Flash Memory Summit

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

Brad Settlemyer

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

[1]Dept. of Computer Science, North Carolina State University

[2]Ultrascale Research Center, Los Alamos National Laboratory

What to do about DRAM?

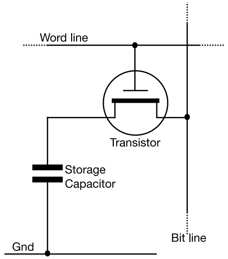
- DRAM scaling and reliability is an issue
  - Last 2 decades: scaled ~33% slower than core count
  - High power consumption (fast refresh and cell count)
  - Reaching density limits
- Memories with higher density than DRAM will allow different design points for exascale computers
  - Fewer nodes to reach higher aggregate memory capacities

What to do about DRAM?

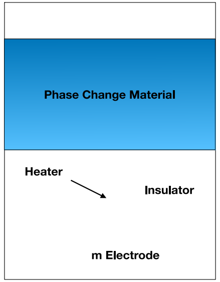
- Memory technologies such as phase change memory (PCM) and spin-transfer torque RAM (STT-RAM)
    - Byte-addressable, non-volatile memory device
    - Higher density
    - Shrinks easier than DRAM
    - Higher write latency
    - Lower write durability
- **Enables scaling the main memory capacity with core count**

Intel's Optane DC Persistent Memory Module

- Based on PCM
- 8x the density of DRAM
- Uses DIMM slots
- Cheaper than DRAM



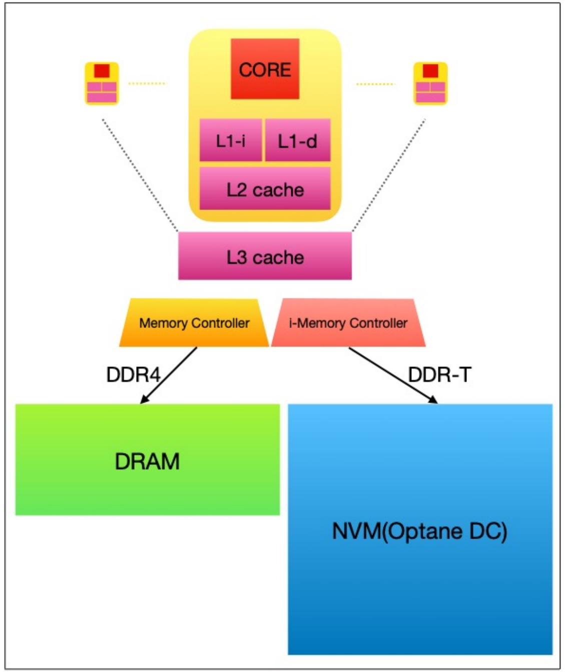
(a) DRAM cell



(b) PCM cell

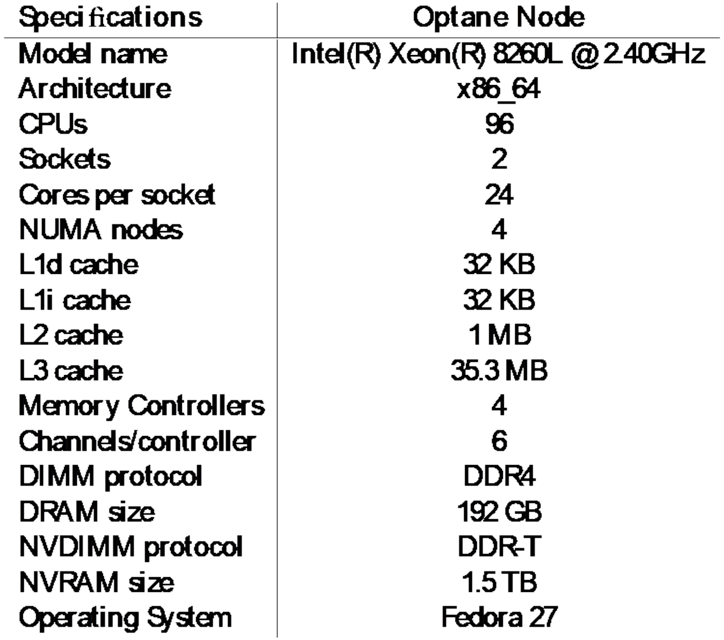
Intel’s Optane DC Persistent Memory Module

- 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

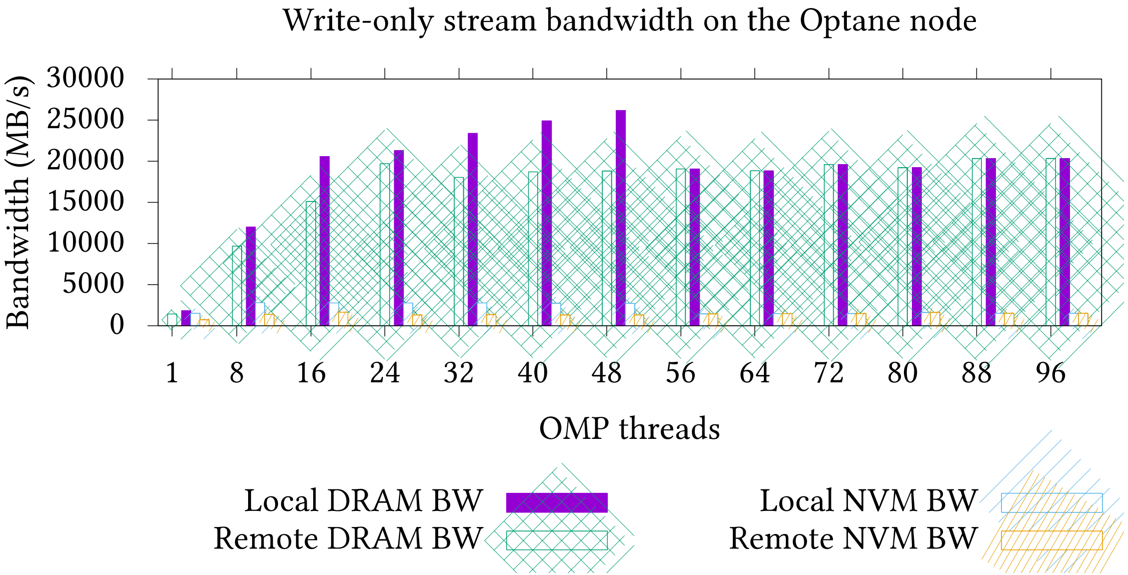
- Single node with Intel’s 48-core Cascade Lake processor
- Benchmarks
  - STREAM-like custom benchmark
  - AMG – multi grid
  - VPIC – particle in cell
  - LULESH - hydrodynamics
  - SNAP – deterministic transport
- Operation Modes
  - DRAM-only
  - Memory-mode
  - Hybrid mode



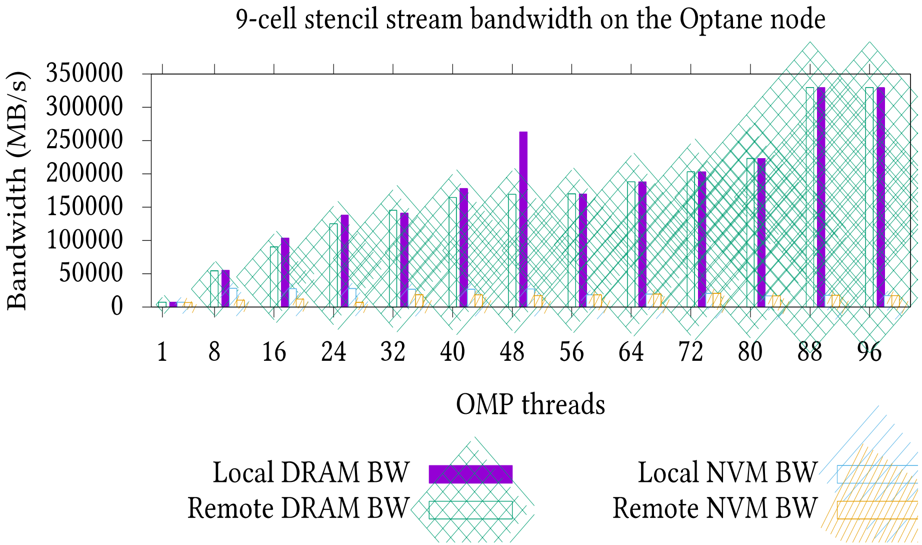
Specifications	Optane Node
Model name	Intel(R) Xeon(R) 8260L @ 2.40GHz
Architecture	x86_64
CPUs	96
Sockets	2
Cores per socket	24
NUMA nodes	4
L1d cache	32 KB
L1i cache	32 KB
L2 cache	1 MB
L3 cache	35.3 MB
Memory Controllers	4
Channels/controller	6
DIMM protocol	DDR4
DRAM size	192 GB
NVDIMM protocol	DDR-T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

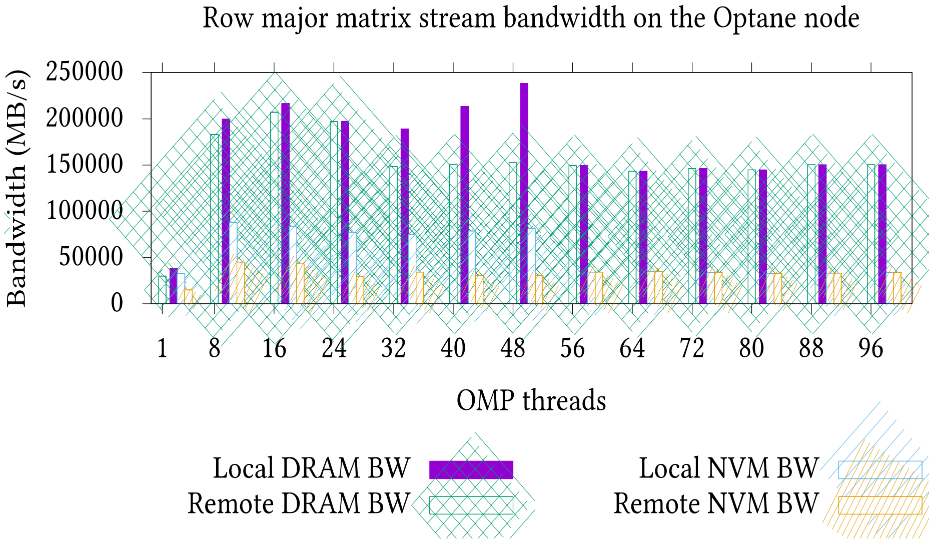
Optane DIMM Raw Performance

- 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



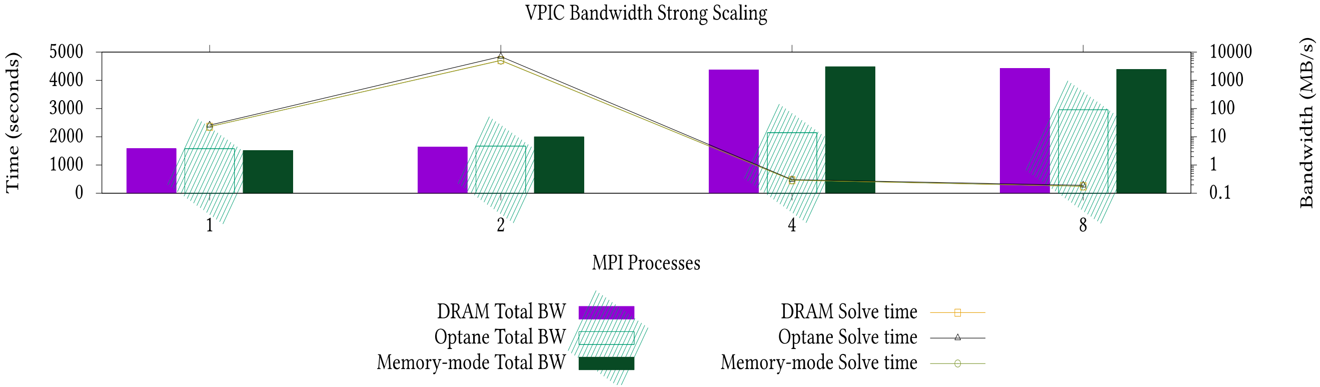
More STREAMS-like Performance

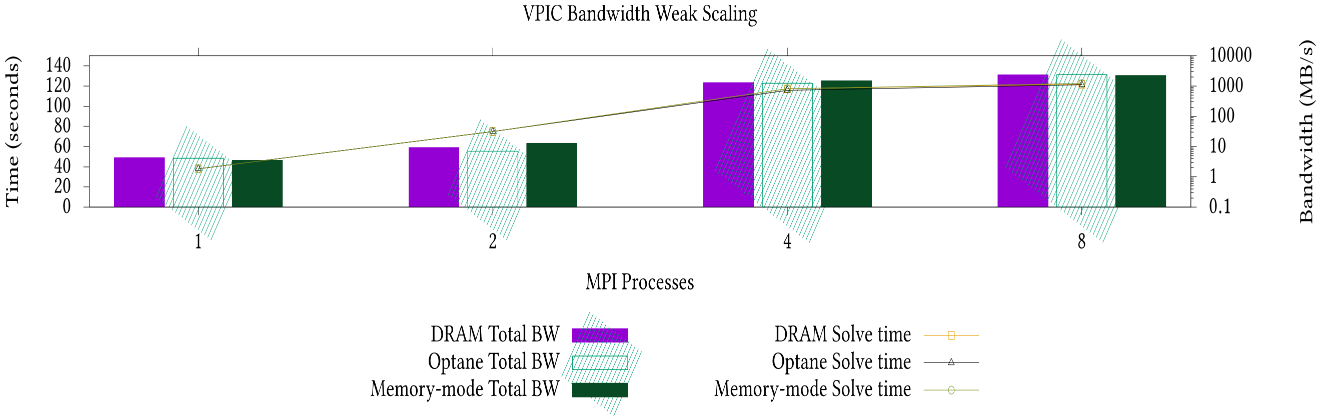

9-cell stencil stream bandwidth on the Optane node


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

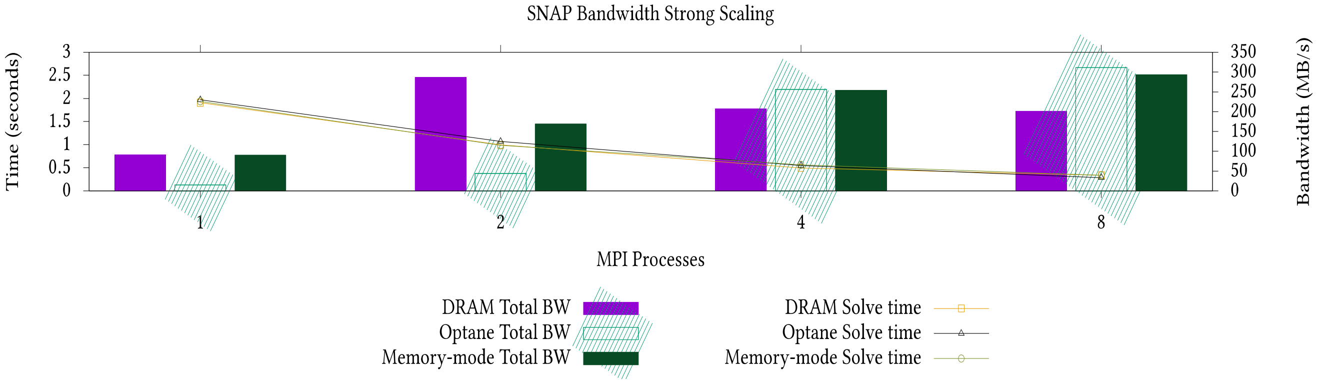


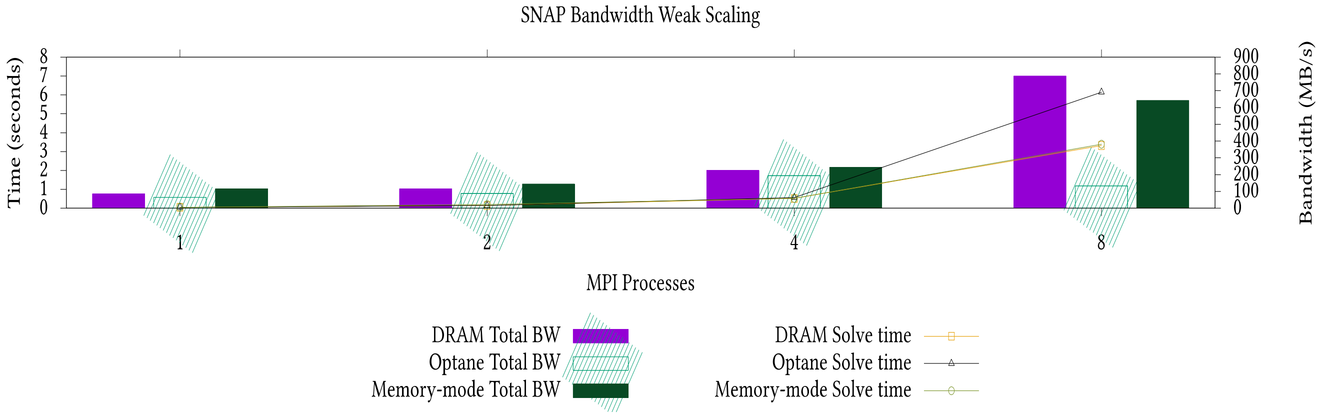


Flash Memory Summit

Performance Evaluation - SNAP

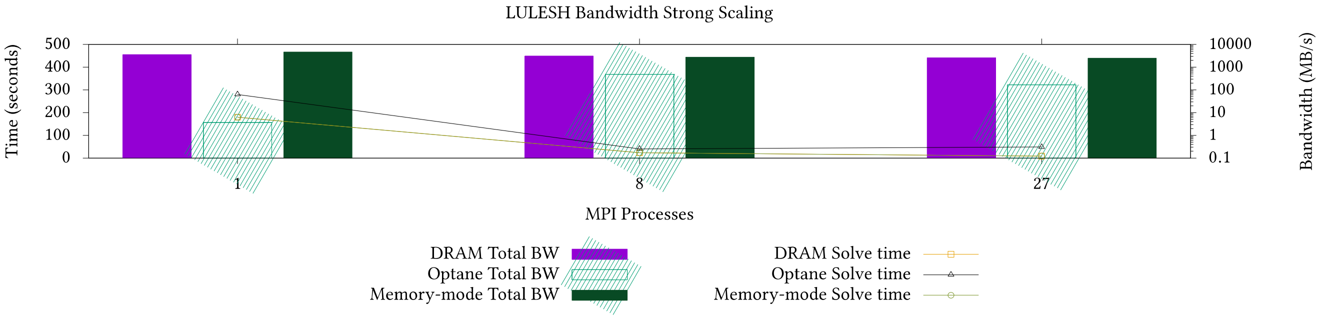
- Particle transport code
- Low overall memory bandwidth requirement
  - Note the absolute scale
- Latency dominant workload
  - Working set size issue
  - Cache/DRAM latency is excellent
  - Optane latency is bad

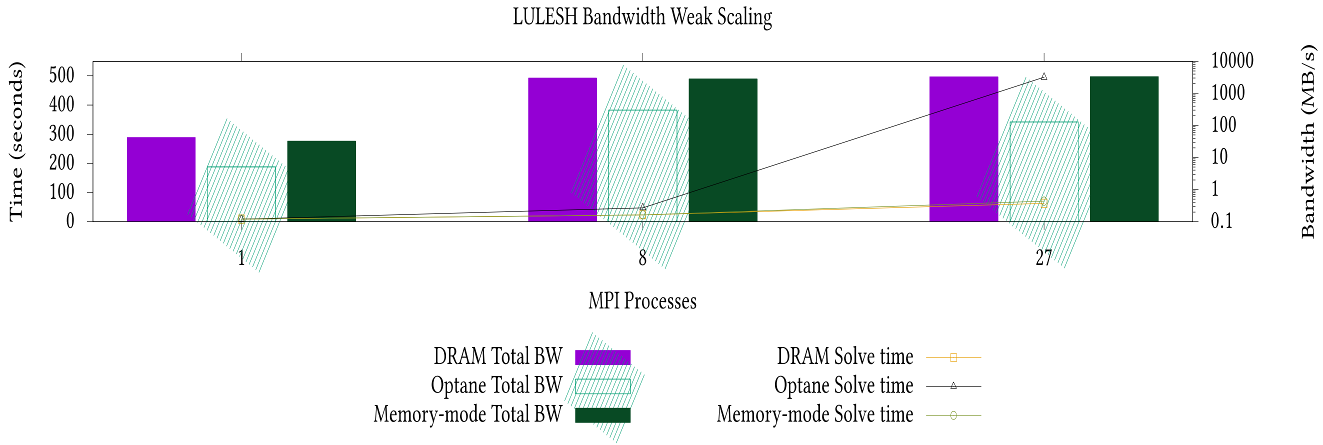

SNAP Bandwidth Strong Scaling


SNAP Bandwidth Weak Scaling

Performance Evaluation - LULESH

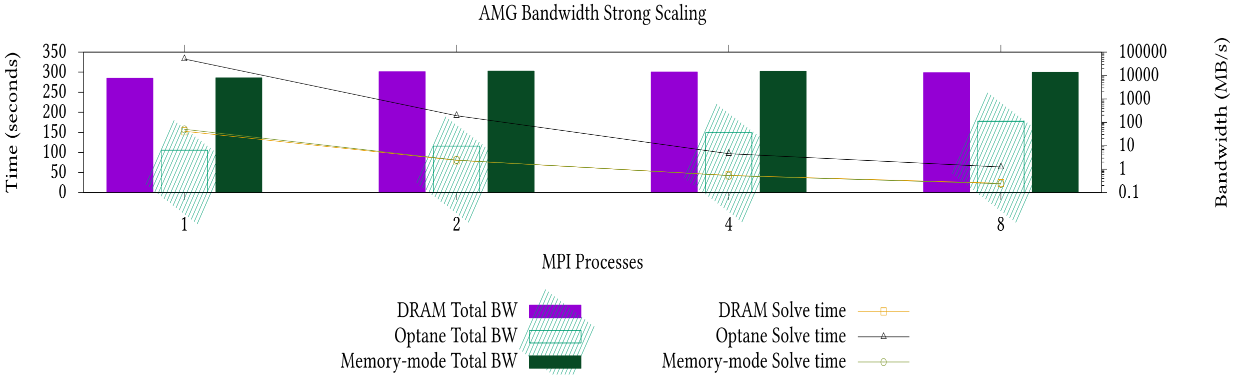
- ALE Simulation
- Strong scaling shows problem fits in L4 cache
- Weak scaling shows what happens as the shared cache capacity is exhausted
- Mixed benefits

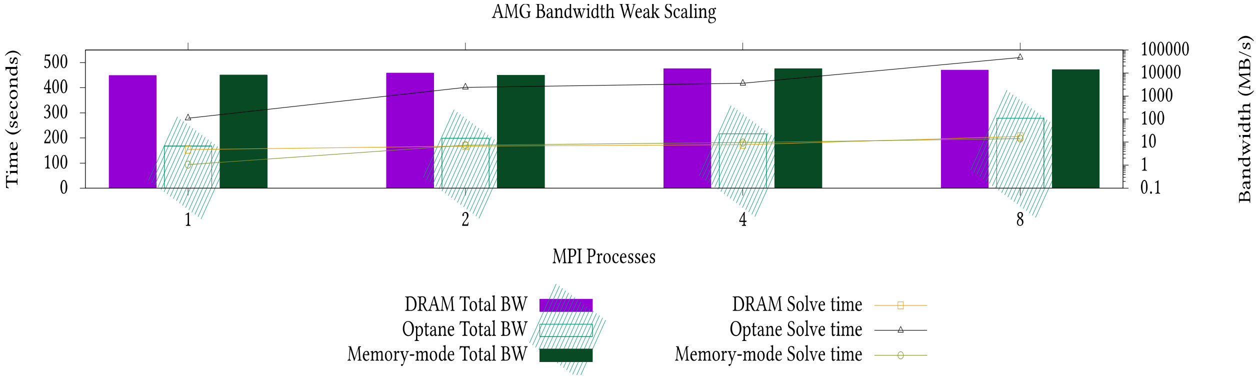

LULESH Bandwidth Strong Scaling


LULESH Bandwidth Weak Scaling

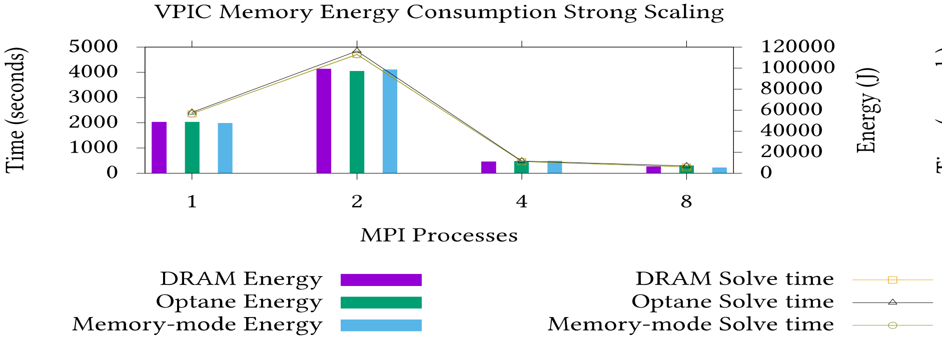
Performance Evaluation - AMG

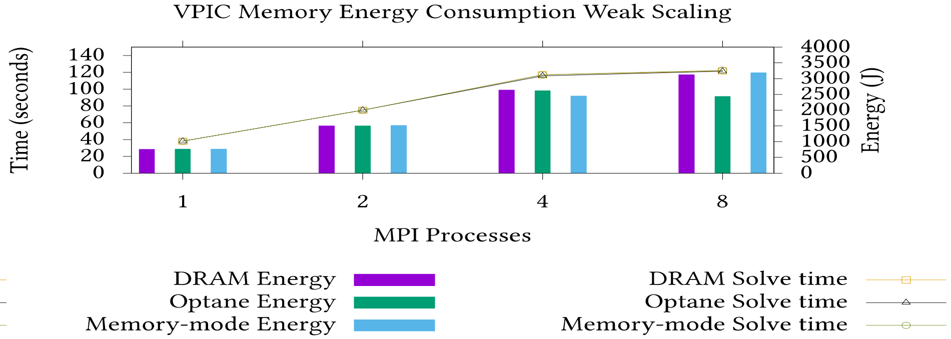
- AMG
  - Algebraic Multi-grid solver
- L4 DRAM achieves similar bandwidth
- Code is bound on memory latency!





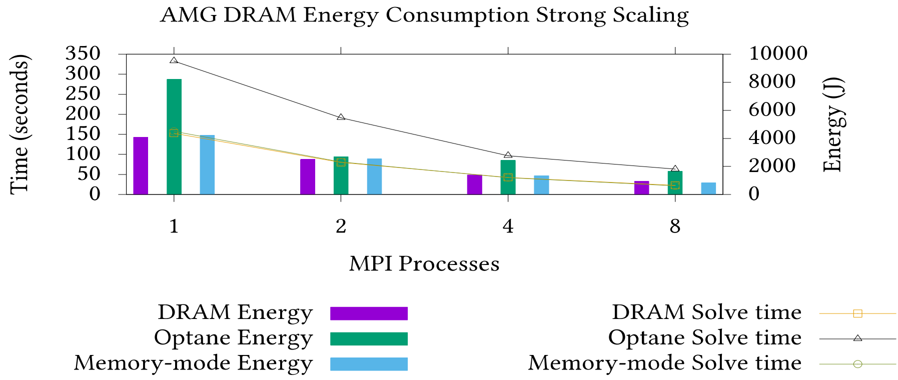
Energy Use: The Good

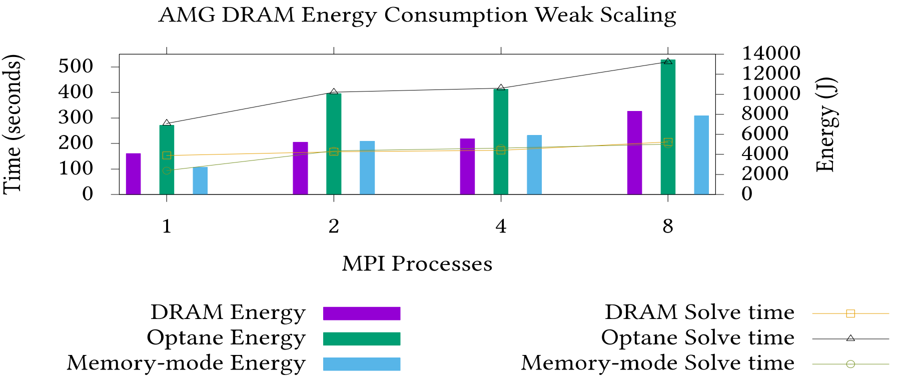

VPIC Memory Energy Consumption Strong Scaling


VPIC Memory Energy Consumption Weak Scaling

- 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


AMG DRAM Energy Consumption Strong Scaling


AMG DRAM Energy Consumption Weak Scaling

- 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 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 …

Flash Memory Summit

Thank you!

- Questions

- Contact Info
  - mlang@lanl.gov