• Assume we have permanently allocated each IR Name and Temp to a register, a stack slot, or (for free variable names) a closure slot.

• So when generating code for an IR operation, need to be able to cope with either register or memory operand in each position.

• X86 instructions are much more constrained:

  mov $s,$d $d$ can be reg or mem; $s$ can be reg, imm, or mem (but not if $d$ is)
  add $s,$d $d := d + s$; $d$ can be reg or mem, $s$ can be reg, imm, or mem (but not if $d$ is)
  sub $s,$d $d := d - s$; $d$ can be reg or mem, $s$ can be reg, imm, or mem (but not if $d$ is)
  imul $s,$d $d := d \times s$; $d$ must be reg, $s$ can be reg, imm, or mem
  idiv $s$ $%rax := %rdx:\%rax \text{ div } s$ and $%rdx := %rdx:\%rax \mod s$;
      $s$ must be reg or mem
  cmp $s_1,$s_2 test sense is “backwards”;
      $s_1$ can be reg, imm, or mem; $s_2$ can be reg or mem (but not if $s_1$ is)

• Can use these utility functions:

  X86.Operand gen_source_operand (IR.Operand rand, int size, boolean mem_ok, boolean imm_ok, X86.Reg temp)
  X86.Operand gen_target_operand (IR.Operand rand, int size, X86.Reg temp)
• Remember to keep track of sizes:

(0) IR.BOOL = X86.B = 1 byte (registers %al, etc.)
(1) IR.INT = X86.L = 4 bytes (registers %eax, etc.)
(2) IR.PTR = X86.Q = 8 bytes (registers %rax, etc.)

Register names must match instruction suffixes, or you get an assembler error. Immediates and memory operands are sized automatically.

• One pattern for generating 2-addr code from 3-addr code:

\[
\begin{align*}
\text{IR.sub } a, b, c & \quad \text{X86.mov } a, c \quad \text{can omit if } a = c \\
& \quad \text{X86.sub } b, c \\
\end{align*}
\]

But what if \( b = c \) ?!

\[
\begin{align*}
\text{IR.sub } a, b, b & \quad \text{X86.mov } a, b \quad \text{can omit if } a = b \\
& \quad \text{X86.sub } b, b \quad \text{Oops!}
\end{align*}
\]

Must be smarter in this case, or dumber in the regular case.

Hint: Use %r10 and %r11 as temporaries within IR instructions.

• There are many possible improvements, especially when dealing with constants or commutative operators.
Choice of evaluation order when linearizing expression trees can affect number of registers needed.

Example: Assume a RISC-like load-store instruction set. If we compute left child first, need 4 regs, but doing right child first needs only 3.

\[(a+b) - ((c+d) - (e+f))\]

- load \(a, r_1\) load \(c, r_1\)
  \[
  \begin{array}{c}
  / \\
  \end{array}
  \]

- load \(b, r_2\) load \(d, r_2\)
  \[
  \begin{array}{c}
  / \\
  \end{array}
  \]

- add \(r_1, r_2, r_1\) add \(r_1, r_2, r_1\)
  \[
  \begin{array}{c}
  + \\
  \end{array}
  \]

- load \(c, r_2\) load \(e, r_2\)
  \[
  \begin{array}{c}
  / \\
  \end{array}
  \]

- load \(d, r_3\) load \(f, r_3\)
  \[
  \begin{array}{c}
  / \\
  \end{array}
  \]

- add \(r_2, r_3, r_2\) add \(r_2, r_3, r_2\)
  \[
  \begin{array}{c}
  a \\
  \end{array}
  \]

- load \(e, r_3\) sub \(r_1, r_2, r_1\)
  \[
  \begin{array}{c}
  + \\
  \end{array}
  \]

- load \(f, r_4\) load \(a, r_2\)
  \[
  \begin{array}{c}
  / \\
  \end{array}
  \]

- load \(r_3, r_4, r_3\) load \(b, r_3\)
  \[
  \begin{array}{c}
  c \\
  \end{array}
  \]

- sub \(r_2, r_3, r_2\) add \(r_2, r_3, r_2\)
  \[
  \begin{array}{c}
  d \\
  \end{array}
  \]

- sub \(r_1, r_2, r_1\) sub \(r_2, r_1, r_1\)
  \[
  \begin{array}{c}
  e \\
  \end{array}
  \]

- f
Key idea (Sethi & Ullman): At each node, first evaluate subtree requiring largest number of registers to evaluate. Can then save result of this evaluation in a register while doing other subtree.

1. Label each node with minimum number of registers needed to evaluate subtree.

   ```plaintext
   risc_label(t)
   if isLeaf(t) then
       t->label = 1  (*depends on machine architecture*)
   else
       label(t->left)
       label(t->right)
       if (t->left->label == t->right->label)
           t->label = t->left->label + 1
       else
           t->label = max(t->left->label, t->right->label)
   ```

2. Use labels to guide order of code emission; emit code for higher-numbered subtree first.
Some machines allow one operand to be a complex expression, while the other must be a register, which also holds the result (“accumulator” style):

```
add 37, r0 ; r0 <- r0 + 37
add [b], r1 ; r1 <- r1 + *b
sub r0, r1 ; r1 <- r1 - r0
```

These machines have different Sethi-Ullman numbering, e.g., right leaves might require no temporary registers at all.

Can use associativity to make trees “less bushy,” e.g.

```
+3
/ \  / \   \
/ \ +2 becomes +2  d
/ \  / \   / \  
+2  +2 c  +2  c
/ \ / \   / \ 
 a  b  c  d  a  b
```
• Extend analysis of register use to program units larger than expressions but still completely analyzable at compile time.

• **Basic Block** = sequence of instructions with single entry & exit.

• If first instruction of BB is executed, so is remainder of block (in order).

• To calculate basic blocks:

  (1) Determine BB leaders ($\rightarrow$):

  (a) First statement in routine
  (b) Target of any jump (conditional or unconditional).
  (c) Statement following any jump.

  (What about subroutine calls?)

  (2) Basic block extends from leader to (but not including) next leader (or end of routine).
prod := 0;
i := 1;
while i <= 20 do
    prod := prod + a[i] * b[i];
i := i + 1
end

→ 1. prod := 0
   2. i := 1
   → 3. if i > 20 goto 14
   → 4. t1 := i * 4
       5. t2 := addr a
       6. t3 := *(t2+t1)
       7. t4 := i * 4
       8. t5 := addr b
       9. t6 := *(t5+t4)
      10. t7 := t3 * t6
      11. prod := prod + t7
      12. i := i + 1
      13. goto 3
→ 14. ---
Can combine code generation with “greedy” register allocation: bring each variable into a register when first needed, and leave it there as long as it’s needed (if possible).

Maintain register descriptors saying which variable is in each register, and address descriptors saying where (in memory and/or a register) each variable is.

For each IR instruction $x := y \text{ op } z$:
1. If $y$ isn’t in a register, load it into a free one, updating descriptors.
2. Similarly for $z$.
3. If $y$ and/or $z$ are no longer live following this instruction, mark their registers as free.
4. Choose a free register for $x$, updating descriptors.
5. Generate instruction $\text{ op } r_y, r_z, r_x$.

For the special case $x := y$, load $y$ into a register, if necessary, and then mark that register as holding $x$ too.
BB Code Generation using Liveness (more)

- Must now be careful not to free a register unless none of its associated variables is live.
- Registers behave like a cache for memory locations.
- Nasty problems for source-level debuggers and dump utilities – where is that variable?!?

Example (assuming risc-like instruction set)...

Source code:
\[ d := (a-b) + (a-c) + (a-c) \]

Live after inst:
\[ a \quad b \quad c \]

IR:
\[
\begin{align*}
  t &:= a - b \quad a \quad c \quad t \\
  u &:= a - c \quad t \quad u \\
  v &:= t + u \quad u \quad v \\
  d &:= v + u \quad d
\end{align*}
\]

<table>
<thead>
<tr>
<th>IR Statement</th>
<th>Code</th>
<th>Regs</th>
<th>Addr</th>
</tr>
</thead>
<tbody>
<tr>
<td>t := a - b</td>
<td>ld a,r0</td>
<td>r0:a</td>
<td>a:r0,mem</td>
</tr>
<tr>
<td></td>
<td>ld b,r1</td>
<td>r1:t</td>
<td>b,c:mem</td>
</tr>
<tr>
<td></td>
<td>sub r0,r1,r1</td>
<td>t:r1</td>
<td></td>
</tr>
<tr>
<td>u := a - c</td>
<td>ld c,r2</td>
<td>r0:u</td>
<td>a,b,c:mem</td>
</tr>
<tr>
<td></td>
<td>sub r0,r2,r0</td>
<td>r1:t</td>
<td>u:r0</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>t:r1</td>
</tr>
<tr>
<td>v := t + u</td>
<td>add r1,r0,r1</td>
<td>r0:u</td>
<td>a,b,c:mem</td>
</tr>
<tr>
<td></td>
<td></td>
<td>r1:v</td>
<td>u:r0</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>v:r1</td>
</tr>
<tr>
<td>d := v + u</td>
<td>add r0,r1,r0</td>
<td>r0:d</td>
<td>a,b,c:mem</td>
</tr>
<tr>
<td></td>
<td>st r0,d</td>
<td></td>
<td>a,b,c:mem</td>
</tr>
<tr>
<td></td>
<td>st r1:v</td>
<td></td>
<td>d:r0,mem</td>
</tr>
</tbody>
</table>
CONTROL-FLOW GRAPHS

To assign registers on a per-procedure basis, need to perform liveness analysis on entire procedure, not just basic blocks.

To analyze the properties of entire procedures with multiple basic blocks, we use a control-flow graph.

In simplest form, control flow graph has one node per statement, and an edge from $n_1$ to $n_2$ if control can ever flow directly from statement 1 to statement 2.

We write $\text{pred}[n]$ for the set of predecessors of node $n$, and $\text{succ}[n]$ for the set of successors.

(In practice, usually build control-flow graphs where each node is a basic block, rather than a single statement.)

Example....
a = 0
L: b = a + 1
    c = c + b
    a = b * 2
if a < N goto L
return c
LIVENESS ANALYSIS USING DATAFLOW ANALYSIS

Working from the future to the past, we can determine the edges over which each variable is live.

In the example:

b is live on 2 → 3 and on 3 → 4.

a is live from on 1 → 2, on 4 → 5, and on 5 → 2 (but not on 2 → 3 → 4).

c is live throughout (including on entry → 1).

Can see that two registers suffice to hold a, b, c.
We can do liveness analysis (and many other analyses) via dataflow analysis.

A node defines a variable if its corresponding statement assigns to it.

A node uses a variable if its corresponding statement mentions that variable in an expression (e.g., on the rhs of assignment).

For any variable $v$, define

- $\text{def}[v] = \text{set of graph nodes that define } v$
- $\text{use}[v] = \text{set of graph nodes that use } v$

Similarly, for any node $n$, define

- $\text{def}[n] = \text{set of variables defined by node } n$
- $\text{use}[v] = \text{set of variables used by node } n$.
A variable is **live** on an edge if there is a directed path from that edge to a **use** of the variable that does not go through any **def**.

A variable is **live-in** at a node if it is live on any in-edge of that node; it is **live-out** if it is live on any out-edge.

Then the following equations hold:

\[
in[n] = use[n] \cup (out[n] - def[n])
\]

\[
out[n] = \bigcup_{s \in succ[n]} in[s]
\]

We want the **least fixed point** of these equations: the smallest *in* and *out* sets such that the equations hold.
We can find this solution by iteration:

- Start with empty sets
- Use equations to add variables to sets, one node at a time.
- Repeat until sets don’t change any more.

Adding additional variables to the sets is safe, as long as the sets still obey the equations, but inaccurately suggests that more live variables exist than actually do.

Again, in practice we normally work on a CFG in which each node is an entire basic block. It is easy to compute the in and out sets of each block by a simple backward pass over its instructions.
For correctness, order in which we take nodes doesn’t matter, but it turns out to be fastest to take them in roughly reverse order:

<table>
<thead>
<tr>
<th>node</th>
<th>use</th>
<th>def</th>
<th>1st out</th>
<th>in</th>
<th>2nd out</th>
<th>in</th>
<th>3rd out</th>
<th>in</th>
</tr>
</thead>
<tbody>
<tr>
<td>6</td>
<td>c</td>
<td></td>
<td>c</td>
<td></td>
<td>c</td>
<td></td>
<td>c</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>a</td>
<td></td>
<td>c</td>
<td>ac</td>
<td>ac</td>
<td>ac</td>
<td>ac</td>
<td>ac</td>
</tr>
<tr>
<td>4</td>
<td>b</td>
<td>a</td>
<td>ac</td>
<td>bc</td>
<td>ac</td>
<td>bc</td>
<td>ac</td>
<td>bc</td>
</tr>
<tr>
<td>3</td>
<td>bc</td>
<td>c</td>
<td>bc</td>
<td>bc</td>
<td>bc</td>
<td>bc</td>
<td>bc</td>
<td>bc</td>
</tr>
<tr>
<td>2</td>
<td>a</td>
<td>b</td>
<td>bc</td>
<td>ac</td>
<td>bc</td>
<td>ac</td>
<td>bc</td>
<td>ac</td>
</tr>
<tr>
<td>1</td>
<td>a</td>
<td></td>
<td>ac</td>
<td>c</td>
<td>ac</td>
<td>c</td>
<td>ac</td>
<td>c</td>
</tr>
</tbody>
</table>

Implementation issues:

- Algorithm always terminates, because each iteration must enlarge at least one set, but sets are limited in size (by total number of variables).
- Time complexity is $O(N^4)$ worst-case, but between $O(N)$ and $O(N^2)$ in practice.
Consider the following graph:

```
1 V
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>a = b*b</td>
<td></td>
</tr>
<tr>
<td>---</td>
<td>---</td>
</tr>
</tbody>
</table>

2 V
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>c = a+b</td>
<td></td>
</tr>
<tr>
<td>---</td>
<td>---</td>
</tr>
</tbody>
</table>

3 V
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>c &gt;= b ?</td>
<td></td>
</tr>
<tr>
<td>---</td>
<td>---</td>
</tr>
</tbody>
</table>

4 V

\  
5 V

<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>return a</td>
<td></td>
</tr>
<tr>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>return c</td>
<td></td>
</tr>
</tbody>
</table>
```

Is a live-out at node 2? It depends on whether control flow ever reaches node 4.
A smart compiler could answer no.

A smarter compiler could answer similar questions about more complicated programs.

But no compiler can ever \textit{always} answer such questions correctly. This is a consequence of the \textit{uncomputability} of the \textit{Halting Problem}.

So we must be content with \textit{static} liveness, which talks about paths of control-flow edges, and is just a \textit{conservative} approximation of \textit{dynamic liveness}, which talks about actual execution paths.
Theorem  There is no program $H$ that takes an input any program $P$ and input $X$, and (without infinite-looping) returns true if $P(X)$ halts and false if $P(X)$ infinite-loops.

Proof  Suppose there were such an $H$. From it, construct the function $F(Y) = \begin{cases} \text{true} & \text{if } H(Y, Y) \\
\text{false} & \text{else}
\end{cases}$

Now consider $F(F)$.

- If $F(F)$ halts, then, by the definition of $H$, $H(F, F)$ is true, so the then clause executes, so $F(F)$ does not halt.
- But, if $F(F)$ loops forever, then $H(F, F)$ is false, so the else clause is taken, so $F(F)$ halts.

Hence $F(F)$ halts if and only if it doesn’t halt.

Since we’ve reached a contradiction, the initial assumption is wrong: there can be no such $H$.  

**REACHABILITY PROBLEM**

**Corollary** No program $H'(P, X, L)$ can tell, for any program $P$, input $X$, and label $L$ within $P$, whether $L$ is ever reached on an execution of $P$ on $X$.

**Proof** If we had $H'$, we could construct $H$. Consider a program transformation $T$ that, from any program $P$ constructs a new program by putting a label $L$ at the end of the program, and changing every `halt` to `goto L`. Then $H(P, X) = H'(T(P), X, L)$. 


Mixing instruction selection and register allocation gets confusing; need a more systematic way to look at the problem.

- Initially generate code assuming an infinite number of “logical” registers; calculate live ranges. E.g., for previous example:

\[
\begin{align*}
\text{ld a}, & \text{t0} ; \text{a:t0 t0} \\
\text{ld b}, & \text{t1} ; \text{b:t1 t0 t1} \\
\text{sub t0}, & \text{t1,t2} ; \text{t:t2 t0 t2} \\
\text{ld c}, & \text{t3} ; \text{c:t3 t0 t2 t3} \\
\text{sub t0}, & \text{t3,t4} ; \text{u:t4 t2 t4} \\
\text{add t2}, & \text{t4,t5} ; \text{v:t5 t4 t5} \\
\text{add t5}, & \text{t4,t6} ; \text{d:t6 t6} \\
\text{st t6}, & \text{d} \\
\end{align*}
\]

Live after instr.

- Build a **register interference graph**, which has
  - a node for each logical register.
  - an edge between two nodes if the corresponding registers are simultaneously live.
Interference Graph Example:

```
  -------------------
  /                   \
 t0 ----- t1  t2
 |   ______________//
 |  / ______/       |
 | /       /        |
 |/        / t3 t4----- t5 t6
```

A **coloring** of a graph is an assignment of colors to nodes such that no two connected nodes have the same color. (Like coloring a map, where nodes=countries and edges connect countries with common border.)

Suppose we have $k$ physical registers available. Then aim is to color interference graph with $k$ or fewer colors. This implies we can allocate logical registers to physical registers without spilling.

In general case, determining whether a graph can be $k$-colored is hard (N.P. Complete, and hence probably exponential).

But a simple heuristic will **usually** find a $k$-coloring if there is one.
1. Choose a node with fewer than \( k \) neighbors.

2. Remove that node. Note that if we can color the resulting graph with \( k \) colors, we can also color the original graph, by giving the deleted node a color different from all its neighbors.

3. Repeat until either

   - there are no nodes with fewer than \( k \) neighbors, in which case we must spill; or

   - the graph is gone, in which case we can color the original graph by adding the deleted nodes back in one at a time and coloring them.

In our example, heuristic finds a 3-coloring. There cannot be a 2-coloring (why not?).

Each “color” corresponds to a physical register, so 3 registers will do for this example.