Memory Consistency Models

Daniel Sanchez
Computer Science and Artificial Intelligence Lab
M.I.T.
Coherence vs Consistency

• Cache coherence makes private caches invisible to software
  – Concerns reads/writes to a single memory location
Coherence vs Consistency

- **Cache coherence** makes private caches invisible to software
  - Concerns reads/writes to a single memory location

- **Memory consistency** models precisely specify how memory behaves with respect to read and write operations from multiple processors
  - Concerns reads/writes to multiple memory locations
Why Consistency Matters

Initial memory contents

\begin{itemize}
\item a: 0
\item flag: 0
\end{itemize}

Processor 1

\begin{itemize}
\item Store (a), 10;
\item Store (flag), 1;
\end{itemize}

Processor 2

\begin{itemize}
\item L: Load r1, (flag);
\item if r_1 == 0 goto L;
\item Load r2, (a);
\end{itemize}

What value does r2 hold after both processors finish running this code?
Why Consistency Matters

*Initial memory contents*

- a: 0
- flag: 0

**Processor 1**

- Store (a), 10;
- Store (flag), 1;

**Processor 2**

- L: Load r1, (flag);
- if r₁ == 0 goto L;
- Load r2, (a);

- What value does r2 hold after both processors finish running this code?

It depends on the order in which processor 2 observes processor 1’s stores!
Why Consistency Matters

*Initial memory contents*

a: 0
flag: 0

**Processor 1**

Store (a), 10;
Store (flag), 1;

**Processor 2**

L: Load r1, (flag);
if r1 == 0 goto L;
Load r2, (a);

- What value does r2 hold after both processors finish running this code?

  It depends on the order in which processor 2 observes processor 1’s stores!

  10 if Store (flag) > Store (a); 0 or 10 otherwise
Sequential Consistency
A Straightforward Memory Model

“A system is sequentially consistent if the result of any execution is the same as if the operations of all the processors were executed in some sequential order, and the operations of each individual processor appear in the order specified by the program”

Leslie Lamport

Sequential Consistency =

arbitrary order-preserving interleaving
of memory references of sequential programs
Sequential Consistency

- In-order instruction execution
- Atomic loads and stores

Processor 1
Store \( (a) \), 10;
Store \( (\text{flag}) \), 1;

Processor 2
L: Load \( r_1, (\text{flag}) \);
if \( r_1 == 0 \) goto L;
Load \( r_2, (a) \);
Sequential Consistency

- In-order instruction execution
- Atomic loads and stores

**Processor 1**
- Store (a), 10;
- Store (flag), 1;

**Processor 2**
- Load r1, (flag);
- if \( r_1 == 0 \) goto L;
- Load r2, (a);
Sequential Consistency

- In-order instruction execution
- Atomic loads and stores

Processor 1
Store \( (a), 10; \)
Store \( (\text{flag}), 1; \)

Processor 2
L:
Load \( r_1, (\text{flag}); \)
if \( r_1 == 0 \) goto L;
Load \( r_2, (a); \)
Sequential Consistency

- In-order instruction execution
- Atomic loads and stores

```plaintext
Processor 1
Store (a), 10;
Store (flag), 1;

Processor 2
L: Load r1, (flag);
if r1 == 0 goto L;
Load r2, (a);
```
Sequential Consistency

- In-order instruction execution
- Atomic loads and stores

**Processor 1**
- Store (a), 10;
- Store (flag), 1;

**Processor 2**
- L: Load r1, (flag);
- if r_1 == 0 goto L;
- Load r2, (a);
Sequential Consistency

- In-order instruction execution
- Atomic loads and stores

*SC is easy to understand, but architects and compiler writers want to violate it for performance*
Memory Model Issues

Architectural optimizations that are correct for uniprocessors often violate sequential consistency and result in a new memory model for multiprocessors.
Consistency Models

• Sequential Consistency
  – All reads and writes in order

• Relaxed Consistency (one or more of the following)
  – Loads may be reordered after loads
    • e.g., PA-RISC, Power, Alpha
  – Loads may be reordered after stores
    • e.g., PA-RISC, Power, Alpha
  – Stores may be reordered after stores
    • e.g., PA-RISC, Power, Alpha, PSO
  – Stores may be reordered after loads
    • e.g., PA-RISC, Power, Alpha, PSO, TSO, x86
  – Other more esoteric characteristics
    • e.g., Alpha
Committed Store Buffers

- CPU can continue execution while earlier committed stores are still propagating through memory system
  - Processor can commit other instructions (including loads and stores) while first store is committing to memory
  - Committed store buffer can be combined with speculative store buffer in an out-of-order CPU

- Local loads can bypass values from buffered stores to same address
Example 1: Store Buffers

<table>
<thead>
<tr>
<th>Process 1</th>
<th>Process 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Store (flag₁), 1;</td>
<td>Store (flag₂), 1;</td>
</tr>
<tr>
<td>Load r₁, (flag₂);</td>
<td>Load r₂, (flag₁);</td>
</tr>
</tbody>
</table>

Initially, all memory locations contain zeros
Example 1: Store Buffers

<table>
<thead>
<tr>
<th>Process 1</th>
<th>Process 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Store (flag₁), 1;</td>
<td>Store (flag₂), 1;</td>
</tr>
<tr>
<td>Load r₁, (flag₂);</td>
<td>Load r₂, (flag₁);</td>
</tr>
</tbody>
</table>

Question: Is it possible that $r₁ = 0$ and $r₂ = 0$?

Initially, all memory locations contain zeros
# Example 1: Store Buffers

<table>
<thead>
<tr>
<th>Process 1</th>
<th>Process 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Store ((\text{flag}_1), 1;)</td>
<td>Store ((\text{flag}_2), 1;)</td>
</tr>
<tr>
<td>Load (r_1, (\text{flag}_2);)</td>
<td>Load (r_2, (\text{flag}_1);)</td>
</tr>
</tbody>
</table>

**Question:** Is it possible that \(r_1=0\) and \(r_2=0\)?

- **Sequential consistency:** No
Example 1: Store Buffers

Process 1

Store (flag₁), 1;
Load r₁, (flag₂);

Process 2

Store (flag₂), 1;
Load r₂, (flag₁);

Question: Is it possible that r₁ = 0 and r₂ = 0?

• Sequential consistency: No

• Suppose Loads can go ahead of Stores waiting in the store buffer: Yes!

Initially, all memory locations contain zeros
Example 1: Store Buffers

<table>
<thead>
<tr>
<th>Process 1</th>
<th>Process 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Store (flag₁), 1;</td>
<td>Store (flag₂), 1;</td>
</tr>
<tr>
<td>Load r₁, (flag₂);</td>
<td>Load r₂, (flag₁);</td>
</tr>
</tbody>
</table>

Question: Is it possible that $r₁=0$ and $r₂=0$?
- Sequential consistency: No
- Suppose Loads can go ahead of Stores waiting in the store buffer: Yes!

Total Store Order (TSO):

Sun SPARC, IBM 370

Initially, all memory locations contain zeros
Example 2: Store-Load Bypassing

<table>
<thead>
<tr>
<th>Process 1</th>
<th>Process 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Store (flag₁), 1;</td>
<td>Store (flag₂), 1;</td>
</tr>
<tr>
<td>Load r₃, (flag₁);</td>
<td>Load r₄, (flag₂);</td>
</tr>
<tr>
<td>Load r₁, (flag₂);</td>
<td>Load r₂, (flag₁);</td>
</tr>
</tbody>
</table>
### Example 2: Store-Load Bypassing

<table>
<thead>
<tr>
<th>Process 1</th>
<th>Process 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Store ((\text{flag}_1)), 1;</td>
<td>Store ((\text{flag}_2)), 1;</td>
</tr>
<tr>
<td>Load (r_3), ((\text{flag}_1));</td>
<td>Load (r_4), ((\text{flag}_2));</td>
</tr>
<tr>
<td>Load (r_1), ((\text{flag}_2));</td>
<td>Load (r_2), ((\text{flag}_1));</td>
</tr>
</tbody>
</table>

**Question:** *Do extra Loads have any effect?*
Example 2: Store-Load Bypassing

<table>
<thead>
<tr>
<th>Process 1</th>
<th>Process 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Store ((flag_1), 1;)</td>
<td>Store ((flag_2), 1;)</td>
</tr>
<tr>
<td>Load (r_3, (flag_1);)</td>
<td>Load (r_4, (flag_2);)</td>
</tr>
<tr>
<td>Load (r_1, (flag_2);)</td>
<td>Load (r_2, (flag_1);)</td>
</tr>
</tbody>
</table>

**Question:** Do extra Loads have any effect?
- **Sequential consistency:** No
Example 2: Store-Load Bypassing

<table>
<thead>
<tr>
<th>Process 1</th>
<th>Process 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Store (flag(_1)), 1;</td>
<td>Store (flag(_2)), 1;</td>
</tr>
<tr>
<td>Load r(_3), (flag(_1));</td>
<td>Load r(_4), (flag(_2));</td>
</tr>
<tr>
<td>Load r(_1), (flag(_2));</td>
<td>Load r(_2), (flag(_1));</td>
</tr>
</tbody>
</table>

Question: Do extra Loads have any effect?
- **Sequential consistency:** No
- Suppose Store-Load bypassing is permitted in the store buffer
Example 2: Store-Load Bypassing

<table>
<thead>
<tr>
<th>Process 1</th>
<th>Process 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Store (flag_1), 1;</td>
<td>Store (flag_2), 1;</td>
</tr>
<tr>
<td>Load r_3, (flag_1);</td>
<td>Load r_4, (flag_2);</td>
</tr>
<tr>
<td>Load r_1, (flag_2);</td>
<td>Load r_2, (flag_1);</td>
</tr>
</tbody>
</table>

Question: Do extra Loads have any effect?
- **Sequential consistency:** No
- Suppose Store-Load bypassing is permitted in the store buffer
  - No effect in Sparc’s TSO model, still not SC
Example 2: Store-Load Bypassing

<table>
<thead>
<tr>
<th>Process 1</th>
<th>Process 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Store (flag₁), 1;</td>
<td>Store (flag₂), 1;</td>
</tr>
<tr>
<td>Load r₃, (flag₁);</td>
<td>Load r₄, (flag₂);</td>
</tr>
<tr>
<td>Load r₁, (flag₂);</td>
<td>Load r₂, (flag₁);</td>
</tr>
</tbody>
</table>

Question: Do extra Loads have any effect?
- **Sequential consistency:** No
- Suppose *Store-Load bypassing is permitted in the store buffer*
  - No effect in Sparc’s TSO model, still not SC
  - In IBM 370, a load cannot return a written value until it is visible to other processors => implicitly adds a memory fence, looks like SC
Interleaved Memory System

- Achieve greater throughput by spreading memory addresses across two or more parallel memory subsystems
  - In snooping system, can have two or more snoops in progress at same time (e.g., Sun UE10K system has four interleaved snooping busses)
  - Greater bandwidth from main memory system as two memory modules can be accessed in parallel
## Example 3: Non-FIFO Store buffers

<table>
<thead>
<tr>
<th>Process 1</th>
<th>Process 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Store (a), 1;</td>
<td>Load $r_1$, (flag);</td>
</tr>
<tr>
<td>Store (flag), 1;</td>
<td>Load $r_2$, (a);</td>
</tr>
</tbody>
</table>
### Example 3: Non-FIFO Store buffers

<table>
<thead>
<tr>
<th>Process 1</th>
<th>Process 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Store (a), 1;</td>
<td>Load $r_1$, (flag);</td>
</tr>
<tr>
<td>Store (flag), 1;</td>
<td>Load $r_2$, (a);</td>
</tr>
</tbody>
</table>

**Question:** Is it possible that $r_1 = 1$ but $r_2 = 0$?
## Example 3: Non-FIFO Store buffers

<table>
<thead>
<tr>
<th>Process 1</th>
<th>Process 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Store (a), 1;</td>
<td>Load $r_1$, (flag);</td>
</tr>
<tr>
<td>Store (flag), 1;</td>
<td>Load $r_2$, (a);</td>
</tr>
</tbody>
</table>

**Question:** Is it possible that $r_1 = 1$ but $r_2 = 0$?

- **Sequential consistency:** No
Example 3: Non-FIFO Store buffers

<table>
<thead>
<tr>
<th>Process 1</th>
<th>Process 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Store (a), 1;</td>
<td>Load $r_1$, (flag);</td>
</tr>
<tr>
<td>Store (flag), 1;</td>
<td>Load $r_2$, (a);</td>
</tr>
</tbody>
</table>

Question: Is it possible that $r_1=1$ but $r_2=0$?
- **Sequential consistency:** No
- **With non-FIFO store buffers:** Yes
### Example 3: Non-FIFO Store buffers

<table>
<thead>
<tr>
<th>Process 1</th>
<th>Process 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Store (a), 1;</td>
<td>Load $r_1$, (flag);</td>
</tr>
<tr>
<td>Store (flag), 1;</td>
<td>Load $r_2$, (a);</td>
</tr>
</tbody>
</table>

**Question:** Is it possible that $r_1=1$ but $r_2=0$?
- **Sequential consistency:** No
- **With non-FIFO store buffers:** Yes

Sparc’s PSO memory model
Example 4: Non-Blocking Caches

\[
\begin{array}{ll}
\text{Process 1} & \text{Process 2} \\
\text{Store (a), 1;} & \text{Load } r_1, (\text{flag}); \\
\text{Store (flag), 1;} & \text{Load } r_2, (a); \\
\end{array}
\]
Example 4: Non-Blocking Caches

<table>
<thead>
<tr>
<th>Process 1</th>
<th>Process 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Store (a), 1;</td>
<td>Load (r_1), (flag);</td>
</tr>
<tr>
<td>Store (flag), 1;</td>
<td>Load (r_2), (a);</td>
</tr>
</tbody>
</table>

Question: Is it possible that \(r_1=1\) but \(r_2=0\)?
**Example 4: Non-Blocking Caches**

<table>
<thead>
<tr>
<th>Process 1</th>
<th>Process 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Store (a), 1;</td>
<td>Load $r_1$, (flag);</td>
</tr>
<tr>
<td>Store (flag), 1;</td>
<td>Load $r_2$, (a);</td>
</tr>
</tbody>
</table>

**Question:** Is it possible that $r_1=1$ but $r_2=0$?

- **Sequential consistency:** No
## Example 4: Non-Blocking Caches

### Process 1
- Store (a), 1;
- Store (flag), 1;

### Process 2
- Load r₁, (flag);
- Load r₂, (a);

**Question:** Is it possible that $r₁=1$ but $r₂=0$?
- **Sequential consistency:** No
- **Assuming stores are ordered:** Yes because Loads can be reordered
### Example 4: Non-Blocking Caches

<table>
<thead>
<tr>
<th>Process 1</th>
<th>Process 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Store (a), 1;</td>
<td>Load $r_1$, (flag);</td>
</tr>
<tr>
<td>Store (flag), 1;</td>
<td>Load $r_2$, (a);</td>
</tr>
</tbody>
</table>

**Question:** Is it possible that $r_1=1$ but $r_2=0$?

- **Sequential consistency:** No
- **Assuming stores are ordered:** Yes because Loads can be reordered

Alpha, Sparc’s RMO, PowerPC’s WO
Example 5: Register Renaming

<table>
<thead>
<tr>
<th>Process 1</th>
<th>Process 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Store (flag₁), ( r_1 );</td>
<td>Store (flag₂), ( r_2 );</td>
</tr>
<tr>
<td>Load ( r_1 ), (flag₂);</td>
<td>Load ( r_2 ), (flag₁);</td>
</tr>
</tbody>
</table>

Initially both \( r_1 \) and \( r_2 \) contain 1.
Example 5: Register Renaming

Initially both \( r_1 \) and \( r_2 \) contain 1.

<table>
<thead>
<tr>
<th>Process 1</th>
<th>Process 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Store ((\text{flag}_1), r_1);</td>
<td>Store ((\text{flag}_2), r_2);</td>
</tr>
<tr>
<td>Load ( r_1 ), ((\text{flag}_2));</td>
<td>Load ( r_2 ), ((\text{flag}_1));</td>
</tr>
</tbody>
</table>

Initially both \( r_1 \) and \( r_2 \) contain 1.

Question: Is it possible that \( r_1 = 0 \) but \( r_2 = 0 \)?
### Example 5: Register Renaming

Initially both $r_1$ and $r_2$ contain 1.

<table>
<thead>
<tr>
<th>Process 1</th>
<th>Process 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Store (flag$_1$), $r_1$;</td>
<td>Store (flag$_2$), $r_2$;</td>
</tr>
<tr>
<td>Load $r_1$, (flag$_2$);</td>
<td>Load $r_2$, (flag$_1$);</td>
</tr>
</tbody>
</table>

**Question:** Is it possible that $r_1=0$ but $r_2=0$?
- **Sequential consistency:** No
## Example 5: Register Renaming

<table>
<thead>
<tr>
<th>Process 1</th>
<th>Process 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Store (flag$_1$), $r_1$;</td>
<td>Store (flag$_2$), $r_2$;</td>
</tr>
<tr>
<td>Load $r_1$, (flag$_2$);</td>
<td>Load $r_2$, (flag$_1$);</td>
</tr>
</tbody>
</table>

Initially both $r_1$ and $r_2$ contain 1.

**Question:** Is it possible that $r_1=0$ but $r_2=0$?

- **Sequential consistency:** No
- **Register renaming:** Yes because it removes anti-dependencies
### Example 5: Register Renaming

Initially both $r_1$ and $r_2$ contain 1.

<table>
<thead>
<tr>
<th>Process 1</th>
<th>Process 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Store (flag$_1$), $r_1$;</td>
<td>Store (flag$_2$), $r_2$;</td>
</tr>
<tr>
<td>Load $r_1$, (flag$_2$);</td>
<td>Load $r_2$, (flag$_1$);</td>
</tr>
</tbody>
</table>

**Question:** Is it possible that $r_1=0$ but $r_2=0$?
- **Sequential consistency:** No
- **Register renaming:** Yes because it removes anti-dependencies
Example 5: Register Renaming

Initially both $r_1$ and $r_2$ contain 1.

**Question:** Is it possible that $r_1=0$ but $r_2=0$?

- **Sequential consistency:** No
- **Register renaming:** Yes because it removes anti-dependencies
Initially both $r_1$ and $r_2$ contain 1.

**Question:** Is it possible that $r_1 = 0$ but $r_2 = 0$?

- **Sequential consistency:** No
- **Register renaming:** Yes because it removes anti-dependencies
Example 5: Register Renaming

Initially both $r_1$ and $r_2$ contain 1.

**Process 1**
- Store (flag$_1$), $r_1$;
- Load $r_1$, (flag$_2$);

**Process 2**
- Store (flag$_2$), $r_2$;
- Load $r_2$, (flag$_1$);

Question: Is it possible that $r_1=0$ but $r_2=0$?

- **Sequential consistency:** No
- **Register renaming:** Yes because it removes anti-dependencies
Example 5: Register Renaming

Initially both \( r_1 \) and \( r_2 \) contain 1.

**Process 1**
- Store (flag\(_1\)), \( r_1 \);
- Load \( r_1 \), (flag\(_2\));

**Process 2**
- Store (flag\(_2\)), \( r_2 \);
- Load \( r_2 \), (flag\(_1\));

Register renaming will eliminate this edge

Question: Is it possible that \( r_1 = 0 \) but \( r_2 = 0 \)?
- **Sequential consistency:** No
- **Register renaming:** Yes because it removes anti-dependencies
Initially both $r_1$ and $r_2$ contain 1.

Question: Is it possible that $r_1=0$ but $r_2=0$?
- **Sequential consistency:** No
- **Register renaming:** Yes because it removes anti-dependencies
Example 6: Speculative Execution

<table>
<thead>
<tr>
<th>Process 1</th>
<th>Process 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Store (a), 1;</td>
<td>L: Load $r_1$, (flag);</td>
</tr>
<tr>
<td>Store (flag), 1;</td>
<td>if $r_1 == 0$ goto L;</td>
</tr>
<tr>
<td></td>
<td>Load $r_2$, (a);</td>
</tr>
</tbody>
</table>
Example 6: Speculative Execution

<table>
<thead>
<tr>
<th>Process 1</th>
<th>Process 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Store (a), 1;</td>
<td>L: Load r_1, (flag);</td>
</tr>
<tr>
<td>Store (flag), 1;</td>
<td>if r_1 == 0 goto L;</td>
</tr>
<tr>
<td></td>
<td>Load r_2, (a);</td>
</tr>
</tbody>
</table>

Question: Is it possible that $r_1=1$ but $r_2=0$?
# Example 6: Speculative Execution

<table>
<thead>
<tr>
<th>Process 1</th>
<th>Process 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Store (a), 1;</td>
<td>L: Load r₁, (flag);</td>
</tr>
<tr>
<td>Store (flag), 1;</td>
<td>if r₁ == 0 goto L;</td>
</tr>
<tr>
<td></td>
<td>Load r₂, (a);</td>
</tr>
</tbody>
</table>

**Question:** Is it possible that $r₁=1$ but $r₂=0$?

- **Sequential consistency:** *No*
Example 6: Speculative Execution

<table>
<thead>
<tr>
<th>Process 1</th>
<th>Process 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Store (a), 1;</td>
<td>L: Load (r_1), (flag);</td>
</tr>
<tr>
<td>Store (flag), 1;</td>
<td>if (r_1 == 0) goto L;</td>
</tr>
<tr>
<td></td>
<td>Load (r_2), (a);</td>
</tr>
</tbody>
</table>

Question: Is it possible that \(r_1 = 1\) but \(r_2 = 0\)?

- **Sequential consistency**: No
- **With speculative loads**: Yes even if the stores are ordered
Example 7: Address Speculation

<table>
<thead>
<tr>
<th>Process 1</th>
<th>Process 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Store (flag₁), ( r₁ );</td>
<td>Store (flag₂), ( r₃ );</td>
</tr>
<tr>
<td>Load ( r₂ ), (flag₂);</td>
<td>Load ( r₄ ), (flag₁);</td>
</tr>
</tbody>
</table>

Initially both \( r₁ \) and \( r₃ \) contain 1.

Flag₁ and flag₂ are registers pointing at memory locations L14-16.
Example 7: Address Speculation

Initially both $r_1$ and $r_3$ contain 1.

<table>
<thead>
<tr>
<th>Process 1</th>
<th>Process 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Store (flag₁), $r_1$;</td>
<td>Store (flag₂), $r_3$;</td>
</tr>
<tr>
<td>Load $r_2$, (flag₂);</td>
<td>Load $r_4$, (flag₁);</td>
</tr>
</tbody>
</table>

Question: Is it possible that $r_2=0$ but $r_4=0$?

Flag₁ and flag₂ are registers pointing at memory locations L14-16.
Example 7: Address Speculation

Initially both $r_1$ and $r_3$ contain 1.

Process 1
Store (flag_1), $r_1$;
Load $r_2$, (flag_2);

Process 2
Store (flag_2), $r_3$;
Load $r_4$, (flag_1);

Question: Is it possible that $r_2=0$ but $r_4=0$?

• Sequential consistency: No

Flag_1 and flag_2 are registers pointing at memory locations.
Example 7: Address Speculation

Initially both \( r_1 \) and \( r_3 \) contain 1.

<table>
<thead>
<tr>
<th>Process 1</th>
<th>Process 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Store (flag(_1)), r(_1);</td>
<td>Store (flag(_2)), r(_3);</td>
</tr>
<tr>
<td>Load r(_2), (flag(_2));</td>
<td>Load r(_4), (flag(_1));</td>
</tr>
</tbody>
</table>

Question: Is it possible that \( r_2 = 0 \) but \( r_4 = 0 \)?

- **Sequential consistency:** No
- **Address speculation:** Yes because it removes the dependencies between the stores and loads

Flag\(_1\) and flag\(_2\) are registers pointing at memory locations L14-16
Example 7: Address Speculation

Initially both $r_1$ and $r_3$ contain 1.

<table>
<thead>
<tr>
<th>Process 1</th>
<th>Process 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Store (flag_1), $r_1$;</td>
<td>Store (flag_2), $r_3$;</td>
</tr>
<tr>
<td>Load $r_2$, (flag_2);</td>
<td>Load $r_4$, (flag_1);</td>
</tr>
</tbody>
</table>

Question: Is it possible that $r_2=0$ but $r_4=0$?

- **Sequential consistency:** No
- **Address speculation:** Yes because it removes the dependencies between the stores and loads

Flag_1 and flag_2 are registers pointing at memory locations L14-16
Initially both $r_1$ and $r_3$ contain 1.

**Process 1**
- Store (flag$_1$), $r_1$;
- Load $r_2$, (flag$_2$);

**Process 2**
- Store (flag$_2$), $r_3$;
- Load $r_4$, (flag$_1$);

Question: Is it possible that $r_2=0$ but $r_4=0$?
- **Sequential consistency**: No
- **Address speculation**: Yes because it removes the dependencies between the stores and loads

Flag$_1$ and flag$_2$ are registers pointing at memory locations.
Example 7: Address Speculation

Initially both $r_1$ and $r_3$ contain 1.

<table>
<thead>
<tr>
<th>Process 1</th>
<th>Process 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Store ($flag_1$), $r_1$; Load $r_2$, ($flag_2$);</td>
<td>Store ($flag_2$), $r_3$; Load $r_4$, ($flag_1$);</td>
</tr>
</tbody>
</table>

Question: Is it possible that $r_2=0$ but $r_4=0$?
- **Sequential consistency:** No
- **Address speculation:** Yes because it removes the dependencies between the stores and loads

Flag$_1$ and flag$_2$ are registers pointing at memory locations $L_{14-16}$
Example 7: Address Speculation

Initially both $r_1$ and $r_3$ contain 1.

<table>
<thead>
<tr>
<th>Process 1</th>
<th>Process 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Store (flag$_1$), $r_1$;</td>
<td>Store (flag$_2$), $r_3$;</td>
</tr>
<tr>
<td>Load $r_2$, (flag$_2$);</td>
<td>Load $r_4$, (flag$_1$);</td>
</tr>
</tbody>
</table>

Question: Is it possible that $r_2=0$ but $r_4=0$?
- **Sequential consistency:** No
- **Address speculation:** Yes because it removes the dependencies between the stores and loads

Flag$_1$ and flag$_2$ are registers pointing at memory locations L14-16
Example 7: Address Speculation

Initially both $r_1$ and $r_3$ contain 1.

**Process 1**
- Store (flag$_1$), $r_1$
- Load $r_2$, (flag$_2$)

**Process 2**
- Store (flag$_2$), $r_3$
- Load $r_4$, (flag$_1$)

**Address speculation will eliminate this edge**

**Question:** Is it possible that $r_2=0$ but $r_4=0$?

- **Sequential consistency:** No
- **Address speculation:** Yes because it removes the dependencies between the stores and loads

Flag$_1$ and flag$_2$ are registers pointing at memory locations L14-16
Example 7: Address Speculation

Initially both $r_1$ and $r_3$ contain 1.

<table>
<thead>
<tr>
<th>Process 1</th>
<th>Process 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Store (flag$_1$), $r_1$;</td>
<td>Store (flag$_2$), $r_3$;</td>
</tr>
<tr>
<td>Load $r_2$, (flag$_2$);</td>
<td>Load $r_4$, (flag$_1$);</td>
</tr>
</tbody>
</table>

Question: Is it possible that $r_2=0$ but $r_4=0$?

- **Sequential consistency:** No
- **Address speculation:** Yes because it removes the dependencies between the stores and loads

Flag$_1$ and flag$_2$ are registers pointing at memory locations.
Example 7: Address Speculation

Initially both \( r_1 \) and \( r_3 \) contain 1.

<table>
<thead>
<tr>
<th>Process 1</th>
<th>Process 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Store (flag(_1)), ( r_1 ); Store (flag(_2)), ( r_3 );</td>
<td>Store (flag(_2)), ( r_3 ); Load ( r_4 ), (flag(_1));</td>
</tr>
<tr>
<td>Load ( r_2 ), (flag(_2));</td>
<td>Load ( r_4 ), (flag(_1));</td>
</tr>
</tbody>
</table>

Address speculation will eliminate this edge

Question: Is it possible that \( r_2 = 0 \) but \( r_4 = 0 \)?

- **Sequential consistency:** No
- **Address speculation:** Yes because it removes the dependencies between the stores and loads

Flag\(_1\) and flag\(_2\) are registers pointing at memory locations.
Example 8: Store Atomicity

<table>
<thead>
<tr>
<th>Process 1</th>
<th>Process 2</th>
<th>Process 3</th>
<th>Process 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Store (a),1;</td>
<td>Store (a),2;</td>
<td>Load r₁, (a);</td>
<td>Load r₃, (a);</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Load r₂, (a);</td>
<td>Load r₄, (a);</td>
</tr>
</tbody>
</table>
Example 8: Store Atomicity

<table>
<thead>
<tr>
<th>Process 1</th>
<th>Process 2</th>
<th>Process 3</th>
<th>Process 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Store (a),1;</td>
<td>Store (a),2;</td>
<td>Load r₁, (a);</td>
<td>Load r₃, (a);</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Load r₂, (a);</td>
<td>Load r₄, (a);</td>
</tr>
</tbody>
</table>

**Question:** Is it possible that \( r₁=1 \) and \( r₂=2 \) but \( r₃=2 \) and \( r₄=1 \) ?
Example 8: Store Atomicity

<table>
<thead>
<tr>
<th>Process 1</th>
<th>Process 2</th>
<th>Process 3</th>
<th>Process 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Store (a),1;</td>
<td>Store (a),2;</td>
<td>Load ( r_1 ), (a);</td>
<td>Load ( r_3 ), (a);</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Load ( r_2 ), (a);</td>
<td>Load ( r_4 ), (a);</td>
</tr>
</tbody>
</table>

Question: Is it possible that \( r_1 = 1 \) and \( r_2 = 2 \) but \( r_3 = 2 \) and \( r_4 = 1 \)?

- **Sequential consistency:** No
Example 8: Store Atomicity

<table>
<thead>
<tr>
<th>Process 1</th>
<th>Process 2</th>
<th>Process 3</th>
<th>Process 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Store (a),1;</td>
<td>Store (a),2;</td>
<td>Load r₁, (a);</td>
<td>Load r₃, (a);</td>
</tr>
<tr>
<td>Load r₂, (a);</td>
<td>Load r₄, (a);</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Question: Is it possible that \( r₁ = 1 \) and \( r₂ = 2 \) but \( r₃ = 2 \) and \( r₄ = 1 \)?

- **Sequential consistency**: No
- **Even if Loads on a processor are ordered, the different ordering of stores can be observed if the Store operation is not atomic.**
Example 9: Causality

<table>
<thead>
<tr>
<th>Process 1</th>
<th>Process 2</th>
<th>Process 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Store (flag₁), 1;</td>
<td>Load r₁, (flag₁);</td>
<td>Load r₂, (flag₂);</td>
</tr>
<tr>
<td></td>
<td>Store (flag₂), 1;</td>
<td>Load r₃, (flag₁);</td>
</tr>
</tbody>
</table>
### Example 9: Causality

<table>
<thead>
<tr>
<th>Process 1</th>
<th>Process 2</th>
<th>Process 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Store ((\text{flag}_1),1;)</td>
<td>Load (r_1, (\text{flag}_1);)</td>
<td>Load (r_2, (\text{flag}_2);)</td>
</tr>
<tr>
<td>Store ((\text{flag}_2),1;)</td>
<td></td>
<td>Load (r_3, (\text{flag}_1);)</td>
</tr>
</tbody>
</table>

**Question:** Is it possible that \(r_1=1\) and \(r_2=1\) but \(r_3=0\) ?
**Example 9: Causality**

<table>
<thead>
<tr>
<th>Process 1</th>
<th>Process 2</th>
<th>Process 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Store (flag(_1)),1;</td>
<td>Load r(_1), (flag(_1));</td>
<td>Load r(_2), (flag(_2));</td>
</tr>
<tr>
<td></td>
<td>Store (flag(_2)),1;</td>
<td>Load r(_3), (flag(_1));</td>
</tr>
</tbody>
</table>

**Question:** Is it possible that \( r_1 = 1 \) and \( r_2 = 1 \) but \( r_3 = 0 \)?

- **Sequential consistency:** No
Example 9: Causality

<table>
<thead>
<tr>
<th>Process 1</th>
<th>Process 2</th>
<th>Process 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Store (flag₁), 1;</td>
<td>Load r₁, (flag₁);</td>
<td>Load r₂, (flag₂);</td>
</tr>
<tr>
<td>Store (flag₂), 1;</td>
<td></td>
<td>Load r₃, (flag₁);</td>
</tr>
</tbody>
</table>

Question: Is it possible that \( r₁ = 1 \) and \( r₂ = 1 \) but \( r₃ = 0 \)?

- **Sequential consistency:** No
- **With load/load reordering:** Yes

Alpha
Weaker Memory Models & Memory Fence Instructions

Architectures with weaker memory models provide memory fence instructions to prevent otherwise permitted reorderings of loads and stores.

Store \( (a_1), r2; \)

Load \( r1, (a_2); \)

Similarly:

The Load and Store can be reordered if \( a_1 \neq a_2 \). Insertion of Fence\(_{wr}\) will disallow this reordering.
Weaker Memory Models & Memory Fence Instructions

- Architectures with weaker memory models provide memory fence instructions to prevent otherwise permitted reorderings of loads and stores.

\[
\text{Store } (a_1), \text{ r2;}
\]
\[
\text{Fence}_\text{wr}
\]
\[
\text{Load } r1, (a_2);
\]

The Load and Store can be reordered if \( a_1 \neq a_2 \). Insertion of Fence\(_{wr} \) will disallow this reordering.

Similarly:
Weaker Memory Models & Memory Fence Instructions

- Architectures with weaker memory models provide memory fence instructions to prevent otherwise permitted reorderings of loads and stores.

\[
\text{Store } (a_1), \text{ r2}; \\
\text{Fence}_{wr}; \\
\text{Load } r1, (a_2);
\]

The Load and Store can be reordered if \( a_1 \neq a_2 \). Insertion of Fence_{wr} will disallow this reordering.

Similarly:

\[
\text{Fence}_{rr}; \quad \text{Fence}_{rw}; \quad \text{Fence}_{ww};
\]
Weaker Memory Models & Memory Fence Instructions

- Architectures with weaker memory models provide memory fence instructions to prevent otherwise permitted reorderings of loads and stores.

\[
\text{Store (}a_1\text{), r2;} \\
\text{Fence}_{wr} \\
\text{Load r1, (}a_2\text{);}
\]

The Load and Store can be reordered if \(a_1 \neq a_2\). Insertion of Fence\(_{wr}\) will disallow this reordering.

Similarly:

- Fence\(_{rr}\)
- Fence\(_{rw}\)
- Fence\(_{ww}\)

SUN’s Sparc: MEMBAR;
MEMBARRR; MEMBARRW; MEMBARWR; MEMBARWW

PowerPC: Sync; EIEIO
Enforcing Ordering using Fences

**Processor 1**
Store (a),10;
Store (flag),1;

**Processor 2**
L: Load $r_1$, (flag);
  if $r_1 == 0$ goto L;
  Load $r_2$, (a);
Enforcing Ordering using Fences

**Processor 1**
- Store (a), 10;
- Store (flag), 1;
- Fence_{ww};
- Store (flag), 1;

**Processor 2**
- L: Load r₁, (flag);
- if r₁ == 0 goto L;
- Load r₂, (a);

**Processor 1**
- Store (a), 10;
- Store (flag), 1;

**Processor 2**
- L: Load r₁, (flag);
- if r₁ == 0 goto L;
- Fence_{rr};
- Load r₂, (a);
Enforcing Ordering using Fences

**Processor 1**
Store (a),10;
Store (flag),1;

**Processor 1**
Store (a),10;
Fence$_{ww}$;
Store (flag),1;

**Processor 2**
L: Load $r_1$, (flag);
if $r_1 == 0$ goto L;
Load $r_2$, (a);

**Processor 2**
L: Load $r_1$, (flag);
if $r_1 == 0$ goto L;
Fence$_{rr}$;
Load $r_2$, (a);

*Weak ordering*
Weaker (Relaxed) Memory Models

- Alpha, Sparc PowerPC, ...
- Store is globally performed
- TSO, PSO, RMO, ...
- Write-buffers
- RMO=W0?
- SMP, DSM, CMP
Weaker (Relaxed) Memory Models

- Hard to understand and remember

Alpha, Sparc, PowerPC, ...

Store is globally performed

TSO, PSO, RMO, ...

Write-buffers

RMO=WQO?

SMP, DSM, CMP
Weaker (Relaxed) Memory Models

- Hard to understand and remember
- Unstable - *Modèle de l’année*
Weaker (Relaxed) Memory Models

- Hard to understand and remember
- Unstable - *Modèle de l’année*
- Abandon weaker memory models in favor of implementing SC
Implementing SC

1. The memory operations of each individual processor appear to all processors in the order the requests are made to the memory

   - Provided by cache coherence, which ensures that all processors observe the same order of loads and stores to an address

2. Any execution is the same as if the operations of all the processors were executed in some sequential order

   - Provided by enforcing a dependence between each memory operation and the following one
SC Data Dependence
SC Data Dependence

- **Stall**
  - Use *in-order execution and blocking caches*
  - Cache coherence plus allowing a processor to have only one request in flight at a time will provide SC
SC Data Dependence

• **Stall**
  - Use in-order execution and blocking caches
    • Cache coherence plus allowing a processor to have only one request in flight at a time will provide SC

• **Change architecture** ⇒ **Relaxed memory models**
  - Use OOO and non-blocking caches
    • Cache coherence and allowing multiple concurrent requests (to different addresses) gives high performance
    • Add fence operations to force ordering when needed
SC Data Dependence

• **Stall**
  - Use in-order execution and blocking caches
  - Cache coherence plus allowing a processor to have only one request in flight at a time will provide SC

• **Change architecture ⇒ Relaxed memory models**
  - Use OOO and non-blocking caches
    - Cache coherence and allowing multiple concurrent requests (to different addresses) gives high performance
    - Add fence operations to force ordering when needed

• **Speculate...**
Sequential Consistency Speculation

- Local load-store ordering uses standard OOO mechanism

- Globally non-speculative stores
  - Stores execute at commit -> stores are in-order!

- Globally speculative loads
  - **Guess** at issue that the memory location used by a load will not change between issue and commit of the instruction
    - this is equivalent to loads happening in-order at commit
      - **Check** at commit by remembering all loads addresses starting at issue and watching for writes to that location.

- **Data Management** for rollback relies on the basic out-of-order speculative data management used for uni-processor rollback and instruction re-execution.
SC Speculative Behavior

CPU A
SC Speculative Behavior

1: ST A

CPU A
SC Speculative Behavior

1: ST A

2: LD A
SC Speculative Behavior

1: ST A
2: LD A

CPU A
SC Speculative Behavior

1: ST A
2: LD A
3: LD A

CPU A
SC Speculative Behavior

1: ST A
2: LD A
3: LD A

CPU A
SC Speculative Behavior

1: ST A
2: LD A
3: LD A
4: ST A

CPU A
SC Speculative Behavior

1: ST A
2: LD A
3: LD A
4: ST A

CPU A
CPU B
SC Speculative Behavior

1: ST A
2: LD A
3: LD A
4: ST A
SC Speculative Behavior

1: ST A
2: LD A
3: LD A
4: ST A
SC Speculative Behavior

1: ST A
2: LD A
3: LD A
4: ST A

CPU A

CPU B

ST A
SC Speculative Behavior

1: ST A
2: LD A
3: LD A
4: ST A

CPU A

CPU B

ST A
SC Speculative Behavior

1: ST A
2: LD A
3: LD A
4: ST A

CPU A

CPU B

ST A
SC Speculative Behavior

1: ST A
2: LD A
3: LD A
4: ST A
SC Speculative Behavior

1: ST A
2: LD A
3: LD A
4: ST A
Properly Synchronized Programs

• Very few programmers do programming that relies on SC; instead, they use higher-level synchronization primitives
  – locks, semaphores, monitors, atomic transactions

• A “properly synchronized program” is one where each shared writable variable is protected (say, by a lock) so that there is no race in updating the variable
  – There is still race to get the lock
  – There is no way to check if a program is properly synchronized

• For properly synchronized programs, instruction reordering does not matter as long as updated values are committed before leaving a locked region
Release Consistency

[Garachorloo 1990]

- Only care about inter-processor memory ordering at thread synchronization points, not in between
- Can treat all synchronization instructions as the only ordering points

...  
Acquire(lock)  // All following loads get most recent written values  
...  Read and write shared data  
Release(lock)  // All preceding writes are globally visible before  
// lock is freed.  
...
Takeaways

- SC is too low level a programming model. High-level programming should be based on critical sections & locks, atomic transactions, monitors, ...

- High-level parallel programming should be oblivious of memory model issues
  - Programmer should not be affected by changes in the memory model

- ISA definition for Load, Store, Memory Fence, synchronization instructions should
  - Be precise
  - Permit maximum flexibility in hardware implementation
  - Permit efficient implementation of high-level parallel constructs
Thank you!

Next Lecture:
On-Chip Networks