Lecture 16: 代码生成及优化-I

Xiaoyuan Xie 谢晓园
xxie@whu.edu.cn
计算机学院E301

代码生成

代码生成器的位置

- **根据中间表示生成代码**
- **代码生成器之前可能有一个优化组件**
- **代码生成器的三个主要任务**
  - 指令选择：选择适当的指令实现IR语句
  - 寄存器分配和指派：把哪个值放在哪个寄存器中
  - 指令排序：按照什么顺序安排指令执行

图 8-1　代码生成器的位置

代码生成器中的问题

- **正确性、易于实现、测试和维护**
- **输入IR的选择**
  - 四元式、三元式、字节代码、堆栈机代码、后缀表示、抽象语法树、DAG、...
- **目标程序**
  - RISC、CISC、可重定向代码、汇编语言
- **指令选择**
  - 影响因素：IR层次、指令集特性、目标代码质量
- **寄存器分配和指派**
- **求值顺序**

指令选择

```
x = y + z;

LD  R0, y        // R0 = y          (load y into register R0)
ADD R0, R0, z    // R0 = R0 + z     (add z to R0)
ST  x, R0        // x = R0          (store R0 into x)
```

```
a = b + c
d = a + e
```

如果变量a不再被使用了呢?

```
LD  R0, b        // R0 = b
ADD R0, R0, c    // R0 = R0 + c
ST  a, R0        // a = R0
LD  R0, a        // R0 = a
ADD R0, R0, e    // R0 = R0 + e
ST  d, R0        // d = R0
```

冗余的指令

指令选择

得考虑
指令代价

```
a = a + 1
```

```
LD   R0, a          // R0 = a
ADD  R0, R0, #1     // R0 = R0 + 1
ST   a, R0          // a = R0
```

```
INC a
```

目标语言 - 一个三地址机器的模型

- **指令集**
  - 加载：LD dst, addr
    - 把地址addr中的内容加载到dst所指寄存器
  - 保存：ST x, r
    - 把寄存器r中的内容保存到x中
  - 计算：OP dst, src1, src2
    - 把src1和scr2中的值运算后将结果存放到dst中
  - 无条件跳转：BR L
    - 控制流转向标号L的指令
  - 条件跳转：Bcond r, L
    - 对r中的值进行测试，如果为真则转向L

目标语言 - 一个三地址机器的模型

- **寻址模式**
  - 变量x：指向分配x的内存位置
  - a(r)：地址是a的左值加上r中的值
  - constant(r)：寄存器中内容加上前面的常数即其地址
  - *r：寄存器r的内容为其地址
  - *constant(r)：r中内容加上常量所指地址中存放的值为其地址
  - 常量#constant

实例

- **x=y-z**
  - LD　　R1, y　　//R1=y
  - LD　R2, z　　//R2=x
  - SUB　R1, R1, R2　　//R1=R1-R2
  - ST　x, R1　　//x=R1
- **一些思考**
  - 如果y已经装载进某个寄存器
  - 如果z已经装载进某个寄存器
  - 如果x再也没有被使用

实例

- **b=a[i]**
  - LD R1, i //R1=i
  - MUL R1, R1, 8 //R1=R1*8
  - LD R2, a(R1) //R2=contents(a+contents(R1))
  - ST b, R2 //b = R2
- **a[j]=c**
  - LD R1, c //R1=c
  - LD R2, j //R2=j
  - MUL R2, R2, 8 //R2=R2*8
  - ST a(R2), R1 //contents(a+contents(R2))=R1

实例

```
x = *p
 LD   R1, p          // R1 = p
 LD   R2, 0(R1)      // R2 = contents(0 + contents(R1))
 ST   x, R2          // x = R2

*p = y
 LD   R1, p          // R1 = p
 LD   R2, y          // R2 = y
 ST   0(R1), R2      // contents(0 + contents(R1)) = R2

if x < y goto L
 LD   R1, x          // R1 = x
 LD   R2, y          // R2 = y
 SUB  R1, R1, R2     // R1 = R1 - R2
 BLTZ R1, M          // if R1 < 0 jump to M
```

程序和指令的代价

- **不同的目的有不同的度量**
  - 最短编译时间、目标程序大小、运行时间、能耗
- **假设：每个指令有固定的代价，设定为1加上运算分量寻址模式的代价**
  - LD R0, R1；代价为1
  - LD R0, M；代价是2
  - LD R1, *100(R2)；代价为2

- **为给定源程序生成最优目标程序**
  - 不可判定

目标代码中的地址

目标代码中的地址

- **如何将IR中的名字转换成为目标代码中的地址**
  - 不同的区域中的名字采用不同的寻址方式
- **如何为过程调用和返回生成代码**
  - 静态分配
  - 栈式分配

活动记录的静态分配

- **活动记录的大小和布局由符号表决定**
- **每个过程静态地分配一个数据区域**
  - 开始位置用staticArea表示
- **过程调用时**
  - 在活动记录中存放返回地址
- **过程调用结束后**
  - 控制权返回

活动记录的静态分配

- **call callee的实现**
  - ST callee.staticArea, #here+20 //存放返回地址
  - BR callee.codeArea
- **callee中的语句return**
  - BR *callee.staticArea

例子

- **三地址代码**

//过程c的代码

action1
call p
action 2
halt

//过程p的代码
action3
return

```
                          // c 的代码
100:  ACTION1             // action1 的代码
120:  ST 364, #140        // 在位置 364 上存放返回地址 140
132:  BR 200              // 调用 p
140:  ACTION2
160:  HALT                // 返回操作系统
      ...
                          // p 的代码
200:  ACTION3
220:  BR *364             // 返回在位置 364 保存的地址处
      ...
                          // 300-363 存放 c 的活动记录
300:                      // 返回地址
304:                      // c 的局部数据
      ...
                          // 364-451 存放 p 的活动记录
364:                      // 返回地址
368:                      // p 的局部数据
```

静态分配会使得语言受到什么限制?

活动记录栈式分配

- **问题**
  - 运行时刻才能知道一个过程的活动记录的位置
- **解决**
  - 活动记录的位置存放在寄存器中
  - 偏移地址

活动记录栈式分配

- **寄存器SP指向栈顶**
- **第一个过程（main）初始化栈区**
- **过程调用指令序列**
  - ADD SP, SP, #caller.recordSize //增加栈指针
  - ST 0(SP), #here+16 //保存返回地址
  - BR callee.codeArea //转移到被调用者
- **过程返回指令序列**
  - BR *0(SP) //被调用者执行，返回调用者
  - SUP SP, SP, #caller.recordSize //调用者减少栈指针值

例子

```
                    // code for m
action1
call q
action2
halt

                    // code for p
action3
return

                    // code for q
action4
call p
action5
call q
action6
call q
return
```

```
                                    // m的代码
100:  LD SP, #600                   // 初始化栈
108:  ACTION1                       // action1的代码
128:  ADD SP, SP, #msize            // 调用指令序列的开始
136:  ST *SP, #152                  // 将返回地址压入栈
144:  BR 300                        // 调用q
152:  SUB SP, SP, #msize            // 恢复SP的值
160:  ACTION12
180:  HALT
      ...
                                    // p的代码
200:  ACTION3
220:  BR *0(SP)                     // 返回
      ...
                                    // q的代码
300:  ACTION4                       // 包含有跳转到456的条件转移指令
320:  ADD SP, SP, #qsize
328:  ST *SP, #344                  // 将返回地址压入栈
336:  BR 200                        // 调用p
344:  SUB SP, SP, #qsize
352:  ACTION5
```

```
372:  ADD SP, SP, #qsize
380:  BR *SP, #396                  // 将返回地址压入栈
388:  BR 300                        // 调用q
396:  SUB SP, SP, #qsize
404:  ACTION6
424:  ADD SP, SP, #qsize
432:  ST *SP, #440                  //将返回地址压入栈
440:  BR 300                        // 调用q
448:  SUB SP, SP, #qsize
456:  BR *0(SP)                     // 返回
      ...
600:                                // 栈区的开始处
```

图 8-6　栈式分配时的目标代码

名字的运行时刻地址

- **在三地址语句中使用名字（实际上是指向符号表条目）来引用变量 （相对地址）**
- **语句x=0**
  - 如果x分配在静态区域，且静态区开始位置为static
    - static[12] = 0　　ST　112　#0
  - 如果x分配在栈区，且相对地址为12，则
    - ST　12(SP)　　#0

基本块和流图

基本块和流图

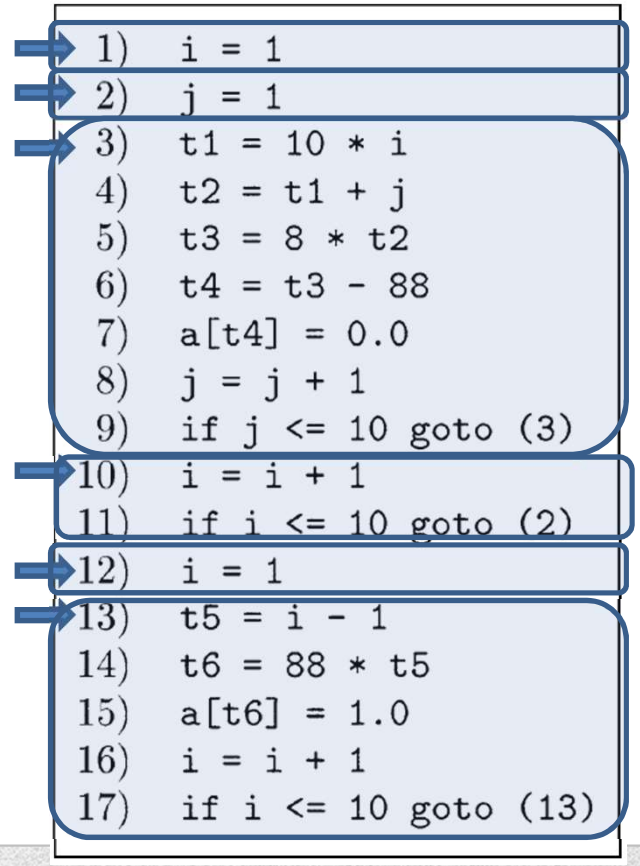
- **中间代码的流图表示法**
  - 中间代码划分成为基本块(basic block)
    - 控制流只能从第一个指令进入
    - 除基本块的最后一个指令外，控制流不会跳转/停机
  - 结点：基本块
  - 边：指明了基本块的执行顺序
- **流图可以作为优化的基础**
  - 指出了基本块之间的控制流
  - 可以根据流图了解到一个值是否会被使用等信息

划分基本块的算法

- **输入：三地址指令序列**
- **输出：基本块的列表**
- **方法：**
  - 确定基本块的首指令
    - 第一个三地址指令
    - 任意一个条件或无条件转移指令的目标指令
    - 紧跟在一个条件/无条件转移指令之后的指令
  - 确定基本块
    - 每个首指令对应于一个基本块：从首指令开始到下一个首指令

基本块划分的例子

- **第一个指令**
  - 1
- **跳转指令的目标**
  - 3、2、13
- **跳转指令的下一条指令**
  - 10、12
- **基本块：**
  - 1-1；2-2；3-9；10-11；
  - 12-12；13-17

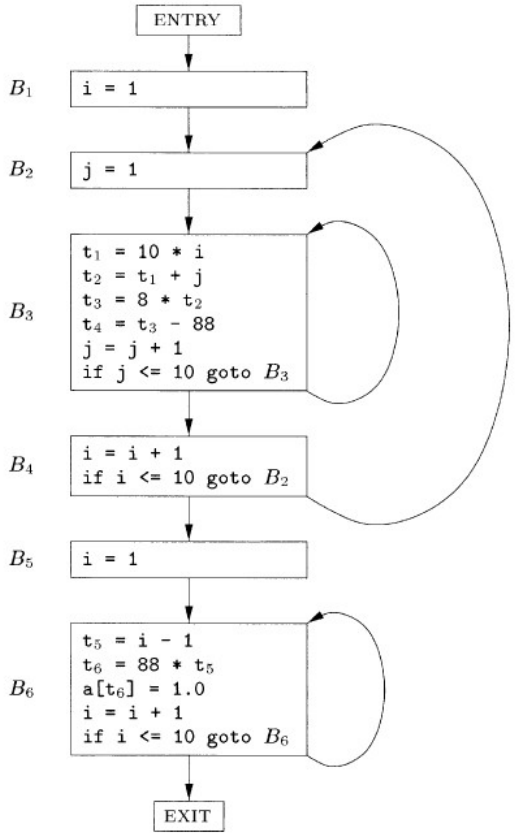


流图的构造

- **流图的顶点是基本块**
- **两个顶点B和C之间有一条有向边当且仅当基本块C的第一个指令可能在B的最后一个指令之后执行**
  - 从B的结尾指令是一条跳转到C的开头的条件/无条件语句
  - 在原来的序列中，C紧跟在B之后，且B的结尾不是无条件跳转语句
- **称B是C的前驱，C是B的后继**
- **入口和出口结点**
  - 流图中额外添加的边，不与中间代码（基本块）对应
  - 入口到第一条指令有一条边
  - 从任何可能最后执行的基本块到出口有一条边

流图的例子

- **因跳转而生成的边**
  - B3→B3
  - B4→B2
  - B6→B6
- **因为顺序而生成的边**
  - 其它

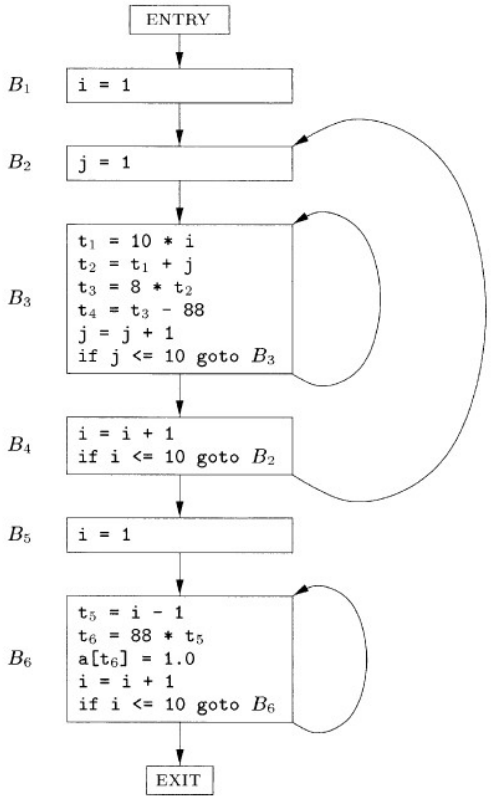


循环

- **程序的大部分运行时间花费在循环上**
  - 循环是识别的重点
- **循环的定义**
  - 循环L是一个结点集合
  - 存在一个循环入口（loop entry）结点，是唯一的、前驱可以在循环L之外的结点，到达其余结点的路径必然先经过这个入口结点
  - 其余结点都存在到达入口结点的非空路径，且路径都在L中

循环的例子

- **循环**
  - $\{B_3\}$
  - $\{B_6\}$
  - $\{B_2,B_3,B_4\}$
- 对于$\{B_2,B_3,B_4\}$的解释
  - B_2为入口结点
  - B_1，B_5，B_6不在循环内的理由
    - 到达B_1可不经过B_2
    - B_5，B_6没有到达B_2的结点



基本块的优化

- **针对基本块的局部优化**
  - 针对基本块的优化可以有很好的效果
  - 基本块中各个指令要么都执行，要么都不执行

基本块的DAG表示

重要

- **基本块可以用DAG表示**
  - 每个变量有对应的DAG的结点，代表初始值
  - 每个语句s有一个相关的结点N，代表计算得到的值
    - N的子结点对应于（其运算分量当前值的）其它语句
    - 结点N的标号是s的运算符
    - N和一组变量关联，表示s是在此基本块内最晚对它们定值的语句
- **输出结点：结点对应的变量在基本块出口处活跃**
- **从DAG，我们可以知道各个变量最后的值和初始值的关系**

回顾：表达式的DAG

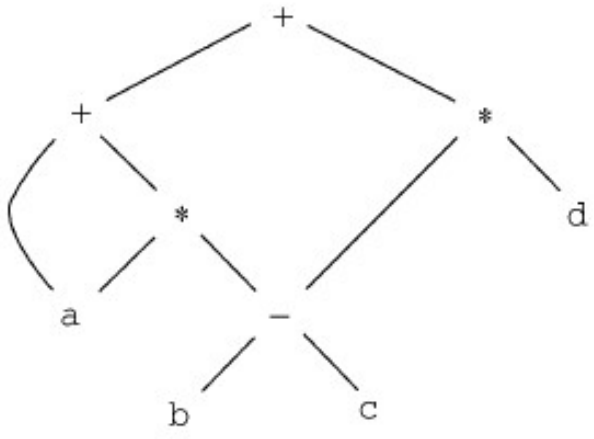
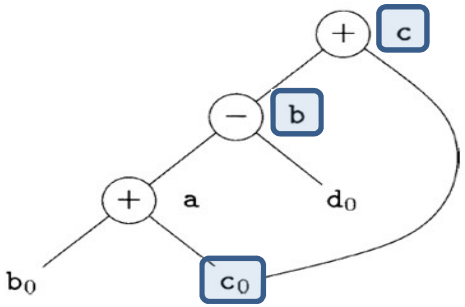


图 6-3　表达式 a + a * (b - c) + (b - c) * d 的 DAG

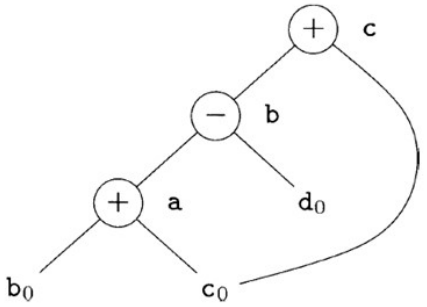
例子

- **指令序列**
  - a = b + c
  - b = a - d
  - c = b + c
- **过程：**
  - 结点b_0，c_0和d_0对应于b，c和d的初始值，它们和相应结点关联；
  - a=b+c: 构造第一个加法结点，a与之关联
  - b=a-d：构造减法结点，b与之关联
  - c=b+c：构造第二个加法结点，c与之关联（注意第一个子结点对应于减法结点）
- **如果还有第四条指令c=a，如何处理DAG?**



DAG的构造

- **为基本块中出现的每个变量建立结点（表示初始值），各变量和相应结点关联**
- **顺序扫描各个三地址指令，进行如下处理**
  - 如果指令为x=y op z
    - 为这个指令建立结点N，标号为op
    - N的子结点为y、z当前关联的结点
    - 令x和N关联
  - 如果指令为x=y
    - 不建立新结点
    - 设y关联到N，那么x现在也关联到N
- **扫描结束后，对于所有在出口处活跃的变量x，将x所关联的结点设置为输出结点**



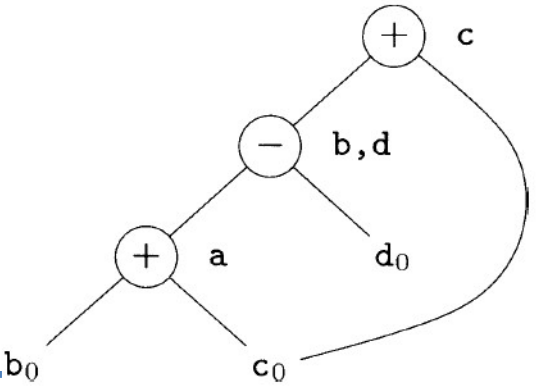
DAG的作用

- **DAG描述了基本块运行时各变量的值（和初始值）之间的关系**
- **我们可以以DAG为基础，对代码进行转换**
  - 消除局部公共子表达式
  - 消除死代码
  - 对语句重新排序
  - 对运算分量的顺序进行重排

局部公共子表达式

- **检测局部公共子表达式**
  - 建立某个结点M之前，首先检查是否存在一个结点N，它和M具有相同的运算符和子结点（顺序也相同）
  - 如果存在，则不需要生成新的结点，用N代表M
- **例如**
  - a=b+c
  - b=a-d
  - c=b+c
  - d=a-d

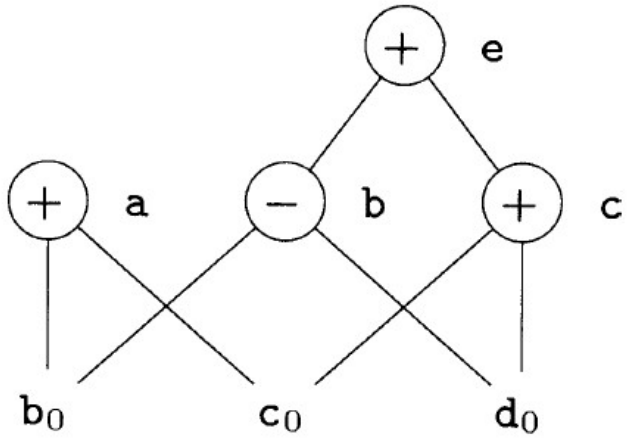
注意：两个b+c实际上并不是公共子表达式



消除死代码

- 在DAG上消除没有附加活跃变量的**根结点**（没有父结点的结点），即消除死代码
- 如果图中c、e不是活跃变量，则可以删除标号为e、c的结点

死代码=？不可达代码



应用代数恒等式的优化

- **消除计算步骤**
  - x+0=0+x=x　　x-0=x
  - x*1=1*x=x　　x/1=x
- **强度消减**
  - x^2=x*x　　2*x=x+x
- **常量合并**
  - 2*3.14可以用6.28替换
- **实现这些优化时，只需要在DAG上寻找特定的模式**

其他一些代数恒等式

- **条件表达式和算术表达式**
  - x>y
  - x-y
- **结合律**
  - a=b+c
  - e=c+d+b

```
a = b + c
t = c + d
e = t + b
```

```
a = b + c
e = a + d
```

  - 如果t只在基本块内使用

数组引用

- **数组引用实例**
  - x=a[i]
  - a[j]=y
  - z=a[i]

  } **a[i]是公共子表达式吗?**

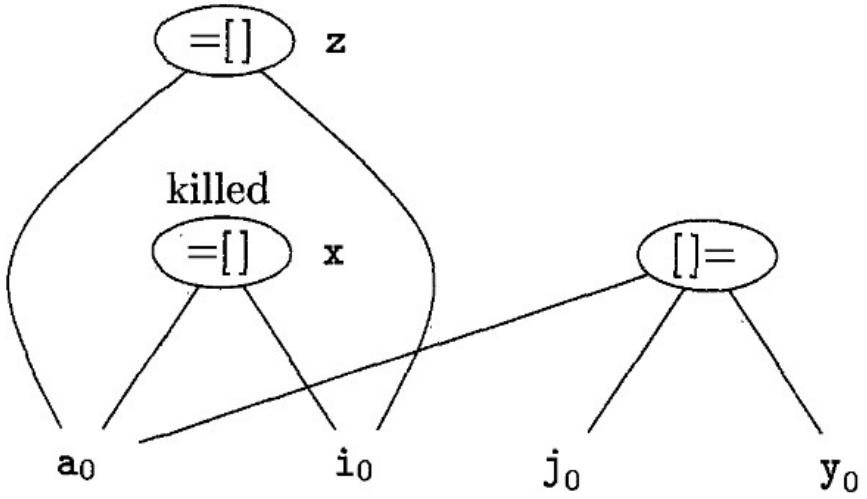
- **注意：a[j]可能改变a[i]的值，因此不能和普通的运算符一样构造相应的结点**

数组引用

- **数组引用的DAG表示**
  - 从数组取值的运算x=a[i]对应于=[]的结点
    - x作为这个结点的标号之一
    - 该节点的左右子节点分别代表数组初始值和下标
  - 对数组赋值（例如a[j]=y）的运算对应于[]=的结点，没有关联的变量、且杀死所有依赖于a的变量
    - 三个子节点分别表示a_0、j和y
  - 数组赋值隐含的意思：将数组作为整体考虑，对数组元素赋值即改变整个数组

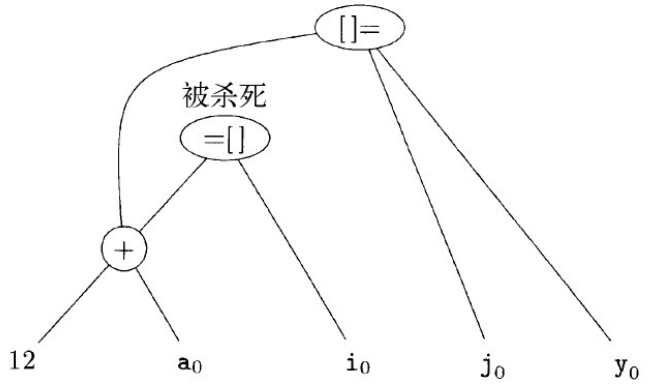
数组引用的DAG表示实例

- **基本块**
  - x = a[i]
  - a[j] = y
  - z = a[i]



数组引用的DAG的例子

- **设a是数组，b是指针**
  - b=12+a
  - x=b[i]
  - b[j]=y
- **注意a是被杀死的结点的孙结点**
- **一个结点被杀死，意味着它不能被复用**
  - 考虑再有指令m=b[i]?

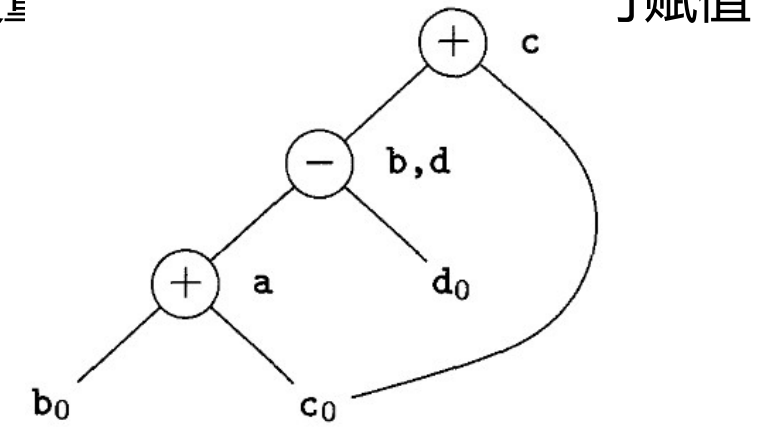


指针赋值/过程调用

- **通过指针进行取值/赋值：x=*p　*q=y。最粗略地估计：**
  - x使用了任意变量，因此无法消除死代码
  - 而*q=y对任意变量赋值，因此杀死了全部其他结点
- **可以通过（全局/局部）指针分析部分解决这个问题**
- **过程调用也类似，必须安全地假设它**
  - 使用了可访问范围内的所有变量
  - 修改了可访问范围内的所有变量

从DAG到基本块

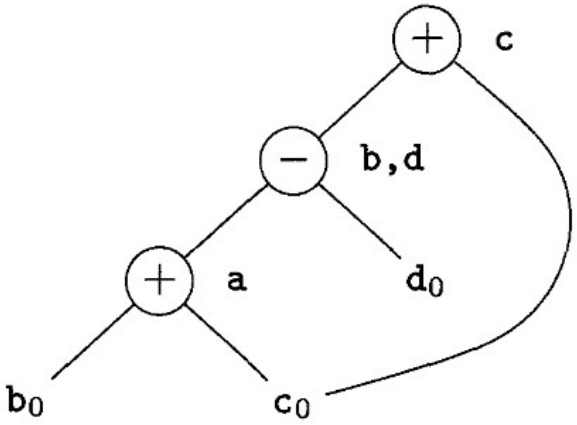
- **重构的方法**
  - 每个结点构造一个三地址语句，计算对应的值
  - 结果应该尽量赋给一个活跃的变量
  - 如果结点有多个关联的变量，则需要用复制语句进行赋值



重组基本块的例子

```
a = b + c
b = a - d
c = b + c
d = a - d
```

```
a = b + c
d = a - d
c = d + c
```



结果应该尽量赋给一个活跃的变量

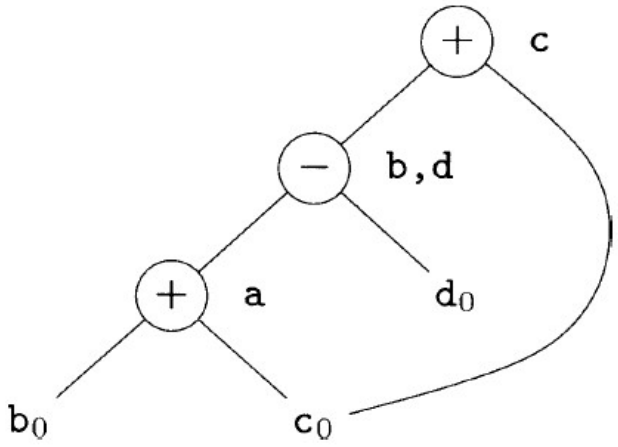
- 如果b在基本块出口不活跃

重组基本块的例子

- **如果b和d都在基本块出口处活跃**

```
a = b + c
b = a - d
c = b + c
d = a - d
```

  - a = b + c
  - d = a - d
  - b = d
  - c = d + c



重组的规则

- **重组时应该注意求值的顺序**
  - 指令的顺序必须遵守DAG中结点的顺序
  - 对数组的赋值必须跟在所有原来在它之前的赋值/求值操作之后
  - 对数组元素的求值必须跟在所有原来在它之前的赋值指令之后
  - 对变量的使用必须跟在所有原来在它之前的过程调用和指针间接赋值之后
  - 任何过程调用或者指针间接赋值必须跟在原来在它之前的变量求值之后
- **总的来说，我们必须保证：如果两个指令之间可能相互影响，那么他们的顺序就不应该改变**

基本块的代码生成器

- **根据三地址指令序列生成机器指令，假设**
  - 每个三地址指令只有一个对应的机器指令
  - 有一组寄存器用于计算基本块内部的值
- **主要目标**
  - 尽量减少加载和保存指令，即最大限度利用寄存器

Thank you!