Parallel Programming & Cluster Computing

Stupid Compiler Tricks

Henry Neeman, University of Oklahoma
Paul Gray, University of Northern Iowa

SC08 Education Program’s Workshop on Parallel & Cluster Computing
Oklahoma Supercomputing Symposium, Monday October 6 2008
Outline

- Dependency Analysis
  - What is Dependency Analysis?
  - Control Dependencies
  - Data Dependencies
- Stupid Compiler Tricks
  - Tricks the Compiler Plays
  - Tricks You Play With the Compiler
  - Profiling
Dependency Analysis
What Is Dependency Analysis?

Dependency analysis describes how different parts of a program affect one another, and how various parts require other parts in order to operate correctly.

A control dependency governs how different sequences of instructions affect each other.

A data dependency governs how different pieces of data affect each other.

Much of this discussion is from references [1] and [5].
Control Dependencies

Every program has a well-defined *flow of control* that moves from instruction to instruction.

This flow can be affected by several kinds of operations:

- Loops
- Branches (if, select case/switch)
- Function/subroutine calls
- I/O (typically implemented as calls)

Dependencies affect **parallelization**!
Branch Dependency

\[ y = 7 \]
\[
{\text{IF}} \ (x \neq 0) \ {\text{THEN}} \\
\quad y = 1.0 / x \\
{\text{END IF}}
\]

Note that \((x \neq 0)\) means “\(x\) not equal to zero.”

The value of \(y\) depends on what the condition \((x \neq 0)\) evaluates to:

- If the condition \((x \neq 0)\) evaluates to \(\text{.TRUE.}\), then \(y\) is set to \(1.0 / x\). (1 divided by \(x\)).
- Otherwise, \(y\) remains 7.
Loop Carried Dependency

DO i = 2, length
  a(i) = a(i-1) + b(i)
END DO

Here, each iteration of the loop depends on the previous:
  iteration i=3 depends on iteration i=2,
  iteration i=4 depends on iteration i=3,
  iteration i=5 depends on iteration i=4, etc.

This is sometimes called a loop carried dependency.
There is no way to execute iteration i until after iteration i-1 has completed, so this loop can’t be parallelized.
Why Do We Care?

**Loops** are the favorite control structures of High Performance Computing, because compilers know how to **optimize** their performance using instruction-level parallelism: superscalar, pipelining and vectorization can give excellent speedup.

**Loop carried dependencies** affect whether a loop can be parallelized, and how much.
Loop or Branch Dependency?

Is this a **loop carried dependency** or a **branch dependency**?

```plaintext
DO i = 1, length
   IF (x(i) /= 0) THEN
      y(i) = 1.0 / x(i)
   END IF
END DO
```
Call Dependency Example

\[ x = 5 \]
\[ y = \text{myfunction}(7) \]
\[ z = 22 \]

The flow of the program is interrupted by the call to \textit{myfunction}, which takes the execution to somewhere else in the program.

It’s similar to a branch dependency.
I/O Dependency

\[ X = a + b \]
\[ PRINT *, x \]
\[ Y = c + d \]

Typically, I/O is implemented by hidden subroutine calls, so we can think of this as equivalent to a call dependency.
Reductions Aren’t Dependencies

array_sum = 0
DO i = 1, length
  array_sum = array_sum + array(i)
END DO

A **reduction** is an operation that converts an array to a scalar.
Other kinds of reductions: product, **.AND.**, **.OR.**, minimum, maximum, index of minimum, index of maximum, number of occurrences of a particular value, etc.

Reductions are so common that hardware and compilers are optimized to handle them.

Also, they aren’t really dependencies, because the order in which the individual operations are performed doesn’t matter.
Data Dependencies

“A data dependence occurs when an instruction is dependent on data from a previous instruction and therefore cannot be moved before the earlier instruction [or executed in parallel].” [6]

\[ a = x + y + \cos(z) ; \]

\[ b = a * c ; \]

The value of \( b \) depends on the value of \( a \), so these two statements must be executed in order.
Output Dependencies

\[
x = a / b;
\]
\[
y = x + 2;
\]
\[
x = d - e;
\]

Notice that \textbf{x} is assigned \textbf{two different values}, but only one of them is retained after these statements are done executing. In this context, the final value of \textbf{x} is the “output.”

Again, we are forced to execute in order.
Why Does Order Matter?

- Dependencies can affect whether we can execute a particular part of the program in **parallel**.
- If we cannot execute that part of the program in parallel, then it’ll be **SLOW**.
Loop Dependency Example

```c
if ((dst == src1) && (dst == src2)) {
    for (index = 1; index < length; index++) {
        dst[index] = dst[index-1] + dst[index];
    }
}
else if (dst == src1) {
    for (index = 1; index < length; index++) {
        dst[index] = dst[index-1] + src2[index];
    }
}
else if (dst == src2) {
    for (index = 1; index < length; index++) {
        dst[index] = src1[index-1] + dst[index];
    }
}
else if (src1 == src2) {
    for (index = 1; index < length; index++) {
        dst[index] = src1[index-1] + src1[index];
    }
}
else {
    for (index = 1; index < length; index++) {
        dst[index] = src1[index-1] + src2[index];
    }
}
```
Loop Dep Example (cont’d)

if ((dst == src1) && (dst == src2)) {
    for (index = 1; index < length; index++) {
        dst[index] = dst[index-1] + dst[index];
    }
} 
else if (dst == src1) {
    for (index = 1; index < length; index++) {
        dst[index] = dst[index-1] + src2[index];
    }
} 
else if (dst == src2) {
    for (index = 1; index < length; index++) {
        dst[index] = src1[index-1] + dst[index];
    }
} 
else if (src1 == src2) {
    for (index = 1; index < length; index++) {
        dst[index] = src1[index-1] + src1[index];
    }
} 
else {
    for (index = 1; index < length; index++) {
        dst[index] = src1[index-1] + src2[index];
    }
}

The various versions of the loop either:

- **do have loop carried dependencies**, or
- **don’t have loop carried dependencies**.

SC08 Parallel & Cluster Computing: Stupid Compiler Tricks  
Oklahoma Supercomputing Symposium, October 6 2008
Loop Dependency Performance

Loop Carried Dependency Performance

Better

MFLOPs

dst=src1+src2 dst=src1+src1 dst=dst+src2 dst=src1+dst dst=dst+dst

Pentium3 500 MHz
POWER4
Pentium4 2GHz
EM64T 3.2 GHz

SC08 Parallel & Cluster Computing: Stupid Compiler Tricks
Oklahoma Supercomputing Symposium, October 6 2008
Stupid Compiler Tricks
Stupid Compiler Tricks

- Tricks Compilers Play
  - Scalar Optimizations
  - Loop Optimizations
  - Inlining
- Tricks You Can Play with Compilers
  - Profiling
  - Hardware counters
The people who design compilers have a lot of experience working with the languages commonly used in High Performance Computing:

- Fortran: 45ish years
- C: 30ish years
- C++: 15ish years, plus C experience

So, they’ve come up with clever ways to make programs run faster.
Tricks Compilers Play
Scalar Optimizations

- Copy Propagation
- Constant Folding
- Dead Code Removal
- Strength Reduction
- Common Subexpression Elimination
- Variable Renaming
- Loop Optimizations

Not every compiler does all of these, so it sometimes can be worth doing these by hand.

Much of this discussion is from [2] and [5].
Copy Propagation

Before

\[ x = y \]
\[ z = 1 + x \]

Has data dependency

Compile

After

\[ x = y \]
\[ z = 1 + y \]

No data dependency
Constant Folding

Before

\[
\begin{align*}
add &= 100 \\
aug &= 200 \\
sum &= add + aug
\end{align*}
\]

After

\[
sum = 300
\]

Notice that \textit{sum} is actually the sum of two constants, so the compiler can precalculate it, eliminating the addition that otherwise would be performed at runtime.
Dead Code Removal

**Before**

```plaintext
var = 5
PRINT *, var
STOP
PRINT *, var * 2
```

**After**

```plaintext
var = 5
PRINT *, var
STOP
```

Since the last statement never executes, the compiler can eliminate it.
Strength Reduction

**Before**

\[
\begin{align*}
x &= y^{**2.0} \\
a &= c / 2.0 
\end{align*}
\]

**After**

\[
\begin{align*}
x &= y \ast y \\
a &= c \ast 0.5 
\end{align*}
\]

Raising one value to the power of another, or dividing, is more expensive than multiplying. If the compiler can tell that the power is a small integer, or that the denominator is a constant, it’ll use multiplication instead.

Note: In Fortran, \texttt{“y ** 2.0”} means “\textit{y to the power 2}.”
Common Subexpression Elimination

Before

d = c * (a / b)
e = (a / b) * 2.0

After

adivb = a / b
d = c * adivb
e = adivb * 2.0

The subexpression \((a / b)\) occurs in both assignment statements, so there’s no point in calculating it twice.
This is typically only worth doing if the common subexpression is expensive to calculate.
Variable Renaming

Before
\[
\begin{align*}
  x &= y \times z \\
  q &= r + x \times 2 \\
  x &= a + b
\end{align*}
\]

After
\[
\begin{align*}
  x0 &= y \times z \\
  q &= r + x0 \times 2 \\
  x &= a + b
\end{align*}
\]

The original code has an output dependency, while the new code doesn’t – but the final value of \( x \) is still correct.
Loop Optimizations

- Hoisting Loop Invariant Code
- Unswitching
- Iteration Peeling
- Index Set Splitting
- Loop Interchange
- Unrolling
- Loop Fusion
- Loop Fission

Not every compiler does all of these, so it sometimes can be worth doing some of these by hand.

Much of this discussion is from [3] and [5].
Hoisting Loop Invariant Code

Code that doesn’t change inside the loop is called **loop invariant**. It doesn’t need to be calculated over and over.

**Before**

\[
\begin{align*}
\text{DO } & i = 1, n \\
a(i) &= b(i) + c \times d \\
e &= g(n) \\
\text{END DO}
\end{align*}
\]

**After**

\[
\begin{align*}
\text{temp} &= c \times d \\
\text{DO } & i = 1, n \\
a(i) &= b(i) + \text{temp} \\
\text{END DO} \\
e &= g(n)
\end{align*}
\]
Unswitching

Before

DO i = 1, n
  DO j = 2, n
    IF (t(i) > 0) THEN
      a(i,j) = a(i,j) * t(i) + b(j)
    ELSE
      a(i,j) = 0.0
    END IF
  END DO
END DO

So, it can migrate outside the j loop.

After

DO i = 1, n
  IF (t(i) > 0) THEN
    DO j = 2, n
      a(i,j) = a(i,j) * t(i) + b(j)
    END DO
  ELSE
    DO j = 2, n
      a(i,j) = 0.0
    END DO
  END IF
END DO

The condition is j-independent.
Iteration Peeling

DO i = 1, n
   IF ((i == 1) .OR. (i == n)) THEN
      x(i) = y(i)
   ELSE
      x(i) = y(i + 1) + y(i - 1)
   END IF
END DO

We can eliminate the IF by *peeling* the weird iterations.

x(1) = y(1)
DO i = 2, n - 1
   x(i) = y(i + 1) + y(i - 1)
END DO
x(n) = y(n)
Index Set Splitting

Before

\[
\text{DO } i = 1, n \\
a(i) = b(i) + c(i) \\
\text{IF } (i > 10) \text{ THEN} \\
d(i) = a(i) + b(i - 10) \\
\text{END IF} \\
\text{END DO}
\]

After

\[
\text{DO } i = 1, 10 \\
a(i) = b(i) + c(i) \\
\text{END DO} \\
\text{DO } i = 11, n \\
a(i) = b(i) + c(i) \\
d(i) = a(i) + b(i - 10) \\
\text{END DO}
\]

Note that this is a generalization of \textit{peeling}. 
Loop Interchange

**Before**

```plaintext
DO i = 1, ni
  DO j = 1, nj
    a(i,j) = b(i,j)
  END DO
END DO
```

**After**

```plaintext
DO j = 1, nj
  DO i = 1, ni
    a(i,j) = b(i,j)
  END DO
END DO
```

Array elements \(a(i,j)\) and \(a(i+1,j)\) are near each other in memory, while \(a(i,j+1)\) may be far, so it makes sense to make the \(i\) loop be the inner loop. (This is reversed in C, C++ and Java.)
Unrolling

**Before**

\[
\begin{align*}
\text{DO } i &= 1, n \\
\text{a}(i) &= \text{a}(i) + \text{b}(i) \\
\text{END DO}
\end{align*}
\]

**After**

\[
\begin{align*}
\text{DO } i &= 1, n, 4 \\
a(i) &= a(i) + b(i) \\
a(i+1) &= a(i+1) + b(i+1) \\
a(i+2) &= a(i+2) + b(i+2) \\
a(i+3) &= a(i+3) + b(i+3) \\
\text{END DO}
\end{align*}
\]

You generally shouldn’t unroll by hand.
Why Do Compilers Unroll?

We saw last time that a loop with a lot of operations gets better performance (up to some point), especially if there are lots of arithmetic operations but few main memory loads and stores.

Unrolling creates multiple operations that typically load from the same, or adjacent, cache lines.

So, an unrolled loop has more operations without increasing the memory accesses by much.

Also, unrolling decreases the number of comparisons on the loop counter variable, and the number of branches to the top of the loop.
Loop Fusion

Before

\[
\text{DO } i = 1, n \\
\quad a(i) = b(i) + 1 \\
\text{END DO} \\
\text{DO } i = 1, n \\
\quad c(i) = a(i) / 2 \\
\text{END DO} \\
\text{DO } i = 1, n \\
\quad d(i) = 1 / c(i) \\
\text{END DO}
\]

After

\[
\text{DO } i = 1, n \\
\quad a(i) = b(i) + 1 \\
\quad c(i) = a(i) / 2 \\
\quad d(i) = 1 / c(i) \\
\text{END DO}
\]

As with unrolling, this has fewer branches. It also has fewer total memory references.
Loop Fission

DO i = 1, n
  a(i) = b(i) + 1
  c(i) = a(i) / 2
  d(i) = 1 / c(i)
END DO !! i = 1, n

DO i = 1, n
  a(i) = b(i) + 1
END DO !! i = 1, n
DO i = 1, n
  c(i) = a(i) / 2
END DO !! i = 1, n
DO i = 1, n
  d(i) = 1 / c(i)
END DO !! i = 1, n

Fission reduces the cache footprint and the number of operations per iteration.
To Fuse or to Fizz?

The question of when to perform fusion versus when to perform fission, like many many optimization questions, is highly dependent on the application, the platform and a lot of other issues that get very, very complicated. Compilers don’t always make the right choices. That’s why it’s important to examine the actual behavior of the executable.
Inlining

Before

DO i = 1, n  
a(i) = func(i)  
END DO
...
REAL FUNCTION func (x)
...
  func = x * 3  
END FUNCTION func

After

DO i = 1, n  
a(i) = i * 3  
END DO

When a function or subroutine is **inlined**, its contents are transferred directly into the calling routine, eliminating the overhead of making the call.
Tricks You Can Play with Compilers
The Joy of Compiler Options

Every compiler has a different set of options that you can set. Among these are options that control single processor optimization: superscalar, pipelining, vectorization, scalar optimizations, loop optimizations, inlining and so on.
Example Compile Lines

- **IBM XL**
  
  ```
  xlf90 -O -qmaxmem=-1 -qarch=auto
  -qtune=auto -qcache=auto -qhot
  ```

- **Intel**
  
  ```
  ifort -O -march=core2 -mtune=core2
  ```

- **Portland Group f90**
  
  ```
  pgf90 -O3 -fastsse -tp core2-64
  ```

- **NAG f95**
  
  ```
  f95 -O4 -Ounsafe -ieee=nonstd
  ```
What Does the Compiler Do?

Example: NAG f95 compiler

\texttt{f95 -O<level> source.f90}

Possible levels are \texttt{-O0}, \texttt{-O1}, \texttt{-O2}, \texttt{-O3}, \texttt{-O4}:

\begin{itemize}
  \item \texttt{-O0} \hspace{1cm} No optimisation. …
  \item \texttt{-O1} \hspace{1cm} Minimal quick optimisation.
  \item \texttt{-O2} \hspace{1cm} Normal optimisation.
  \item \texttt{-O3} \hspace{1cm} Further optimisation.
  \item \texttt{-O4} \hspace{1cm} Maximal optimisation.\textsuperscript{[4]}
\end{itemize}

The man page is pretty cryptic.
Arithmetic Operation Speeds

Ordered Arithmetic Operations

SC08 Parallel & Cluster Computing: Stupid Compiler Tricks
Oklahoma Supercomputing Symposium, October 6 2008
Optimization Performance

![Graph showing performance comparison between different operations for Pentium3 NAG O0, Pentium3 NAG O4, Pentium3 Vast no opt, and Pentium3 Vast opt.]
More Optimized Performance

Performance

<table>
<thead>
<tr>
<th>Operation</th>
<th>MFLOP/s</th>
</tr>
</thead>
<tbody>
<tr>
<td>rnam</td>
<td></td>
</tr>
<tr>
<td>imam</td>
<td></td>
</tr>
<tr>
<td>rmad</td>
<td></td>
</tr>
<tr>
<td>imad</td>
<td></td>
</tr>
<tr>
<td>rdot</td>
<td></td>
</tr>
<tr>
<td>reuc</td>
<td></td>
</tr>
<tr>
<td>rlot8</td>
<td></td>
</tr>
<tr>
<td>rlot10</td>
<td></td>
</tr>
<tr>
<td>rlot12</td>
<td></td>
</tr>
<tr>
<td>rlot16</td>
<td></td>
</tr>
<tr>
<td>rlot20</td>
<td></td>
</tr>
<tr>
<td>rlot24</td>
<td></td>
</tr>
</tbody>
</table>

- Pentium3 NAG O0
- Pentium3 NAG 04
- Pentium3 VAST no opt
- Pentium3 VAST opt
Profiling
Profiling

Profiling means collecting data about how a program executes. The two major kinds of profiling are:

- Subroutine profiling
- Hardware timing
Subroutine Profiling

Subroutine profiling means finding out how much time is spent in each routine.

The 90-10 Rule: Typically, a program spends 90% of its runtime in 10% of the code.

Subroutine profiling tells you what parts of the program to spend time optimizing and what parts you can ignore.

Specifically, at regular intervals (e.g., every millisecond), the program takes note of what instruction it’s currently on.
Profiling Example

On IBM pSeries systems:

\texttt{xlf90 -O -g -pg ...}

The \texttt{-g -pg} options tell the compiler to set the executable up to collect profiling information.

Running the executable generates a file named \texttt{gmon.out}, which contains the profiling information.
Profiling Example (cont’d)

When the run has completed, a file named `gmon.out` has been generated.

Then:

```
gprof executable
```

produces a list of all of the routines and how much time was spent in each.
### Profiling Result

<table>
<thead>
<tr>
<th>% cumulative</th>
<th>time</th>
<th>cumulative seconds</th>
<th>self seconds</th>
<th>calls</th>
<th>self ms/call</th>
<th>total ms/call</th>
<th>name</th>
</tr>
</thead>
<tbody>
<tr>
<td>27.6</td>
<td>52.72</td>
<td>52.72</td>
<td>480000</td>
<td>0.11</td>
<td>0.11</td>
<td>longwave_ [5]</td>
<td></td>
</tr>
<tr>
<td>24.3</td>
<td>99.06</td>
<td>46.35</td>
<td>897</td>
<td>51.67</td>
<td>51.67</td>
<td>mpdata3_ [8]</td>
<td></td>
</tr>
<tr>
<td>7.9</td>
<td>114.19</td>
<td>15.13</td>
<td>300</td>
<td>50.43</td>
<td>50.43</td>
<td>turb_ [9]</td>
<td></td>
</tr>
<tr>
<td>7.2</td>
<td>127.94</td>
<td>13.75</td>
<td>299</td>
<td>45.98</td>
<td>45.98</td>
<td>turb_scalar_ [10]</td>
<td></td>
</tr>
<tr>
<td>4.7</td>
<td>136.91</td>
<td>8.96</td>
<td>300</td>
<td>29.88</td>
<td>29.88</td>
<td>advect2_z_ [12]</td>
<td></td>
</tr>
<tr>
<td>4.1</td>
<td>144.79</td>
<td>7.88</td>
<td>300</td>
<td>26.27</td>
<td>31.52</td>
<td>cloud_ [11]</td>
<td></td>
</tr>
<tr>
<td>3.9</td>
<td>152.22</td>
<td>7.43</td>
<td>300</td>
<td>24.77</td>
<td>212.36</td>
<td>radiation_ [3]</td>
<td></td>
</tr>
<tr>
<td>2.3</td>
<td>156.65</td>
<td>4.43</td>
<td>897</td>
<td>4.94</td>
<td>56.61</td>
<td>smlr_ [7]</td>
<td></td>
</tr>
<tr>
<td>2.2</td>
<td>160.77</td>
<td>4.12</td>
<td>300</td>
<td>13.73</td>
<td>24.39</td>
<td>tke_full_ [13]</td>
<td></td>
</tr>
<tr>
<td>1.7</td>
<td>163.97</td>
<td>3.20</td>
<td>300</td>
<td>10.66</td>
<td>10.66</td>
<td>shear_prod_ [15]</td>
<td></td>
</tr>
<tr>
<td>1.5</td>
<td>166.79</td>
<td>2.82</td>
<td>300</td>
<td>9.40</td>
<td>9.40</td>
<td>rhs_ [16]</td>
<td></td>
</tr>
<tr>
<td>1.4</td>
<td>169.53</td>
<td>2.74</td>
<td>300</td>
<td>9.13</td>
<td>9.13</td>
<td>advect2_xy_ [17]</td>
<td></td>
</tr>
<tr>
<td>1.3</td>
<td>172.00</td>
<td>2.47</td>
<td>300</td>
<td>8.23</td>
<td>15.33</td>
<td>poisson_ [14]</td>
<td></td>
</tr>
<tr>
<td>1.2</td>
<td>174.27</td>
<td>2.27</td>
<td>480000</td>
<td>0.00</td>
<td>0.12</td>
<td>long_wave_ [4]</td>
<td></td>
</tr>
<tr>
<td>1.0</td>
<td>176.13</td>
<td>1.86</td>
<td>299</td>
<td>6.22</td>
<td>177.45</td>
<td>advect_scalar_ [6]</td>
<td></td>
</tr>
<tr>
<td>0.9</td>
<td>177.94</td>
<td>1.81</td>
<td>300</td>
<td>6.04</td>
<td>6.04</td>
<td>buoy_ [19]</td>
<td></td>
</tr>
</tbody>
</table>

...
To Learn More Supercomputing

http://www.oscer.ou.edu/education.php
Thanks for your attention!

Questions?
References