Until today, we assumed that instructions execute sequentially one after another in a von Neumann style of execution. However, today’s processors do not execute instructions in this model; architectural “enhancements” allow processors to execute code faster:

- pipelining
- multiple function units
- Instruction scheduling refers to re-ordering instructions in a program to exploit Instruction Level Parallelism (ILP) and improve performance.
- Instruction scheduling is still an active area of research because of the difficulty of the problem (NP-complete) and the changing natures of processors.

In the von Neumann model of execution an instruction starts only after its predecessor completes.

This is not a very efficient model of execution.

- von Neumann bottleneck or the memory wall.

Instruction Scheduling

- In the von Neumann model of execution an instruction starts only after its predecessor completes.

Each of these stages completes its operation in one cycle (shorter the the cycle in the von Neumann model).

An instruction still takes the same time (maybe a little more due to pipelining) to execute.

Almost all processors today use instructions pipelines to allow overlap of instructions (Pentium 4 has a 20 stage pipeline!!!).

The execution of an instruction is divided into stages; each stage is performed by a separate part of the processor.

- F: Fetch instruction from cache or memory.
- D: Decode instruction.
- E: Execute. ALU operation or address calculation.
- M: Memory access.
- W: Write back result into register.

Each of these stages completes its operation in one cycle (shorter the the cycle in the von Neumann model).

An instruction still takes the same time (maybe a little more due to pipelining) to execute.

However, we overlap these stages in time to complete an instruction every cycle.

Structural Hazards
- two instructions need the same resource at the same time
- memory or functional units in a superscalar.

Data Hazards
- an instructions needs the results of a previous instruction
- solved by forwarding and/or stalling
- cache miss?

Control Hazards
- jump & branch address not known until later in pipeline
- solved by delay slot and/or prediction.
Jump/Branch Delay Slot(s)

- Control hazards, i.e. jump/branch instructions.
- Unconditional jump address available only after Decode.
- Conditional branch address available only after Execute.

<table>
<thead>
<tr>
<th>Instruction</th>
<th>F</th>
<th>D</th>
<th>E</th>
<th>M</th>
<th>W</th>
</tr>
</thead>
<tbody>
<tr>
<td>jump/branch</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>instr 2</td>
<td>F</td>
<td>D</td>
<td>E</td>
<td>M</td>
<td>W</td>
</tr>
<tr>
<td>instr 3</td>
<td>F</td>
<td>D</td>
<td>E</td>
<td>M</td>
<td>W</td>
</tr>
<tr>
<td>instr 4</td>
<td>F</td>
<td>D</td>
<td>E</td>
<td>M</td>
<td>W</td>
</tr>
</tbody>
</table>

Another option is to insert a no-op instructions (software).

<table>
<thead>
<tr>
<th>Instruction</th>
<th>F</th>
<th>D</th>
<th>E</th>
<th>M</th>
<th>W</th>
</tr>
</thead>
<tbody>
<tr>
<td>jump</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>nop</td>
<td>F</td>
<td>D</td>
<td>E</td>
<td>M</td>
<td>W</td>
</tr>
<tr>
<td>instr 2</td>
<td>F</td>
<td>D</td>
<td>E</td>
<td>M</td>
<td>W</td>
</tr>
</tbody>
</table>

Both degrade performance!

A better solution is to make the branch take effect only after the delay slots.
- One or two instructions always get executed after the branch but before the branching takes effect.

<table>
<thead>
<tr>
<th>Instruction</th>
<th>F</th>
<th>D</th>
<th>E</th>
<th>M</th>
<th>W</th>
</tr>
</thead>
<tbody>
<tr>
<td>bra</td>
<td>F</td>
<td>D</td>
<td>E</td>
<td>M</td>
<td>W</td>
</tr>
<tr>
<td>instr x</td>
<td>F</td>
<td>D</td>
<td>E</td>
<td>M</td>
<td>W</td>
</tr>
<tr>
<td>instr y</td>
<td>F</td>
<td>D</td>
<td>E</td>
<td>M</td>
<td>W</td>
</tr>
<tr>
<td>instr 2</td>
<td>F</td>
<td>D</td>
<td>E</td>
<td>M</td>
<td>W</td>
</tr>
<tr>
<td>instr 3</td>
<td>F</td>
<td>D</td>
<td>E</td>
<td>M</td>
<td>W</td>
</tr>
</tbody>
</table>

What instruction(s) to use?

Branch Prediction

- Current processors will speculatively execute at conditional branches.
  - If a branch direction is correctly guessed, great!
  - If not, the pipeline is flushed before instructions commit (WB).
- Why not let compiler schedule?
  - The average number of instructions per basic block in typical C code is about 5 instructions.
  - Branches are not statically predictable
  - What happens if you have a 20 stage pipeline?

Data Hazards

- \( r1 = r2 + r3 \)
- \( r4 = r1 + r1 \)

\( F \quad D \quad E \quad M \quad W \)

\( [r2] \) available here

- \( r1 = [r2] \)
- \( r4 = r1 + r1 \)

\( F \quad D \quad E \quad M \quad W \)

\( [r4] \) available here
To schedule a basic block, you need to determine scheduling constraints and express these using a dependence graph. For a basic block, this graph is a DAG. Each node is a machine instruction and the edges are the dependencies between instructions.

### Flow (True) Dependencies
- A flow dependence exists if an instruction $I_1$ writes to a register or location that $I_2$ uses.
- This is written $I_1 \delta^f I_2$

Flow dependencies are true dependencies, that is these dependencies are necessary to transmit information between statements.

### Anti Dependencies
- An anti dependence exists if an instruction $I_1$ uses a register that $I_2$ changes.
- This is written $I_1 \delta^a I_2$

Anti dependencies are false dependencies, that is they arise due to the reuse of memory locations.

### Output Dependencies
- An output dependence exists if an instruction $I_1$ writes to register that $I_2$ also writes to.
- This is written $I_1 \delta^o I_2$

Output dependencies are also false dependencies, that is they arise due to the reuse of memory locations.

### List Scheduling Algorithm - Example

1. load R1, b
2. load R2, c
3. add R2,R1
4. store a, R2
5. load R3, e
6. load R4,f
7. sub R3,R4
8. store d,R3

Assume, that loaded values are available after 2 cycles (from beginning of load instruction). So, there is need for an extra cycle after 2 and after 6. In the absence of instruction scheduling, NOPs must be inserted.

Step 1: construct a dependence graph of the basic block. (The edges are weighted with the latency of the instruction).
Step 2: use the dependence graph to determine instructions that can execute; insert on a list, called the Ready list.
Step 3: use the dependence graph and the Ready list to schedule an instruction that causes the smallest possible stall; update the Ready list. Repeat until Ready list is empty!
List Scheduling Algorithm - Example

1. load R1, b
2. load R2, c
3. add R2, R1
4. store a, R2
5. load R3, e
6. load R4, f
7. sub R3, R4
8. store d, R3

\[ a = b + c \]
\[ d = e - f \]

We're done. Now have a schedule that requires no stalls and no NOPs.

Superscalars, i.e. multiple functional units

- Almost all modern processors are superscalars
  - have multiple functional units
  - Intel 486: 1 pipeline; Pentium: 2 pipelines; Pentium 4: up to 6 instructions per clock cycle.
- Need to model the CPU as accurately as possible
  - which instructions can execute simultaneously
  - relative delay of different types of instructions
- Can use a Greedy / Ready list method
  - not always optimal, nontrivial scheduling is NP

<table>
<thead>
<tr>
<th>IntOp</th>
<th>IntMem</th>
<th>IntOp</th>
<th>IntMem</th>
</tr>
</thead>
<tbody>
<tr>
<td>FltOp</td>
<td>Flt,d</td>
<td>FltOp</td>
<td>FltOp</td>
</tr>
<tr>
<td>IntOp</td>
<td>IntLd</td>
<td>IntOp</td>
<td>IntLd</td>
</tr>
</tbody>
</table>

(with FltOp \(\rightarrow\) IntLd)

Trace Scheduling

- Basic blocks typically contain a small number of instructions.
- With many function units, we may not be able to keep all the units busy with just the instructions of a basic block.
- Trace scheduling allows block scheduling across basic blocks.
- The basic idea is to dynamically determine which blocks are executed more frequently. The set of such basic blocks is called a trace.

The trace is then scheduled as a single basic block.
- Blocks that are not part of the trace must be modified to restore program semantics if/when execution goes off-trace.
Instruction Scheduling for Loops

- Loop bodies are typically too small to produce a schedule that exploits all resources.
- But, most of execution is spent in loops.
- Need ways to schedule loops:
  - Loop unrolling.
  - Software pipelining.
- Focus on the main ideas; the details are considerable.

Loop Example

- Machine parameters:
  - 1 memory unit capable of either a load or a store. Each operation takes 2 cycles. No delay slots.
  - One multiplier unit. A multiply operation takes 3 cycles.
  - One adder unit. An add operation takes 2 cycles.
  - The adder and the multiplier are capable of performing a branch operation in 2 cycles.
  - All units are pipelined, allowing the initiation of an operation per one clock cycle.

Loop:

for i = 1 to N
    a[i] = a[i] * b
    r6 = *(r2) (ld)
    r6 = r6 * r3 (mul)
    *(r2) = r6 (st)
    r2 = r2 + 4 (add)
    if (r2 <= r5) go to L1 (ble)

Block Scheduling

L1:

```
r6 = *(r2) (ld)
r6 = r6 * r3 (mul)
r2 = r2 + 4 (add)
if (r2 <= r5) go to L1 (ble)
```

Multiple issue.
Add takes 2 cycles; latency = 1
Multiply takes 3 cycles; latency = 2

```
mul R2,R0,R1
add R3,R3,R2
add R4,R0,R1
add R5,R5,R4
```

MUL
ADD

Loop unrolling: replicate the loop body.

```
Loop: load R6, (R1)
mul R6,R6,R3
store (R1), R6
add R1,R1,#4
load R6, (R1)
mul R6,R6,R3
store (R1), R6
add R1,R1,#4
cmp R1,R5
cmp R1,R5
gle Loop
```

Multiple issue.
Add takes 2 cycles; latency = 1; RI = 1.
Multiply takes 3 cycles; Latency = 2; RI = 1.

```
mul R2,R0,R1
add R3,R3,R2
add R4,R0,R1
add R5,R5,R4
add R3,R3,R2
```

MUL
ADD

```
mul R2,R0,R1
add R3,R3,R2
add R4,R0,R1
add R5,R5,R4
```

MUL
ADD

```
mul R2,R0,R1
add R3,R3,R2
add R4,R0,R1
add R5,R5,R4
```

MUL
ADD

```
mul R2,R0,R1
add R3,R3,R2
add R4,R0,R1
add R5,R5,R4
```

MUL
ADD

```
mul R2,R0,R1
add R3,R3,R2
add R4,R0,R1
add R5,R5,R4
```

MUL
ADD

```
mul R2,R0,R1
add R3,R3,R2
add R4,R0,R1
add R5,R5,R4
```

MUL
ADD

```
mul R2,R0,R1
add R3,R3,R2
add R4,R0,R1
add R5,R5,R4
```

MUL
ADD
Loop Unrolling

L1:
- \( r_6 = *(r_2) \) (ld)
- \( r_6 = r_6 \times r_3 \) (mul)
- \(*r_2 = r_6 \) (st)
- \( r_2 = r_2 + 4 \) (add)
- if \( r_2 < r_5 \) go to L1 (ble)

Register Re-naming

L1:
- \( r_6 = *(r_2) \) (ld)
- \( r_6 = r_6 \times r_3 \) (mul)
- \(*r_2 = r_6 \) (st)
- \( r_2 = r_2 + 4 \) (add)
- if \( r_1 < r_5 \) go to L1 (ble)

Software Pipelining

- Software pipelining: overlap multiple iterations of a loop to fully utilize hardware resources.
- Find the steady-state window so that:
  - all the instructions of the loop body are executed
  - but from different iterations

Software Pipelining

<table>
<thead>
<tr>
<th>Iteration 1</th>
<th>Iteration 2</th>
<th>Iteration 3</th>
<th>Iteration 4</th>
<th>Iteration 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>SOFTWARE PIPELINED ITERATION</td>
<td>SOFTWARE PIPELINED ITERATION</td>
<td>SOFTWARE PIPELINED ITERATION</td>
<td>SOFTWARE PIPELINED ITERATION</td>
<td>SOFTWARE PIPELINED ITERATION</td>
</tr>
</tbody>
</table>

PRELUDE

POSTLUDE
Software Pipelining: An Example

Iteration 1
LD F0, 0(R1)
ADDD F4, F0, F2
SD 0(R1), F4
SUBI R1, R1, #8
BNEZ R1, Loop

Iteration 2
LD F0, 0(R1)
ADDD F4, F0, F2
SD 0(R1), F4

Iteration 3
SUBI R1, R1, #8
BNEZ R1, Loop

Iteration 4
LD F0, 0(R1)
ADDD F4, F0, F2
SD 0(R1), F4

Iteration 5
SUBI R1, R1, #8
BNEZ R1, Loop

Software Pipelining: An Example

Loop Unrolling and Software Pipelining

- Loop Unrolling
  - helps uncover Instruction Level Parallelism (ILP)
  - reduces looping overhead (increment and branch)
  - generates a lot of code, copies of the loop body

- Software Pipelining
  - also helps uncover ILP
  - does not reduce looping overhead
  - loop body is always executing at top speed
  - usually uses less code space

- Both require that number of iterations is known
- If unroll factor does not evenly divide iterations, the extra iterations must be caught by a pre- or post-amble
Emerging Architectures

- Simultaneous Multithreading (SMT)
  - Execute multiple threads simultaneously
  - Keeps functional units busy
  - Example: Intel Pentium IV with HyperThreading (HT)
  - Sun: 2 cores both with SMT (an SMT CMP)

EPIC / VLIW Architectures

- EPIC – Explicitly Parallel Instruction Computing (Itanium / IA64)
- VLIW – Very Long Instruction Word (Transmeta)
- Compiler explicitly packages independent instructions
- Hardware does not due reordering
- Chip can run at higher clock rates

Some may live, while some may die
Which one will compilers work best with?

VLIW -vs- Superscalar

- VLIW
  - use very long multi-operation instructions
  - the instruction specifies what each functional unit is to do
  - expects dependence free instructions
  - compiler must explicitly detect and schedule independent instructions

- Superscalar
  - uses traditional sequential operations
  - processor fetches multiple instructions per cycle
  - detects dependencies and schedules accordingly
  - has dynamic information available
  - compiler can help by placing independent operations close to each other

Problems with VLIW

- Compiler must statically determine dependencies
- Compiler must have very detailed model of architecture
  - number and type of functional units
  - delays for each operation
  - memory delays
  - latencies are very important
- A new generation with more units or different latencies means recompile
- EPIC (Itanium / IA64) tries to address some of these
  - compiler expresses parallelism; hardware schedules ops
  - no fixed length to instructions, just a number of bundles
  - relationship of one bundle to another is expressed

Predication: Branches are Bad

- Provide predicate registers and predicated instructions
- Can set a predicate register to true or false using a comparison instruction
- Most instructions can be predicated so that they only commit if their predicate is true
- Can schedule and execute across multiple directions of a branch and only valid instructions will commit

```c
if (m == n) {
    a = a + b;
} else {
    b = b + 1;
}
```

```asm
cmp.eq p1, p2 = r1, r2
(p1) add r1 = r1, r2
(p2) add r2 = r2, 1
```

Speculation

- Control Speculation
  - move an instruction above a branch instruction
  - may be done if it is always safe
  - or can be done speculatively

```asm
ld8 r1 = [r4]
ld8 r2 = [r5]
add r3 = r1, r2
add r3 = r1, r2
(p1) br.cond label
```

```asm
ld8 s r2 = [r5]
add r3 = r1, r2
add r3 = r1, r2
(p1) br.cond label
chk.s r3, fixupcode
```

- EPIC speculatively loads a value, if the load fails a NAT bit is set
- NAT bits propagate through all other uses
- the chk.s instructions checks the NAT bit, if set it calls the fixupcode

CSE 427: Computer Architectures
Speculation

- Data Speculation
  - move a load before a store that may be aliased
  - may be done if it is always safe
  - or can be done speculatively

- the Advanced Load Address Table (ALAT) checks for collisions

EPIC / Itanium Processor Family

- Prediction and Speculation
- Special support for software pipelining
  - rotating registers
  - special epilogue counters
- Performance????
  - need more compiler work
  - a big change and it’s still early
  - does ok on floating-point
  - profiling may help
  - runtime techniques may help
- What are the issues?

We’ve now covered optimizations found in most commercial compilers

- Compilers improve performance dramatically
- The optimizations in this class improve code
- A little test: compile gcc using cc

CC Optimization Levels

-xO1 Does basic local optimization (peephole).

-xO2 Does basic local and global optimization.
  This is induction variable elimination, local and global common subexpression elimination, algebraic simplification, copy propagation, constant propagation, loop-invariant optimization, register allocation, basic block merging, tail recursion elimination, dead code elimination, tail call elimination and complex expression expansion.

-xO3 Performs like -xO2 but, also optimizes references or definitions for external variables. Loop unrolling and software pipelining are also performed. In general, the -xO3 level results in increased code size. Does not deal with pointer disambiguation.

-xO4 Performs like -xO3 but, also does automatic inlining of functions contained in the same file; this usually improves execution speed. The -xO4 level does trace the effects of pointer assignments. In general, the -xO4 level results in increased code size.

-xO5 Generates the highest level of optimization. Uses optimization algorithms that take more compilation time or that do not have as high a certainty of improving execution time.

-fast: -O4 plus some other specific flags…
Performance of CC Opt Levels

% Imp: -xO0 → -xO1 = 41%; -xO1 → -xO2 = 20%; -xO2 → -xO3 = 11%