Performance characterization of a DRAM-NVM hybrid memory architecture for HPC applications using Intel Optane DC Persistent Memory Modules

Brad Settlemyer

1Onkar Patil, 2Latchesar Ionkov, 2Jason Lee, 1Frank Mueller, 2Michael Lang

1Dept. of Computer Science, North Carolina State University
2Ultrascale Research Center, Los Alamos National Laboratory
What to do about DRAM?

- DRAM scaling and reliability is an issue
  - Last 2 decades: scaled ~33% slower than core count
  - High power consumption (fast refresh and cell count)
  - Reaching density limits

- Memories with higher density than DRAM will allow different design points for exascale computers
  - Fewer nodes to reach higher aggregate memory capacities
What to do about DRAM?

- Memory technologies such as phase change memory (PCM) and spin-transfer torque RAM (STT-RAM)
  - Byte-addressable, non-volatile memory device
  - Higher density
  - Shrinks easier than DRAM
  - Higher write latency
  - Lower write durability

- Enables scaling the main memory capacity with core count
Intel’s Optane DC Persistent Memory Module

- Based on PCM
- 8x the density of DRAM
- Uses DIMM slots
- Cheaper than DRAM
Intel’s Optane DC Persistent Memory Module

- Memory interface uses DDR-T protocol via the i-Memory Controller

- Modes of operation
  - Memory mode
    - DRAM is L4 cache for Optane
  - App-direct mode
    - Optane is a block device
  - Mixed mode
    - Mem mode + App direct
  - Hybrid mode
    - Optane extends DRAM address space
Evaluation Platform

- Single node with Intel’s 48-core Cascade Lake processor
- Benchmarks
  - STREAM-like custom benchmark
  - AMG – multi grid
  - VPIC – particle in cell
  - LULESH - hydrodynamics
  - SNAP – deterministic transport
- Operation Modes
  - DRAM-only
  - Memory-mode
  - Hybrid mode

<table>
<thead>
<tr>
<th>Specifications</th>
<th>Optane Node</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model name</td>
<td>Intel(R) Xeon(R) 8260L @ 2.40GHz</td>
</tr>
<tr>
<td>Architecture</td>
<td>x86_64</td>
</tr>
<tr>
<td>CPUs</td>
<td>96</td>
</tr>
<tr>
<td>Sockets</td>
<td>2</td>
</tr>
<tr>
<td>Cores per socket</td>
<td>24</td>
</tr>
<tr>
<td>NUMA nodes</td>
<td>4</td>
</tr>
<tr>
<td>L1d cache</td>
<td>32 KB</td>
</tr>
<tr>
<td>L1i cache</td>
<td>32 KB</td>
</tr>
<tr>
<td>L2 cache</td>
<td>1 MB</td>
</tr>
<tr>
<td>L3 cache</td>
<td>35.3 MB</td>
</tr>
<tr>
<td>Memory Controllers</td>
<td>4</td>
</tr>
<tr>
<td>Channels/controller</td>
<td>6</td>
</tr>
<tr>
<td>DIMM protocol</td>
<td>DDR4</td>
</tr>
<tr>
<td>DRAM size</td>
<td>192 GB</td>
</tr>
<tr>
<td>NVDIMM protocol</td>
<td>DDR-T</td>
</tr>
<tr>
<td>NVRAM size</td>
<td>1.5 TB</td>
</tr>
<tr>
<td>Operating System</td>
<td>Fedora 27</td>
</tr>
</tbody>
</table>
Optane DIMM Raw Performance

- Streams observed in HPC applications
  - Linear arrays and matrices
  - Different access patterns
  - Measured bandwidth
- Executed on all NUMA nodes and all CPU sets
  - Local vs Remote

![Write-only stream bandwidth on the Optane node](image)
More STREAMS-like Performance

9-cell stencil stream bandwidth on the Optane node

Row major matrix stream bandwidth on the Optane node
Performance Evaluation (VPIC)

- Vectorized Particle-In-Cell Code
  - This code is known to scale well and perform well with HT
- Optane delivers excellent performance
- VPIC uses CPU cache hierarchy effectively

---

**VPIC Bandwidth Strong Scaling**

- DRAM Total BW
- Optane Total BW
- Memory-mode Total BW

**VPIC Bandwidth Weak Scaling**

- DRAM Total BW
- Optane Total BW
- Memory-mode Total BW

---

**MPI Processes**

- DRAM Solve time
- Optane Solve time
- Memory-mode Solve time
Performance Evaluation - SNAP

- Particle transport code
- Low overall memory bandwidth requirement
  - Note the absolute scale
- Latency dominant workload
  - Working set size issue
  - Cache/DRAM latency is excellent
  - Optane latency is bad
Performance Evaluation - LULESH

- ALE Simulation
- Strong scaling shows problem fits in L4 cache
- Weak scaling shows what happens as the shared cache capacity is exhausted
- Mixed benefits

![Diagram of LULESH Bandwidth Strong Scaling](image)

![Diagram of LULESH Bandwidth Weak Scaling](image)
Performance Evaluation - AMG

- AMG
  - Algebraic Multi-grid solver
- L4 DRAM achieves similar bandwidth
- Code is bound on memory latency!
Energy Use: The Good

- Optane-only and Optane w/L4 DRAM similar performance, power
- No free lunch for cache bound codes (performance = energy)
Energy Use: The not so Good

- Optane-only is both slower and uses more energy
  - Idle power is dominating energy use
- Optane w/ l4 DRAM
  - Similar performance, similar bandwidth, similar energy use
- No free lunch for bandwidth bound codes (performance = energy)
Future Work

- Exploit capacity to reduce network/compute (memoization)
- Identify needed changes to existing cache hierarchy
- Identify strategies for leveraging Optane to fit energy budgets
- Compiler-based analysis and profiling information to optimize the use of NVDIMMs for various applications
- Designing HPC platforms that use Optane efficiently
  - Trade network energy for optane capacity?
Conclusion

- DRAM Caching appears to just work for bandwidth?!  
  - But codes that are memory latency bound still struggle!

- Slower byte-addressable memory device hampers performance of memory-bound HPC applications  
  - Higher access latencies  
  - Lower memory bandwidth

- Energy efficiency is complicated …  
  - You may lose performance due to excess idling  
  - But maybe you can reduce network …
Thank you!

- Questions

- Contact Info
  - mlang@lanl.gov