Treadmarks: Distributed Shared Memory on Standard Workstations and Operating Systems

P. Keleher, A. Cox, S. Dwarkadas, and W. Zwaenepoel
The Winter Usenix Conference 1994

2008-22952
Jun Lee
DSM (distributed shared memory)

- A **software system** for parallel computation
  - Shares distributed memories
  - Easier programming
    - Provide a single global address space
DSM (distributed shared memory)

- No widely available DSM implementations
  - In-house research platforms
  - Kernel modifications
  - Poor performance
    - Imitating consistency protocols of hardware
    - False sharing
Treadmarks

- **Objectives**
  - Commercially available workstations and OS
    - Standard Unix system on DECstation
  - Efficient user-level DSM implementation
    - Reduce communication overhead

- **Design**
  - LRC (lazy release consistency)
  - Multiple writer protocol
  - Lazy diff creation
Consistency protocol (SC)

- **Sequential Consistency**
  - Every write visible “immediately”
  - Single writer
Consistency protocol (SC)

- Sequential Consistency
  - Every write visible “immediately”
  - Single writer

Big problem with page size granularity
Consistency protocol (SC)

- **Sequential Consistency**
  - Every write visible “immediately”
  - Single writer

```
Page X
a b c
```

```
P0
W(x0):a
W(x2):c
```

```
P1
W(x1):b
W(x3):d
```

```
Page X
a b c d
```

- False sharing
Consistency protocol (RC)

- **Release Consistency**
  - Relaxed memory consistency model
    - delay making its changes visible to other processors until certain synchronization accesses occurs
  - Synchronization points
    - Acquire(), Release() (similar to locks, barriers)
  - Two types
    - ERC (eager), LRC (lazy)
Consistency protocol (RC)

- **Release Consistency**
  - Acquire() and release() are sequentially consistent
    - Release() is performed after all previous operations have completed
    - Operations are performed after previous acquire() have been performed

- Acquire() and release() pair between conflicting accesses
  - SC and RC produce the same results.
Consistency protocol (RC)

- ERC
  - Write information is delivered at the release point
Consistency protocol (RC)

- ERC
  - Write information is delivered at the release point
Consistency protocol (RC)

- **LRC**
  - The delivery is postponed until the acquire
  - Fewer messages than ERC
Consistency protocol (RC)

- ERC vs. LRC

**ERC**

<table>
<thead>
<tr>
<th></th>
<th>P0</th>
<th>P1</th>
<th>P2</th>
<th>P3</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Acquire(L)</td>
<td>R(a):0</td>
<td>R(a):0</td>
<td>R(a):0</td>
</tr>
<tr>
<td></td>
<td>W(a):1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Release(L)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**LRC**

<table>
<thead>
<tr>
<th></th>
<th>P0</th>
<th>P1</th>
<th>P2</th>
<th>P3</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Acquire(L)</td>
<td>R(a):1</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>R(a):0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Release(L)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Consistency protocol (RC)

- ERC vs. LRC
Multiple writer protocol

Page X
a c

W(x0):a
Acquire(L)
W(x2):c
Release(L)
P0

Page X
ab c
W(x1):b
Acquire(L)
P1

Page X
abcd
W(x3):?

False sharing
Multiple writer protocol

Page X

P0

W(x0):a

Acquire(L)

W(x2):c

Release(L)

P1

W(x1):b

Acquire(L)

W(x3):?

W(x3):d

Release(L)

Page X

a b c d

Page X

abc
c

W(x0):a

W(x2):c

False sharing

Page X

abcd

W(x1):b

W(x3):d
Twin and Diff

P0

Page X

\[
\begin{array}{cc}
\text{a} & \text{c} \\
\text{Twin X} & \text{Diff} \\
\end{array}
\]

\[
\begin{align*}
W(x0):a \\
\text{Acquire(L)} \\
W(x2):c \\
\text{Release(L)} \\
\end{align*}
\]

P1

Page X

\[
\begin{array}{cccc}
\text{a} & \text{b} & \text{c} & \text{d} \\
\text{Twin X} & \text{Diff} \\
\end{array}
\]

\[
\begin{align*}
W(x1):b \\
\text{Acquire(L)} \\
W(x3):? \\
W(x3):d \\
\text{Release(L)} \\
\end{align*}
\]
Twin and Diff

Page X

P0

W(x0):a

Acquire(L)

Twin X

Diff

W(x2):c

Release(L)

P1

W(x1):b

Acquire(L)

Twin X

Diff

interval

W(x3):?

W(x3):d

Release(L)

Page X

a b c d

a b c
Implementation
Etc.

- **Lock & barrier**
  - Statically assigned manager

- **Garbage collection**
  - reclaim the space used by write notice records, interval records, and diffs
  - Triggered when the free space drops below a threshold
Evaluation

- Experimental Environment
  - 8 DECstation-5000/240
  - connected to a 100-Mbps ATM LAN and a 10-Mbps Ethernet

- Applications
  - Water – molecular dynamics simulation
  - Jacobi – Successive Over-Relaxation
  - TSP – branch & bound algorithm to solve the traveling salesman problem
  - Quicksort – using bubblesort to sort subarray of less than 1K element
  - ILINK – genetic linkage analysis
Evaluation

Execution statistics

<table>
<thead>
<tr>
<th></th>
<th>Water</th>
<th>Jacobi</th>
<th>TSP</th>
<th>Quicksort</th>
<th>ILINK</th>
</tr>
</thead>
<tbody>
<tr>
<td>Input</td>
<td>343 mols 5 steps</td>
<td>2000x1000 floats</td>
<td>19-city tour</td>
<td>256000 integers</td>
<td>CLP</td>
</tr>
<tr>
<td>Time (secs)</td>
<td>15.0</td>
<td>32.0</td>
<td>43.8</td>
<td>13.1</td>
<td>1113</td>
</tr>
<tr>
<td>Barriers/sec</td>
<td>2.5</td>
<td>6.3</td>
<td>0</td>
<td>0.4</td>
<td>0.4</td>
</tr>
<tr>
<td>Locks/sec</td>
<td>589.4</td>
<td>0</td>
<td>16.1</td>
<td>53.9</td>
<td>0</td>
</tr>
<tr>
<td>_msgs/sec</td>
<td>2238</td>
<td>334</td>
<td>404</td>
<td>703</td>
<td>456</td>
</tr>
<tr>
<td>Kbytes/sec</td>
<td>798</td>
<td>415</td>
<td>121</td>
<td>788</td>
<td>164</td>
</tr>
</tbody>
</table>

Speedup

![Graph showing speedup with lines for Water, TSP, Quicksort, and ILINK]
Evaluation

Execution time breakdown

![Bar chart showing execution time breakdown for Water, Jacobi, TSP, Qsort, and ILINK.
- Computation
- TreadMarks
- Unix
- Idle Time]
Evaluation

Unix overhead breakdown

TreadMarks overhead breakdown
Evaluation

Execution time breakdown for Water
Evaluation

ERC vs. LRC

Speedup

Message rate

Lazy  Eager

Lazy  Eager
Evaluation

ERC vs. LRC

Data rate

Diff creation rate
Implementation

P0 side | P1 side
---|---
| | 
| | 
| | 
| | 
| | 
| | 
| | 
| | 
| | 

Pages

<table>
<thead>
<tr>
<th>...</th>
<th>...</th>
<th>P</th>
<th>...</th>
</tr>
</thead>
</table>

Time stamp

0 0 0
Implementation

<table>
<thead>
<tr>
<th>Pages</th>
<th>P0 side</th>
<th>P1 side</th>
</tr>
</thead>
<tbody>
<tr>
<td>...</td>
<td>Acq(L)</td>
<td>W(a)</td>
</tr>
<tr>
<td>a</td>
<td>W.N</td>
<td>Rel(L)</td>
</tr>
<tr>
<td>...</td>
<td>W(b)</td>
<td></td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Time stamp</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1 0 0</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Pages</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>...</td>
<td></td>
<td></td>
</tr>
<tr>
<td>b</td>
<td>W.N</td>
<td></td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Time stamp</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>0 0 0</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Implementation

**P0 side**
- twin
- W.N
- P0
- Acq(L)
- W(a)
- Rel(L)
- W(b)
- Acq(L)

**P1 side**
- twin
- W.N
- P1

**Pages**
- …
- a
- …
- …

**Time stamp**
- P0: 200
- P1: 000
Implementation

<table>
<thead>
<tr>
<th>P0 side</th>
<th>P1 side</th>
</tr>
</thead>
<tbody>
<tr>
<td>twin</td>
<td>twin</td>
</tr>
<tr>
<td>pages</td>
<td>pages</td>
</tr>
<tr>
<td>. . .</td>
<td>. . .</td>
</tr>
<tr>
<td>a</td>
<td>b</td>
</tr>
<tr>
<td>. . .</td>
<td>. . .</td>
</tr>
<tr>
<td>W.N</td>
<td>W.N</td>
</tr>
<tr>
<td>W(a)</td>
<td>W(b)</td>
</tr>
<tr>
<td>Acq(L)</td>
<td>Rel(L)</td>
</tr>
<tr>
<td>Acq(L)</td>
<td></td>
</tr>
<tr>
<td>time stamp</td>
<td>time stamp</td>
</tr>
<tr>
<td>2 0 0</td>
<td>0 0 0</td>
</tr>
</tbody>
</table>

Diagram: Diagram showing the flow of operations between P0 and P1 sides, including Acquire (Acq), Release (Rel), Write Notice (W.N), and process array (Proc Array).
Implementation

P0 side
- P0
- Acq(L)
- W(a)
- Rel(L)
- W(b)
- Acq(L)
- W(c)

P1 side
- P1
- W.N

Twin
- P
- W.N

Pages
- ....
- a
- ...
- ...

Time stamp
- 200

Twin
- W.N

Pages
- ....
- b
- ...

Time stamp
- 110

Twin
- W.N
- W.N

P0
- W.N

P1
- W.N

Diff
- b
- a
Implementation

<table>
<thead>
<tr>
<th>Pages</th>
<th>P0 side</th>
<th>P1 side</th>
<th>twin</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>Acq(L)</td>
<td>W(b)</td>
<td>a</td>
</tr>
<tr>
<td>P</td>
<td>W(a)</td>
<td>Rel(L)</td>
<td>b</td>
</tr>
<tr>
<td>...</td>
<td>Acq(L)</td>
<td></td>
<td>c</td>
</tr>
<tr>
<td>...</td>
<td>W(c)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Time stamp:
- P0 side: 200
- P1 side: 110
- Twin side: a b
Implementation

P0 side

<table>
<thead>
<tr>
<th>pages</th>
<th>Acq(L)</th>
<th>W(a)</th>
<th>Rel(L)</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>⋮</td>
<td>⋮</td>
<td>⋮</td>
<td>⋮</td>
</tr>
<tr>
<td>P</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

P1 side

<table>
<thead>
<tr>
<th>pages</th>
<th>W(b)</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td></td>
</tr>
<tr>
<td>b</td>
<td></td>
</tr>
<tr>
<td>⋮</td>
<td>⋮</td>
</tr>
</tbody>
</table>

time stamp

200

200
Implementation

P1 side

P2 side

P0

P1

Acq(L)

a

b

twin

pages

\[ \ldots \]

\[ a \ b \ c \]

\[ \ldots \]

\[ \ldots \]

time stamp

\[ 1 \ 2 \ 0 \]

\[ 1 \ 1 \ 0 \]

\[ 0 \ 0 \ 0 \]

\[ 1 \ 0 \ 0 \]

\[ \ldots \]

\[ P \]

\[ \ldots \]

\[ \ldots \]

\[ \ldots \]

time stamp

\[ 0 \ 0 \ 0 \]
Implementation

```
+----------------+     +----------------+
| twin           |     | twin            |
| a b            |     | a b             |
+----------------+     +----------------+
   | . . .          |     | . . .           |
   | a b c          |     | a b            |
   | . . .          |     | . . .           |
+----------------+     +----------------+
```

**P1 side**

```
| . . . | W.N | P1 | W.N | P0 | W.N | P1 |
| . . . |     |    |     |    |     |
| a b   | diff| diff|     |    |     |
| . . . |     |    |     |    |     |
```

**P2 side**

```
| . . . |
| W.N  |
| P1   |
| P0   |
| P1   |

```

**Pages**

```
  Acq(L)
```

**Timestamp**

```
1 2 0
```

```
1 1 0 0 0
```

```
1 0 0
```

```
0 1 . . .
```

```
```

**Diagram**

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```
Thank you!

Any questions?