Using Non Volatile Memory to build energy and cost efficient clusters

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

[1]Department of Computer Science, North Carolina State University
[2]Ultrascale Research Center, Los Alamos National Laboratory

Motivation

- Currently, in the Petascale era
  - Summit supercomputer[1]
    - 4608 nodes
    - 2.67 PB of DDR4 DRAM+HBM2 main memory
    - 13 MW peak power consumption
- Further scaling of problem sizes is limited by the cost to build and operate larger HPC machines

Hypothesis

- Using Non-Volatile Memory as an extension of main memory
  - Run larger problem sizes on fewer nodes
  - Reduce the power consumption
  - Reduce the operating and initial setup cost of the cluster

Intel's Optane DC NVDIMMS

- High capacity, byte-addressable memory
- Based on Phase Change Memory (PCM), cheaper per bit than DRAM
- 8x more denser and 6x slower than DRAM, lower power
- Operates in Memory, App-Direct, Flat mode

Experimental Setup

- Single node with Intel's Optane DC PMMs
  - Cascade Lake processor, 192GB DRAM+1.5TB of NVM
- 4 nodes with Skylake processors
  - 377 GB/node connected by Mellanox EDR 100GB/s switches
- Power meters to collect usage numbers at node and rack level, Likwid to collect DIMMs and processor energy
- Benchmarks: VPIC
  - "lpi" input, 0.3 to 1 TB, memory agnostic allocation, MPI

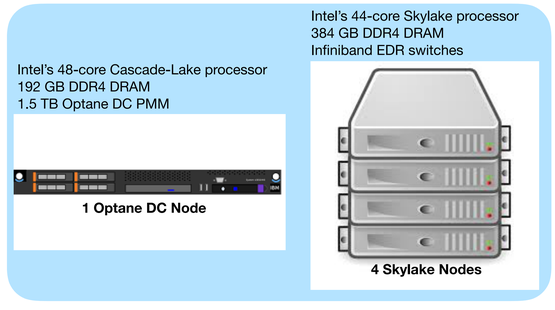

Figure: Experimental setup. Left: Single node with Optane DC PMMs Right: 4 nodes with equivalent amount of DRAM

Results

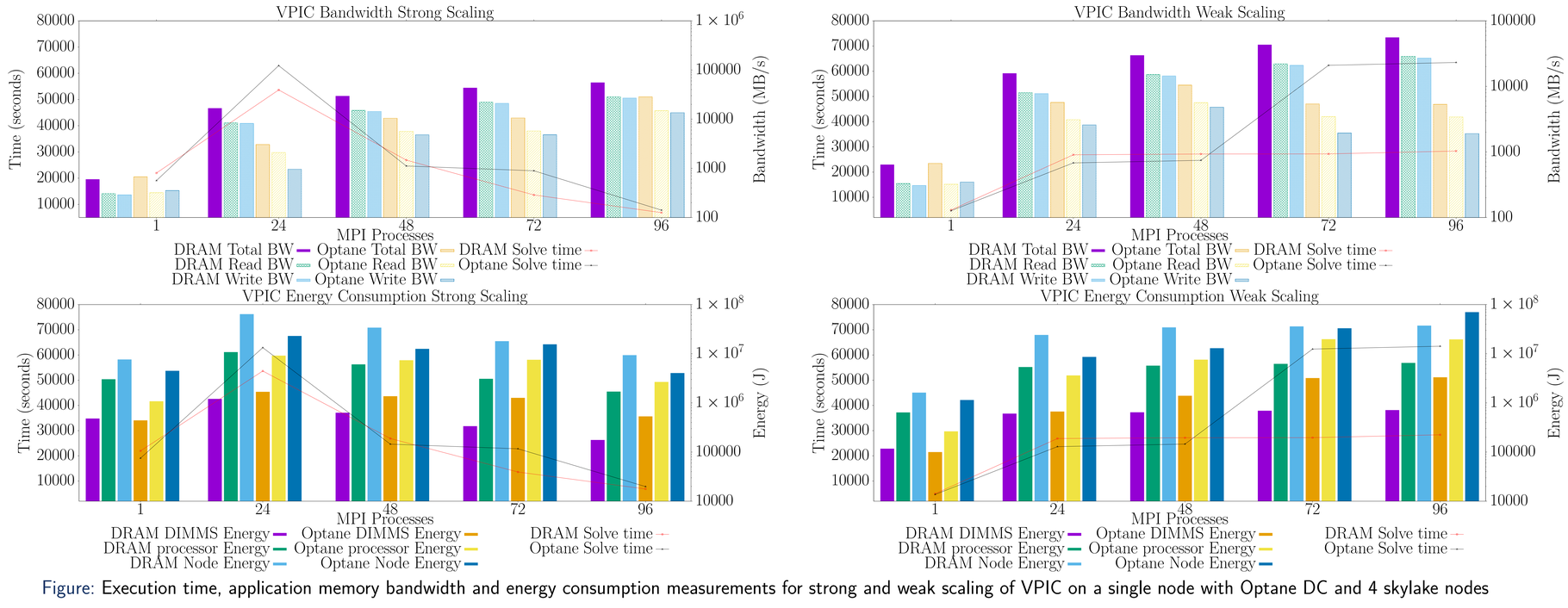

Figure: Execution time, application memory bandwidth and energy consumption measurements for strong and weak scaling of VPIC on a single node with Optane DC and 4 skylake nodes

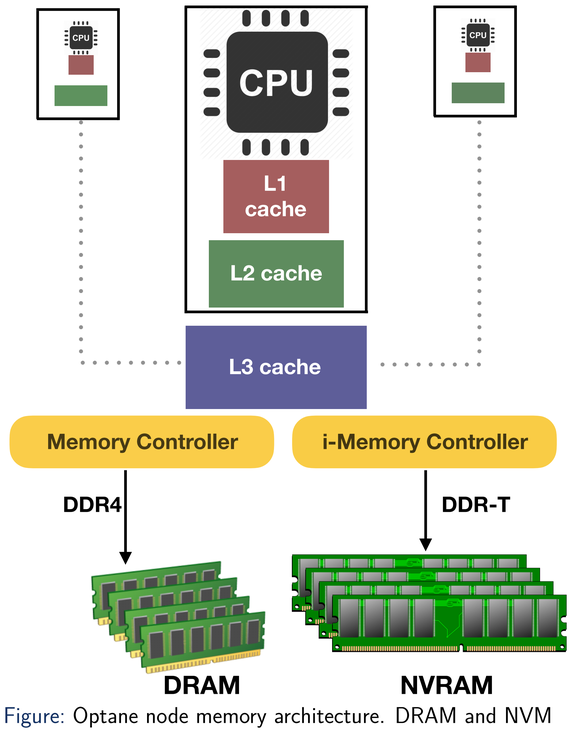

Figure: Optane node memory architecture. DRAM and NVM are part of the same address space using a small kernel patch. Optane DC PMMS are connected to a different memory controller than the DRAM DIMMS and use a 256 byte cache line access [2].

Observations

- Strong scaling
  - Similar execution time for both Optane and Skylake configurations
  - Memory bandwidth has minimal effect
  - Up to 3x energy savings for Optane compared to Skylake
- Weak scaling
  - Optane has lower execution time than Skylake before oversubscription
  - Auxiliary components of Skylake consume large amount of energy
  - Up to 3x energy savings for Optane compared to Skylake
- VPIC aligns data accesses with the cache line size optimizing its cache hits
- Not affected by the memory latency of the underlying memory technology

Inference

- Energy benefits achieved due to low power consumption of Optane DC
- Compute-bound applications benefit from more aggregated memory on fewer nodes
- Lowers the acquisition and operational cost of the cluster

Conclusion

- NVM can help in building future HPC systems
  - At lower acquisitional and operating cost
  - In a space and energy efficient manner
  - Compute on larger problem sizes with fewer nodes

References

[1] J. Hines, "Stepping up to summit," *Computing in science & engineering.*, vol. 20, no. 2, pp. 78–82, 2018.

[2] O. Patil, L. Ionkov, J. Lee, F. Mueller, and M. Lang, "Performance characterization of a dram-nvm hybrid memory architecture for hpc applications using intel optane dc persistent memory modules," in *Proceedings of the fifth ACM/IEEE International Symposium on Memory Systems*, pp. 288–303, ACM/IEEE, 2019.

Contact Information

- NCSU: https://arcb.csc.ncsu.edu/mueller/hetmem.html
- USRC: https://usrc.lanl.gov
- Email: opatil@ncsu.edu
- LA-UR-19-27887