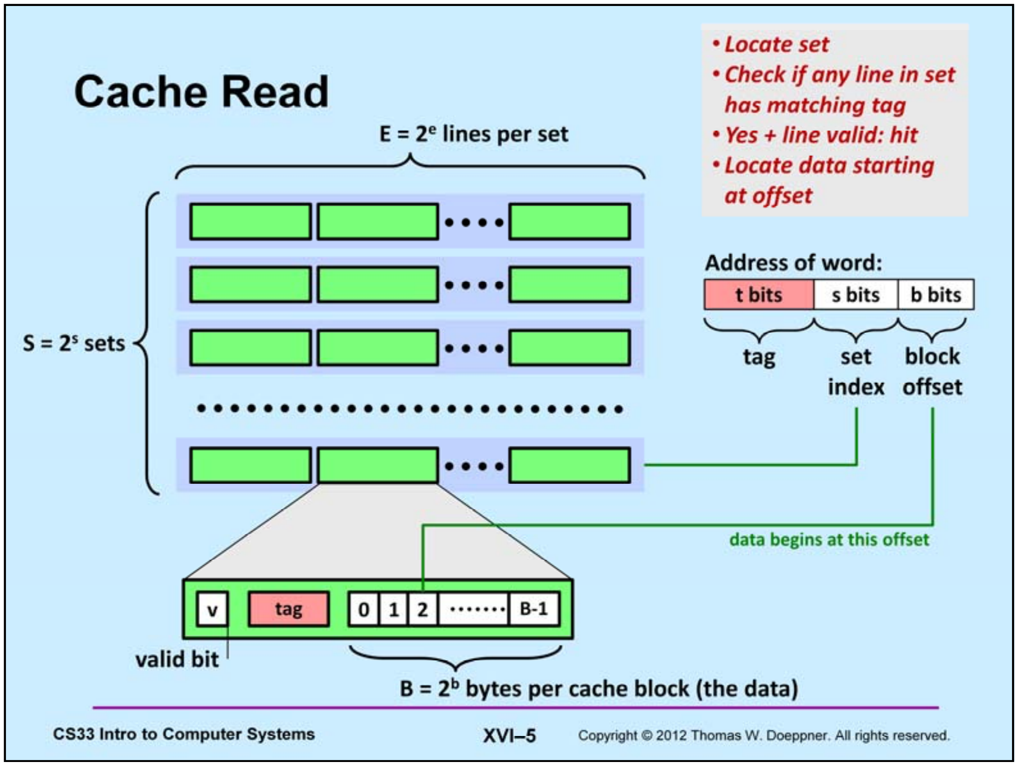
- Cache memories are small, fast SRAM-based memories managed automatically in hardware
  - hold frequently accessed blocks of main memory
- CPU looks first for data in caches (e.g., L1, L2, and L3), then in main memory
- Typical system structure:



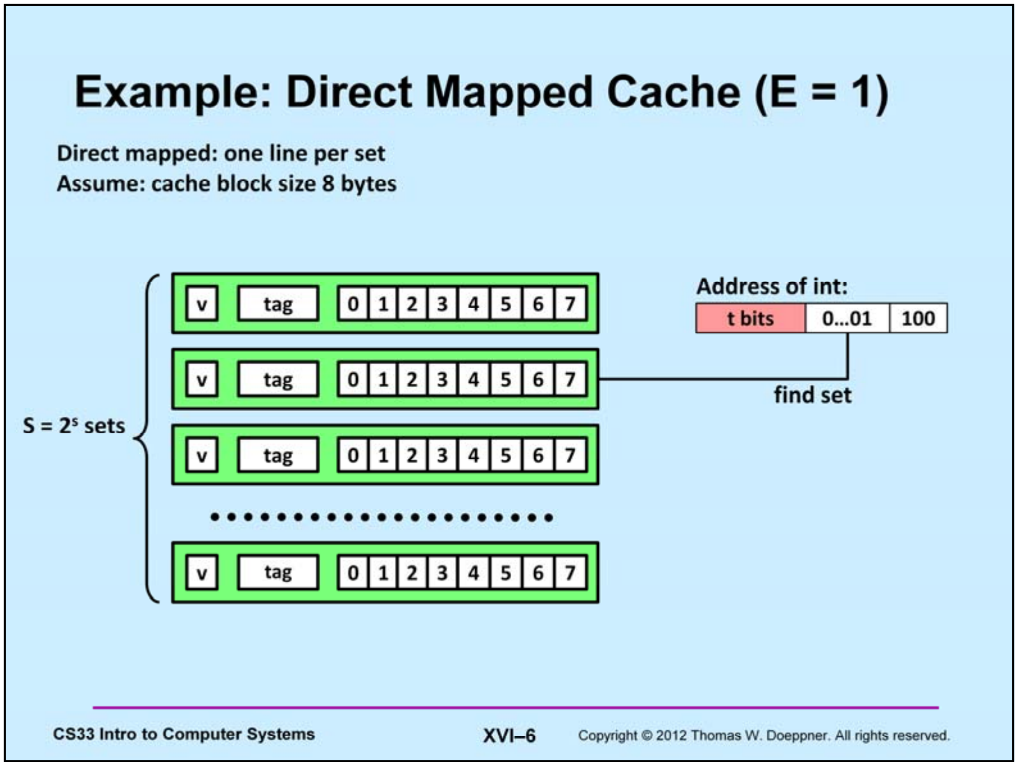
Supplied by CMU.



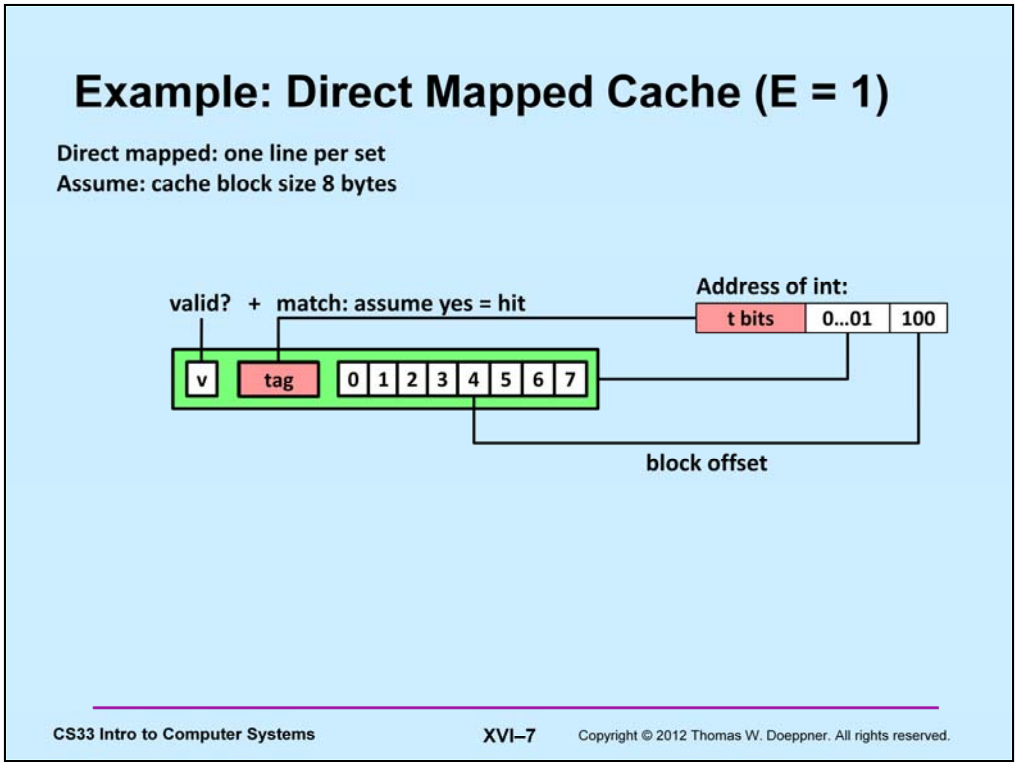
Supplied by CMU.



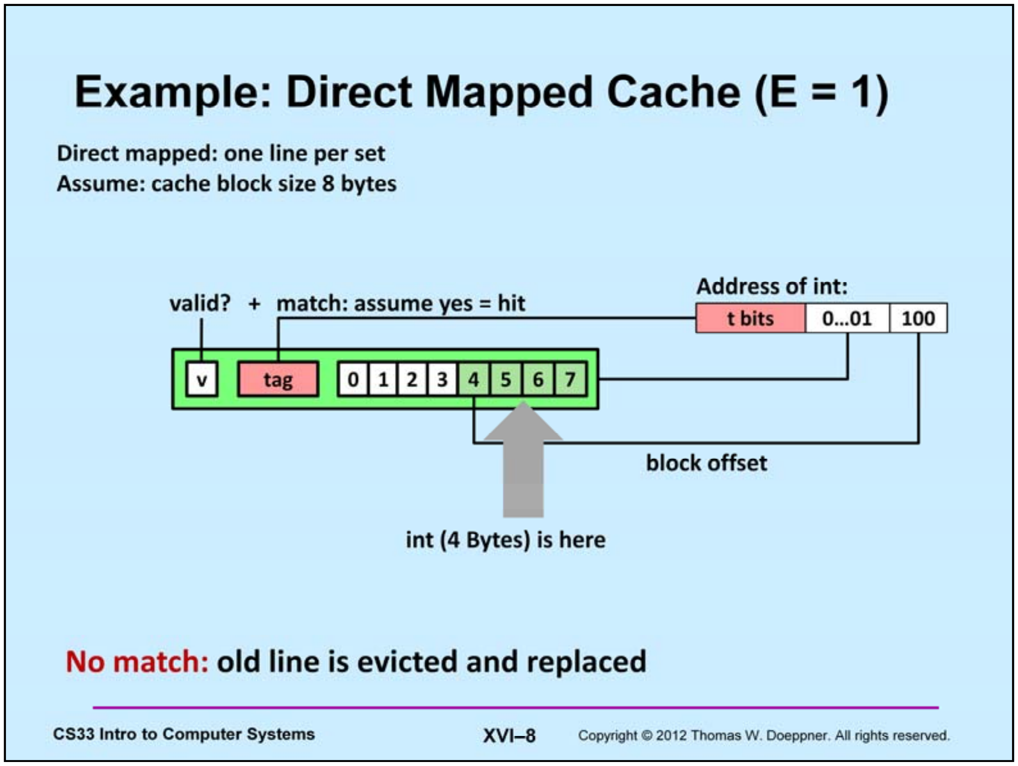
Supplied by CMU.



Supplied by CMU.



Supplied by CMU.



Supplied by CMU.

Direct-Mapped Cache Simulation

t=1	s=2	b=1
x	xx	x

M=16 byte addresses, B=2 bytes/block,
S=4 sets, E=1 Blocks/set

Address trace (reads, one byte per read):

0	$[0\underline{00}0_2]$,	miss
1	$[0\underline{00}1_2]$,	hit
7	$[0\underline{11}1_2]$,	miss
8	$[1\underline{00}0_2]$,	miss
0	$[0\underline{00}0_2]$	miss

	v	Tag	Block
Set 0	1	0	M[0-1]
Set 1			
Set 2			
Set 3	1	0	M[6-7]

Supplied by CMU.

A Higher-Level Example

Ignore the variables sum, i, j

```
int sum_array_rows(double a[16][16])
{
    int i, j;
    double sum = 0;

    for (i = 0; i < 16; i++)
        for (j = 0; j < 16; j++)
            sum += a[i][j];
    return sum;
}
```

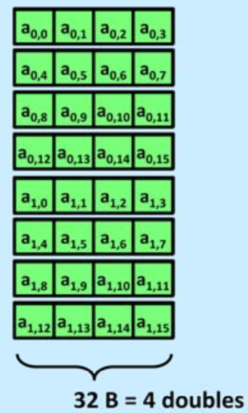
```
int sum_array_cols(double a[16][16])
{
    int i, j;
    double sum = 0;

    for (j = 0; i < 16; i++)
        for (i = 0; j < 16; j++)
            sum += a[i][j];
    return sum;
}
```

Supplied by CMU.

A Higher-Level Example

```
int sum_array_rows(double a[16][16])
{
    int i, j;
    double sum = 0;

    for (i = 0; i < 16; i++)
        for (j = 0; j < 16; j++)
            sum += a[i][j];
    return sum;
}
```

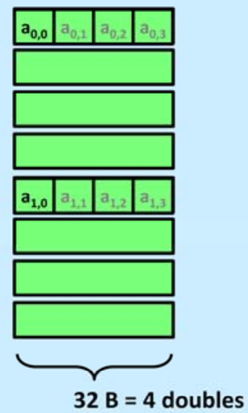
```
int sum_array_cols(double a[16][16])
{
    int i, j;
    double sum = 0;

    for (j = 0; i < 16; i++)
        for (i = 0; j < 16; j++)
            sum += a[i][j];
    return sum;
}
```

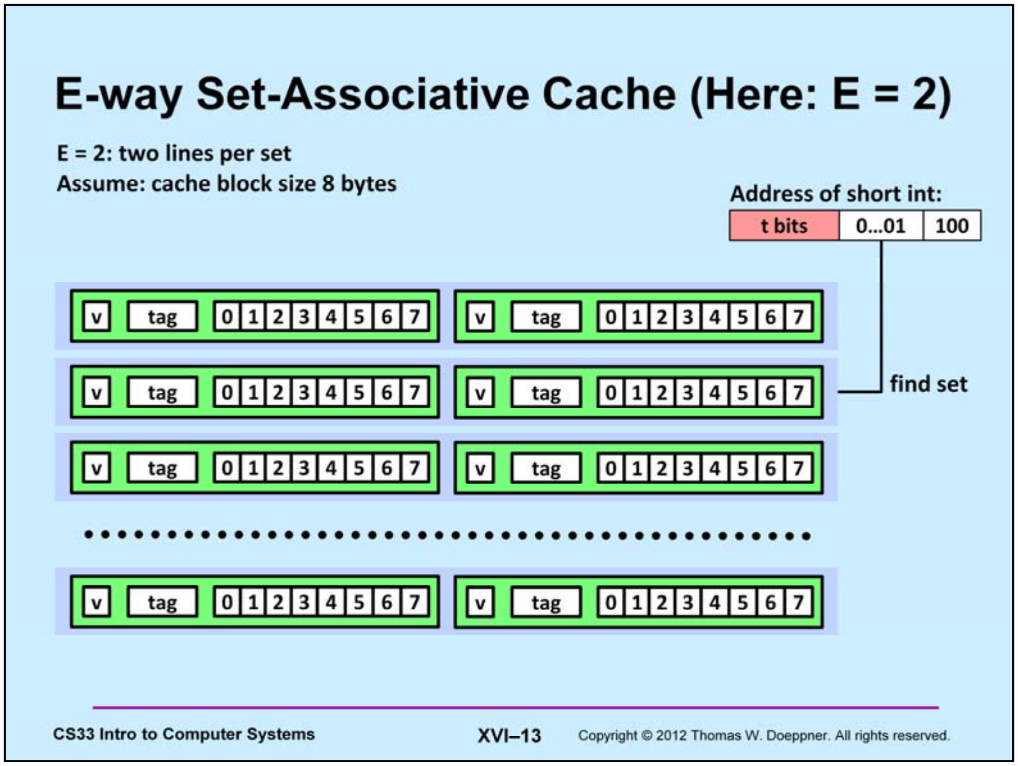


A Higher-Level Example

```c
int sum_array_rows(double a[16][16])
{
    int i, j;
    double sum = 0;

    for (i = 0; i < 16; i++)
        for (j = 0; j < 16; j++)
            sum += a[i][j];
    return sum;
}
```

```c
int sum_array_cols(double a[16][16])
{
    int i, j;
    double sum = 0;

    for (j = 0; i < 16; i++)
        for (i = 0; j < 16; j++)
            sum += a[i][j];
    return sum;
}
```

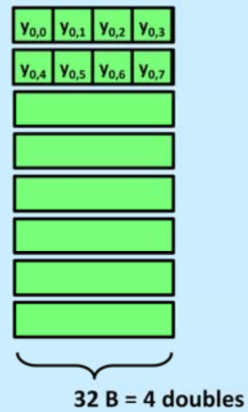




Supplied by CMU.

Conflict Misses

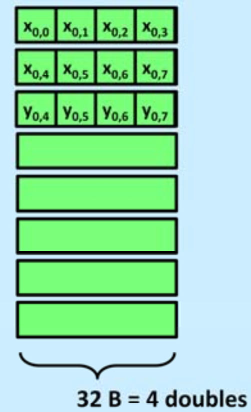
```
double dotprod(double x[8], double y[8]) {
  double sum = 0.0;
  int i;

  for (i=0; i<8; i++)
    sum += x[i] * y[i];

  return sum;
}
```



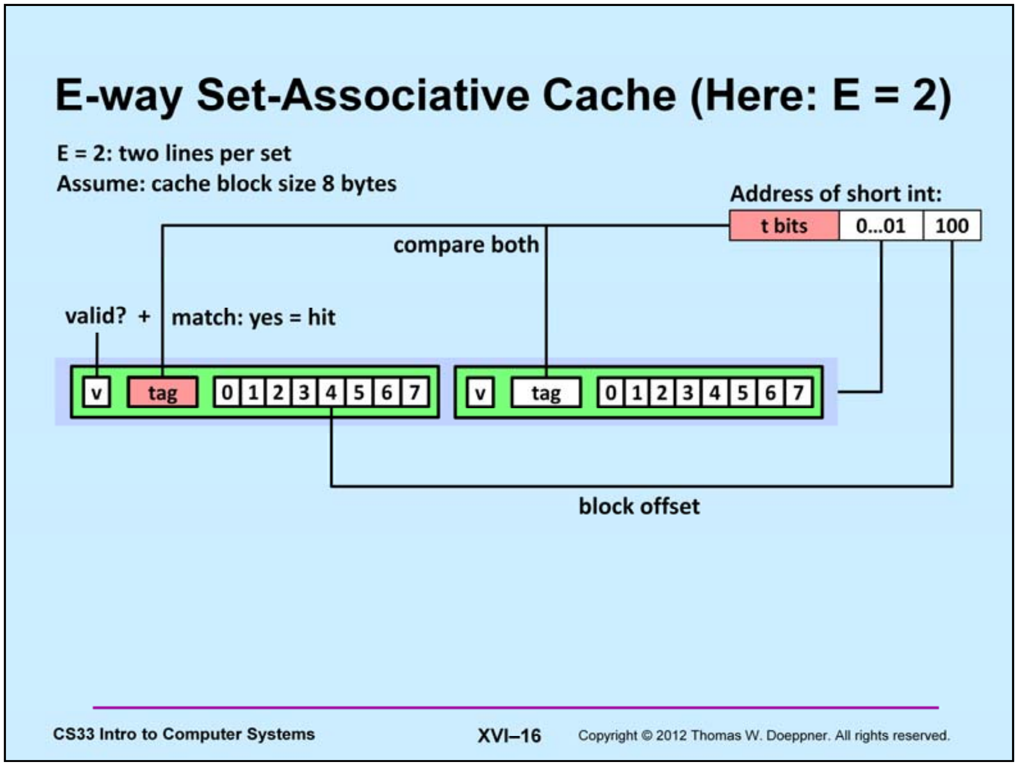
If arrays x and y have the same alignment, i.e., both start in the same cache set, then each access to an element of y replaces the cache line containing the corresponding element of x, and vice versa. The result is that loop is executed very slowly — each access to either array results in a conflict miss.

Conflict Misses

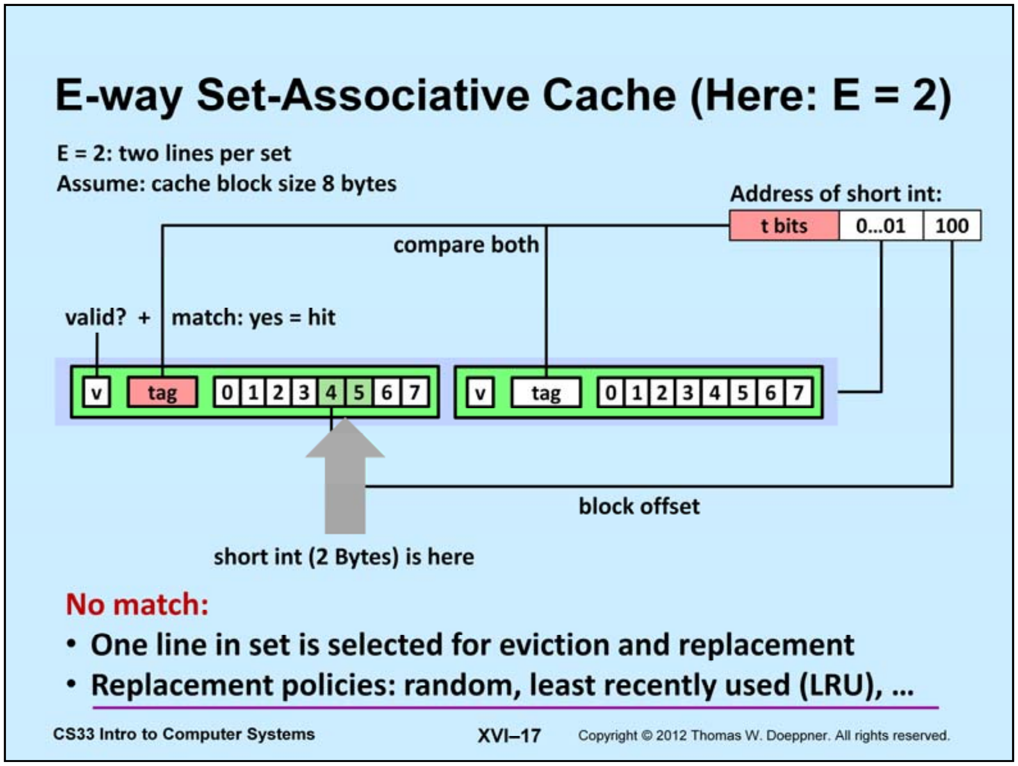
```
double dotprod(double x[8], double y[8]) {
  double sum = 0.0;
  int i;

  for (i=0; i<8; i++)
    sum += x[i] * y[i];

  return sum;
}
```



However, if the two arrays start in different cache sets, then the loop executes quickly — there is a cache miss on just every fourth access to each array.



Supplied by CMU.



Supplied by CMU.

2-Way Set-Associative Cache Simulation

t=2	s=1	b=1
xx	x	x

M=16 byte addresses, B=2 bytes/block,
S=2 sets, E=2 blocks/set

Address trace (reads, one byte per read):

0	**$[00\underline{0}0_2]$,**	miss
1	**$[00\underline{0}1_2]$,**	hit
7	**$[01\underline{1}1_2]$,**	miss
8	**$[10\underline{0}0_2]$,**	miss
0	**$[00\underline{0}0_2]$**	hit

	v	**Tag**	**Block**
Set 0	1	00	M[0-1]
	1	10	M[8-9]
Set 1	1	01	M[6-7]
	0		

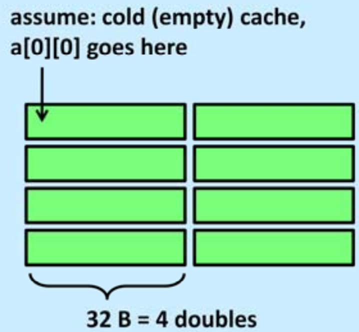
Supplied by CMU.

A Higher-Level Example

Ignore the variables sum, i, j

```
int sum_array_rows(double a[16][16])
{
    int i, j;
    double sum = 0;

    for (i = 0; i < 16; i++)
        for (j = 0; j < 16; j++)
            sum += a[i][j];
    return sum;
}
```

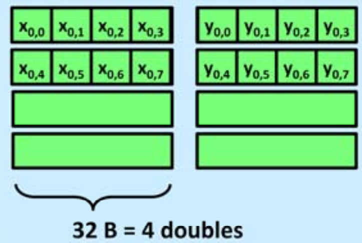
```
int sum_array_rows(double a[16][16])
{
    int i, j;
    double sum = 0;

    for (j = 0; i < 16; i++)
        for (i = 0; j < 16; j++)
            sum += a[i][j];
    return sum;
}
```



Supplied by CMU.

Conflict Misses

```
double dotprod(double x[8], double y[8]) {
  double sum = 0.0;
  int i;

  for (i=0; i<8; i++)
    sum += x[i] * y[i];

  return sum;
}
```

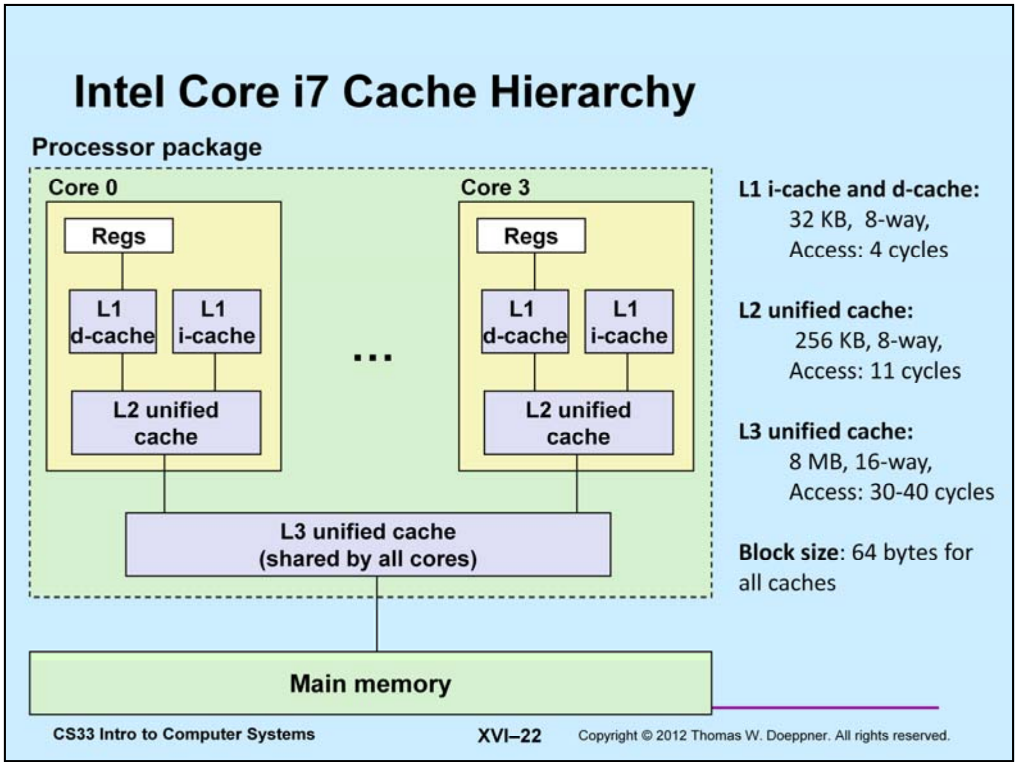


With a 2-way set-associative cache, our dot-product example runs quickly even if the two arrays have the same alignment.

What About Writes?

- Multiple copies of data exist:
  - L1, L2, main memory, disk
- What to do on a write-hit?
  - write-through (write immediately to memory)
  - write-back (defer write to memory until replacement of line)
    - need a dirty bit (line different from memory or not)
- What to do on a write-miss?
  - write-allocate (load into cache, update line in cache)
    - good if more writes to the location follow
  - no-write-allocate (writes immediately to memory)
- Typical
  - write-through + no-write-allocate
  - write-back + write-allocate

Supplied by CMU.



Supplied by CMU.

The L3 cache is known as the *last-level cache* (LLC) in the Intel documentation.

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 size, etc.
- **Hit time**
  - time to deliver a line in the cache to the processor
    - » includes time to determine whether the line is in the cache
  - typical numbers:
    - » 1-2 clock cycle for L1
    - » 5-20 clock cycles for L2
- **Miss penalty**
  - additional time required because of a miss
    - » typically 50-200 cycles for main memory (trend: increasing!)

Supplied by CMU.

Let’s Think About Those Numbers

- Huge difference between a hit and a miss
  - could be 100x, if just L1 and main memory
- Would you believe 99% hits is twice as good as 97%?
  - consider:
    cache hit time of 1 cycle
    miss penalty of 100 cycles
  - average access time:
    97% hits: .97 * 1 cycle + 0.03 * 100 cycles ≈ 4 cycles
    99% hits: .99 * 1 cycle + 0.01 * 100 cycles ≈ 2 cycles
- This is why “miss rate” is used instead of “hit rate”

Supplied by CMU.

Writing Cache-Friendly Code

- **Make the common case go fast**
  - focus on the inner loops of the core functions
- **Minimize the misses in the inner loops**
  - repeated references to variables are good (temporal locality)
  - stride-1 reference patterns are good (spatial locality)

Key idea: our qualitative notion of locality is quantified through our understanding of cache memories

Supplied by CMU.

Today

- Cache organization and operation
- Performance impact of caches
  - The memory mountain
  - Rearranging loops to improve spatial locality
  - Using blocking to improve temporal locality

Supplied by CMU.

The Memory Mountain

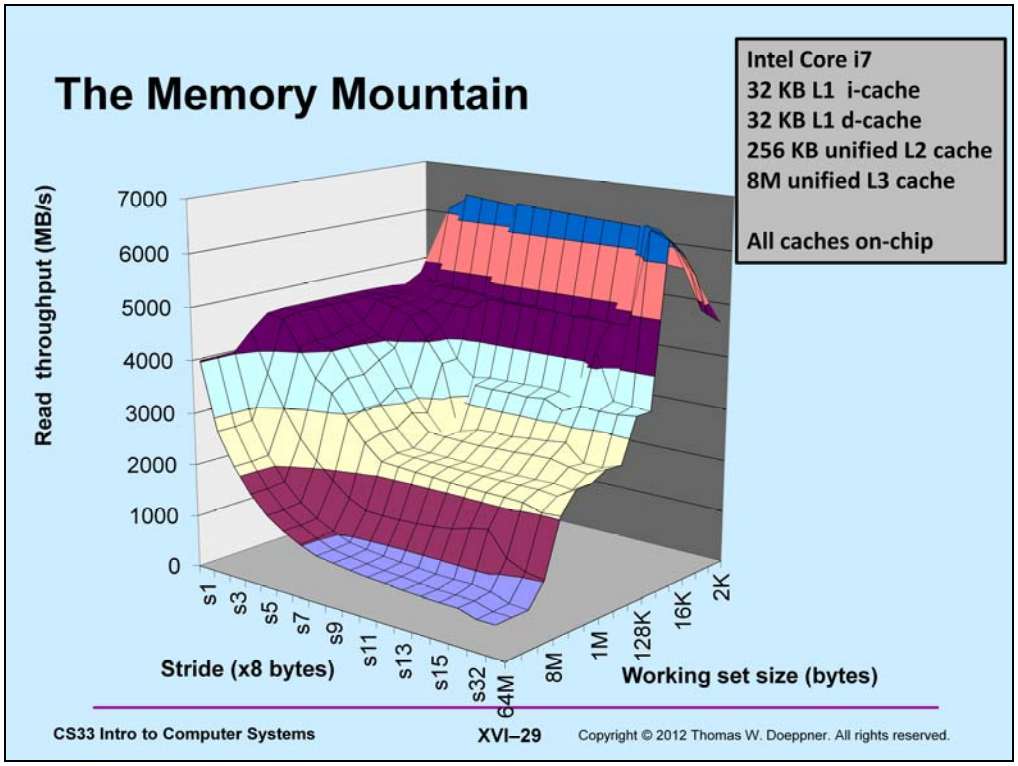
- **Read throughput** (read bandwidth)
  - number of bytes read from memory per second (MB/s)
- **Memory mountain:** measured read throughput as a function of spatial and temporal locality
  - compact way to characterize memory system performance

Supplied by CMU.

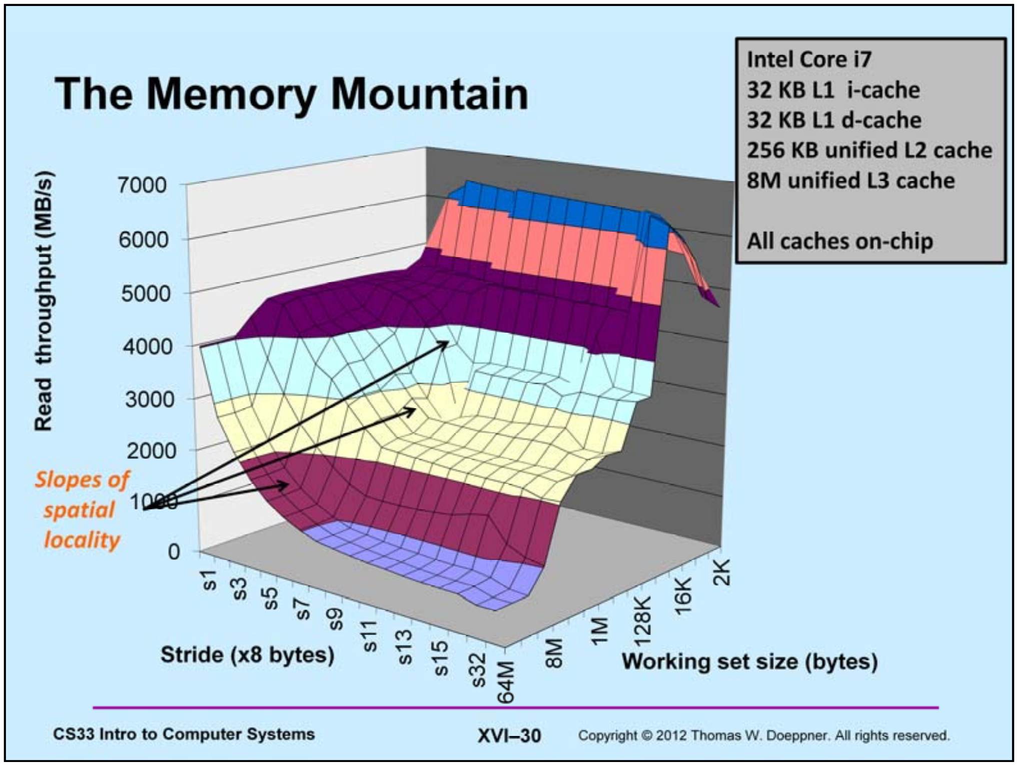
Memory Mountain Test Function

```
/* The test function */
void test(int elems, int stride) {
    int i, result = 0;
    volatile int sink;

    for (i = 0; i < elems; i += stride)
        result += data[i];
    sink = result; /* So compiler doesn't optimize away the loop */
}

/* Run test(elems, stride) and return read throughput (MB/s) */
double run(int size, int stride, double Mhz)
{
    double cycles;
    int elems = size / sizeof(int);

    test(elems, stride);                     /* warm up the cache */
    cycles = fcyc2(test, elems, stride, 0);  /* call test(elems,stride) */
    return (size / stride) / (cycles / Mhz); /* convert cycles to MB/s */
}
```

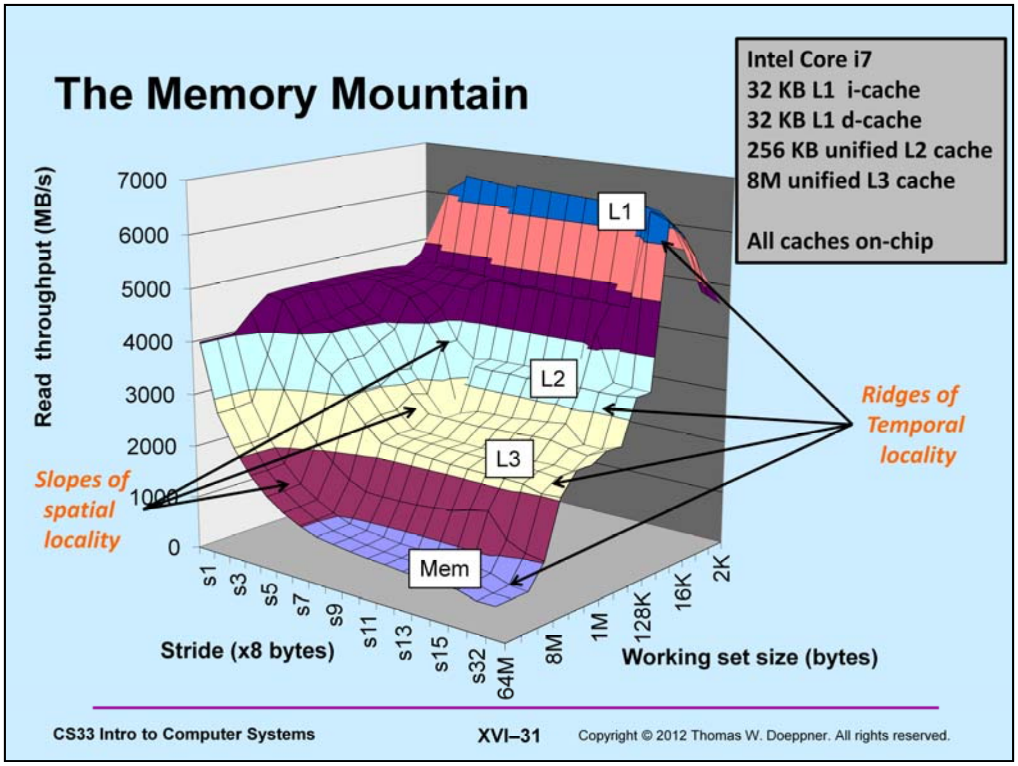

Supplied by CMU.



Supplied by CMU.



Supplied by CMU.



Supplied by CMU.

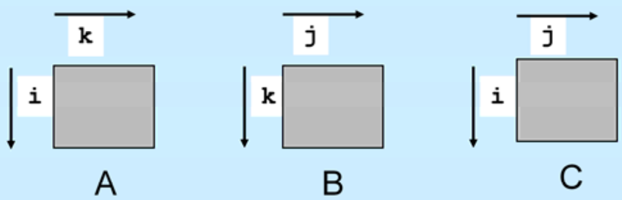
Today

- Cache organization and operation
- Performance impact of caches
  - The memory mountain
  - **Rearranging loops to improve spatial locality**
  - Using blocking to improve temporal locality

Supplied by CMU.

Miss-Rate Analysis for Matrix Multiply

- **Assume:**
  - line size = 32B (big enough for four 64-bit words)
  - matrix dimension (N) is very large
    - » approximate 1/N as 0.0
  - cache is not even big enough to hold multiple rows
- **Analysis method:**
  - look at access pattern of inner loop



Supplied by CMU.

Matrix Multiplication Example

- **Description:**
  - multiply N x N matrices
  - $O(N^3)$ total operations
  - N reads per source element
  - N values summed per destination
    - » but may be able to hold in register

Variable `sum` *held in register*

```
/* 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;
  }
}
```


Supplied by CMU.

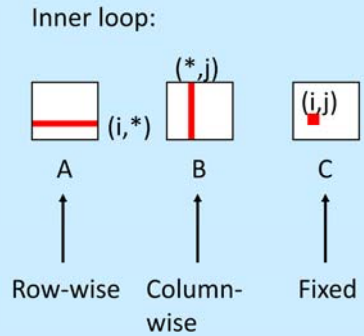
Layout of C Arrays in Memory (review)

- C arrays allocated in row-major order
  - each row in contiguous memory locations
- Stepping through columns in one row:
  - ```
    for (i = 0; i < N; i++)
        sum += a[0][i];
    ```
  - accesses successive elements
  - if block size (B) > 4 bytes, exploit spatial locality
    - » compulsory miss rate = 4 bytes / B
- Stepping through rows in one column:
  - ```
    for (i = 0; i < n; i++)
        sum += a[i][0];
    ```
  - accesses distant elements
  - no spatial locality!
    - » compulsory miss rate = 1 (i.e. 100%)

Supplied by CMU.

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;
  }
}
```



Misses per inner loop iteration:

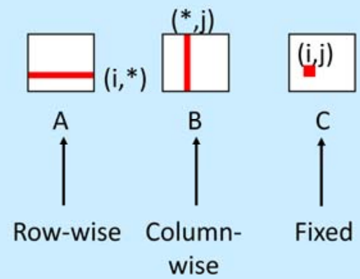
A	B	C
0.25	1.0	0.0

Supplied by CMU.

Matrix Multiplication (jik)

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

Inner loop:



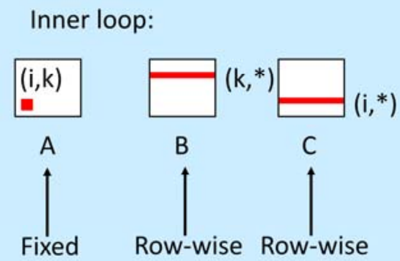
Misses per inner loop iteration:

A	B	C
0.25	1.0	0.0

Supplied by CMU.

Matrix Multiplication (kij)

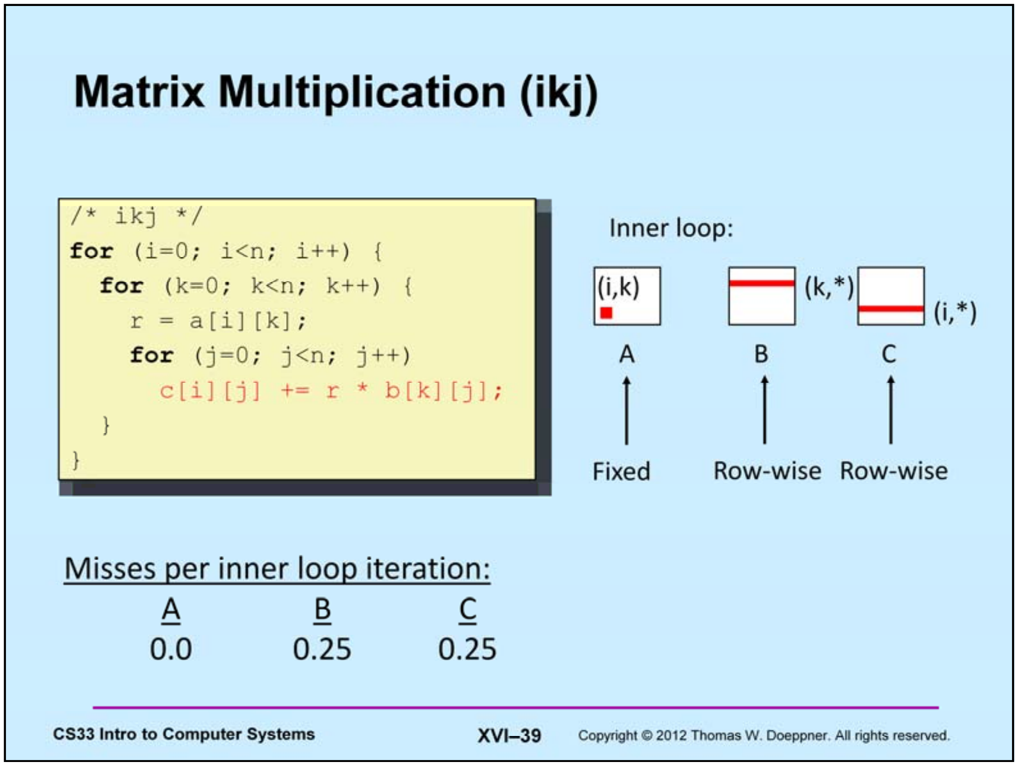
```
/* kij */
for (k=0; k<n; k++) {
  for (i=0; i<n; i++) {
    r = a[i][k];
    for (j=0; j<n; j++)
      c[i][j] += r * b[k][j];
  }
}
```



Misses per inner loop iteration:

A	B	C
0.0	0.25	0.25

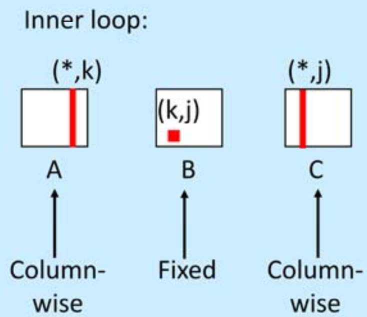
Supplied by CMU.



Supplied by CMU.

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;
  }
}
```



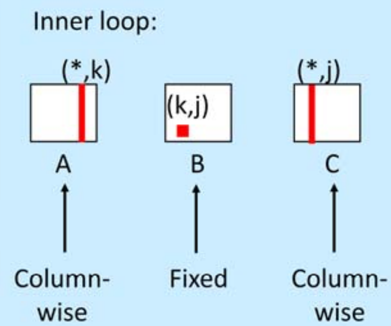
Misses per inner loop iteration:

A	B	C
1.0	0.0	1.0

Supplied by CMU.

Matrix Multiplication (kji)

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



Misses per inner loop iteration:

A	B	C
1.0	0.0	1.0

Supplied by CMU.

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**

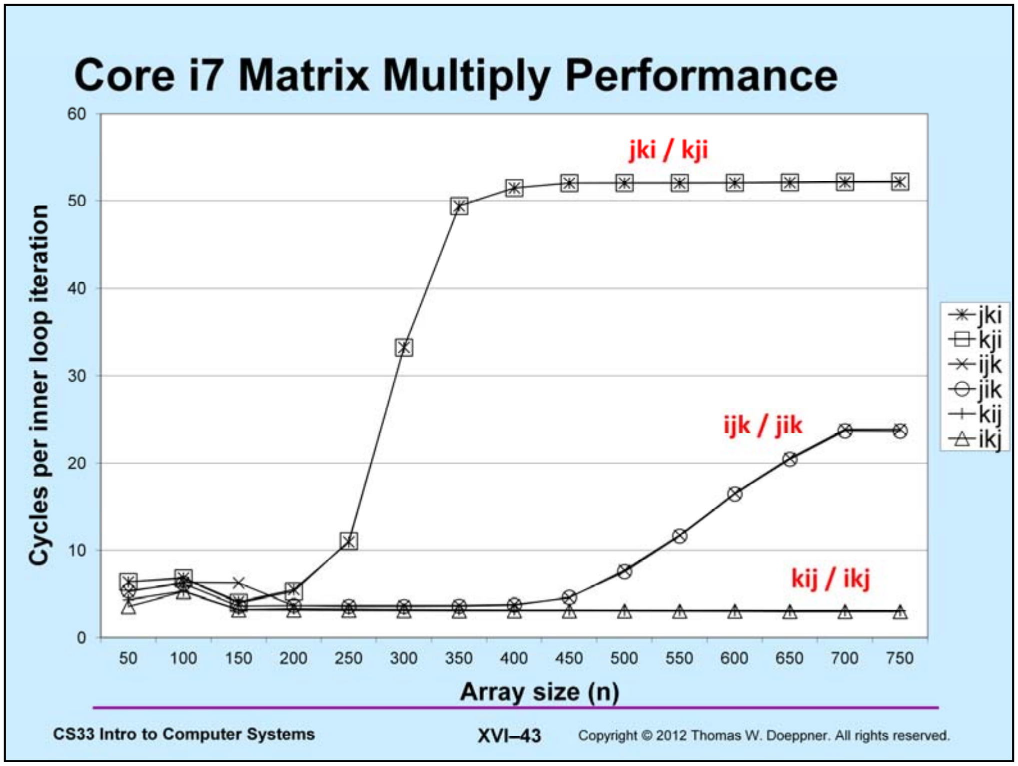
```
for (k=0; k<n; k++)
 for (i=0; i<n; i++) {
  r = a[i][k];
  for (j=0; j<n; j++)
   c[i][j] += r * b[k][j];
  }
```

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 = **2.0**

Supplied by CMU.



Supplied by CMU.

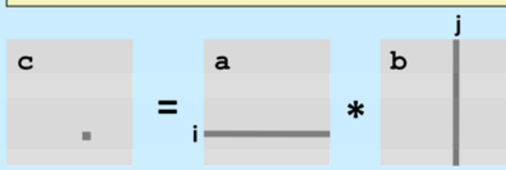
Today

- Cache organization and operation
- Performance impact of caches
  - The memory mountain
  - Rearranging loops to improve spatial locality
  - **Using blocking to improve temporal locality**

Supplied by CMU.

Example: Matrix Multiplication

```
c = (double *) calloc(sizeof(double), n*n);

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

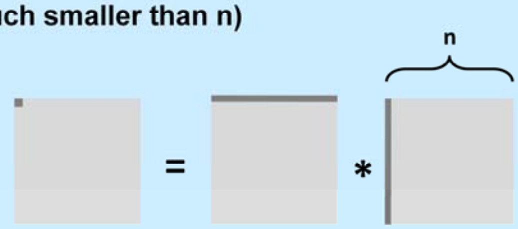


Supplied by CMU.

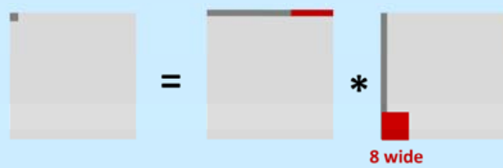
Cache-Miss Analysis

- **Assume:**
  - matrix elements are doubles
  - cache block = 8 doubles
  - cache size C << n (much smaller than n)

- **First iteration:**
  - n/8 + n = 9n/8 misses



  - afterwards in cache: (schematic)

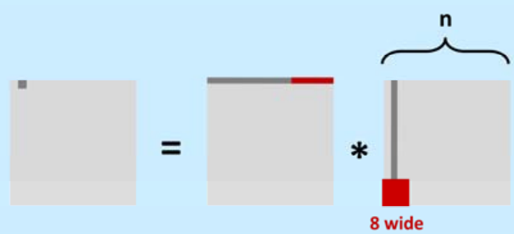


Supplied by CMU.

Cache-Miss Analysis

- **Assume:**
  - matrix elements are doubles
  - cache block = 8 doubles
  - cache size C << n (much smaller than n)

- **Second iteration:**
  - again:
    $n/8 + n = 9n/8$ misses

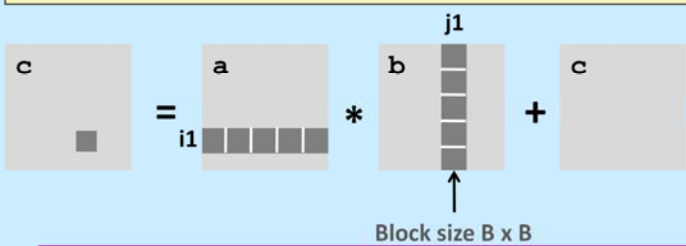


- **Total misses:**
  - $9n/8 * n^2 = (9/8) * n^3$

Supplied by CMU.

Blocked Matrix Multiplication

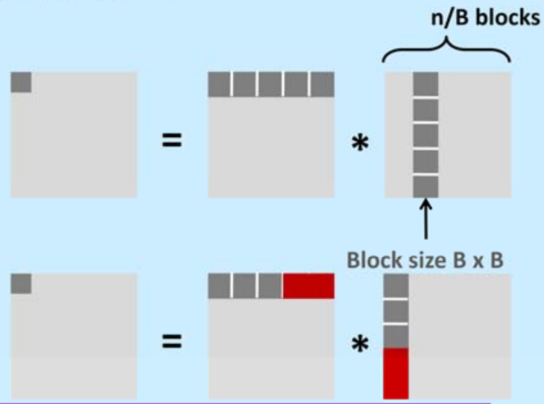
```
c = (double *) calloc(sizeof(double), n*n);

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



Supplied by CMU.

Cache-Miss Analysis

- **Assume:**
  - cache block = 8 doubles
  - cache size C << n (much smaller than n)
  - three blocks ■ fit into cache: $3B^2 < C$
- **First (block) iteration:**
  - $B^2/8$ misses for each block
  - $2n/B * B^2/8 = nB/4$ (omitting matrix c)
  - afterwards in cache (schematic)

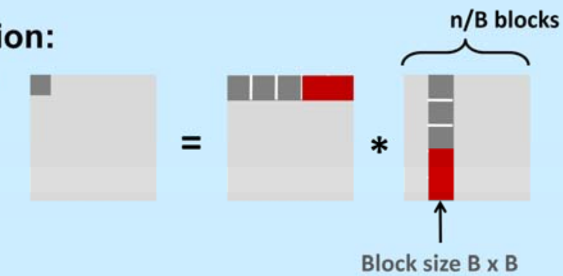


Supplied by CMU.

Cache-Miss Analysis

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

- **Second (block) iteration:**
  - same as first iteration
  - $2n/B * B^2/8 = nB/4$



- **Total misses:**
  - $nB/4 * (n/B)^2 = n^3/(4B)$

Supplied by CMU.

Summary

- No blocking: $(9/8) * n^3$
- Blocking: $1/(4B) * n^3$

- Suggest largest possible block size B, but limit $3B^2 < C$!

- Reason for dramatic difference:
  - matrix multiplication has inherent temporal locality:
    - » input data: $3n^2$, computation $2n^3$
    - » every array elements used O(n) times!
  - but program has to be written properly

Supplied by CMU.

Concluding Observations

- Programmer can optimize for cache performance
  - how data structures are organized
  - how data are accessed
    - » nested loop structure
    - » blocking is a general technique
- All systems favor “cache-friendly code”
  - getting absolute optimum performance is very platform specific
    - » cache sizes, line sizes, associativities, etc.
  - can get most of the advantage with generic code
    - » keep working set reasonably small (temporal locality)
    - » use small strides (spatial locality)

Supplied by CMU.