Hardware Support for ACID Transactions in Persistent Memory Systems

Arpit Joshi

ARM Research Summit, 2018
Collaborators

• Vijay Nagarajan
• Marcelo Cintra
• Stratis Viglas
Persistent Memory is here…
Persistent Memory is here...

Intel Displays 512GB Optane DC Persistent Memory DIMMs

by Paul Alcorn May 31, 2018 at 2:02 AM

Intel held its Memory and Storage day today at its Santa Clara headquarters to announce its Optane DC Persistent Memory DIMMs. The new DIMMs slot into the DRAM interface, just like a normal stick of RAM, but come in three capacities of 128, 256, and 512GB. That's a massive capacity increase compared to the industry-leading 128GB DDR4 memory sticks. Intel designed the DIMMs to bridge both the performance and pricing gap between storage and memory, so the new DIMMs should land at much lower price points than typical DRAM.

Intel teases Optane DIMMS, but you may need a new Xeon first

128GB, 256GB and 512GB modules offered as new storage tier below RAM, above SSD

By Simon Sharwood, APAC Editor 31 May 2018 at 03:50

Intel Launches Optane DIMMs Up To 512GB: Apache Pass Is Here!

by Ian Cutress & Billy Talla on May 30, 2018 2:15 PM BST

Posted in: Intel, DDR4, 3D XPoint, Optane, NVMe, Persistent Memory

Intel's new Optane DC persistent memory DIMM. (Credit: AnandTech)
Persistent Memory Systems
Persistent Memory Systems

• Persistent Memory
  - Non-volatility over the memory bus
  - Load/Store interface to persistent data
Persistent Memory Systems

- Persistent Memory
  - Non-volatility over the memory bus
  - Load/Store interface to persistent data
Persistent Memory Systems

- **Persistent Memory**
  - Non-volatility over the memory bus
  - Load/Store interface to persistent data

- **Crash Consistency**
  - Is the persistent state consistent?
  - Programming Model: ACID Transactions
Persistent Memory Systems

“Ensuring failure atomicity for all this computation without failure-atomic transactions is practically infeasible, if not impossible.”
Marathe et al. [HotStorage’17]

Persistent Memory Systems

Crash Consistency
- Is the persistent state consistent?
- Programming Model: ACID Transactions
Persistent Memory Systems

“Ensuring failure atomicity for all this computation without failure-atomic transactions is practically infeasible, if not impossible.”

Marathe et al. [HotStorage’17]

How fast can we support ACID?
ACID Transactions

L1 → LLC → Persistent Memory
ACID Transactions

Atomic Visibility

Persistent Memory
ACID Transactions

![Diagram of memory levels]

- **L1**
- **LLC**
- **Persistent Memory**

**Atomic Visibility**

**Atomic Durability**
ACID Transactions

Atomic Visibility

Locks
STM
HTM

Persistent Memory

Atomic Durability

L1

L1

LLC
Atomic Visibility: HTM
Atomic Visibility: HTM

- Commercial HTMs [Intel, IBM]
Atomic Visibility: HTM

• Commercial HTMs [Intel, IBM]
  - Version Management: read/write sets in L1 cache

<table>
<thead>
<tr>
<th>Cache Line</th>
<th>R</th>
<th>W</th>
</tr>
</thead>
<tbody>
<tr>
<td>A = 15</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>B = 20</td>
<td>1</td>
<td></td>
</tr>
</tbody>
</table>
Atomic Visibility: HTM

- **Commercial HTMs** [Intel, IBM]
  - **Version Management**: read/write sets in L1 cache
  - **Conflict Detection**: piggy back on the coherence protocol
Atomic Visibility: HTM

- **Commercial HTMs** [Intel, IBM]
  - **Version Management**: read/write sets in L1 cache
  - **Conflict Detection**: piggy back on the coherence protocol
  - **Commit**: make updates non-speculative

<table>
<thead>
<tr>
<th>Cache Line</th>
<th>R</th>
<th>W</th>
</tr>
</thead>
<tbody>
<tr>
<td>A = 15</td>
<td></td>
<td></td>
</tr>
<tr>
<td>B = 20</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Atomic Visibility: HTM

• Commercial HTMs [Intel, IBM]
  - **Version Management**: read/write sets in L1 cache
  - **Conflict Detection**: piggy back on the coherence protocol
  - **Commit**: make updates non-speculative
  - **Abort**: invalidate write set
Atomic Visibility: HTM

- **Commercial HTMs** [Intel, IBM]
  - **Version Management**: read/write sets in L1 cache
  - **Conflict Detection**: piggy back on the coherence protocol
  - **Commit**: make updates non-speculative
  - **Abort**: invalidate write set

Write-sets in commercial HTMs limited by the size of the L1 cache.
Atomic Durability: Logging
Atomic Durability: Logging

- Logging for durability [Doshi’16, Joshi’17, Shin’17, Ogleari’18]
Atomic Durability: Logging

- **Logging for durability** [Doshi’16, Joshi’17, Shin’17, Ogleari’18]
  - Write a log entry for every update

<table>
<thead>
<tr>
<th>Persistent Memory</th>
<th>Transaction Log</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>In-place Values</strong></td>
<td><strong>Transaction Log</strong></td>
</tr>
<tr>
<td>A = 10</td>
<td>A = 15</td>
</tr>
<tr>
<td>B = 20</td>
<td>B = 25</td>
</tr>
<tr>
<td>C = 30</td>
<td></td>
</tr>
</tbody>
</table>
**Atomic Durability: Logging**

- **Logging for durability** [Doshi’16, Joshi’17, Shin’17, Ogleari’18]
  - Write a log entry for every update
  - **Commit**: Update the values in-place

<table>
<thead>
<tr>
<th>Persistent Memory</th>
<th>In-place Values</th>
<th>Transaction Log</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td><strong>A = 15</strong></td>
<td></td>
</tr>
<tr>
<td></td>
<td><strong>B = 25</strong></td>
<td></td>
</tr>
<tr>
<td></td>
<td><strong>C = 30</strong></td>
<td></td>
</tr>
</tbody>
</table>
Atomic Durability: Logging

- **Logging for durability** [Doshi’16, Joshi’17, Shin’17, Ogleari’18]
  - Write a log entry for every update
  - **Commit**: Update the values in-place
  - **Abort**: Undo any in-place updates
Atomic Durability: Logging

• **Logging for durability** [Doshi’16, Joshi’17, Shin’17, Ogleari’18]
  - Write a log entry for every update
  - **Commit**: Update the values in-place
  - **Abort**: Undo any in-place updates

---

**In-place updates in the critical path of commit**

**High memory write bandwidth requirement**
ACID = HTM + Logging

Goals:
- Support fast commits
- Minimise memory bandwidth consumption
- Extend the supported transaction size
- Maintain the simplicity of commercial HTMs
DHTM: Durable Hardware Transactional Memory

Log Writes
DHTM: Durable Hardware Transactional Memory

Commercial HTM + Hardware Redo Log
DHTM: Durable Hardware Transactional Memory

- Commercial HTM + Hardware Redo Log
  - H/W Redo Log + Log Buffer
    - Reduced memory bandwidth
    - Fast commits

Persistant Memory
DHTM: Durable Hardware Transactional Memory

Commercial HTM + Hardware Redo Log
- H/W Redo Log + Log Buffer
  - Reduced memory bandwidth
  - Fast commits
- H/W Log + Sticky State
  - Extended transaction size to the LLC
  - Simplicity of commercial HTM
DHTM: Log Buffer

- Redo Log Bandwidth Problem
DHTM: Log Buffer

• Redo Log Bandwidth Problem
  - write a log entry for every store
DHTM: Log Buffer

- **Redo Log Bandwidth Problem**
  - write a log entry for every store
  - multiple stores create multiple log entries
DHTM: Log Buffer

• Redo Log Bandwidth Problem
  - write a log entry for every store
  - multiple stores create multiple log entries

• Solution: Log Buffer
DHTM: Log Buffer

- **Redo Log Bandwidth Problem**
  - write a log entry for every store
  - multiple stores create multiple log entries

- **Solution: Log Buffer**
  - track cache lines being modified
DHTM: Log Buffer

- **Redo Log Bandwidth Problem**
  - write a log entry for every store
  - multiple stores create multiple log entries

- **Solution: Log Buffer**
  - track cache lines being modified
  - multiple writes coalesced in a log entry
DHTM: Log Buffer

- **Redo Log Bandwidth Problem**
  - write a log entry for every store
  - multiple stores create multiple log entries

- **Solution: Log Buffer**
  - track cache lines being modified
  - multiple writes coalesced in a log entry
  - log entry written to persistent memory on eviction from log buffer
DHTM: Transaction States
DHTM: Transaction States

Begin Transaction

Active
DHTM: Transaction States

- Begin Transaction
- Active
- End Transaction & Log Records Persisted
- Commit
DHTM: Transaction States

Begin Transaction

Active

End Transaction & Log Records Persisted

Commit

In-place Data Persisted

Commit Complete
DHTM: Transaction States

- **Active**
  - Begin Transaction
  - End Transaction & Log Records Persisted

- **Commit**
  - In-place Data Persisted

- **Abort**
  - Conflict

- **Commit Complete**
**DHTM: Commit Example**

**L1 Cache**

<table>
<thead>
<tr>
<th>Cache Line</th>
<th>R</th>
<th>W</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**State**

<table>
<thead>
<tr>
<th>Log Buffer</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
</tr>
</tbody>
</table>

**Persistent Memory**

**In-place Values**

<p>| |</p>
<table>
<thead>
<tr>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>A = 10</td>
</tr>
<tr>
<td>B = 20</td>
</tr>
<tr>
<td>C = 30</td>
</tr>
</tbody>
</table>

**Transaction Log**

**Begin_Transaction**

- Write (A=15)
- Read (B)
- Write (B=25)

**End_Transaction**
DHTM: Commit Example

**L1 Cache**

<table>
<thead>
<tr>
<th>Cache Line</th>
<th>R</th>
<th>W</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**State**

**Active**

**Log Buffer**

**Persistent Memory**

**In-place Values**

<p>| |</p>
<table>
<thead>
<tr>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>A = 10</td>
</tr>
<tr>
<td>B = 20</td>
</tr>
<tr>
<td>C = 30</td>
</tr>
</tbody>
</table>

**Transaction Log**

**Begin_Transaction**

- Write (A=15)
- Read (B)
- Write (B=25)

**End_Transaction**
DHTM: Commit Example

**L1 Cache**

<table>
<thead>
<tr>
<th>Cache Line</th>
<th>R</th>
<th>W</th>
</tr>
</thead>
<tbody>
<tr>
<td>A = 15</td>
<td></td>
<td>1</td>
</tr>
</tbody>
</table>

**State**

- Active

**Log Buffer**

- A

**Persistent Memory**

**In-place Values**

- A = 10
- B = 20
- C = 30

**Transaction Log**

**Begin_Transaction**

- Write (A=15)
- Read (B)
- Write (B=25)

**End_Transaction**
DHTM: Commit Example

L1 Cache

<table>
<thead>
<tr>
<th>Cache Line</th>
<th>R</th>
<th>W</th>
</tr>
</thead>
<tbody>
<tr>
<td>A = 15</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>B = 20</td>
<td></td>
<td>1</td>
</tr>
</tbody>
</table>

State

Active

Log Buffer

A

Persistent Memory

In-place Values

<table>
<thead>
<tr>
<th>A = 10</th>
</tr>
</thead>
<tbody>
<tr>
<td>B = 20</td>
</tr>
<tr>
<td>C = 30</td>
</tr>
</tbody>
</table>

Transaction Log

Begin_Transaction

Write (A=15)

Read (B)

Write (B=25)

End_Transaction
DHTM: Commit Example

L1 Cache

<table>
<thead>
<tr>
<th>Cache Line</th>
<th>R</th>
<th>W</th>
</tr>
</thead>
<tbody>
<tr>
<td>A = 15</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>B = 25</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

State

Active

Log Buffer

B

Persistent Memory

Transaction Log

<table>
<thead>
<tr>
<th>In-place Values</th>
<th>Transaction Log</th>
</tr>
</thead>
<tbody>
<tr>
<td>A = 10</td>
<td>A = 15</td>
</tr>
<tr>
<td>B = 20</td>
<td></td>
</tr>
<tr>
<td>C = 30</td>
<td></td>
</tr>
</tbody>
</table>

Begin_Transaction

Write (A=15)

Read (B)

Write (B=25)

End_Transaction
DHTM: Commit Example

**L1 Cache**

<table>
<thead>
<tr>
<th>Cache Line</th>
<th>R</th>
<th>W</th>
</tr>
</thead>
<tbody>
<tr>
<td>A = 15</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>B = 25</td>
<td></td>
<td>1</td>
</tr>
</tbody>
</table>

**Commit**

**Log Buffer**

**Transaction Log**

<table>
<thead>
<tr>
<th>In-place Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>A = 10</td>
</tr>
<tr>
<td>B = 20</td>
</tr>
<tr>
<td>C = 30</td>
</tr>
</tbody>
</table>

**Persistent Memory**

**Begin_Transaction**

Write (A=15)

Read (B)

Write (B=25)

**End_Transaction**
DHTM: Commit Example

**L1 Cache**

- **Cache Line**
  - A = 15
  - B = 25

**State**

- **Commit Complete**

**Log Buffer**

**Persistent Memory**

- **In-place Values**
  - A = 15
  - B = 25
  - C = 30

**Transaction Log**

- A = 15
- B = 25
- Commit
- Complete

**Log Buffer**

**Begin_Transaction**

- Write (A=15)
- Read (B)
- Write (B=25)

**End_Transaction**
DHTM: Supporting Overflow
DHTM: Supporting Overflow

• Problems with Overflow:
DHTM: Supporting Overflow

• Problems with Overflow:
  - Version Management:
    - global operation on write-set on a commit/abort
    - overhead infeasible in larger caches (beyond L1)
DHTM: Supporting Overflow

• Problems with Overflow:
  - Version Management:
    - global operation on write-set on a commit/abort
    - overhead infeasible in larger caches (beyond L1)
  - Conflict Detection:
    - additional metadata to detect conflicts
    - increased complexity due to NACK based protocols
DHTM: Supporting Overflow
DHTM: Supporting Overflow

• Solution
DHTM: Supporting Overflow

• Solution
  - Version Management:
  - Overflow List
DHTM: Supporting Overflow

• Solution
  - Version Management:
    - Overflow List
DHTM: Supporting Overflow

• Solution
  - Version Management:
  - Overflow List
DHTM: Supporting Overflow

• **Solution**
  - **Version Management:**
    - Overflow List
  - **Conflict Detection:**
    - maintain sticky state on overflow (similar to LogTM)
    - avoid NACK by restricting overflow to LLC
Evaluation

<table>
<thead>
<tr>
<th></th>
<th>Atomic Visibility</th>
<th>Atomic Durability</th>
</tr>
</thead>
<tbody>
<tr>
<td>ATOM</td>
<td>Locks</td>
<td>Hardware Undo Log</td>
</tr>
<tr>
<td>LogTM+ATOM</td>
<td>HTM (LogTM)</td>
<td>Hardware Undo Log</td>
</tr>
<tr>
<td>DHTM</td>
<td>HTM</td>
<td>Hardware Redo Log (Log Buffer)</td>
</tr>
</tbody>
</table>

**System Configuration**

- We evaluate an 8-core machine with a 2-level cache hierarchy
- HTM’s implement (first) writer wins conflict resolution policy
Evaluation
Evaluation

![Evaluation Diagram]

- ATOM
- LogTM+ATOM
- DHTM
Evaluation

![Diagram showing evaluation results for various data structures: queue, hash, sdg, sps, btree, rbtree, gmean. The x-axis represents different data structures, and the y-axis represents performance metrics. The bars indicate performance measures for ATOM, LogTM+ATOM, and DHTM.](image-url)
Evaluation
Evaluation

![Bar chart showing evaluation results for different data structures. The chart compares ATOM, LogTM+ATOM, and DHTM performance across queue, hash, sdg, sps, btree, rbtree, and gmean. The gmean shows a 26% improvement.]
Evaluation

![Graph showing evaluation results for different data structures]

- queue
- hash
- sdg
- sps
- btree
- rbtree
- gmean

Legend:
- ATOM
- LogTM+ATOM
- DHTM

17% improvement
Conclusion

• Persistent memory systems require crash consistency
• ACID Transactions: widely used crash consistency mechanism
• DHTM: ACID transactions in hardware
  - Atomic Visibility: commercial HTM
  - Atomic Durability: bandwidth optimized hardware redo log
  - Leverage hardware logging to extend transaction size unto LLC
Hardware Support for ACID Transactions in Persistent Memory Systems

Arpit Joshi

ARM Research Summit, 2018
BACK UP
Evaluation
Evaluation

[Graph showing evaluation results for TPC-C and TATP with ATOM and DHTM categories]
Evaluation

![Bar chart comparing TPC-C and TATP]

- **ATOM**
- **DHTM**

- **TPC-C** shows significantly higher performance than **TATP**.
Unbounded HTMs

- Do not support atomic durability
- Retrofitting atomic durability to existing mechanisms not optimal (eg. LogTM+ATOM)
- Complex NACK based coherence protocols for conflict detection
DHTM: Abort Example

**L1 Cache**

<table>
<thead>
<tr>
<th>Cache Line</th>
<th>R</th>
<th>W</th>
</tr>
</thead>
<tbody>
<tr>
<td>A = 15</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>B = 25</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

**State**

Active

**Log Buffer**

B

**Persistent Memory**

**In-place Values**

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>A = 10</td>
<td></td>
</tr>
<tr>
<td>B = 20</td>
<td></td>
</tr>
<tr>
<td>C = 30</td>
<td></td>
</tr>
</tbody>
</table>

**Transaction Log**

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>A = 15</td>
<td></td>
</tr>
</tbody>
</table>

**Begin_Transaction**

Write (A=15)

Read (B)

Write (B=25)

**End_Transaction**
DHTM: Abort Example

**L1 Cache**

<table>
<thead>
<tr>
<th>Cache Line</th>
<th>R</th>
<th>W</th>
</tr>
</thead>
<tbody>
<tr>
<td>A = 15</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>B = 25</td>
<td></td>
<td>1</td>
</tr>
</tbody>
</table>

**Persistent Memory**

**In-place Values**

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>A = 10</td>
<td></td>
<td></td>
</tr>
<tr>
<td>B = 20</td>
<td></td>
<td></td>
</tr>
<tr>
<td>C = 30</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**State**

Active

**Log Buffer**

B

**Transaction Log**

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>A = 15</td>
<td></td>
</tr>
</tbody>
</table>

**Begin_Transaction**

Write (A=15)

Read (B)

Write (B=25)

**End_Transaction**
DHTM: Abort Example

L1 Cache

State

Abort

Log Buffer

Persistent Memory

In-place Values

Transaction Log

A = 15

Abort

A = 10

B = 20

C = 30

Begin Transaction
Write (A=15)
Read (B)
Write (B=25)
End Transaction

Abort
Hardware Support for ACID Transactions in Persistent Memory Systems

Arpit Joshi

ARM Research Summit, 2018