

Recap

- Cache memory organization and operation
- Memory mountain

Learning Goals

- Understand how we can optimize our code to improve efficiency and speed
- Learn about the optimizations GCC can perform

Plan for Today

- Writing cache-friendly code
- Optimization

Disclaimer: Slides for this lecture were borrowed from

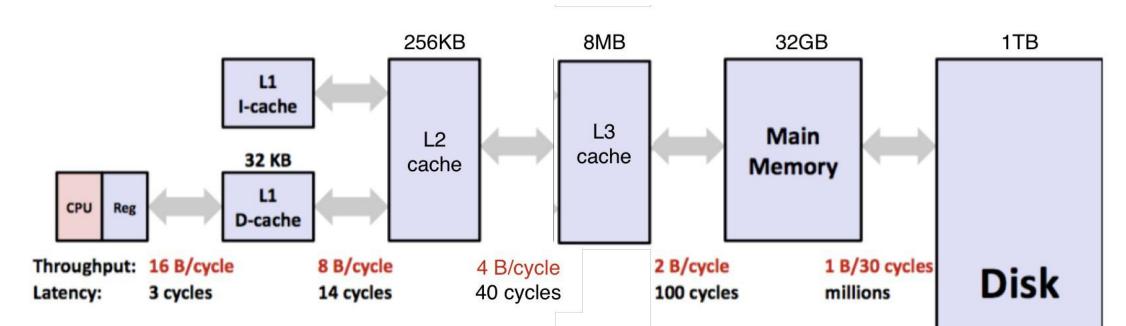
- —Nick Troccoli's Stanford CS107 class
- —Ashley Taylor's Stanford CS106B class

Plan for Today

- Writing cache-friendly code
- Optimization

Caching

- Processor speed is not the only bottleneck in program performance memory access is perhaps even more of a bottleneck!
- Memory exists in levels and goes from *really fast* (registers) to *really slow* (disk).
- As data is more frequently used, it ends up in faster and faster memory.



Caching

All caching depends on locality.

Temporal locality

- Repeat access to the same data tends to be co-located in TIME
- Intuitively: things I have used recently, I am likely to use again soon

Spatial locality

- Related data tends to be co-located in SPACE
- Intuitively: data that is near a used item is more likely to also be accessed

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)

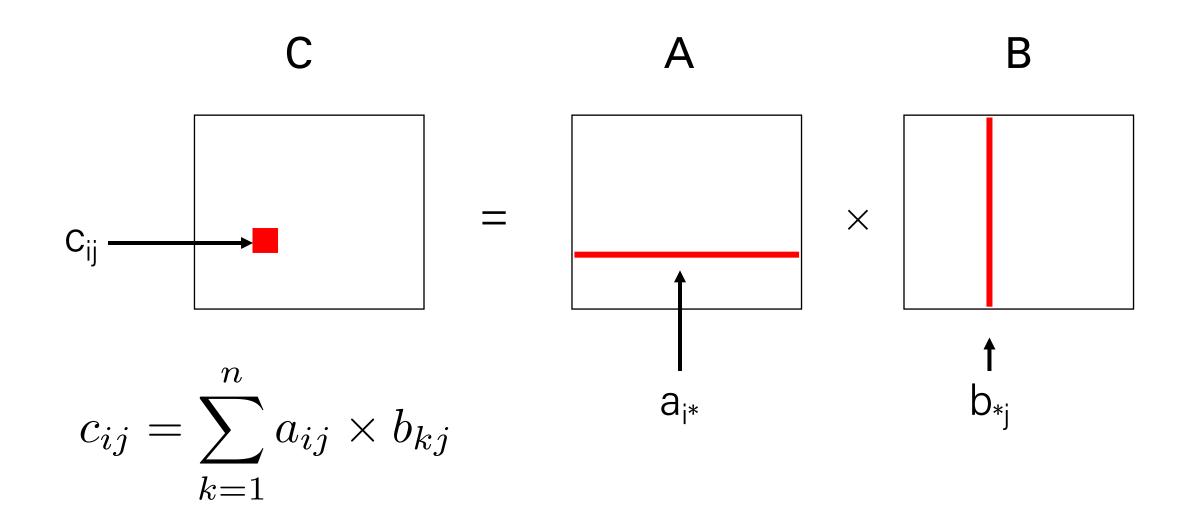
Key idea: Our qualitative notion of locality is quantified through our understanding of cache memories

Lecture Plan

- Writing cache-friendly code
 - Rearranging loops to improve spatial locality
 - Using blocking to improve temporal locality
- Optimization

Example: Matrix Multiplication

Matrix Multiplication Example



Matrix Multiplication Example

- Description:
 - Multiply N x N matrices
 - Matrix elements are doubles (8 bytes)
 - O(N³) total operations
 - N reads per source element
 - N values summed per destination
 - but may be able to hold 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;
}
</pre>
```

matmult/mm.c

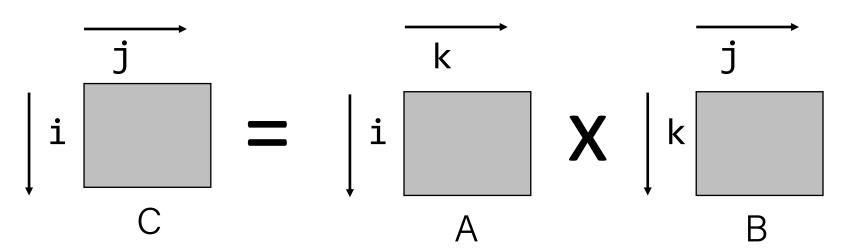
Miss Rate Analysis for Matrix Multiply

Assume

- Block size = 32B (big enough for four doubles)
- 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



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) > sizeof(aij) bytes, exploit spatial locality miss rate = sizeof(aij) / B

Stepping through rows in one column:

```
for (i = 0; i < n; i++)
sum += a[i][0];
```

- -accesses distant elements
- no spatial locality!miss rate = 1 (i.e. 100%)

Matrix Multiplication (ijk)

```
/* ijk */
                                            Inner loop:
for (i=0; i<n; i++) {
  for (j=0; j<n; j++) {
    sum = 0.0;
                                                 (i,*)
                                                      В
    for (k=0; k<n; k++)
      sum += a[i][k] * b[k][j];
    c[i][j] = sum;
                                           Row-wise Column-
                                                             Fixed
                                                   wise
                                 matmult/mm.c
```

Misses per inner loop iteration:

<u>A</u> <u>B</u> <u>C</u> 0.25 1.0 0.0

Matrix Multiplication (jik)

```
/* jik */
                                            Inner loop:
for (j=0; j<n; j++) {
  for (i=0; i<n; i++) {
                                                             (i,j)
    sum = 0.0;
    for (k=0; k<n; k++)
      sum += a[i][k] * b[k][j];
    c[i][j] = sum;
                                            Row-wise Column-
                                                             Fixed
                                                     wise
                                 matmult/mm.c
```

Misses per inner loop iteration:

<u>A</u>	<u>B</u>	<u>C</u>
0.25	1.0	0.0

Matrix Multiplication (kij)

0.25

0.0

0.25

```
/* kij */
                                               Inner loop:
for (k=0; k<n; k++) {
  for (i=0; i<n; i++) {
                                                           (k,*)
                                              (i,k)
                                                                    (i,*)
    r = a[i][k];
    for (j=0; j<n; j++)
       c[i][j] += r * b[k][j];
                                                     Row-wise Row-wise
                                              Fixed
                                   matmult/mm.c
Misses per inner loop iteration:
```

Matrix Multiplication (ikj)

Misses per inner loop iteration:

0.0

<u>B</u>

0.25

0.25

```
/* ikj */
                                              Inner loop:
for (i=0; i<n; i++) {
  for (k=0; k<n; k++) {
                                                          (k,*)
                                             (i,k)
                                                                  (i,*)
    r = a[i][k];
    for (j=0; j<n; j++)
      c[i][j] += r * b[k][j];
                                                    Row-wise Row-wise
                                            Fixed
                                  matmult/mm.c
```

Matrix Multiplication (jki)

```
/* jki */
                                              Inner loop:
for (j=0; j<n; j++) {
                                               (*,k)
                                                              (*,j)
  for (k=0; k<n; k++) {
                                                      (k,j)
    r = b[k][j];
    for (i=0; i<n; i++)
       c[i][j] += a[i][k] * r;
                                                             Column-
                                            Column-
                                                      Fixed
                                              wise
                                                              wise
                                  matmult/mm.c
```

Misses per inner loop iteration:

<u>A</u>	<u>B</u>	<u>C</u>
1.0	0.0	1.0

Matrix Multiplication (kji)

```
/* kji */
                                             Inner loop:
for (k=0; k<n; k++) {
                                                (*,k)
  for (j=0; j<n; j++) {
                                                       (k,j)
    r = b[k][j];
    for (i=0; i<n; i++)
      c[i][j] += a[i][k] * r;
                                                       Fixed
                                                              Column-
                                             Column-
                                              wise
                                                                wise
                                 matmult/mm.c
```

Misses per inner loop iteration:

<u>A</u>	<u>B</u>	<u>C</u>
1.0	0.0	1.0

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;
}
}</pre>
```

```
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];
}</pre>
```

```
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;
  }
}</pre>
```

ijk (& jik):

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

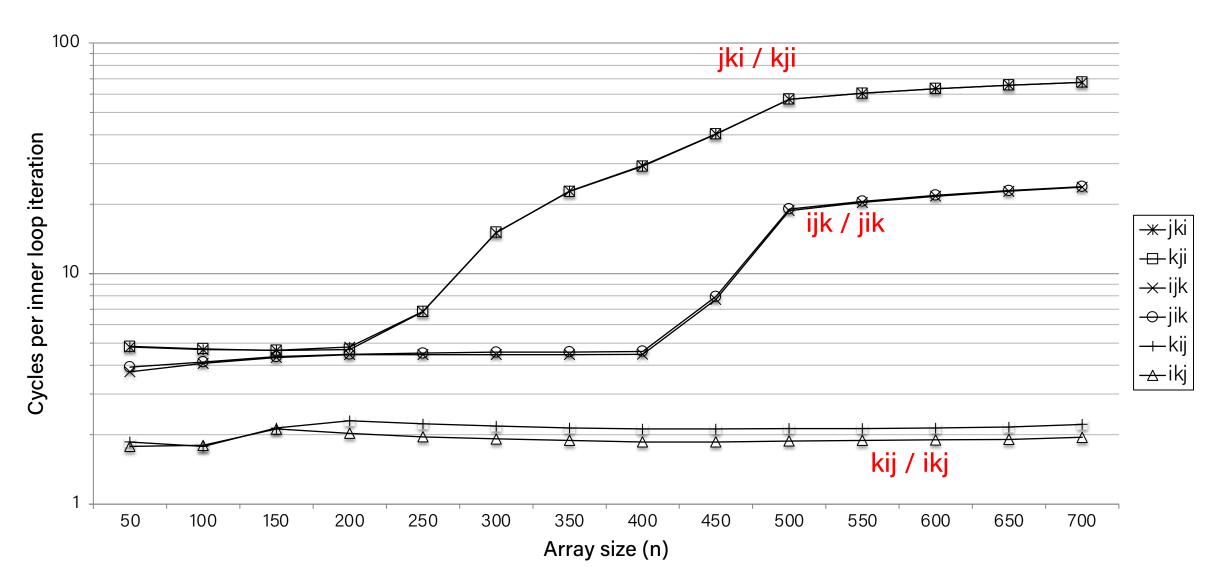
```
kij (& ikj):
```

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

```
jki (& kji):
```

- 2 loads, 1 store
- misses/iter = 2.0

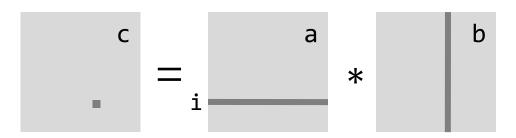
Core i7 Matrix Multiply Performance



Lecture Plan

- Writing cache-friendly code
 - Rearranging loops to improve spatial locality
 - Using blocking to improve temporal locality
- Optimization

Example: Matrix Multiplication



Cache Miss Analysis

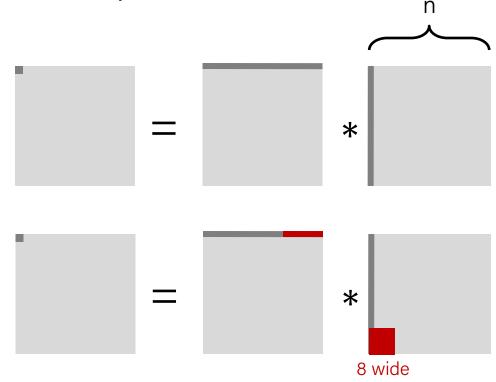
Assume

- Matrix elements are doubles
- Cache block = 8 doubles
- Cache size C << n (much smaller than n)</p>

First iteration:

- n/8 + n = 9n/8 misses

Afterwards in cache: (schematic)



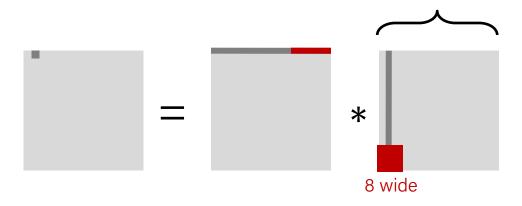
Cache Miss Analysis

Assume

- Matrix elements are doubles
- Cache block = 8 doubles
- Cache size C << n (much smaller than n)</p>

Second iteration:

- Again: n/8 + n = 9n/8 misses

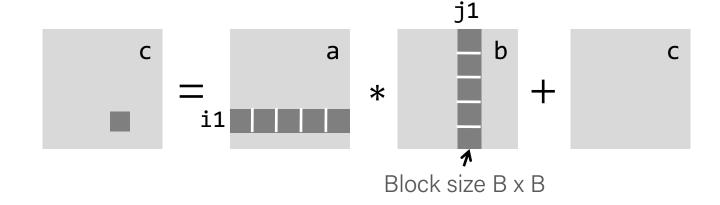


Total misses:

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

Blocked Matrix Multiplication

matmult/bmm.c



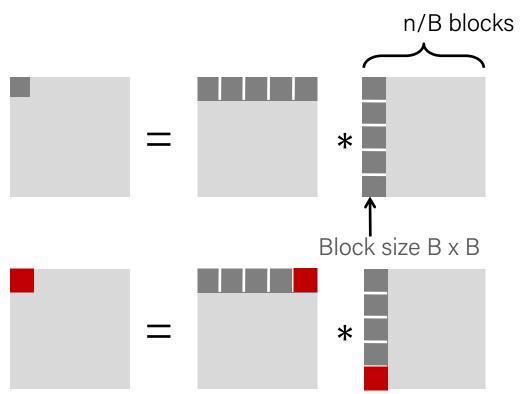
Cache Miss Analysis

Assume

- Cache block = 8 doubles
- Cache size C << n (much smaller than n)
- Three blocks fit into cache: 3B² < C

First (block) iteration:

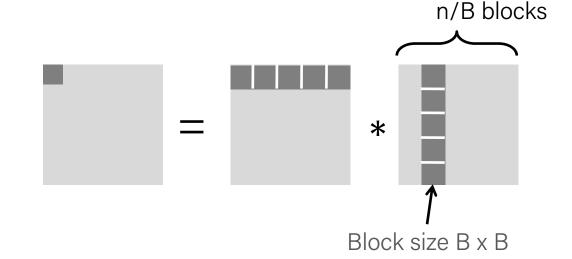
- B²/8 misses for each block
- -2n/B * B²/8 = nB/4 (omitting matrix c)
- Afterwards in cache (schematic)



Cache Miss Analysis

- Assume:
 - Cache block = 8 doubles
 - Cache size C << n (much smaller than n)
 - Three blocks fit into cache: 3B² < C

- Second (block) iteration:
 - Same as first iteration
 - -2n/B * B²/8 = nB/4



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

Blocking Summary

- **No blocking:** (9/8) * n³
- **Blocking:** 1/(4B) * n³

Suggest largest possible block size B, but limit 3B² < C!

- Reason for dramatic difference:
 - Matrix multiplication has inherent temporal locality:
 - Input data: 3n², computation 2n³
 - Every array elements used O(n) times!
 - But program has to be written properly

Naïve vs. Blocked Matrix Multiplication

Naïve Multiplication

Blocked Multiplication



 \approx 1,020,000 cache misses

≈ 90,000 cache misses

Recap

- Cache memories can have significant performance impact
- You can write your programs to exploit this!
 - Focus on the inner loops, where bulk of computations and memory accesses occur.
 - Try to maximize spatial locality by reading data objects with sequentially with stride 1.
 - Try to maximize temporal locality by using a data object as often as possible once it's read from memory.

Lecture Plan

- Writing cache-friendly codes
- Optimization
 - What is optimization?
 - GCC Optimization
 - Limitations of GCC Optimization
 - Caching revisited

Optimization

- Optimization is the task of making your program faster or more efficient with space or time. You already know explorations of efficiency with Big-O notation!
- Targeted, intentional optimizations to alleviate bottlenecks can result in big gains. But it's important to only work to optimize where necessary.

Optimization

Most of what you need to do with optimization can be summarized by:

- If doing something seldom and only on small inputs, do whatever is simplest to code, understand, and debug
- 2) If doing things thing a lot, or on big inputs, make the primary algorithm's Big-O cost reasonable
- 3) Let gcc do its magic from there
- 4) Optimize explicitly as a last resort

GCC Optimization

- Today, we'll be comparing two levels of optimization in the gcc compiler:
 - gcc -00 // mostly just literal translation of C
 - gcc -02 // enable nearly all reasonable optimizations
 - (we use -0g, like -00 but with less needless use of the stack)
- There are other custom and more aggressive levels of optimization, e.g.:

```
- -03 //more aggressive than -02, trade size for speed- -0s //optimize for size
```

- -Ofast //disregard standards compliance (!!)
- Exhaustive list of gcc optimization-related flags:
 - https://gcc.gnu.org/onlinedocs/gcc/Optimize-Options.html

Example: Matrix Multiplication

Here's a standard matrix multiply, a triply-nested for loop:

```
void mmm(double a[][DIM], double b[][DIM], double c[][DIM], int n) {
    for (int i = 0; i < n; i++) {
        for (int j = 0; j < n; j++) {
            for (int k = 0; k < n; k++) {
                c[i][j] += a[i][k] * b[k][j];
            }
        }
    }
}</pre>
```

```
./mult // -00 (no optimization)
matrix multiply 25^2: cycles 0.43M
matrix multiply 50^2: cycles 3.02M
matrix multiply 100^2: cycles 24.82M
```

```
./mult_opt // -02 (with optimization)
matrix multiply 25^2: cycles 0.13M (opt)
matrix multiply 50^2: cycles 0.66M (opt)
matrix multiply 100^2: cycles 5.55M (opt)
```

- Constant Folding
- Common Sub-expression Elimination
- Dead Code
- Strength Reduction
- Code Motion
- Tail Recursion
- Loop Unrolling

Optimizations may target one or more of:

- Static instruction count
- Dynamic instruction count
- Cycle count / execution time

- Constant Folding
- Common Sub-expression Elimination
- Dead Code
- Strength Reduction
- Code Motion
- Tail Recursion
- Loop Unrolling

Constant Folding

Constant Folding pre-calculates constants at compile-time where possible.

int seconds = 60 * 60 * 24 * n_days;

What is the consequence of this for you as a programmer? What should you do differently or the same knowing that compilers can do this for you?



Constant Folding

```
int fold(int param) {
    char arr[5];
    int a = 0x107;
    int b = a * sizeof(arr);
    int c = sqrt(2.0);
    return a * param + (a + 0x15 / c + strlen("Hello") * b - 0x37) / 4;
}
```

Constant Folding: Before (-00)

```
00000000000400626 <fold>:
  400626:
                                                     %rbp
                  55
                                              push
                  53
  400627:
                                              push
                                                     %rbx
                  48 83 ec 08
                                                     $0x8,%rsp
  400628:
                                              sub
                  89 fd
  40062c:
                                                     %edi,%ebp
                                              mov
                                                     0xda(%rip),%xmm0
  40062e:
                  f2 0f 10 05 da 00 00
                                              movsd
  400635:
                  00
  400636:
                  e8 d5 fe ff ff
                                              callq 400510 <sqrt@plt>
  40063b:
                  f2 0f 2c c8
                                              cvttsd2si %xmm0,%ecx
  40063f:
                  69 ed 07 01 00 00
                                              imul
                                                     $0x107,%ebp,%ebp
  400645:
                  b8 15 00 00 00
                                                     $0x15,%eax
                                              mov
  40064a:
                  99
                                              cltd
  40064b:
                  f7 f9
                                              idiv
                                                     %ecx
  40064d:
                  8d 98 07 01 00 00
                                              lea
                                                     0x107(%rax),%ebx
                  bf 04 07 40 00
                                                     $0x400704,%edi
  400653:
                                              mov
  400658:
                  e8 93 fe ff ff
                                              callq
                                                     4004f0 <strlen@plt>
                                                     $0x523,%rax,%rax
  40065d:
                  48 69 c0 23 05 00 00
                                              imul
  400664:
                  48 63 db
                                              movslq %ebx,%rbx
  400667:
                  48 8d 44 18 c9
                                                     -0x37(%rax, %rbx, 1), %rax
                                              lea
                  48 c1 e8 02
  40066c:
                                              shr
                                                     $0x2,%rax
  400670:
                  01 e8
                                                     %ebp,%eax
                                              add
  400672:
                  48 83 c4 08
                                                     $0x8,%rsp
                                              add
  400676:
                                                     %rbx
                  5b
                                              pop
  400677:
                                                     %rbp
                  5d
                                              pop
  400678:
                  c3
                                              retq
```

Constant Folding: After (-O2)

00000000004004f0 <fold>:

4004f0: 69 c7 07 01 00 00 imul \$0x107,%edi,%eax

4004f6: 05 a5 06 00 00 add \$0x6a5,%eax

4004fb: c3 retq

4004fc: 0f 1f 40 00 nopl 0x0(%rax)

- Constant Folding
- Common Sub-expression Elimination
- Dead Code
- Strength Reduction
- Code Motion
- Tail Recursion
- Loop Unrolling

Common Sub-Expression Elimination

Common Sub-Expression Elimination prevents the recalculation of the same thing many times by doing it once and saving the result.

```
int a = (param2 + 0x201);
int b = param1 * (param2 + 0x201) + a;
return a * (param2 + 0x201) + b * (param2 + 0x201);
```

Common Sub-Expression Elimination

Common Sub-Expression Elimination prevents the recalculation of the same thing many times by doing it once and saving the result.

```
This optimization is
 int a = (param2 + 0x201);
                                                 done even at -00!
int b = param1 * (param2 + 0x201) + a;
return a * (param2 + 0x201) + b * (param2 + 0x201);
00000000004004f0 <subexp>:
                                 add
  4004f0: 81 c6 07 01 00 00
                                        $0x201,%esi
                                        %esi,%edi
  4004f6: Of af fe
                                 imul
 4004f9: 8d 04 77
                                 lea
                                        (%rdi,%rsi,2),%eax
 4004fc: Of af c6
                                 imul
                                        %esi,%eax
 4004ff:
                                 retq
         c3
```

- Constant Folding
- Common Sub-expression Elimination
- Dead Code
- Strength Reduction
- Code Motion
- Tail Recursion
- Loop Unrolling

Dead Code

Dead code elimination removes code that doesn't serve a purpose:

```
if (param1 < param2 && param1 > param2) {
    printf("This test can never be true!\n");
// Empty for loop
for (int i = 0; i < 1000; i++);
// If/else that does the same operation in both cases
if (param1 == param2) {
    param1++;
} else {
    param1++;
// If/else that more trickily does the same operation in both cases
if (param1 == 0) {
    return 0;
} else {
    return param1;
```

Dead Code: Before (-O0)

```
00000000004004d6 <dead code>:
            b8 00 00 00 00
 4004d6:
                                            $0x0,%eax
                                     mov
 4004db:
         eb 03
                                            4004e0 <dead code+0xa>
                                     jmp
 4004dd:
         83 c0 01
                                     add
                                            $0x1,%eax
                                            $0x3e7,%eax
 4004e0: 3d e7 03 00 00
                                     cmp
         7e f6
                                     jle
                                            4004dd <dead code+0x7>
 4004e5:
         39 f7
                                            %esi,%edi
 4004e7:
                                     cmp
                                            4004f0 <dead_code+0x1a>
 4004e9:
         75 05
                                     jne
         8d 47 01
                                            0x1(%rdi),%eax
 4004eb:
                                     lea
                                            4004f3 <dead code+0x1d>
 4004ee:
         eb 03
                                     jmp
                                            0x1(%rdi),%eax
 4004f0:
         8d 47 01
                                     lea
 4004f3:
            f3 c3
                                     repz retq
```

Dead Code: After (-O2)

- Constant Folding
- Common Sub-expression Elimination
- Dead Code
- Strength Reduction
- Code Motion
- Tail Recursion
- Loop Unrolling

Strength Reduction

Strength reduction changes divide to multiply, multiply to add/shift, and mod to AND to avoid using instructions that cost many cycles (multiply and divide).

```
int a = param2 * 32;
int b = a * 7;
int c = b / 3;
int d = param2 \% 2;
for (int i = 0; i <= param2; i++) {
    c += param1[i] + 0x107 * i;
return c + d;
```

Strength Reduction: After (-O3)

```
unsigned udiv19(unsigned arg) {
return arg / 19;
}
```

```
udiv19(unsigned int):
    mov eax, edi
    mov edx, 2938661835
    imul rax, rdx
    shr rax, 32
    sub edi, eax
    shr edi
    add eax, edi
    shr eax, 4
    ret
```

https://godbolt.org/z/Wq8ra3

What really happens here?

$$a \cdot \frac{1}{19} \approx \frac{a \cdot \frac{2938661835}{2^{32}} + \frac{a - a \cdot \frac{2938661835}{2^{32}}}{2^{4}}}{2^{4}}$$

$$a \cdot \frac{1}{19} \approx \left(a \cdot 2938661835 \cdot 2^{-32} + \left(a - a \cdot 2938661835 \cdot 2^{-32}\right) \cdot 2^{-1}\right) \cdot 2^{-4}$$

$$a \cdot \frac{1}{19} \approx a \cdot \frac{7233629131}{137438953472}$$

- Constant Folding
- Common Sub-expression Elimination
- Dead Code
- Strength Reduction
- Code Motion
- Tail Recursion
- Loop Unrolling

Code Motion

Code motion moves code outside of a loop if possible.

```
for (int i = 0; i < n; i++) {
   sum += arr[i] + foo * (bar + 3);
}</pre>
```

Common subexpression elimination deals with expressions that appear multiple times in the code. Here, the expression appears once, but is calculated each loop iteration.

- Constant Folding
- Common Sub-expression Elimination
- Dead Code
- Strength Reduction
- Code Motion
- Tail Recursion
- Loop Unrolling

Tail Recursion

Tail recursion is an example of where GCC can identify recursive patterns that can be more efficiently implemented iteratively.

```
long factorial(int n) {
   if (n <= 1) {
      return 1;
   }
   else return n * factorial(n - 1);
}</pre>
```

Tail Recursion

Tail recursion: When a recursive call is made as the <u>final</u> action of a recursive function.

```
long factorial(int n) {
   if (n <= 1) {
      return 1;
   }
   else return n * factorial(n - 1);
}</pre>
```

Tail-recursive factorial

```
// returns n!, or 1 * 2 * 3 * 4 * ... * n.
long factorial(int n, long accum = 1) {
   if (n <= 1) {
      return accum;
   }
   else return factorial(n - 1, accum * n);
}</pre>
```

 Tail recursive solutions often end up passing partial computations as parameters that would otherwise be computed after the recursive call

Non-recursive factorial

```
// returns n!, or 1 * 2 * 3 * 4 * ... * n.
long factorial(int n) {
   long accum = 1;
   for (int i = 1; i <= n; i++) {
       accum *= i;
   }
   return accum;
}</pre>
```

- Sometimes looking at the non-recursive version of a function can help you find the tail recursive solution
 - Often looks more like the non-recursive version, with a variable or parameter keeping track of partial computations
 - Loop is replaced by a recursive call

- Constant Folding
- Common Sub-expression Elimination
- Dead Code
- Strength Reduction
- Code Motion
- Tail Recursion
- Loop Unrolling

Loop Unrolling

Loop Unrolling: Do **n** loop iterations' worth of work per actual loop iteration, so we save ourselves from doing the loop overhead (test and jump) every time, and instead incur overhead only every n-th time.

```
for (int i = 0; i <= n - 4; i += 4) {
    sum += arr[i];
    sum += arr[i + 1];
    sum += arr[i + 2];
    sum += arr[i + 3];
} // after the loop handle any leftovers</pre>
```

Limitations of GCC Optimization

GCC can't optimize everything! You ultimately may know more than GCC does.

```
int char_sum(char *s) {
    int sum = 0;
    for (size_t i = 0; i < strlen(s); i++) {
        sum += s[i];
    }
    return sum;
}</pre>
```

What is the bottleneck? **strlen called for every character**What can GCC do? **code motion - pull strlen out of loop**

Limitations of GCC Optimization

GCC can't optimize everything! You ultimately may know more than GCC does.

```
void lower1(char *s) {
    for (size_t i = 0; i < strlen(s); i++) {
        if (s[i] >= 'A' && s[i] <= 'Z') {
            s[i] -= ('A' - 'a');
        }
    }
}</pre>
```

What is the bottleneck? What can GCC do?

strlen called for every character nothing! s is changing, so GCC doesn't know if length is constant across iterations. But we know its length doesn't change.

Optimizing Your Code

- Explore various optimizations you can make to your code to reduce instruction count and runtime.
 - More efficient Big-O for your algorithms
 - Explore other ways to reduce instruction count
 - Look for hotspots using callgrind
 - Optimize using –02
 - And more...

Optimizing Your Code

- Explore various optimizations you can make to your code to reduce instruction count and runtime.
 - More efficient Big-O for your algorithms
 - Explore other ways to reduce instruction count
 - Look for hotspots using callgrind
 - Optimize using –02
 - And more...

Compiler Optimizations

Why not always just compile with -02?

- Difficult to debug optimized executables only optimize when complete
- Optimizations may not always improve your program. The compiler does its best, but may not work, or slow things down, etc. Experiment to see what works best!

Why should we bother saving repeated calculations in variables if the compiler has common subexpression elimination?

• The compiler may not always be able to optimize every instance. Plus, it can help reduce redundancy!

Course Evaluations

- I hope you can take the time to fill out the end-semester COMP201 course evaluation.
- I sincerely appreciate any feedback you have about the course and read every piece of feedback we receive.
- I am always looking for ways to improve!

Q&A

