Key Questions

Q1. How do we find independent instructions to fetch/execute?

Q2. How do we enable more compiler optimizations?
   e.g., common subexpression elimination, constant propagation, dead code elimination, redundancy elimination, ...

Q3. How do we increase the instruction fetch rate?
   i.e., have the ability to fetch more instructions per cycle

A: Enabling the compiler to optimize across a larger number of instructions that will be executed straight line (without branches getting in the way) eases all of the above

VLIW (Very Long Instruction Word)

- Simple hardware with multiple function units
  - Reduced hardware complexity
  - Little or no scheduling done in hardware, e.g., in-order
  - Hopefully, faster clock and less power
- Compiler required to group and schedule instructions (compare to OoO superscalar)
  - Predicated instructions to help with scheduling (trace, etc.)
  - More registers (for software pipelining, etc.)
- Example machines:
  - Multiflow, Cydra 5 (8-16 ops per VLIW)
  - IA-64 (3 ops per bundle)
  - TMS32xxx (5+ ops per VLIW)
  - Crusoe (4 ops per VLIW)
Comparison between SS ↔ VLIW

From Mark Smotherman, "Understanding EPIC Architectures and Implementations"
EPIC – Intel IA-64 Architecture

- Gets rid of lock-step execution of instructions within a VLIW instruction
- Idea: More ISA support for static scheduling and parallelization
  - Specify dependencies within and between VLIW instructions (explicitly parallel)

+ No lock-step execution
+ Static reordering of stores and loads + dynamic checking
  - Hardware needs to perform dependency checking (albeit aided by software)
  - Other disadvantages of VLIW still exist


IA-64 Instructions

- IA-64 “Bundle” (~EPIC Instruction)
  - Total of 128 bits
  - Contains three IA-64 instructions
  - Template bits in each bundle specify dependencies within a bundle

<table>
<thead>
<tr>
<th>Instruction 2</th>
<th>Instruction 1</th>
<th>Instruction 3</th>
<th>Template</th>
</tr>
</thead>
<tbody>
<tr>
<td>41 bits</td>
<td>41 bits</td>
<td>41 bits</td>
<td>5 bits</td>
</tr>
</tbody>
</table>

- Op | Register 1 | Register 2 | Register 3 | Predicate |
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>14 bits</td>
<td>7 bits</td>
<td>7 bits</td>
<td>7 bits</td>
<td>6 bits</td>
</tr>
</tbody>
</table>

IA-64 Instruction

- Fixed-length 41 bits long
- Contains three 7-bit register specifiers
- Contains a 6-bit field for specifying one of the 64 one-bit predicate registers

IA-64 Instruction Bundles and Groups

- Groups of instructions can be executed safely in parallel
  - Marked by “stop bits”

- Bundles are for packaging
  - Groups can span multiple bundles
    - Alleviates recompilation need somewhat

VLIW: Finding Independent Operations

- Within a basic block, there is limited instruction-level parallelism
- To find multiple instructions to be executed in parallel, the compiler needs to consider multiple basic blocks

- Problem: Moving an instruction above a branch is unsafe because instruction is not guaranteed to be executed

- Idea: Enlarge blocks at compile time by finding the frequently-executed paths
  - Trace scheduling
  - Superblock scheduling
  - Hyperblock scheduling
  - Software Pipelining

It's all about the compiler and how to schedule the instructions to maximize parallelism
List Scheduling: For 1 basic block

- Assign priority to each instruction
- Initialize ready list that holds all ready instructions
  - Ready = data ready and can be scheduled
- Choose one ready instruction \( I \) from ready list with the highest priority
  - Possibly using tie-breaking heuristics
- Insert \( I \) into schedule
  - Making sure resource constraints are satisfied
- Add those instructions whose precedence constraints are now satisfied into the ready list

Data Precedence Graph

Instruction Prioritization Heuristics

- Number of descendants in precedence graph
- Maximum latency from root node of precedence graph
- Length of operation latency
- Ranking of paths based on importance
- Combination of above

VLIW List Scheduling

- Assign Priorities
- Compute Data Ready List - all operations whose predecessors have been scheduled.
- Select from DRL in priority order while checking resource constraints
- Add newly ready operations to DRL and repeat for next instruction
Extending the scheduling domain

- Basic block is too small to get any real parallelism
- How to extend the basic block?
  - Why do we have basic blocks in the first place?
  - Loops
    - Loop unrolling
    - Software pipelining
  - Non-loops
    - Will almost always involve some speculation
    - And, thus, profiling may be very important

Safety and Legality in Code Motion

- Two characteristics of speculative code motion:
  - Safety: whether or not spurious exceptions may occur
  - Legality: whether or not result will be always correct
- Four possible types of code motion:

Code Movement Constraints

- Downward
  - When moving an operation from a BB to one of its dest BB’s, all the other dest basic blocks should still be able to use the result of the operation
  - the other source BB’s of the dest BB should not be disturbed
- Upward
  - When moving an operation from a BB to its source BB’s
    - register values required by the other dest BB’s must not be destroyed
    - the movement must not cause new exceptions

Trace Scheduling

- Trace: A frequently executed path in the control-flow graph (has multiple side entrances and multiple side exits)
- Idea: Find independent operations within a trace to pack into VLIW instructions.
  - Traces determined via profiling
  - Compiler adds fix-up code for correctness (if a side entrance or side exit of a trace is exercised at runtime, corresponding fix-up code is executed)
Trace Scheduling Idea

- There may be conditional branches from the middle of the trace (side exits) and transitions from other traces into the middle of the trace (side entrances).
- These control-flow transitions are ignored during trace scheduling.
- After scheduling, fix-up/bookkeeping code is inserted to ensure the correct execution of off-trace code.

Trace Scheduling (III)

What bookkeeping is required when Instr 1 is moved below the side entrance in the trace?

Trace Scheduling (IV)
Trace Scheduling (V)

What bookkeeping is required when Instr 5 moves above the side entrance in the trace?

Trace Scheduling Fixup Code Issues

- Sometimes need to copy instructions more than once to ensure correctness on all paths (see C below)

 Trace Scheduling Overview

- Trace Selection
  - select seed block (the highest frequency basic block)
  - extend trace (along the highest frequency edges)
  - forward (successor of the last block of the trace)
  - backward (predecessor of the first block of the trace)
  - don’t cross loop back edge
  - bound max_trace_length heuristically

- Trace Scheduling
  - build data precedence graph for a whole trace
  - perform list scheduling and allocate registers
  - add compensation code to maintain semantic correctness

- Speculative Code Motion (upward)
  - move an instruction above a branch if safe
Trace Scheduling Example (V)

```
fdiv f1, f2, f3
beq r1, $0
ld  r2, 0(r3)
add  r2, r2, 4
beq  r2, $0
fsub  f2,  f2,  f6
st.d  f2, 0(r8)
```

**Trace Scheduling Tradeoffs**

- **Advantages**
  - Enables the finding of more independent instructions → fewer NOPs in a VLIW instruction

- **Disadvantages**
  - Profile dependent
  - What if dynamic path deviates from trace → lots of NOPs in the VLIW instructions
  - Code bloat and additional fix-up code executed
  - Due to side entrances and side exits
  - Infrequent paths interfere with the frequent path
  - Effectiveness depends on the bias of branches
    - Unbiased branches → smaller traces → less opportunity for finding independent instructions

**Superblock Scheduling**

- Trace: multiple entry, multiple exit block
- Superblock: single-entry, multiple exit block
  - A trace with side entrances are eliminated
  - Infrequent paths do not interfere with the frequent path
- More optimization/scheduling opportunity than traces
- Eliminates “difficult” bookkeeping due to side entrances

**Can You Do This with a Trace?**

Original Code

```
opA: mul r1,r2,3
99
7

opB: add r2,r2,1
99
7

opC: mul r3,r2,3
1
```

Code After Superblock Formation

```
opA: mul r1,r2,3
99

opB: add r2,r2,1
99

opC: mul r3,r2,3
```

Code After Common Subexpression Elimination

```
opA: mul r1,r2,3
99

opB: add r2,r2,1

opC: mov r3,r1
ABC'

opA: mul r1,r2,3
```
Superblock Scheduling Shortcomings

-- Still profile-dependent

-- No single frequently executed path if there is an unbiased branch
   -- Reduces the size of superblocks

-- Code bloat and additional fix-up code executed
   -- Due to side exits

Hyperblock Scheduling

- Idea: Use predication support to eliminate unbiased branches and increase the size of superblocks

- Hyperblock: A single-entry, multiple-exit block with internal control flow eliminated using predication (if-conversion)

Advantages
+ Reduces the effect of unbiased branches on scheduled block size

Disadvantages
-- Requires predicated execution support
-- All disadvantages of predicated execution

Hyperblock Formation (I)

- Hyperblock formation
  1. Block selection
  2. Tail duplication
  3. If-conversion

- Block selection
  - Select subset of BBs for inclusion in HB
  - Difficult problem
  - Weighted cost/benefit function
    - Height overhead
    - Resource overhead
    - Dependency overhead
    - Branch elimination benefit
    - Weighted by frequency


Hyperblock Formation (II)

Tail duplication same as with Superblock formation
Hyperblock Formation (III)

If-convert (predicate) intra-hyperblock branches

Can We Do Better?

- Hyperblock still
  - Profile dependent
  - Requires fix-up code
  - And, requires predication support

- Single-entry, single-exit enlarged blocks
  - Block-structured ISA
    - Optimizes multiple paths (can use predication to enlarge blocks)
    - No need for fix-up code (duplication instead of fixup)

Non-Faulting Loads and Exception Propagation

- *ld.s* fetches speculatively from memory
  - i.e. any exception due to *ld.s* is suppressed
- If *ld.s r1* did not cause an exception then *chk.s r1* is a NOP, else a branch is taken (to execute some compensation code)

Non-Faulting Loads and Exception Propagation in IA-64

- Load data can be speculatively consumed prior to check
- “speculation” status is propagated with speculated data
- Any instruction that uses a speculative result also becomes speculative itself (i.e. suppressed exceptions)
- *chk.s* checks the entire dataflow sequence for exceptions
Aggressive ST-LD Reordering in IA-64

- *ld.a* starts the monitoring of any store to the same address as the advanced load
- If no aliasing has occurred since *ld.a*, *ld.c* is a NOP
- If aliasing has occurred, *ld.c* re-loads from memory

Summary and Questions

- Trace, superblock, hyperblock, block-structured ISA
- How many entries, how many exits does each of them have?
  - What are the corresponding benefits and downsides?
- What are the common benefits?
  - Enable and enlarge the scope of code optimizations
  - Reduce fetch breaks; increase fetch rate
- What are the common downsides?
  - Code bloat (code size increase)
  - Wasted work if control flow deviates from enlarged block’s path

What about loops?

- Unrolling
- Software pipelining
Loop Unrolling

- Idea: Replicate loop body multiple times within an iteration
  + Reduces loop maintenance overhead
    - Induction variable increment or loop condition test
  + Enlarges basic block (and analysis scope)
    - Enables code optimization and scheduling opportunities
  -- What if iteration count not a multiple of unroll factor? (need extra code to detect this)
  -- Increases code size

Software Pipelining

- Software pipelining is an instruction scheduling technique that reorders the instructions in a loop.
  - Possibly moving instructions from one iteration to the previous or the next iteration.
  - Very large improvements in running time are possible.
- The first serious approach to software pipelining was presented by Aiken & Nicolau.
  - Impractical as it ignores resource hazards (focusing only on data-dependence constraints).
    - But sparked a large amount of follow-on research.

Goal of SP

- Increase distance between dependent operations by moving destination operation to a later iteration

Assume all have latency of 2

A: \( a \leftarrow \text{ld} \ [d] \)
B: \( b \leftarrow a \ast a \)
C: \( \text{st} \ [d], b \)
D: \( d \leftarrow d + 4 \)

Can we decrease the latency?

- Lets unroll

A: \( a \leftarrow \text{ld} \ [d] \)
B: \( b \leftarrow a \ast a \)
C: \( \text{st} \ [d], b \)
D: \( d \leftarrow d + 4 \)
A1: \( a \leftarrow \text{ld} \ [d] \)
B1: \( b \leftarrow a \ast a \)
C1: \( \text{st} \ [d], b \)
D1: \( d \leftarrow d + 4 \)
Rename variables

A: a ← ld [d]
B: b ← a * a
C: c ← st [d], b
D: d1 ← d + 4

A1: a1 ← ld [d1]
B1: b1 ← a1 * a1
C1: c ← st [d1], b1
D1: d ← d1 + 4

Schedule

A: a ← ld [d]
B: b ← a * a
C: c ← st [d], b
D: d1 ← d + 4

A1: a1 ← ld [d1]
B1: b1 ← a1 * a1
C1: c ← st [d1], b1
D1: d ← d1 + 4

Unroll Some More

A: a ← ld [d]
B: b ← a * a
C: c ← st [d], b
D: d1 ← d + 4

A1: a1 ← ld [d1]
B1: b1 ← a1 * a1
C1: c ← st [d1], b1
D1: d ← d1 + 4

A2: a2 ← ld [d2]
B2: b2 ← a2 * a2
C2: c ← st [d2], b2
D2: d ← d2 + 4

Unroll Some More

A: a ← ld [d]
B: b ← a * a
C: c ← st [d], b
D: d1 ← d + 4

A1: a1 ← ld [d1]
B1: b1 ← a1 * a1
C1: c ← st [d1], b1
D1: d ← d1 + 4

A2: a2 ← ld [d2]
B2: b2 ← a2 * a2
C2: c ← st [d2], b2
D2: d ← d2 + 4

A3: a3 ← ld [d3]
B3: b3 ← a3 * a3
C3: c ← st [d3], b3
D3: d ← d3 + 4
One More Time

Can Rearrange

Rearrange

A: a ← ld [d]
B: b ← a * a
C: c ← st [d], b
D: d ← d + 4
A1: a1 ← ld [d1]
B1: b1 ← a1 * a1
C1: c1 ← st [d1], b1
D1: d1 ← d1 + 4
A2: a2 ← ld [d2]
B2: b2 ← a2 * a2
C2: c2 ← st [d2], b2
D2: d ← d2 + 4

Rearrange

A: a ← ld [d]
B: b ← a * a
C: c ← st [d], b
D: d ← d + 4
A1: a1 ← ld [d1]
B1: b1 ← a1 * a1
C1: c1 ← st [d1], b1
D1: d1 ← d1 + 4
A2: a2 ← ld [d2]
B2: b2 ← a2 * a2
C2: c2 ← st [d2], b2
D2: d ← d2 + 4
Goal of SP

- Increase distance between dependent operations by moving destination operation to a later iteration
- But also, to uncover ILP across iteration boundaries!

Example

Assume operating on a infinite wide machine
Example

Assume operating on an infinite wide machine

\[
\begin{align*}
& A_0 \\
& A_1 \\
& A_i \\
& B_0 \\
& B_i \\
& C_i \\
& B_{i-1} \\
& C_{i-1} \\
& C_i \\
\end{align*}
\]

Prolog

loop body

epilog

Dealing with exit conditions

for (i=0; i<N; i++)
{
    i=0
    if (i >= N) goto done
    A_i
    B_i
    C_i
    if (i+1 == N) goto last
    i=1
    A_i
    B_i
    C_i
    if (i+2 == N) goto epilog
    i=2
    A_i
    B_i
    C_i
}

loop:
A_i
B_i
C_i
i++
if (i < N) goto loop
epilog:
\[
B_i \\
C_i \\
\]
last:
c_i
done:

Loop Unrolling V. SP

For SuperScalar or VLIW
- Loop Unrolling reduces loop overhead
- Software Pipelining reduces fill/drain
- Best is if you combine them

VLIW

- Depends on the compiler
  - As often is the case: compiler algs developed for VLIW are relevant to superscalar, e.g., software pipelining.
  - Why wouldn't SS dynamically "software pipeline?"
- As always: Is there enough statically knowable parallelism?
- What about wasted FUs? Code bloat?
- Many DSPs are VLIW. Why?