Loops are Key

- Loops are extremely important
  - the "90-10" rule
- Loop optimization involves
  - understanding control-flow structure
  - Understanding data-dependence information
  - sensitivity to side-effecting operations
  - extra care in some transformations such as register spilling

Common loop optimizations

- Hoisting of loop-invariant computations
  - pre-compute before entering the loop
- Elimination of induction variables
  - change p=i*w+b to p=b, p+=w, when w, b invariant
- Loop unrolling
  - to improve scheduling of the loop body
- Software pipelining
  - To improve scheduling of the loop body
- Loop permutation
  - to improve cache memory performance

Finding Loops

- To optimize loops, we need to find them!
- Could use source language loop information in the abstract syntax tree...
- BUT:
  - There are multiple source loop constructs: for, while, do-while, even goto in C
  - Want IR to support different languages
  - Ideally, we want a single concept of a loop so all have same analysis, same optimizations
- Solution: dismantle source-level constructs, then re-find loops from fundamentals
Finding Loops

• To optimize loops, we need to find them!

• Specifically:
  - loop-header node(s)
    - nodes in a loop that have immediate predecessors not in the loop
  - back edge(s)
    - control-flow edges to previously executed nodes
  - all nodes in the loop body

Control-flow analysis

• Many languages have goto and other complex control, so loops can be hard to find in general

• Determining the control structure of a program is called control-flow analysis

• Based on the notion of dominators

Dominators

• \( a \text{ dom } b \)
  - node \( a \) dominates \( b \) if every possible execution path from entry to \( b \) includes \( a \)

• \( a \text{ sdom } b \)
  - \( a \) strictly dominates \( b \) if \( a \) dom \( b \) and \( a \neq b \)

• \( a \text{ idom } b \)
  - \( a \) immediately dominates \( b \) if \( a \) sdom \( b \), AND there is no \( c \) such that \( a \) sdom \( c \) and \( c \) sdom \( b \)

Back edges and loop headers

• A control-flow edge from node B3 to B2 is a back edge if B2 dom B3

• Furthermore, in that case node B2 is a loop header

B3

\[ j = j + 2 \]
\[ k = \text{true} \]
\[ i = i + 1 \]

B2

\[ i \leq n \]

Entry

B1

\[ k = \text{false} \]
\[ i = 1 \]
\[ j = 2 \]

..k..
Natural loop

- Consider a back edge from node n to node h
- The natural loop of \( n \rightarrow h \) is the set of nodes \( L \) such that for all \( x \in L \):
  - \( h \) dom \( x \) and
  - there is a path from \( x \) to \( n \) not containing \( h \)

Examples

Simple example:

(often it's more complicated, since a FOR loop found in the source code might need an if/then guard)

Examples

Try this:

```
for (...) {
  if {
    ...
  } else {
    ...
    if (x) {
      e;
      break;
    }
  }
}
```
Examples

for (...) {
    if {
        ...
    } else {
        ...
    }
    if (x) {
        e;
        break;
    }
}

Examples

• Yes, it can happen in C

Examples

lexically, in loop, but not in natural loop

Natural Loops

One loop per header..

What are the natural loops?

- (1, 2, 3)
- (0, 1, 2)
- {}
Nested Loops

- Unless two natural loops have the same header, they are either disjoint or nested within each other.
- If A and B are loops (sets of blocks) with headers a and b such that $a \neq b$ and $b \in A$
  - $B \subseteq A$
  - Loop B is nested within A
- Loop B is the inner loop
- Can compute the loop-nest tree

General Loops

- A more general looping structure is a **strongly connected component** of the control flow graph
  - subgraph $<N_{scc}, E_{scc}>$ such that
    - every block in $N_{scc}$ is reachable from every other node using only edges in $E_{scc}$

Reducible Flow Graphs

There is a special class of flow graphs, called reducible flow graphs, for which several code-optimizations are especially easy to perform.

In reducible flow graphs loops are unambiguously defined and dominators can be efficiently computed.

Reducible flow graphs

**Definition:** A flow graph $G$ is reducible iff we can partition the edges into two disjoint groups, forward edges and back edges, with the following two properties.

1. The forward edges form an acyclic graph in which every node can be reached from the initial node of $G$.
2. The back edges consist only of edges whose heads dominate their tails.

This flow graph has no back edges. Thus, it would be reducible if the entire graph were acyclic, which is not the case.
**Alternative definition**

- **Definition**: A flow graph $G$ is reducible if we can repeatedly collapse (reduce) together blocks ($x, y$) where $x$ is the only predecessor of $y$ (ignoring self loops) until we are left with a single node.

**Properties of Reducible Flow Graphs**

- In a reducible flow graph, **all loops are natural loops**.
- Can use DFS to find loops.
- Many analyses are more efficient
  - polynomial versus exponential

**Good News**

- Most flow graphs are reducible.
- Languages can prohibit irreducibility
  - goto free C
  - Java
- Programmers usually don’t use such constructs even if they’re available
  - >90% of old Fortran code reducible

**Dealing with Irreducibility**

- Don’t
- Can split nodes and duplicate code to get reducible graph
  - possible exponential blowup
- Other techniques...
Loop optimizations: Hoisting of loop-invariant computations

Loop-invariant computations

• A definition
  \[ t = x \text{ op } y \]
in a loop is (conservatively) loop-invariant if
  - \( x \) and \( y \) are constants, or
  - all reaching definitions of \( x \) and \( y \) are outside the loop, or
  - only one definition reaches \( x \) (or \( y \)), and that definition is loop-invariant
    • so keep marking iteratively

• Be careful:

  \[ t = \text{expr}; \]
  \[
  \text{for} () \{ \\
    s = t * 2; \\
    t = \text{loop\_invariant\_expr}; \\
    x = t + 2; \\
    …
  \}
  \]

  • Even though \( t \)'s two reaching expressions are each invariant, \( s \) is not invariant…

Hoisting

• In order to “hoist” a loop-invariant computation out of a loop, we need a place to put it

  • We could copy it to all immediate predecessors (except along the back-edge) of the loop header...

  • …But we can avoid code duplication by inserting a new block, called the pre-header
Hoisting conditions

- For a loop-invariant definition
  \[ d: t = x \text{ op } y \]
- we can hoist \( d \) into the loop's pre-header only if
  1. \( d \)'s block dominates all loop exits at which \( t \) is live-out, and
  2. \( d \) is only the only definition of \( t \) in the loop, and
  3. \( t \) is not live-out of the pre-header

We need to be careful...

- All hoisting conditions must be satisfied!

\[
\begin{align*}
\text{L0:} & \quad t = 0 \\
\text{L1:} & \quad i = i + 1 \\
& \quad t = a \ast b \\
& \quad M[i] = t \\
& \quad \text{if } i < N \text{ goto L1} \\
\text{L2:} & \quad x = t
\end{align*}
\]

\[
\begin{align*}
\text{L0:} & \quad t = 0 \\
\text{L1:} & \quad i = i + 1 \\
& \quad t = a \ast b \\
& \quad M[i] = t \\
& \quad \text{if } i < N \text{ goto L1} \\
\text{L2:} & \quad x = t
\end{align*}
\]

OK

\[
\begin{align*}
\text{L0:} & \quad t = 0 \\
\text{L1:} & \quad i = i + 1 \\
& \quad t = a \ast b \\
& \quad M[i] = t \\
& \quad \text{if } i < N \text{ goto L1} \\
\text{L2:} & \quad x = t
\end{align*}
\]

violates 1,3

violates 2
We need to be careful...

- All hoisting conditions must be satisfied!

```
L0:  t = 0
L1:  if i>=N goto L2
     i = i+1
     t = a * b
     M[i] = t
     if i<N goto L1
L2:  x = t
```

OK violations 1,3

```
L0:  t = 0
L1:  if i>=N goto L2
     i = i+1
     t = a * b
     M[i] = t
     goto L1
L2:  x = t
```

violates 2

Loop optimizations:
Induction-variable Strength reduction

The basic idea of IVE

- Suppose we have a loop variable
  - i initially 0; each iteration i = i + 1

- and a variable that linearly depends on it:
  \[ x = i \cdot c1 + c2 \]

- In such cases, we can try to
  - initialize \( x = i_0 \cdot c1 + c2 \) (execute once)
  - increment \( x \) by \( c1 \) each iteration

Simple Example of IVE

```
for i = 0 to n
    a[i] = 0
H:
    if i >= n goto exit
    j <- i \cdot 4
    k <- j + a
    M[k] <- 0
    i <- i + 1
    goto H
```

Clearly, \( j \) & \( k \) do not need to be computed anew each time since they are related to \( i \) and \( i \) changes linearly.
Simple Example of IVE

\[ i \leftarrow 0 \]
\[ j' \leftarrow 0 \]
\[ k' \leftarrow a \]

H:
\[ \text{if } i \geq n \text{ goto exit} \]
\[ j \leftarrow i \times 4 \]
\[ k \leftarrow j + a \]
\[ M[k] \leftarrow 0 \]
\[ i \leftarrow i + 1 \]
\[ j' \leftarrow j' + 4 \]
\[ k' \leftarrow k' + 4 \]
\[ \text{goto H} \]

But, then we don't even need \( j \) (or \( j' \))

Simple Example of IVE

Rewrite comparison

\[ i \leftarrow 0 \]
\[ k' \leftarrow a \]

H:
\[ \text{if } i \geq n \text{ goto exit} \]
\[ k \leftarrow k' \]
\[ M[k] \leftarrow 0 \]
\[ i \leftarrow i + 1 \]
\[ j' \leftarrow j' + 4 \]
\[ k' \leftarrow k' + 4 \]
\[ \text{goto H} \]

But, \( a + (n \times 4) \) is loop invariant

Simple Example of IVE

Invariant code motion on \( a + (n \times 4) \)

\[ i \leftarrow 0 \]
\[ k' \leftarrow a \]

H:
\[ \text{if } k' \geq a + (n \times 4) \text{ goto exit} \]
\[ k \leftarrow k' \]
\[ M[k] \leftarrow 0 \]
\[ i \leftarrow i + 1 \]
\[ j' \leftarrow j' + 4 \]
\[ k' \leftarrow k' + 4 \]
\[ \text{goto H} \]

now, we do copy propagation and eliminate \( k \)
Simple Example of IVE

Copy propagation

\[ k' \leftarrow a \]
\[ n' \leftarrow a + (n \times 4) \]

H:

\[ \text{if } k' \geq n' \text{ goto exit} \]
\[ k \leftarrow k' \]
\[ M[k] \leftarrow 0 \]
\[ k' \leftarrow k' + 4 \]
\[ \text{goto H} \]

Voila!

What we did

- identified induction variables \((i,j,k)\)
- strength reduction (changed \(*\) into \(+\))
- dead-code elimination \((j \leftarrow j')\)
- useless-variable elimination \((j' \leftarrow j' + 4)\)
  (This can also be done with ADCE)
- loop invariant identification & code-motion
- almost useless-variable elimination \((i)\)
- copy propagation

Simple Example of IVE

Compare original and result of IVE

\[
\begin{align*}
&k' \leftarrow a \\
n' \leftarrow a + (n \times 4) \\
&H:\ \\
&\text{if } k' \geq n' \text{ goto exit} \\
&j \leftarrow i \times 4 \\
&k \leftarrow j + a \\
&M[k'] \leftarrow 0 \\
&k' \leftarrow k' + 4 \\
&\text{goto H} \\
\end{align*}
\]

Voila!

Is it faster?

- On some hardware, adds are much faster than multiplies
- Furthermore, one fewer value is computed,
  - thus potentially saving a register
  - and decreasing the possibility of spilling
Loop preparation

• Before attempting IVE, it is best to first perform:
  - constant propagation & constant folding
  - copy propagation
  - loop-invariant hoisting

How to do it, step 1

• First, find the basic IVs
  - scan loop body for defs of the form
    \[ x = x + c \text{ or } x = x - c \]
    where \( c \) is loop-invariant
  - record these basic IVs as
    \[ x = (x, 1, c) \]
    - this represents the IV: \( x = x \times 1 + c \)

Representing IVs

• Characterize all induction variables by:
  \((\text{base-variable, offset, multiple})\)
  - where the offset and multiple are loop-invariant
• IOW, after an induction variable is defined it equals:
  \[ \text{offset} + \text{multiple} \times \text{base-variable} \]

How to do it, step 2

• Scan for derived IVs of the form
  \[ k = i \times c_1 + c_2 \]
  - where \( i \) is a basic IV,
    this is the only def of \( k \) in the loop, and
    \( c_1 \) and \( c_2 \) are loop invariant
• We say \( k \) is in the family of \( i \)
• Record as \( k = (i, c_1, c_2) \)
How to do it, step 3

• Iterate, looking for derived IVs of the form
  \[ k = j \cdot c_1 + c_2 \]
  - where IV \( j = (i, a, b) \), and
  - this is the only def of \( k \) in the loop, and
  - there is no def of \( i \) between the def of \( j \) and
    the def of \( k \)
  - \( c_1 \) and \( c_2 \) are loop invariant
• Record as \( k = (i, a \cdot c_1, b \cdot c_1 + c_2) \)

Finding the IVs

• Maintain three tables: basic & maybe & other
• Find basic Ivvs:
  - Scan stmts. If var \( \not\in \) maybe, and of proper form, put into basic. Otherwise, put var in other and remove from maybe.
• Find compound Ivvs:
  - If var defined more than once, put into other
  - For all stmts of proper form that use a basic IV

Simple Example of IVE

\[
\begin{align*}
i & \leftarrow 0 \\
H: \quad & \text{if } i \geq n \text{ goto exit} \\
j & \leftarrow i \cdot 4 \\
k & \leftarrow j + a \\
M[k] & \leftarrow 0 \\
i & \leftarrow i + 1 \\
\text{goto } H \\
i: \quad (i, 1, 1) \text{ i.e., } i = 1 + 1 \cdot i \\
j: \quad (i, 0, 4) \text{ i.e., } j = 0 + 4 \cdot i \\
k: \quad (i, a, 4) \text{ i.e., } k = a + 4 \cdot i
\end{align*}
\]

So, \( j \) & \( k \) are in family of \( i \)

How to do it, step 4

• This is the strength reduction step
• For an induction variable \( k = (i, c_1, c_2) \)
  - initialize \( k = i \cdot c_1 + c_2 \) in the preheader
  - replace \( k \)'s def in the loop by
    \[ k = k + c_1 \]
  - make sure to do this after \( i \)'s def

» FIX THIS SLIDE
How to do it, step 5

- This is the comparison rewriting step

- For an induction variable $k = (i, a_k, b_k)$
  - If $k$ used only in definition and comparison
    - There exists another variable, $j$, in the same class and is not "useless" and $j = (i, a_j, b_j)$
  - Rewrite $k < n$ as
    $j < (b_j/b_k)(n-a_k)+a_j$

- Note: since they are in same class:
  $(j-a_j)/b_j = (k-a_k)/b_k$

Notes

- Are the $c_1, c_2$ constant, or just invariant?
  - if constant, then you can keep folding them:
    they’re always a constant even for derived IVs
  - otherwise, they can be expressions of loop-invariant variables

- But if constant, can find IVs of the type
  $x = i/b$
  and know that it’s legal, if $b$ evenly divides the stride...

Is it faster? (2)

- On some hardware, adds are much faster than multiplies
  - But...not always a win!
    - Constant multiplies might otherwise be reduced to shifts/adds that result in even better code than IVE
    - Scaling of addresses ($i*4$) might come for free on your processor’s address modes
  - So maybe: only convert $i*c_1+c_2$ when $c_1$ is loop invariant but not a constant

Common loop optimizations

- Hoisting of loop-invariant computations
  - pre-compute before entering the loop
- Elimination of induction variables
  - change $p=i*w+b$ to $p=b, p+=w$, when $w, b$ invariant
- Loop unrolling
  - to improve scheduling of the loop body
- Software pipelining
  - To improve scheduling of the loop body
- Loop permutation
  - to improve cache memory performance
Dependencies in Loops

- Loop independent data dependence occurs between accesses in the same loop iteration.
- Loop-carried data dependence occurs between accesses across different loop iterations.
- There is data dependence between access a at iteration i-k and access b at iteration i when:
  - a and b access the same memory location
  - There is a path from a to b
  - Either a or b is a write

Defining Dependencies

- Flow Dependence $W \Rightarrow R \delta_f$
- Anti-Dependence $R \Rightarrow W \delta_a$
- Output Dependence $W \Rightarrow W \delta_o$

Example Dependencies

S1) $a=0$;
S2) $b=a$;
S3) $c=a+d+e$;
S4) $d=b$;
S5) $b=5+e$;

These are scalar dependencies. The same idea holds for memory accesses.

Data Dependence in Loops

- Dependence can flow across iterations of the loop.
- Dependence information is annotated with iteration information.
- If dependence is across iterations it is loop carried otherwise loop independent.

```
for (i=0; i<n; i++) {
    A[i] = B[i];
    B[i+1] = A[i];
}
```
Data Dependence in Loops

- Dependence can flow across iterations of the loop.
- Dependence information is annotated with iteration information.
- If dependence is across iterations it is loop carried otherwise loop independent.

\[\text{for (i=0; i<n; i++)} \]
\[\text{A}[i] = B[i]; \]
\[\text{B}[i+1] = A[i]; \]

\(\delta\) loop carried

Distance/Direction of the dependence is also important.

Iteration Space

Every iteration generates a point in an n-dimensional space, where n is the depth of the loop nest.

\[\text{for (i=0; i<n; i++)} \]
\[\text{for (j=0; j<4; j++)} \]

Distance Vector

\[\text{for (i=0; i<n; i++)} \]
\[\text{for (j=0; j<4; j++)} \]

Distance vector is the difference between the target and source iterations.

\[d = l_i - l_{i-1}\]

Exactly the distance of the dependence, i.e.,

\[l_i + d = l_i\]
Example of Distance Vectors

for (i=0; i<n; i++)
    for (j=0; j<m; j++){
        B[i,j+1] = ;  // B[i,j+1] = value
        C[i+1,j] = ;  // C[i+1,j] = value
    }

Data Dependences

Loop carried: between two statements instances in two different iterations of a loop.
Loop independent: between two statements instances in the same loop iteration.

Lexically forward: the source comes before the target. Lexically backward: otherwise.

The right-hand side of an assignment is considered to precede the left-hand side.

Review of Linear Algebra

Lexicographic Order

Two n-vectors $\mathbf{a}$ and $\mathbf{b}$ are equal, $\mathbf{a} = \mathbf{b}$, if $a_i = b_i, 1 \leq i \leq n$.

We say that $\mathbf{a}$ is less than $\mathbf{b}$, $\mathbf{a} < \mathbf{b}$, if $a_i < b_i, 1 \leq i \leq n$.

We say that $\mathbf{a}$ is lexicographically less than $\mathbf{b}$, at level $j$, $\mathbf{a} \prec_j \mathbf{b}$, if $a_i = b_i, 1 \leq i < j$ and $a_j < b_j$.

We say that $\mathbf{a}$ is lexicographically less than $\mathbf{b}$, $\mathbf{a} \prec \mathbf{b}$, if there is a $j, 1 \leq j \leq n$, such that $\mathbf{a} \prec_j \mathbf{b}$. 

```text
A yields: [0, 0]
B yields: [0, 1]
C yields: [1, -1]
```
Lexicographic Order
Example of vectors

Consider the vectors \( \mathbf{a} \) and \( \mathbf{b} \) below:

\[
\mathbf{a} = \begin{bmatrix}
1 \\
-1 \\
0 \\
2
\end{bmatrix} \quad \mathbf{b} = \begin{bmatrix}
1 \\
-1 \\
1 \\
-1
\end{bmatrix}
\]

We say that \( \mathbf{a} \) is lexicographically less than \( \mathbf{b} \) at level 3, \( \mathbf{a} \prec \mathbf{b} \), or simply that \( \mathbf{a} \prec \mathbf{b} \).

Both \( \mathbf{a} \) and \( \mathbf{b} \) are lexicographically positive because \( 0 \prec \mathbf{a} \), and \( 0 \prec \mathbf{b} \).

Properties of Lexicographic Order

Let \( n \geq 1 \), and \( i, j, k \) denote arbitrary vectors in \( \mathbb{R}^n \).

1. For each \( u \) in \( 1 \leq u \leq n \), the relation \( \prec_u \) in \( \mathbb{R}^n \) is irreflexive and transitive.

2. The \( n \) relations \( \prec_u \) are pairwise disjoint: 
   \( i \prec_j j \) and \( i \prec_v j \) imply that \( u = v \).

3. If \( i \neq j \), there is a unique integer \( u \) such that \( 1 \leq u \leq n \) and exactly one of the following two conditions holds:
   \( i \prec_j j \) or \( j \prec_i i \).

4. \( i \prec_j j \) and \( j \prec_v k \) together imply that \( i \prec_w k \), where \( w = \min(u,v) \).

Data Dependence in Loops

An Example

Find the dependence relations due to the array \( X \) in the program below:

\[
\begin{align*}
(S_1) & \quad \text{for } i = 2 \text{ to } 9 \text{ do} \\
(S_2) & \quad X[i] = Y[i] + Z[i] \\
(S_3) & \quad A[i] = X[i-1] + 1 \\
(S_4) & \quad \text{end for}
\end{align*}
\]

Solution

To find the data dependence relations in a simple loop, we can unroll the loop and see which statement instances depend on which others:

\[
\begin{align*}
i = 2 && i = 3 && i = 4 \\
\end{align*}
\]

There is a loop-carried, lexically forward, flow dependence from \( S_2 \) to \( S_3 \).