Compiler Optimization
Overview of Optimizations

Compiler optimization is

- to generate **better** code
- not to generate **optimal** code
  - it is an NP-complete problem

What is a **better** version?

- Same result
- Better one or more of the followings
  - Execution time
  - Memory usage
  - Energy/power consumption
  - Network messages
  - Other criteria
Types of Optimizations

- Compiler optimization is essentially transformation
  - Delete something
  - Add something
  - Move something
  - Modify something

- Transform code or data?
  - Data-related optimizations
  - Code-related optimizations
Data-Related Optimizations
Data-Related Optimizations

- Seeks to improve cache behavior by
  - changing the location and representation of data or code
  - exploiting knowledge of memory hierarchy layout
    (is inherently machine-dependent)

- Change code layout

```c
f() {
    ... call h();
}
g() {
    ...
}
h() {
    ...
}
```

OR

```c
f() {
    code of f()
    code of g()
    code of h()
}
```

OR

```c
f() {
    code of f()
    code of h()
    code of g()
}
Which Code Layout is Better?

Assume
- data cache has one N-word line
- the size of each function is N/2-word long
- access sequence is “g, f, h, f, h, f, h”

<table>
<thead>
<tr>
<th>code of f()</th>
<th>code of g()</th>
</tr>
</thead>
<tbody>
<tr>
<td>code of h()</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>cache</th>
</tr>
</thead>
<tbody>
<tr>
<td>code of f()</td>
</tr>
<tr>
<td>code of g()</td>
</tr>
</tbody>
</table>

6 cache misses

6 cache misses

2 cache misses
Data Layout Optimization

- Change the variable declaration order and/or the width

```
struct S {
    int x1;
    int x2[200];
    int x3;
} obj[100];
...
... obj[i].x1 + obj[i].x3
```

```
struct S {
    int x1;
    int x3;
    int x2[200];
} obj[100];
...
... obj[i].x1 + obj[i].x3
```

```
int flag1[200];
int flag2[200];
...
flag1[i] = 0 or 1;
...
flag2[i] = 0 or 1;
... flag1[i] + flag2[i]
```

```
char flag1[200]; // 8-bit, fast access
bit flag2[200]; // 1-bit, need bit operation
...
flag1[i] = 0 or 1;
...
flag2[i] = 0 or 1;
... flag1[i] + flag2[i]
```
Code-Related Optimizations
Code-Related Optimizations

- **Modifying code**  e.g. **strength reduction**
  
  \[
  A = 2a; \quad \equiv \quad A = a \ll 1;
  \]

- **Deleting code**  e.g. **dead code elimination**
  
  \[
  A = 2; \ A = y; \quad \equiv \quad A = y;
  \]

- **Moving code**  e.g. **code motion**
  
  \[
  A = x \cdot y; \ B = A + 1; \ C = y; \quad \equiv \quad A = x \cdot y; \ C = y; \ B = A + 1;
  \]

- **Inserting code**  e.g. **data prefetching**
  
  \[
  \text{while} \ (p \neq \text{NULL})
  \{
  \quad \text{...} \ p = p \rightarrow \text{next}; \}
  
  \equiv
  
  \text{while} \ (p \neq \text{NULL})
  \{
  \quad \text{prefetch} (p \rightarrow \text{next}); \ \text{...} \ p = p \rightarrow \text{next}; \}
  \]
Optimization Categories

- Optimize at what representation level?
  - Source code level
  - IR level
  - Binary code level

- Optimize for specific machine?
  - Machine independent — typically at IR or source level
  - Machine dependent — typically at machine code level

- Optimize across control flow?
  - Local optimization — scope within straight line code
  - Global optimization — scope across control structures

- Optimize across procedures?
  - Intra-procedural — scope within individual procedure
  - Inter-procedural — scope across different procedures
    (Analyze callee to optimize caller and vice versa)
Local Optimizations
Local Optimizations

- Optimizations where the scope includes no control flow
  - Limited in scope but can still do useful things

- **Strength Reduction**
  - The idea is to replace expensive operations (multiplication, division) by less expensive operations (add, sub, shift, mov)
  - Some are redundant and thus can be deleted
    - e.g. \( x = x + 0; y = y \times 1; \)
  - Some can be simplified
    - e.g. \( x = x \times 0; y = y \times 8; \)
      - can be replaced by \( x = 0; y = y \ll 3; \)
  - Is also machine-dependent since it uses knowledge about the underlying machine (e.g. multiplication is expensive)
More Local Optimizations

Constant folding

- Operations on constants can be computed at compile time.
- In general, if \( x = y \text{ op } z \) and \( y \) and \( z \) are constants, then compute at compile time and replace.

Example:

```c
#define LEN 100
x = 2 * LEN;
if (LEN < 0) print(”error”);
```

Can be transformed to ...

```c
x = 200;
if (false) print(”error”);
```

- Is machine-independent since it is beneficial regardless of machine.
Global Optimizations and Control Flow Analysis
Global Optimizations and Control Flow

- Global optimization include more powerful optimizations
  - Can go across control structures
  - In effect, scope of optimization is the entire function (hence the name global)
  - E.g. Global Constant Propagation (GCP):
    - Replace variables with constants if value is known
      \[ X = 7; \]
      \[
      \ldots
      \]
      \[ Y = X + 3; \text{ // Can be replaced by } Y = 10; \]
    - Needs knowledge of control flow
      (Whether evaluation at point A happens before point B)

- Global optimization requires control flow analysis
  - **Control flow analysis**: Compiler analysis that determines flow of control during execution of a function
  - Constructs a control flow graph that describes the flow
A basic block is a maximal sequence of instructions that
- Except the first instruction, there are no other labels;
- Except the last instruction, there are no jumps;

Therefore,
- Can only jump into the beginning of a block
- Can only jump out at the end of a block
- All instructions of block execute or none at all

Basic blocks are basic units of control flow which cannot be divided any further
A control flow graph is a directed graph in which

- Nodes are basic blocks
- Edges represent flow of execution
  - Control statements such as if-then-else, while-loop, for-loop introduce control flow edges

CFG is widely used to represent a program
CFG is widely used for program analysis, especially for global analysis/optimization
Example

L1; t := 2 * x;
    w := t + y;
    if (w < 0) goto L3
L2: ...
    ...
L3: w := -w
    ...

L1; t := 2 * x;
    w := t + y;
    if (w < 0) goto L3
L2: ...
    ...
L3: w := -w
    ...

no

yes
Construction of CFG

- **Step 1**: partition code into basic blocks
  - Identify **leader** instructions that are
    - the first instruction of a program, or
    - target instructions of jump instructions, or
    - instructions immediately following jump instructions
  - A basic block consists of a leader instruction and its subsequent instruction before the next leader

- **Step 2**: add an edge between basic blocks B1 and B2 if
  - there exist a jump from B1 to B2, or
  - B2 follows B1, and B1 does not end with unconditional jump
    - B1 ends with a conditional jump
    - B1 ends with a non-jump instruction (B2 is a target of a jump)
Example

01. A=4
02. T1=A*B
03. L1: T2=T1/C
04: if (T2<W) goto L2
05: M=T1*K
06: T3=M+1
07: L2: H=I
08: M=T3-H
09: if (T3>0) goto L3
10: goto L1
11: L3: halt
Example

01. \textbf{A}=4
02. \textbf{T1}=A*B
03. \textbf{L1}: \textbf{T2}=T1/C
04: if (T2<W) goto L2
05: \textbf{M}=T1*K
06: \textbf{T3}=M+1
07: \textbf{L2}: \textbf{H}=I
08: \textbf{M}=T3-H
09: if (T3>0) goto L3
10: goto L1
11: \textbf{L3}: \textbf{halt}
Global Optimizations

- Extend optimizations to flow of control, i.e. CFG

How do we know it is OK to globally propagate constants?
In particular, there are situations that prohibit this optimization.

To replace \( x \) by a constant \( C \) correctly, we must know:

- **Along all paths**, the last assignment to \( X \) is \( "X:=C" \)
- **All paths** often include branches and even loops
  - Usually it is not trivial
Many compiler optimizations depend on knowing some property $X$ at a particular point in program execution.

- Need to prove at that point property $X$ holds along all paths.

Optimizations Need to be Conservative

- Property $X$ often involves data flow of program.
- E.g. Global Constant Propagation (GCP): $X = 7$;
  
  $...$
  
  $Y = X + 3; // Replace by Y = 10, if X didn't change$

- Needs knowledge of data flow, as well as control flow (Whether data flow is interrupted between points A and B).
Optimizations Need to be Conservative

Many compiler optimizations depend on knowing some property X at a particular point in program execution

- Need to prove at that point property X holds along all paths
- Need to be conservative to ensure correctness
  - An optimization is enabled only when X is definitely true
  - If not sure if it is true or not, it is safe to say don’t know
  - If you don’t know, you don’t do the optimization
  - May lose opt. opportunities but guarantees correctness

Property X often involves data flow of program
- E.g. Global Constant Propagation (GCP):
  X = 7;
  ...
  Y = X + 3; // Replace by Y = 10, if X didn’t change

Needs knowledge of data flow, as well as control flow (Whether data flow is interrupted between points A and B)
Optimizations Need to be Conservative

- Many compiler optimizations depend on knowing some property $X$ at a particular point in program execution
  - Need to prove at that point property $X$ holds along all paths
  - Need to be **conservative** to ensure correctness
    - An optimization is enabled only when $X$ is definitely true
    - If not sure if it is true or not, it is safe to say **don’t know**
    - If you **don’t know**, you don’t do the optimization
    - May lose opt. opportunities but guarantees correctness

- Property $X$ often involves data flow of program
  - E.g. Global Constant Propagation (GCP):
    
    ```
    X = 7;
    ...
    Y = X + 3; // Replace by Y = 10, if X didn’t change
    ```
  - Needs knowledge of data flow, as well as control flow
    (Whether data flow is interrupted between points A and B)
Global Optimizations and Data Flow Analysis
Dataflow Analysis Framework

Dataflow analysis: Compiler analysis that determines what values get propagated from point A to point B

- Requires CFG since values flow through control flow edges
Dataflow Analysis Framework

- **Dataflow analysis**: Compiler analysis that determines what values get propagated from point A to point B
  - Requires CFG since values flow through control flow edges

- **Dataflow analysis framework**: Framework for dataflow analysis that guarantees optimizations are conservative

- Defined by: \( \{D, V, \land, F: V \rightarrow V\} \)
  - \(D\): Direction of propagation (forwards or backwards)
  - \(V\): Set of values (depends on analyzed property)
    - Value for GCP: set of variables with constant values
  - \(\land\): Meet operator \((V \land V \rightarrow V)\)
    - Defines behavior when values meet at control flow merges
  - \(F\): Transfer function \(F: V \rightarrow V\)
    - Defines what happens to value within a basic block
Dataflow Analysis Framework

- **Dataflow analysis**: Compiler analysis that determines what values get propagated from point A to point B
  - Requires CFG since values flow through control flow edges

- **Dataflow analysis framework**: Framework for dataflow analysis that guarantees optimizations are **conservative**
  - Defined by: \{D, V, \&, F: V \rightarrow V\}
  - D: Direction of propagation (forwards or backwards)
  - V: Set of values (depends on analyzed property)
    - Value for GCP: set of variables with constant values
  - \&: Meet operator (V \& V \rightarrow V)
    - Defines behavior when values meet at control flow merges
  - F: Transfer function F: V \rightarrow V
    - Defines what happens to value within a basic block

- **Goal**: To assign V for every point in program
  \rightarrow aids optimization
Rather than bore you with math, let’s learn by example: **global constant propagation (GCP)**

**What is Global Constant Propagation?**
- At compile time, if the value of a variable is a constant, replace the variable with the constant
- “Global” means we substitute across basic blocks and control flow
- At compile time, we don’t know which path is taken
Rather than bore you with math, let’s learn by example: global constant propagation (GCP)

What is Global Constant Propagation?
- At compile time, if the value of a variable is a constant, replace the variable with the constant.
- “Global” means we substitute across basic blocks and control flow.
- At compile time, we don’t know which path is taken.

Compiler can apply a dataflow analysis framework to the problem to ensure conservative optimization.
- What is $D$, $V$, $\land$, $F : V \rightarrow V$ in this context?
What is V?

- Definition: Set of values in property under analysis
- Property for GCP:
  - What are the variables with constant values?
  - And what are their values at the given point?
What is $V$?

- **Definition:** Set of values in property under analysis
- **Property for GCP:**
  - What are the variables with constant values?
  - And what are their values at the given point?
- A given variable can be in one of the following states:
  - $x=1$, $x=2$, ... // defined a constant
  - $x=*$ // not defined yet
  - $x=#$ // don’t know (not provably constant)
What is \( V \)?

- **Definition:** Set of values in property under analysis

- **Property for GCP:**
  - What are the variables with constant values?
  - And what are their values at the given point?

- **A given variable can be in one of the following states:**
  - \( x=1, x=2, \ldots \) // defined a constant
  - \( x=* \) // not defined yet
  - \( x=# \) // don’t know (not provably constant)

- **\( V \) for GCP:** Set of values where each value is the set of variables and their respective states.

- **Examples of values in \( V \):** \{\( x=*, y=10, z=# \), \( x=1, y=#, z=5 \)\}

- **Goal for GCP is to assign a value to each point in program**
What is $\land$?

- $\land$: Meet operator ($V \land V \to V$)
  - Defines behavior when values meet at control flow merges
  - Given
    - $V_{\text{in}}(B)$ — value at the entry of basic block $B$
    - $V_{\text{out}}(B)$ — value at the exit of basic block $B$
  - $V_{\text{in}}(B) = \land V_{\text{out}}(P)$ for each $P$, where $P$ is a predecessor of $B$

- Example of $\land$ operator for GCP:
  \{x=\ast, y=2, z=3\} \land \{x=1, y=2, z=10\} = \{x=1, y=2, z=\#\}

- Relationship between values in $V$ given by a meet operator is called a **Semilattice**
Semilattice for GCP (when there is one variable):

\[ \{x=\ast\} \]

\[ \{x=0\} \]
\[ \{x=1\} \]
\[ \{x=-1\} \]
\[ \{x=\#\} \]

\( \wedge \) operator is defined by the Greatest Lower Bound (GLB) between two values

- \( \{x=\ast\} \wedge \{x=1\} = \{x=1\} \)
- \( \{x=0\} \wedge \{x=1\} = \{x=\#\} \)

Downward direction is always the conservative choice

In effect, GLB is the least conservative but correct choice

Semilattice essentially defines what meet operator means
In a semilattice, there are two special values: $\top$ and $\bot$. 

- $\top$: Called Top Value
  - Initial value when analysis begins
  - For GCP: $\{x=*, y=*, z=*\}$
  - Value is refined in the course of analysis

- $\bot$: Called Bottom Value
  - Value which can be refined no further
  - For GCP: $\{x=\#, y=\#, z=\#\}$
  - Meaning: none of the variables are provably constant

Analysis iteratively refines values until they stabilize somewhere between $\top$ and $\bot$. 
In a semilattice, there are two special values: $\top$ and $\bot$

$\top$: Called **Top Value**
- Initial value when analysis begins
- For GCP: \{x=*, y=*, z=*\}
- Value is refined in the course of analysis
In a semilattice, there are two special values: $\top$ and $\bot$

$\top$: Called **Top Value**
  - Initial value when analysis begins
  - For GCP: $\{x=*, y=*, z=\}$
  - Value is refined in the course of analysis

$\bot$: Called **Bottom Value**
  - Value which can be refined no further
  - For GCP: $\{x=\#, y=\#, z=\#\}$
  - Meaning: none of the variables are provably constant

Analysis iteratively refines values until they stabilize somewhere between $\top$ and $\bot$
What is F?

F: Transfer function (F: V → V)
- Defines what happens to value within a basic block
- Given
  - \( V_{in}(B) \) — value at the entry of basic block B
  - \( V_{out}(B) \) — value at the exit of basic block B
- \( V_{out}(B) = F( V_{in}(B) ) \)

F for GCP:
\[
V_{out}(B) = ( V_{in}(B) - DEF_{\nu}(B) ) \cup DEF_{c}(B)
\]

where \( DEF_{\nu}(B) \) contains variable definitions in B
\( DEF_{c}(B) \) contains constant definitions in B

Easier to reason about if you treat each individual statement as a basic block
There are two modes of propagation: \( \mathbf{F} \) and \( \wedge \)

- Function \( \mathbf{F} \) — propagates values through basic blocks
  - Variables in \( \text{DEF}_v \) are set to \#.
  - Variables in \( \text{DEF}_c \) are set to constant value.

- \( \wedge \) operator — propagates values through CFG edges
  - Merges values from multiple predecessor blocks.
**What is D?**

- **D**: Direction of propagation (forwards or backwards)

---

**Forward Analysis**

- $V_{in}$
- $V_{out}$

**Backward Analysis**

- $V_{in}$
- $V_{out}$

---

**Diagram:**

- Nodes and edges representing the propagation direction in both forward and backward analysis.
What is D?

- Values are propagated forward: **Forward Analysis**
- Values are propagated backward: **Backward Analysis**
- GCP is an example of a Forward Analysis
  - Starting from a constant definition, the ‘constantness’ of a variable propagates forward through CFG
- We will see an example of Backward Analysis soon
In this example, constants can be propagated to $X+1$, $2*X$.
In this example, constants can be propagated to \( X+1 \), \( 2*X \)

\[
\begin{align*}
X &= * \\
X &= 3; \\
& \text{if } (B>0) \\
Y &= Z+W; \\
X &= 4; \\
Y &= 0; \\
X &= X+1 \\
A &= 2*X;
\end{align*}
\]
Example GCP without Loop

In this example, constants can be propagated to $X+1$, $2*X$.
In this example, constants can be propagated to $X+1$, $2X$
In this example, constants can be propagated to $X+1$, $2*X$.
In this example, constants can be propagated to $X+1$, $2*X$
In this example, constants can be propagated to $X+1$, $2*X$.

Example GCP without Loop

```
X:=3;
if (B>0)
  Y:=Z+W;
X:=4;
Y:=0;
X:=X+1
A:=2*X;
```
In this example, constants can be propagated to $X+1$, $2*X$
In this example, loop prevents any constant propagation

\[\begin{align*}
X &:= 3; \\
&\text{if } (B > 0) \\
Y &:= Z + W; \\
X &:= 4; \\
Y &:= 0; \\
X &:= X + 1 \\
A &:= 2 \times X;
\end{align*}\]
In this example, loop prevents any constant propagation.

\[ X = \ast \]

\[
\begin{align*}
X &:= 3; \\
\text{if } (B > 0) &:
\end{align*}
\]

\[
\begin{align*}
Y &:= Z + W; \\
X &:= 4;
\end{align*}
\]

\[
\begin{align*}
Y &:= 0; \\
X &:= X + 1
\end{align*}
\]

\[
A := 2 \times X;
\]
In this example, loop prevents any constant propagation

```plaintext
X:=3;
if (B>0)

Y:=Z+W;
X:=4;

X:=4;
Y:=0;
X:=X+1

A:=2*X;
```

```plaintext
X=*  
X=3
```
In this example, loop prevents any constant propagation.
In this example, loop prevents any constant propagation.
In this example, loop prevents any constant propagation.

\begin{itemize}
\item \texttt{X=3;}
\item \texttt{if (B>0)}
\item \texttt{Y:=Z+W; X:=4;}
\item \texttt{X:=4; Y:=0; X:=X+1}
\item \texttt{A:=2*X;}
\end{itemize}
In this example, loop prevents any constant propagation.
In this example, loop prevents any constant propagation.

```
X := 3;
if (B > 0)
  Y := Z + W;
  X := 4;
  Y := 0;
  X := X + 1
A := 2 * X;
```

**NOT SO FAST!**
Example GCP with Loop

In this example, loop prevents any constant propagation

\[\text{X}:=3;\]
\[\text{if } (B>0)\]
\[\text{Y}:=Z+W;\]
\[\text{X}:=4;\]
\[\text{X}:=4;\]
\[\text{Y}:=0;\]
\[\text{X}:=X+1;\]
\[\text{A}:=2*X;\]
In this example, loop prevents any constant propagation.

Example GCP with Loop

X := 3;
if (B > 0)
Y := Z + W;
X := 4;
Y := 0;
X := X + 1
A := 2 * X;
In this example, loop prevents any constant propagation.
In this example, loop prevents any constant propagation:

\[
\begin{align*}
X & := 3; \\
\text{if } (B>0) \\
Y & := Z + W; \\
X & := 4; \\
X & := X + 1 \\
Y & := 0; \\
X & := X + 1 \\
A & := 2 \times X;
\end{align*}
\]
In this example, loop prevents any constant propagation.
Forward Analysis Algorithm

- **Pseudocode for Forward Analysis**
  
  for (each basic block B) $V_{out}(B) = \top$
  
  while (changes to any $V_{out}(B)$ occur)
      
      for (each basic block B)
          
          $V_{in}(B) = \land P \text{ is a predecessor of } B \ V_{out}(P)$
          
          $V_{out}(B) = F(V_{in}(B))$

- ∧ and $F$ defined differently for each type of analysis

- **Will forward analysis for GCP eventually stop?**
  
  - If there are loops, we may go through the loop many times
  
  - Is there a possibility of an infinite loop?
Termination Problem

- Existence of $\perp$ value ensures termination
  - Values start from $\top$
  - Values can only go down in the semilattice
  - Any values can change at most twice in our example
    ... from * to C, and from C to #

- The maximal number of steps is $O(\text{program}_\text{size})$
Once constants have been globally propagated, we would like to eliminate the dead code.

```
x:=3;
if (b>0)
  y:=z+w;
y:=0;
z:=2*x;
```
Once constants have been globally propagated, we would like to eliminate the dead code.

\[
x := 3; \\
\text{if } (b > 0) \\
y := z + w; \\
y := 0; \\
z := 2 \times 3;
\]
Once constants have been globally propagated, we would like to eliminate the dead code.

```
x:=3;
if (b>0)
    y:=z+w;
y:=0;
z:=2*3;
```
A **dead statement** calculates a value that is not used later.
Otherwise, it is a **live statement**.

In the example,
the 1st statement is dead,
the 2nd statement is live.
Global Liveness Analysis (GLA)

- A variable $X$ is live at statement $S$ if
  - There exists a statement $S_2$ after $S$ that uses $X$
  - There is a path from $S$ to $S_2$
  - There is no intervening assignment to $X$ between $S$ and $S_2$
Liveness Analysis

- Global Liveness Analysis (GLA)
  - A variable X is live at statement S if
    - There exists a statement S2 after S that uses X
    - There is a path from S to S2
    - There is no intervening assignment to X between S and S2

- Again a dataflow analysis framework can be applied to the problem
  - What is $D$, $V$, $\land$, $F: V \rightarrow V$ in this context?

- What is $D$?
  - Liveness Analysis is a **Backward Analysis**
    - Starting from a use, the ‘liveness’ of a variable propagates backward through CFG
  - Changes direction of $\land$ operator and transfer function
Forward and Backward Analysis Again

Forward Analysis

\[ V_{in} \rightarrow V_{out} \]

Backward Analysis

\[ V_{out} \leftarrow V_{in} \]
What is V?

Definition: Set of values in property under analysis
- \( V \) for GLA: Each value is a set of live variables
- Example values: \( \{x, y, z\}, \{y\} \)

\( \top \): initial value at the beginning
- \( \top \) for GLA = {} 
- Start with assumption that no variables are live

\( \bot \): the don’t know value
- \( \bot \) for GLA = \{all variables in function\}
- Meaning: none of the variables are provably dead
What is $\wedge$?

- $\wedge$: Meet operator ($V \wedge V \rightarrow V$) for backward analysis
  - Defines behavior when values meet at control flow merges
  - Given
    - $V_{in}(B)$ — value at the entry of basic block $B$
    - $V_{out}(B)$ — value at the exit of basic block $B$
  - $V_{out}(B) = \wedge V_{in}(S)$ for each $S$, where $S$ is successor of $B$
  - Note the reversal in direction! GLA is a backward analysis.

- $\wedge$ operator for GLA:
  - Meet operator is a simple union $\cup$
  - Example: $\{x, y\} \wedge \{y, z\} = \{x, y\} \cup \{y, z\} = \{x, y, z\}$
  - Union operation monotonically increases set, hence values form a semilattice from $\top$ to $\bot$
What is F?

- **F**: Transfer function \((F: V \rightarrow V)\) for backward analysis
  - Defines what happens to value within a basic block
  - Given
    - \(V_{in}(B)\) — value at the entry of basic block \(B\)
    - \(V_{out}(B)\) — value at the exit of basic block \(B\)
  - \(V_{in}(B) = F(V_{out}(B))\)
  - Again note the reversal in direction!

- **F** for GLA:
  \[
  V_{in}(B) = (V_{out}(B) - \text{DEF}(B)) \cup \text{USE}(B)
  \]
  where \(\text{DEF}(B)\) contains variable definitions in \(B\)
  and \(\text{USE}(B)\) contains variable uses in \(B\)

- Easier to reason about if you treat each individual statement as a basic block
Liveness Example

b = b + c

a = d + 1;
Liveness Example

\[ b = b + c \]

\[ a = d + 1; \]
Liveness Example

\[
b = b + c
\]
\[
a = d + 1;
\]

\[
V_{in}(B1) \quad V_{out}(B1)
\]
\[
V_{in}(B2) \quad V_{out}(B2)
\]
\[
V_{in}(B3) \quad V_{out}(B3) = \{a, b\}
\]
Liveness Example

```
b = b + c
```

V_in(B2)  

V_out(B2)  

V_in(B3)  

V_out(B3) = \{a, b\}

Two sets:  
DEF = \{a\}  
USE = \{d\}
Liveness Example

Two sets:
- DEF = \{a\}
- USE = \{d\}
Liveness Example

Liveness Example

b=b+c

a=d+1;

\( V_{\text{in}}(B1) \)

\( V_{\text{out}}(B1) \)

\( V_{\text{in}}(B2) = \{b, c\} \)

\( V_{\text{out}}(B2) \)

\( V_{\text{in}}(B3) = \{b, d\} \)

\( V_{\text{out}}(B3) = \{a, b\} \)

Two sets:
- DEF = \{a\}
- USE = \{d\}
Liveness Example

\[
\begin{align*}
\text{V}_{\text{out}}(B1) &= \{b, c, d\} \\
\text{V}_{\text{in}}(B1) &= \{b, c\} \\
\text{V}_{\text{out}}(B2) &= \{b\} \\
\text{V}_{\text{in}}(B2) &= \{b, c\} \\
\text{V}_{\text{out}}(B3) &= \{a, b\} \\
\text{V}_{\text{in}}(B3) &= \{b, d\} \\
\end{align*}
\]

Two sets:
- DEF = \{a\}
- USE = \{d\}
Backward Analysis Algorithm

- Pseudocode for Backward Analysis
  for (each basic block B) \( V_{in}(B) = \top \);
  while (changes to any \( V_{in}(B) \) occur)
    for (each basic block B) {
      \( V_{out}(B) = \bigwedge S \text{ is a successor of } B \ V_{in}(S) \)
      \( V_{in}(B) = F(V_{out}(B)) \)
    }

- Note the reversal in direction compared to forward analysis

- Will backward analysis for GLA eventually stop?
  - Again existence of \( \bot \) value ensures termination
  - Value can change N times, where N is the number of variables used in function
  - The maximal number of steps is \( O(\text{program\_size} \times N) \)
Comparison of GCP and GLA

- **D**: Direction of propagation
  - GCP: Forward
  - GLA: Backward

- **V**: Set of values propagated
  - GCP: Whether each variable is constant, and if so the value
  - GLA: Set of live variables

- **∧**: Meet operator
  - GCP: Defined by semilattice (Top $\rightarrow$ Constant $\rightarrow$ Bottom)
  - GLA: Simply the set union operator

- **F**: Transfer function
  - GCP: Subtract variable definitions, add constant definitions
  - GLA: Add variable uses, subtract variable definitions
Global dead code elimination is based on global liveness analysis (GLA)

- Dead code detection
  - A statement $x = \ldots$ is dead code if $x$ is dead after this statement
  - Dead statement can be deleted from the program

Global register allocation is also based on GLA

- Live variables should be placed in registers
- Registers holding dead variables can be reused
What is Register Allocation?

- Process of assigning (a large number of) variables to (a small number of) CPU registers
- Registers are fast
  - access to memory: 100s of cycles
  - access to cache: a few to 10s of cycles
  - access to registers: 1 cycle
- But registers are limited in number
  - x86: 8 regs, MIPS: 32 regs, ARM: 32 regs ...
- Goals of register allocation:
  - Keep frequently accessed variables in registers
  - Keep variables in registers only as long as they are live
Local Register Allocation

Allocate registers basic block by basic block

- Makes allocation decisions on a per-block basis
- Hence the prefix ‘local’
- Uses results of Global Liveness Analysis

Requires only a single scan through each basic block

- Keeps track of two tables:
  - Register table: which regs are currently allocated and where
  - Address table: location(s) where each variable is stored
    (locations can be: register, stack memory, global memory)

- For every use of variable:
  - If variable is already in reg, no action
  - If not, allocate reg to variable from available regs
  - If no available regs, select reg for displacement
Local Register Allocation

- **Which register should be displaced?**
  - Register whose value is no longer live (given by GLA)
  - Register whose value has a copy in another location
  - These registers can be safely recycled
  - Otherwise the register needs to be spilled

- **Spill**: storing register back into own memory location
  - Generate store instruction to memory on assignment
  - Generate load instruction from memory on use
  - Own memory location can be in
    - Stack memory: local variables, temporary variables
    - Global memory: global variables

- At the end of basic block all live registers are spilled
  - Makes all registers available for next basic block allocation
    (Gives allocator clean slate for next basic block)
  - Can be source of inefficiency due to unnecessary spills
    → Addressed by Global Register Allocation
Global Register Allocation

- Allocates registers across basic blocks
- Relies on Global Liveness Analysis just like local register allocation
- Three popular register allocation algorithms
  1. Graph coloring allocator
  2. Linear scan allocator
  3. ILP (Integer Linear Programming) allocator
Algorithm steps:

1. Identify live range interference using GLA
2. Build register interference graph
3. Attempt K-coloring of the graph
   - K is the number of available registers
4. If none found, modify the program, rebuild graph until K-coloring can be obtained
   - Insert spill code to the program
Live Range Interference

- **Live Range**: Set of program points where a variable is live
  - Two live ranges interfere if there is an overlap
  - Vars with interfering ranges cannot reside in the same register
Live Range Interference

- **Live Range**: Set of program points where a variable is live
  - Two live ranges interfere if there is an overlap
  - Vars with interfering ranges cannot reside in the same register

```
x := ...
y := ...;
; := ...x...
```

```plaintext
; := ...
; := ...
```
Live Range Interference

- **Live Range**: Set of program points where a variable is live
  - Two live ranges interfere if there is an overlap
  - Vars with interfering ranges cannot reside in the same register

We annotate each program point (between two statements) to explicitly show the interference.
Example of GLA and interfering live ranges

```
a := b + c;
d := -a;
e := d + f;
f := 2 * e;
b := d + e;
e := e - 1;

b := f + c;
```
Register Interference Graph

Construct **Register Interference Graph (RIG)** such that

- Each node represents a variable
- An edge between two nodes \( V_1 \) and \( V_2 \) represents an interference in live ranges

Based on RIG,

- Two variables can be allocated in the same register if there is no edge between them
- Otherwise, they cannot be allocated in the same register
In the RIG for our example:

- b, c cannot be in the same register
- a, b, d can be in the same register
Allocating Registers using Graph Coloring

Graph coloring is a theoretical problem where ... 
- A coloring of a graph is an assignment of colors to nodes such that nodes connected by an edge have different colors
- A graph is k-colorable if it has a coloring with k colors

Problem of register allocation in RIG maps to graph coloring problem
- Instead of assigning k-colors, we need to assign k registers
- K is the number of available machine registers
- If the graph is k-colorable, we have a register assignment that uses no more than k registers
This is an coloring of our example RIG using 4 colors

- There is no solution with less than 4 colors
Using the coloring result, map it back to the code

\begin{align*}
  a &:= b + c; \\
  d &:= -a; \\
  e &:= d + f; \\
  f &:= 2 \times e; \\
  b &:= d + e; \\
  e &:= e - 1; \\
  b &:= f + c;
\end{align*}
After Register Allocation

Using the coloring result, map it back to the code

\[ a \rightarrow R1 \]
\[ b \rightarrow R2 \]
\[ c \rightarrow R3 \]
\[ d \rightarrow R2 \]
\[ e \rightarrow R1 \]
\[ f \rightarrow R4 \]

- \[ f := 2 \cdot e; \]
- \[ b := d + e; \]
- \[ e := e - 1; \]
- \[ b := f + c; \]
- \[ R1 := R2 + R3; \]
- \[ R2 := -R1; \]
- \[ R1 := R2 + R4; \]
After Register Allocation

Using the coloring result, map it back to the code

\[
\begin{align*}
a &= R1 \\
b &= R2 \\
c &= R3 \\
d &= R2 \\
e &= R1 \\
f &= R4 \\
\end{align*}
\]

\[
\begin{align*}
R1 &= R2 + R3; \\
R2 &= -R1; \\
R1 &= R2 + R4; \\
R4 &= 2*R1; \\
b &= d + e; \\
e &= e - 1; \\
b &= f + c;
\end{align*}
\]
Using the coloring result, map it back to the code:

- $a \rightarrow R1$
- $b \rightarrow R2$
- $c \rightarrow R3$
- $d \rightarrow R2$
- $e \rightarrow R1$
- $f \rightarrow R4$

Diagram:

- $b := f + c$
- $R1 := R2 + R4$
- $R2 := R2 + R1$
- $R1 := R1 - 1$
- $R2 := R2 + R3$
- $R2 := -R1$

Code:

- $R4 := 2 \times R1$
- $R1 := R2 + R4$
Using the coloring result, map it back to the code

a–R1
b–R2
c–R3
d–R2
e–R1
f–R4

R1:=R2+R3;
R2:=-R1;
R1:=R2+R4;

R4:=2*R1;

R2:=R4+R3;

R2:=R2+R1;
R1:=R1-1;

R2:=R2+R4;
How is Graph Coloring Performed?

- For graph $G$ and $k>2$, determining whether $G$ is $k$-colorable is NP complete
  - Problem of $k$-register allocation is NP complete
  - In practice: use heuristical polynomial algorithm that gives close to optimal allocations most of the time
  - Chaitin’s graph coloring is a popular heuristical algorithm
    - Most backends of GCC use Chaitin’s algorithm by default

- What if $k$-register allocation does not exist?
  - Spill a register to memory to reduce RIG and try again
Observation: for a $k$-coloring problem, a node with $k-1$ neighbors can always be colored, no matter what...
**Observation:** for a \( k \)-coloring problem, a node with \( k-1 \) neighbors can always be colored, no matter what
Observation: for a $k$-coloring problem, a node with $k-1$ neighbors can always be colored, no matter what...
Observation: for a $k$-coloring problem, a node with $k-1$ neighbors can always be colored, no matter what
Observation: for a k-coloring problem, a node with k-1 neighbors can always be colored, no matter what...
Corollary: Given graph $G$ for a $k$-coloring problem

- Let $G'$ be the graph after removing a node with fewer than $k$ neighbors
- If $G'$ can be $k$-colored then $G$ can be $k$-colored

Insight: Solving for $G'$ is easier than solving for $G$, so solve for $G'$ instead of $G$

Algorithm

- Phase 1: Repeat until there are no nodes left
  - Pick a node $V$ with fewer than $k$ neighbors
  - Put $V$ on a stack and remove it and its associated edges from the graph
- Phase 2: Assign colors to nodes on the stack in LIFO order
  - Pick a color that is different from its neighbors
  - Such a color is guaranteed to exist due to corollary
    (Analogous to coloring $G$ after adding removed node to $G'$)
Chaitin’s algorithm applied to our example where k=4

Stack={}

CS2210: Compiler Construction
Compiler Optimization
Chaitin’s algorithm applied to our example where $k=4$
Chaitin’s algorithm applied to our example where $k=4$

Stack=$\{a\}$
Chaitin’s algorithm applied to our example where $k=4$

Stack={$a,d$}
Chaitin’s Graph Coloring Example

Chaitin’s algorithm applied to our example where \( k=4 \)

Stack={a,d}
Chaitin’s algorithm applied to our example where $k=4$
Chaitin’s algorithm applied to our example where $k=4$

Stack={$a, d, b$}
Chaitin’s algorithm applied to our example where $k=4$

Stack={$a,d,b,c$}
Chaitin’s algorithm applied to our example where \( k=4 \)

Stack={a, d, b, c}
Chaitin’s Graph Coloring Example

Chaitin’s algorithm applied to our example where $k=4$

Stack=$\{a,d,b,c,e\}$

Diagram:

```
  f
 /|
/  |
e
```
Chaitin’s algorithm applied to our example where $k=4$

Stack=$\{a,d,b,c,e\}$
Starting assigning colors to \( f, e, b, c, d, a \)
Is Chaitin’s Graph Coloring Optimal?

- According to Chaitin’s algorithm:
  Every node has 3 outgoing edges, thus it is not 3-colorable
Is Chaitin’s Graph Coloring Optimal?

- According to Chaitin’s algorithm: Every node has 3 outgoing edges, thus it is not 3-colorable

- However, it is 3-colorable as you can see above
- Chaitin’s algorithm is not optimal
What if Coloring Fails?

- Spill the variable to memory
  - a spilled variable temporarily \textit{lives} in memory
  - e.g. to color the previous graph using 3 colors
    - spill “f” into memory
What if Coloring Fails?

- Spill the variable to memory
  - a spilled variable temporarily **lives** in memory
  - e.g. to color the previous graph using 3 colors
    - spill “f” into memory
What if Coloring Fails?

- Spill the variable to memory
  - a spilled variable temporarily **lives** in memory
  - e.g. to color the previous graph using 3 colors
    - spill “f” into memory
What if Coloring Fails?

- Spill the variable to memory
  - a spilled variable temporarily \textbf{lives} in memory
  - e.g. to color the previous graph using 3 colors
    - spill “f” into memory
What if Coloring Fails?

- Spill the variable to memory
  - a spilled variable temporarily *lives* in memory
  - e.g. to color the previous graph using 3 colors
    - spill “f” into memory

![Graph diagram]

- f
- e
- d
- c
On-line compilers need to generate binary code quickly
- Just-in-time compilation
- Interactive environments e.g. IDE

In these cases, it is beneficial to sacrifice code performance a bit for quicker compilation
- A faster allocation algorithm
- Not sacrificing too much in code quality

Proposed in following publication:
- Poletto, M., Sarkar, V., "Linear scan register allocation", in ACM Transactions on Programming Languages and Systems (TOPLAS), 1999
Linear Scan Register Allocation

- Layout the code in a certain linear order
- Do a single scan to allocate register for each **live interval**

```plaintext
A
B
C
D
E
```

**scan order**
Layout the code in a certain linear order
Do a single scan to allocate register for each live interval

Live Intervals

Allocate greedily at each numbered point in program
A and D may be allocated to same register
Linear Scan Register Allocation

- Layout the code in a certain linear order
- Do a single scan to allocate register for each live interval

Allocate greedily at each numbered point in program

- A and D may be allocated to same register
Linear Scan and Live Intervals

- **Live Interval**: Smallest range of code containing live ranges

  - Live range of \( a = \{B1, B3\} \), \( b = \{B2, B4\} \)

  - If code layout is “B1,B3,B2,B4”, only 1 register is enough
    - Live interval of \( a = \{B1, B3\} \), \( b = \{B2, B4\} \)

  - If code layout is “B1,B2,B3,B4”, then need 2 registers
    - Live interval of \( a = \{B1, B2, B3\} \), \( b = \{B2, B3, B4\} \)
Linear Scan Algorithm

Linear scan RA consists of four steps

S1. Order all instructions in linear fashion
   - Order affects quality of allocation but not correctness

S2. Calculate the set of live intervals
   - Each variable is given a live interval

S3. Greedily allocate register to each interval in order
   - If a register is available then allocation is possible
   - If a register is not available then an already allocated register is chosen (register spill occurs)

S4. Rewrite the code according to the allocation
   - CPU registers replace temporary or program variables
   - Spill code is generated
Register Allocation Time Comparison

- Usage Counts, Linear Scan, and Graph Coloring shown
- Linear Scan allocation is always faster than Graph Coloring
ILP-based Register Allocation

- Uses linear programming to find the “optimal” register allocation
- Idea and steps:
  1. Convert RA problem to a ILP problem
  2. Solve ILP problem using widely known ILP solvers
  3. Map the ILP solution back to register assignment
- Major problem that restricts its wide adoption
  - ILP problem is NP-hard
  - Solving ILP problem is slow → does not scale well to large programs
What is Integer Linear Programming (ILP)?

- **Integer Linear Programming (ILP)**
  - **Variables:**
    - a, b
  - **Constraints:**
    - \(0 \leq a \leq 10\)
    - \(0 \leq b \leq 29\)
    - \(a + b \leq 36\)
  - **Goal function**
    - minimize \(f(a,b) = 3a + 4b\)

- It is trivial if a and b can take real values
- It is NP hard if a and b can only take integer values
How to Convert Register Allocation to ILP?

☐ An example

\[ \ldots \]
\[ (10) \ldots = b + a ; \]
\[ \ldots \]

➢ Want to know to which register b should be allocated i.e.
\[ \text{load } Rx, \text{ addr}(b) \]

☐ Let us form an ILP problem

➢ assume there are four free registers R1, R2, R3, R4

S1: Define variables

\[ V_{\text{var(location)}}^{R_i} \] — we allocate var at location to Ri

\[ V_{b(10)}^{R_1}, V_{b(10)}^{R_2}, V_{b(10)}^{R_3}, V_{b(10)}^{R_4} \]
Converting Register Allocation to ILP

**S2:** Constraints: clearly there are constraints for these variables

- \( V_{var(location)}^{R_i} \) only takes value 0 or 1
  - 0 — not allocate to that register at the place
  - 1 — is allocated to that register at the place

- Any register can hold only one variable at any place
  \( V_{b(10)}^{R_1} + V_{a(10)}^{R_1} \leq 1 \)

- Any variable just needs to take one register
  \( V_{b(10)}^{R_1} + V_{b(10)}^{R_2} + V_{b(10)}^{R_3} + V_{b(10)}^{R_4} = 1 \)

- and many more ...

**S3:** Define goal function

- to minimize memory operations
  \( f_{cost} = (\sum V_{v(mem.p)}^{R_i}) \times \text{LOAD}_{cost} + \ldots \) (store cost)
Conclusion

- Good Register Allocation is crucial to code quality
  - Accesses to memory are costly, even with caches
  - Even with only a handful of program variables, intermediate values introduce many more temporary variables adding to register pressure

- Different algorithms make different trade-offs between allocation time (compiler performance) and code quality (application performance)
Instruction Selection
Instruction selection is the task to select appropriate machine instructions to implement the operations in the intermediate representation (IR).

- Very important for CISC machines, and machines with special purpose instructions (MMX)
- X86, ARM, DSP, ...

There are many semantically equivalent instruction sequences

- How to find the “minimal cost” sequence?
### Some Instruction Patterns

<table>
<thead>
<tr>
<th>Name</th>
<th>Effect</th>
<th>Trees</th>
</tr>
</thead>
<tbody>
<tr>
<td>ADD</td>
<td>( d_i \leftarrow d_j + d_k )</td>
<td>( \text{d} + )</td>
</tr>
<tr>
<td>MUL</td>
<td>( d_i \leftarrow d_j \times d_k )</td>
<td>( \text{d} \times )</td>
</tr>
<tr>
<td>SUB</td>
<td>( d_i \leftarrow d_j - d_k )</td>
<td>( \text{d} - )</td>
</tr>
<tr>
<td>DIV</td>
<td>( d_i \leftarrow d_j / d_k )</td>
<td>( \text{d} \div )</td>
</tr>
<tr>
<td>ADDI</td>
<td>( d_i \leftarrow d_j + c )</td>
<td>( \text{d} + ) ( \text{CONST} ) ( \text{d} )</td>
</tr>
<tr>
<td>SUBI</td>
<td>( d_i \leftarrow d_j - c )</td>
<td>( \text{d} ) ( \text{CONST} )</td>
</tr>
<tr>
<td>MOVEA</td>
<td>( d_j \leftarrow a_i )</td>
<td>( \text{d} ) ( \text{a} )</td>
</tr>
<tr>
<td>MOVED</td>
<td>( a_j \leftarrow d_i )</td>
<td>( \text{a} ) ( \text{d} )</td>
</tr>
<tr>
<td>LOAD</td>
<td>( d_i \leftarrow M[a_j + c] )</td>
<td>( \text{d} ) ( \text{MEM} ) ( \text{CONST} ) ( \text{a} )</td>
</tr>
<tr>
<td>STORE</td>
<td>( M[a_j + c] \leftarrow d_i )</td>
<td>( \text{MOVE} ) ( \text{MEM} ) ( \text{d} ) ( \text{MEM} ) ( \text{CONST} ) ( \text{d} )</td>
</tr>
<tr>
<td>MOVEM</td>
<td>( M[a_j] \leftarrow M[a_i] )</td>
<td>( \text{MOVE} ) ( \text{MEM} ) ( \text{MEM} )</td>
</tr>
</tbody>
</table>
A Parse Tree to be Tiled

```
MOVE
  /\       /\         
MEM  MEM  MEM
  |     |     |       
+     +     +       
  /\   /\   /\       
MEM  MEM  MEM
  |     |     |       
+     +     +       
      /\     /\     
  temp_i const const
      /\     /\       
FP  const FP  const
```
A Parse Tree to be Tiled

```
    MOVE
     /\  \
    MEM MEM
   /   \
  +     +
 /     /
MEM   MEM
 /     \
+     *
/     /
MEM temp_i
/     /
+     const
FP const
```
A Parse Tree to be Tiled

MOVE

MEM

MEM

+

*

FP

const

2: load R, M[fp+a]

temp_i

const

1:

FP

const
A Parse Tree to be Tiled

1: FP

2: load R, M[fp+a]

3: MEM

MOVE

MEM

MEM

+ +

FP

const

const

*
A Parse Tree to be Tiled

MOVE

MEM

MEM

+

+

*

FP

const

1: load R, M[fp+a]

2: load R, M[fp+a]

3: addi R2, R0+4

4: addi R2, R0+4
A Parse Tree to be Tiled

1: \text{const} \quad 2: \text{load } R, M[fp+a] \quad 3: \text{add} R2, R0+4 \quad 4: \text{addi } R2, R0+4 \quad 5: \text{mul } R2, \text{R}\times R2

\text{MEM} \quad \text{MEM}
A Parse Tree to be Tiled

1: \text{const} 
2: \text{load R, M[fp+a]} 
3: \text{temp} 
4: \text{addi R2, R0+4} 
5: \text{mul R2, Ri} \times \text{R2} 
6: \text{add R1, R1+R2} 

MOVE

MEM

MEM

MEM

FP

const

const

\text{temp}
A Parse Tree to be Tiled

1: FP
2: load R, M[fp+a] temp;
3: const
4: addi R2, R0+4 const
5: mul R2, Ri \times R2
6: add R1, R1+R2
7: MEM

MOVE

MEM -> MEM
A Parse Tree to be Tiled

1: const

2: load R, M[fp+a]

3: temp

4: addi R2, R0+4

5: mul R2, Ri \times R2

6: add R1, R1+R2

7: 

8: addi R2, fp+x

MEM

MEM

MOVE

FP

const

const
A Parse Tree to be Tiled

1: load R, M[fp+a]
2: add R1, R1+R2
3: temp
4: addi R2, R0+4
5: mul R2, Ri × R2
6: add R1, R1+R2
7: 
8: addi R2, fp+x
9: MoveMem M[R1], M[R2]
A Parse Tree to be Tiled

1: const
2: load R, M[fp+a]
3: const
4: addi R2, R0+4
5: mul R2, Ri \times R2
6: add R1, R1+R2
7: +

MOVE

MEM

MEM
A Parse Tree to be Tiled

1: \text{const}

2: \text{load } R, M[fp+a]

3: \text{temp;}

4: addi R2, R0+4

5: \text{mul } R2, Ri \times R2

6: add R1, R1+R2

7: \text{temp;}

8: \text{load } R2, M[fp+x]

MEM

MOVE

MEM

MEM
A Parse Tree to be Tiled

1: const
2: load R, M[fp+a]
3: temp
4: addi R2, R0+4
5: mul R2, Ri \times R2
6: add R1, R1+R2
7: 
8: load R2, M[fp+x]
9: store M[R1+0], R2
The END !