Static Compiler Optimization Techniques

• We examined the following static ISA/compiler techniques aimed at improving pipelined CPU performance:
  – Static pipeline scheduling.
  – Loop unrolling.
  – Static branch prediction.
  – Static multiple instruction issue: VLIW.
  – Conditional or predicted instructions/predication.
  – Static speculation

• Here we examine two additional static compiler-based techniques:

1. **Loop-Level Parallelism (LLP) analysis:**
   - Detecting and enhancing loop iteration parallelism
     - Greatest Common Divisor (GCD) test.

2. **Software pipelining (Symbolic loop unrolling).**

• In addition a brief introduction to vector processing (Appendix G) is included to emphasize the importance/origin of LLP analysis.

(3rd Edition: Chapter 4.4, vector processing: Appendix G)
Data Parallelism & Loop Level Parallelism (LLP)

- **Data Parallelism**: Similar independent/parallel computations on different elements of arrays that usually result in independent (or parallel) loop iterations when such computations are implemented as sequential programs.

- A common way to increase parallelism among instructions is to exploit data parallelism among independent iterations of a loop (e.g., exploit Loop Level Parallelism, LLP).
  - One method covered earlier to accomplish this is by **unrolling the loop** either statically by the compiler, or dynamically by hardware, which increases the size of the basic block present. This resulting larger basic block provides more instructions that can be scheduled or re-ordered by the compiler/hardware to eliminate more stall cycles.

- The following loop has parallel loop iterations since computations in each iteration are data parallel and are performed on different elements of the arrays.

  ```c
  for (i=1; i<=1000; i=i+1;)
  x[i] = x[i] + y[i];
  ```

- In supercomputing applications, data parallelism/LLP has been traditionally exploited by vector ISAs/processors, utilizing vector instructions
  - Vector instructions operate on a number of data items (vectors) producing a vector of elements not just a single result value. The above loop might require just four such instructions.
Loop Unrolling Example

When scheduled for pipeline

Loop:  
L.D  F0, 0(R1)  
L.D  F6,-8 (R1)  
L.D  F10, -16(R1)  
L.D  F14, -24(R1)  
ADD.D  F4, F0, F2  
ADD.D  F8, F6, F2  
ADD.D  F12, F10, F2  
ADD.D  F16, F14, F2  
S.D  F4, 0(R1)  
S.D  F8, -8(R1)  
DADDUI  R1, R1,# -32  
S.D  F12, 16(R1),F12  
BNE  R1,R2, Loop  
S.D  F16, 8(R1), F16  ;8-32 = -24

Note: Independent Loop Iterations Resulting from data parallel operations on elements of array X

The execution time of the loop has dropped to 14 cycles, or 14/4 = 3.5 clock cycles per element compared to 7 before scheduling and 6 when scheduled but unrolled.

Unrolling the loop exposed more computations that can be scheduled to minimize stalls by increasing the size of the basic block from 5 instructions in the original loop to 14 instructions in the unrolled loop.

Loop unrolling exploits data parallelism among independent iterations of a loop

Larger Basic Block → More ILP

The execution time of the loop has dropped to 14 cycles, or 14/4 = 3.5 clock cycles per element compared to 7 before scheduling and 6 when scheduled but unrolled.

Speedup = 6/3.5 = 1.7

Unrolling the loop exposed more computations that can be scheduled to minimize stalls by increasing the size of the basic block from 5 instructions in the original loop to 14 instructions in the unrolled loop.

Loop unrolling exploits data parallelism among independent iterations of a loop

for (i=1000; i>0; i=i-1)

x[i] = x[i] + s;

Note: Independent Loop Iterations Resulting from data parallel operations on elements of array X

Usually: Data Parallelism → LLP

i.e more ILP exposed

Larger Basic Block → More ILP

Exposed

EECC551 - Shaaban
Loop-Level Parallelism (LLP) Analysis

• Loop-Level Parallelism (LLP) analysis focuses on whether data accesses in later iterations of a loop are data dependent on data values produced in earlier iterations and possibly making loop iterations independent (parallel).

  e.g. in for (i=1; i <= 1000; i++)

  \[
  x[i] = x[i] + s;
  \]

  the computation in each iteration is independent of the previous iterations and the loop is thus parallel. The use of \( X[i] \) twice is within a single iteration.

  \[\Rightarrow \text{Thus loop iterations are parallel (or independent from each other).}\]

• Loop-carried Data Dependence: A data dependence between different loop iterations (data produced in an earlier iteration used in a later one).

• Not Loop-carried Data Dependence: Data dependence within the same loop iteration.

  • LLP analysis is important in software optimizations such as loop unrolling since it usually requires loop iterations to be independent (and in vector processing).
  
  • LLP analysis is normally done at the source code level or close to it since assembly language and target machine code generation introduces loop-carried name dependence in the registers used in the loop.
        
        – Instruction level parallelism (ILP) analysis, on the other hand, is usually done when instructions are generated by the compiler.

LLP Analysis Example 1

- In the loop:

```c
for (i=1; i<=100; i=i+1) {
    A[i+1] = A[i] + C[i]; /* S1 */
    B[i+1] = B[i] + A[i+1];} /* S2 */
```

(Where A, B, C are distinct non-overlapping arrays)

- **S2** uses the value A[i+1], computed by S1 in the same iteration. This data dependence is within the same iteration (not a loop-carried data dependence).
  
  i.e. S1 → S2 on A[i+1] Not loop-carried data dependence

  ⇒ does not prevent loop iteration parallelism.

- **S1** uses a value computed by S1 in the earlier iteration, since iteration i computes A[i+1] read in iteration i+1 (loop-carried dependence, prevents parallelism). The same applies for **S2** for B[i] and B[i+1]
  
  i.e. S1 → S1 on A[i] Loop-carried data dependence
  S2 → S2 on B[i] Loop-carried data dependence

  ⇒ These two data dependencies are loop-carried spanning more than one iteration (two iterations) preventing loop parallelism.

In this example the loop carried dependencies form two dependency chains starting from the very first iteration and ending at the last iteration.
LLP Analysis Example 2

- In the loop:
  
  \[
  \text{for (i=1; i<=100; i=i+1) } \{
  \begin{align*}
  A[i] &= A[i] + B[i]; \quad /* S1 */ \\
  B[i+1] &= C[i] + D[i]; \quad /* S2 */
  \end{align*}
  \}
  \]

  - S1 uses the value B[i] computed by S2 in the previous iteration (loop-carried dependence)
  
  - This dependence is not circular:
    - S1 depends on S2 but S2 does not depend on S1.
  
  - Can be made parallel by replacing the code with the following:
    
    \[
    \begin{align*}
    &\text{for (i=1; i<=99; i=i+1) } \{
    &B[i+1] = C[i] + D[i]; \\
    &A[i+1] = A[i+1] + B[i+1]; \\
    \}
    \end{align*}
    \]

    - Loop Start-up code

    - Parallel loop iterations (data parallelism in computation exposed in loop code)

    - Loop Completion code

And does not form a data dependence chain
LLP Analysis Example 2

Original Loop:

```
for (i=1; i<=100; i=i+1) {
    A[i] = A[i] + B[i]; /* S1 */
    B[i+1] = C[i] + D[i]; /* S2 */
}
```

Modified Parallel Loop:

(One less iteration)

```
for (i=1; i<=99; i=i+1) {
    B[i+1] = C[i] + D[i];
    A[i+1] = A[i+1] + B[i+1];
}
B[101] = C[100] + D[100];
```

Loop Start-up code

```
```
ILP Compiler Support: Loop-Carried Dependence Detection

To detect loop-carried dependence in a loop, the Greatest Common Divisor (GCD) test can be used by the compiler, which is based on the following:

- If an array element with index: \( a \times i + b \) is stored and element: \( c \times i + d \) of the same array is loaded later where index runs from \( m \) to \( n \), a dependence exists if the following two conditions hold:
  
  1. There are two iteration indices, \( j \) and \( k \), \( m \leq j, k \leq n \) (within iteration limits)
  2. The loop stores into an array element indexed by:
     
     \[ a \times j + b \]
     
     Produce or write (store) element with this Index
     
     and later loads from the same array the element indexed by:
     
     \[ c \times k + d \]
     
     Later read (load) element with this index

Thus:

\[ a \times j + b = c \times k + d \]

Index of element written (stored) earlier

Index of element read (loaded) later

Here \( a, b, c, d \) are constants

For access to elements of an array
The Greatest Common Divisor (GCD) Test

- If a loop carried dependence exists, then:

  \[ \text{GCD}(c, a) \text{ must divide } (d-b) \]

The GCD test is sufficient to guarantee no loop carried dependence.

However, there are cases where GCD test succeeds but no dependence exits because GCD test does not take loop bounds into account.

Example:

```c
for(i=1; i<=100; i=i+1) {
    x[2*i+3] = x[2*i] * 5.0;
}
```

\[ a = 2 \quad b = 3 \quad c = 2 \quad d = 0 \]

\[ \text{GCD}(a, c) = 2 \]

\[ d - b = -3 \]

\[ 2 \text{ does not divide } -3 \Rightarrow \text{No loop carried dependence possible.} \]
Showing Example Loop Iterations to Be Independent

```c
for(i=1; i<=100; i=i+1) {
    x[2*i+3] = x[2*i] * 5.0;
}
```

<table>
<thead>
<tr>
<th>Iteration i</th>
<th>Index of x loaded</th>
<th>Index of x stored</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>5</td>
</tr>
<tr>
<td>2</td>
<td>4</td>
<td>7</td>
</tr>
<tr>
<td>3</td>
<td>6</td>
<td>9</td>
</tr>
<tr>
<td>4</td>
<td>8</td>
<td>11</td>
</tr>
<tr>
<td>5</td>
<td>10</td>
<td>13</td>
</tr>
<tr>
<td>6</td>
<td>12</td>
<td>15</td>
</tr>
<tr>
<td>7</td>
<td>14</td>
<td>17</td>
</tr>
</tbody>
</table>

Index of element stored: \( a \times i + b \)
Index of element loaded: \( c \times i + d \)

\[ a = 2 \quad b = 3 \quad c = 2 \quad d = 0 \]

\[ \text{GCD}(a, c) = 2 \]
\[ d - b = -3 \]

2 does not divide -3

⇒ No dependence possible.

What if \( \text{GCD} (a, c) \) divided \( d - b \)?

For example from last slide

---

EECC551 - Shaaban

#10  Spring 2012  lec#7  4-16-2012
ILP Compiler Support: Software Pipelining (Symbolic Loop Unrolling)

- A compiler technique where loops are reorganized:
  - Each new iteration is made from instructions selected from a number of independent iterations of the original loop.
  - The instructions are selected to separate dependent instructions within the original loop iteration.
  - No actual loop-unrolling is performed.

- Requires:
  - A software equivalent to the Tomasulo approach?
  - Additional start-up code to execute code left out from the first original loop iterations.
  - Additional finish code to execute instructions left out from the last original loop iterations.

Why?

This static optimization is done at machine code level.

i.e parallel iterations

By one or more iterations

4th Edition: Appendix G.3 (3rd Edition: Chapter 4.4)
Software Pipelining (Symbolic Loop Unrolling)

New loop iteration body is made from instructions selected from a number of independent iterations of the original loop.

Purpose: Separate dependent instructions by one or more loop iterations.

A software-pipelined loop chooses instructions from different loop iterations, thus separating the dependent instructions within one iteration of the original loop.

4th Edition: Appendix G.3 (3rd Edition: Chapter 4.4)
**Software Pipelining (Symbolic Loop Unrolling) Example**

Show a software-pipelined version of the code:

Before: Unrolled 3 times

1. \( \text{L.D } F_0,0\text{ (R1)} \)
2. \( \text{ADD.D } F_4,F_0,F_2 \)
3. \( \text{S.D } F_4,0\text{ (R1)} \)
4. \( \text{L.D } F_0,-8\text{ (R1)} \)
5. \( \text{ADD.D } F_4,F_0,F_2 \)
6. \( \text{S.D } F_4,-8\text{ (R1)} \)
7. \( \text{L.D } F_0,-16\text{ (R1)} \)
8. \( \text{ADD.D } F_4,F_0,F_2 \)
9. \( \text{S.D } F_4,-16\text{ (R1)} \)
10. \( \text{DADDUI } R_1,R_1,#-8 \)
11. \( \text{BNE } R_1,R_2,\text{LOOP} \)

3 times because chain of dependence of length 3 instructions exist in body of original loop i.e. L.D \( \rightarrow \) ADD.D \( \rightarrow \) S.D

After: Software Pipelined Version

1. \( \text{L.D } F_0,0\text{ (R1)} \)
2. \( \text{ADD.D } F_4,F_0,F_2 \)
3. \( \text{L.D } F_0,-8\text{ (R1)} \); Stores \( M[i] \)
4. \( \text{ADD.D } F_4,F_0,F_2 \); Adds to \( M[i-1] \)
5. \( \text{L.D } F_0,-16\text{ (R1)} \); Loads \( M[i-2] \)
6. \( \text{DADDUI } R_1,R_1,#-8 \)
7. \( \text{BNE } R_1,R_2,\text{LOOP} \)
8. \( \text{ADD.D } F_4,F_0,F_2 \)
9. \( \text{S.D } F_4,0\text{ (R1)} \)
10. \( \text{ADD.D } F_4,F_0,F_2 \)
11. \( \text{S.D } F_4,-8\text{ (R1)} \)

2 fewer loop iterations

No actual loop unrolling is done (do not rename registers)
Software Pipelining Example Illustrated

Assuming 6 original iterations (for illustration purposes):

\[
\begin{align*}
L.D & \quad F0,0 (R1) \\
ADD.D & \quad F4,F0,F2 \\
S.D & \quad F4,0 (R1)
\end{align*}
\]

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>L.D</td>
<td>L.D</td>
<td>L.D</td>
<td>L.D</td>
<td>L.D</td>
<td>L.D</td>
</tr>
<tr>
<td></td>
<td>ADD.D</td>
<td>ADD.D</td>
<td>ADD.D</td>
<td>ADD.D</td>
<td>ADD.D</td>
<td>ADD.D</td>
</tr>
<tr>
<td></td>
<td>S.D</td>
<td>S.D</td>
<td>S.D</td>
<td>S.D</td>
<td>S.D</td>
<td>S.D</td>
</tr>
</tbody>
</table>

4 Software Pipelined loop iterations (2 fewer iterations)

Loop Body of software Pipelined Version
• **Limits to conventional exploitation of ILP:**

1) *Pipelined clock rate*: Increasing clock rate requires deeper pipelines with longer pipeline latency which increases the CPI increase (longer branch penalty, other hazards).

2) *Instruction Issue Rate*: Limited instruction level parallelism (ILP) reduces actual instruction issue/completion rate. (vertical & horizontal waste)

3) *Cache hit rate*: Data-intensive scientific programs have very large data sets accessed with poor locality; others have continuous data streams (multimedia) and hence poor locality. (poor memory latency hiding).

4) *Data Parallelism*: Poor exploitation of data parallelism present in many scientific and multimedia applications, where similar independent computations are performed on large arrays of data (Limited ISA, hardware support).

• As a result, actual achieved performance is much less than peak potential performance and low computational energy efficiency (computations/watt)
Flynn’s 1972 Classification of Computer Architecture

- **SISD**
  - Single Instruction stream over a Single Data stream (SISD): Conventional sequential machines (e.g. single-threaded processors: Superscalar, VLIW ..).

- **SIMD**
  - Single Instruction stream over Multiple Data streams (SIMD): Vector computers, array of synchronized processing elements. (exploit data parallelism) AKA Data Parallel Systems

- **MISD**
  - Multiple Instruction streams and a Single Data stream (MISD): Systolic arrays for pipelined execution.

- **MIMD**
  - Multiple Instruction streams over Multiple Data streams (MIMD): Parallel computers:
    - Shared memory multiprocessors (e.g. SMP, CMP, NUMA, SMT)
    - Multicomputers: Unshared distributed memory, message-passing used instead (e.g. Computer Clusters)
Vector Processing

- Vector processing exploits data parallelism by performing the same computation on linear arrays of numbers "vectors" using one instruction.
- The maximum number of elements in a vector supported by a vector ISA is referred to as the Maximum Vector Length (MVL).

Scalar ISA (RISC or CISC)

Scalar (1 operation)

```
Add.d F3, F1, F2
```

Vector ISA

Vector (N operations)

```
addv.d v3, v1, v2
```

- Up to Maximum Vector Length (MVL)
- Typical MVL = 64 (Cray)

Appendix F (4th) Appendix G (3rd)
Properties of Vector Processors/ISAs

• Each result in a vector operation is independent of previous results (Data Parallelism, LLP exploited)
  => Multiple pipelined Functional units (lanes) usually used, vector compiler ensures no dependencies between computations on elements of a single vector instruction
  => higher clock rate (less complexity)

• Vector instructions access memory with known patterns
  => Highly interleaved memory with multiple banks used to provide the high bandwidth needed and hide memory latency.
  => Amortize memory latency of over many vector elements
  => No (data) caches usually used. (Do use instruction cache)

• A single vector instruction implies a large number of computations (replacing loops or reducing number of iterations needed)
  By a factor of MVL
  => Fewer instructions fetched/executed.
  => Reduces branches and branch problems (control hazards) in pipelines.

As if loop-unrolling by default MVL times?
Changes to Scalar Processor to Run Vector Instructions

- A vector processor typically consists of an ordinary pipelined scalar unit plus a vector unit.
- The scalar unit is basically not different than advanced pipelined CPUs, commercial vector machines have included both out-of-order scalar units (NEC SX/5) and VLIW scalar units (Fujitsu VPP5000).
- Computations that don’t run in vector mode don’t have high ILP, so can make scalar CPU simple (e.g in-order).
- The vector unit supports a vector ISA including decoding of vector instructions which includes:
  - Vector functional units.
  - ISA vector register bank, vector control registers (vector length, mask)
  - Vector memory Load-Store Units (LSUs).
  - Multi-banked main memory (to support the high data bandwidth needed, data cache not usually used)
- Send scalar registers to vector unit (for vector-scalar ops).
- Synchronization for results back from vector register, including exceptions.
Basic Types of Vector Architecture (ISAs)

• Types of architecture for vector ISAs/processors:
  – **Memory-memory vector ISAs/processors:**
    All vector operations are memory to memory
  – **Vector-register ISAs/processors:**
    All vector operations between vector registers (except load and store)
    • Vector equivalent of load-store architectures (ISAs)
    • Includes all vector machines since the late 1980 Cray, Convex, Fujitsu, Hitachi, NEC
Basic Structure of Vector Register Architecture (Vector MIPS)

1. Pipelined Vector Functional Units
2. Vector registers
3. Vector Load-Store Units (LSUs)
4. Multi-Banked memory for bandwidth and latency-hiding

- Each Vector Register has MVL elements (each 64 bits)
- VLR: Vector Length Register
- VM: Vector Mask Register

- Typical MVL = 64 (Cray)
- MVL range 64-4096 (4K)

Appendix F (4th) Appendix G (3rd)
## Example Vector-Register Architectures

<table>
<thead>
<tr>
<th>Processor (year)</th>
<th>Clock rate (MHz)</th>
<th>Vector registers</th>
<th>Elements per register (64-bit elements)</th>
<th>Vector arithmetic units</th>
<th>Vector load-store units</th>
<th>Lanes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cray-1 (1976)</td>
<td>80</td>
<td>8</td>
<td>64</td>
<td>6: FP add, FP multiply, FP reciprocal, integer add, logical, shift</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Cray X-MP (1983)</td>
<td>118</td>
<td>8</td>
<td>64</td>
<td>8: FP add, FP multiply, FP reciprocal, integer add, 2 logical, shift, population count/parity</td>
<td>2 loads 1 store</td>
<td>1</td>
</tr>
<tr>
<td>Cray Y-MP (1988)</td>
<td>166</td>
<td>8</td>
<td>64</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cray-2 (1985)</td>
<td>244</td>
<td>8</td>
<td>64</td>
<td>5: FP add, FP multiply, FP reciprocal/sqrt, integer add/shift/population count, logical</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Fujitsu VP100/VP200 (1982)</td>
<td>133</td>
<td>8–256</td>
<td>32–1024</td>
<td>3: FP or integer add/logical, multiply, divide</td>
<td>2 (VP100) 1 (VP200)</td>
<td></td>
</tr>
<tr>
<td>Hitachi S810/S820 (1983)</td>
<td>71</td>
<td>32</td>
<td>256</td>
<td>4: FP multiply-add, FP multiply/divide-add unit, 2 integer add/logical</td>
<td>3 loads 1 store 1 (S810) 2 (S820)</td>
<td></td>
</tr>
<tr>
<td>Convex C-1 (1985)</td>
<td>10</td>
<td>8</td>
<td>128</td>
<td>2: FP or integer multiply/divide, add/logical</td>
<td>1 (64 bit) 2 (32 bit)</td>
<td></td>
</tr>
<tr>
<td>NEC SX/2 (1985)</td>
<td>167</td>
<td>8 + 32</td>
<td>256</td>
<td>4: FP multiply/divide, FP add, integer add/logical, shift</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>Cray C90 (1991)</td>
<td>240</td>
<td>8</td>
<td>128</td>
<td>8: FP add, FP multiply, FP reciprocal, integer add, 2 logical, shift, population count/parity</td>
<td>2 loads 1 store</td>
<td>2</td>
</tr>
<tr>
<td>Cray T90 (1995)</td>
<td>460</td>
<td>8</td>
<td>128</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NEC SX/5 (1998)</td>
<td>312</td>
<td>8 + 64</td>
<td>512</td>
<td>4: FP or integer add/shift, multiply, divide, logical</td>
<td>1</td>
<td>16</td>
</tr>
<tr>
<td>Fujitsu VPP5000 (1999)</td>
<td>300</td>
<td>8–256</td>
<td>128–4096</td>
<td>3: FP or integer multiply, add/logical, divide</td>
<td>1 load 1 store</td>
<td>16</td>
</tr>
<tr>
<td>Cray SV1 (1998)</td>
<td>300</td>
<td>8</td>
<td>64</td>
<td>8: FP add, FP multiply, FP reciprocal, integer add, 2 logical, shift, population count/parity</td>
<td>1 load-store 2 1 load 8 (MSP)</td>
<td></td>
</tr>
<tr>
<td>SV1ex (2001)</td>
<td>500</td>
<td>8</td>
<td>64</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>VMIPS (2001)</td>
<td>500</td>
<td>8</td>
<td>64</td>
<td>5: FP multiply, FP divide, FP add, integer add/shift, logical</td>
<td>1 load-store 1</td>
<td></td>
</tr>
</tbody>
</table>

VMIPS = Vector MIPS

---

<table>
<thead>
<tr>
<th>Appendix F (4th)</th>
<th>Appendix G (3rd)</th>
</tr>
</thead>
</table>
## The VMIPS Vector FP Instructions

<table>
<thead>
<tr>
<th>Instruction</th>
<th>Operands</th>
<th>Function</th>
</tr>
</thead>
<tbody>
<tr>
<td>ADDV.D</td>
<td>V1, V2, V3</td>
<td>Add elements of V2 and V3, then put each result in V1.</td>
</tr>
<tr>
<td>ADDVS.D</td>
<td>V1, V2, F0</td>
<td>Add F0 to each element of V2, then put each result in V1.</td>
</tr>
<tr>
<td>SUBV.D</td>
<td>V1, V2, V3</td>
<td>Subtract elements of V3 from V2, then put each result in V1.</td>
</tr>
<tr>
<td>SUBVS.D</td>
<td>V1, V2, F0</td>
<td>Subtract F0 from elements of V2, then put each result in V1.</td>
</tr>
<tr>
<td>SUBSV.D</td>
<td>V1, F0, V2</td>
<td>Subtract elements of V2 from F0, then put each result in V1.</td>
</tr>
<tr>
<td>MULV.D</td>
<td>V1, V2, V3</td>
<td>Multiply elements of V2 and V3, then put each result in V1.</td>
</tr>
<tr>
<td>MULVS.D</td>
<td>V1, V2, F0</td>
<td>Multiply each element of V2 by F0, then put each result in V1.</td>
</tr>
<tr>
<td>DIVV.D</td>
<td>V1, V2, V3</td>
<td>Divide elements of V2 by V3, then put each result in V1.</td>
</tr>
<tr>
<td>DIVVS.D</td>
<td>V1, V2, F0</td>
<td>Divide elements of V2 by F0, then put each result in V1.</td>
</tr>
<tr>
<td>DIVSV.D</td>
<td>V1, F0, V2</td>
<td>Divide F0 by elements of V2, then put each result in V1.</td>
</tr>
</tbody>
</table>

### Vector Memory
- **LV** V1, R1: Load vector register V1 from memory starting at address R1.
- **SV** R1, V1: Store vector register V1 into memory starting at address R1.
- **LVWS** V1, (R1, R2): Load V1 from address at R1 with stride in R2, i.e., R1+1×R2.
- **SVWS** (R1, R2), V1: Store V1 from address at R1 with stride in R2, i.e., R1+1×R2.
- **LVI** V1, (R1+V2): Load V1 with vector whose elements are at R1+V2(1), i.e., V2 is an index.
- **SVI** (R1+V2), V1: Store V1 to vector whose elements are at R1+V2(1), i.e., V2 is an index.

### Vector Index
- **CVI** V1, R1: Create an index vector by storing the values 0, 1×R1, 2×R1, ..., 63×R1 into V1.

### Vector Mask
- **S--V.D** V1, V2: Compare the elements (EQ, NE, GT, LT, GE, LE) in V1 and V2. If condition is true, put 1 in the corresponding bit vector; otherwise put 0. Put resulting bit vector in vector-mask register (VM). The instruction S--VS.D performs the same compare but using a scalar value as one operand.

### Vector Length
- **POP** R1, VM: Count the 1s in the vector-mask register and store count in R1.
- **CVM** R1: Set the vector-mask register to all 1s.
- **MTC1** VLR, R1: Move contents of R1 to the vector-length register.
- **MFC1** R1, VLR: Move the contents of the vector-length register to R1.
- **MVMT** VM, F0: Move contents of F0 to the vector-mask register.
- **MVFM** F0, VM: Move contents of vector-mask register to F0.

### Vector Control Registers:
- VM = Vector Mask
- VLR = Vector Length Register

---

**VMIPS = Vector MIPS**

8 Vector Registers
- V0-V7
- MVL = 64
  - (Similar to Cray)
### DAXPY (Y = a * X + Y)

Assuming vectors X, Y are length 64 = MVL

#### Scalar vs. Vector

<table>
<thead>
<tr>
<th>Instruction</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>L.D F0,a</td>
<td>load scalar a</td>
</tr>
<tr>
<td>LV V1,Rx</td>
<td>load vector X</td>
</tr>
<tr>
<td>MUL.D F2,F0,F2</td>
<td>vector-scalar mult.</td>
</tr>
<tr>
<td>L.D 0(Rx)</td>
<td>load X(i)</td>
</tr>
<tr>
<td>MUL.D F2,F0,F2</td>
<td>a*X(i)</td>
</tr>
<tr>
<td>L.D 0(Ry)</td>
<td>load Y(i)</td>
</tr>
<tr>
<td>ADD.D F4,F2,F4</td>
<td>a*X(i) + Y(i)</td>
</tr>
<tr>
<td>S.D F4,0(Ry)</td>
<td>store into Y(i)</td>
</tr>
<tr>
<td>DADDIU R4,Rx,#512</td>
<td>last address to load</td>
</tr>
<tr>
<td>DSUBU R20,R4,Rx</td>
<td>compute bound</td>
</tr>
<tr>
<td>BNEZ R20,loop</td>
<td>check if done</td>
</tr>
<tr>
<td>L.D F0,a</td>
<td>load scalar a</td>
</tr>
<tr>
<td>LV V1,Rx</td>
<td>load vector X</td>
</tr>
<tr>
<td>MULVS.D V2,V1,F0</td>
<td>vector-scalar mult.</td>
</tr>
<tr>
<td>LV V3,Ry</td>
<td>load vector Y</td>
</tr>
<tr>
<td>ADDV.D V4,V2,V3</td>
<td>add</td>
</tr>
<tr>
<td>SV Ry,V4</td>
<td>store the result</td>
</tr>
</tbody>
</table>

#### Unroll?

What does loop unrolling accomplish?

- **Calculation**:
  - Scalar: 578 (2+9*64) vs. 321 (1+5*64) ops (1.8X)
  - Vector: 578 (2+9*64) vs. 6 instructions (96X)
  - 64 operation vectors + no loop overhead

#### Vector Control Registers:

- **VM = Vector Mask**
- **VLR = Vector Length Register**

---

**Scalar Vs. Vector Code**

As if the scalar loop code was unrolled MVL = 64 times: Every vector instruction replaces 64 scalar instructions.
Vector/SIMD/Multimedia Scalar ISA Extensions

- Vector or Multimedia ISA Extensions: Limited vector instructions added to scalar RISC/CISC ISAs with MVL = 2-8
  - Why? Improved exploitation of data parallelism in scalar ISAs/processors
- Example: Intel MMX: 57 new x86 instructions (1st since 386)
  - similar to Intel 860, Mot. 88110, HP PA-71000LC, UltraSPARC ...
  - 3 integer vector element types: 8 8-bit (MVL =8), 4 16-bit (MVL =4) , 2 32-bit (MVL =2) in packed in 64 bit registers
  - reuse 8 FP registers (FP and MMX cannot mix)
    - short vector: load, add, store 8, 8-bit operands
  - Claim: overall speedup 1.5 to 2X for multimedia applications (2D/3D graphics, audio, video, speech …)
- Intel SSE (Streaming SIMD Extensions) adds support for FP with MVL =2 to MMX
- Intel SSE2 Adds support of FP with MVL = 4 (4 single FP in 128 bit registers), 2 double FP MVL = 2, to SSE

Major Issue: Efficiently meeting the increased data memory bandwidth requirements of such instructions