Leveraging Stencil Computation Performance with Temporal Blocking using Large Cache Capacity on AMD EPYC™ 7003 Processors with AMD 3D V-Cache™ Technology

جامعة الملك عبدالله للعلوم والتقنية
King Abdullah University of Science and Technology

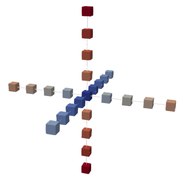
Long Qu, Hatem Ltaief and David Keyes
Extreme Computing Research Center, King Abdullah University of Science and Technology

Motivation

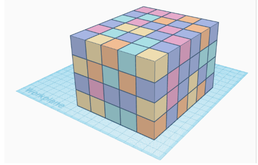
- In structured grid finite-difference, finite-volume, and finite-element discretizations of partial differential equation conservation laws, regular stencil computations constitute the main core kernel in many temporally explicit approaches for such problems.
- For various blocking dimensions, the Spatial Blocking (SB) approach enables data reuse within multiple cache levels. However, the straightforward generalization of SB to manycore architectures, with each core owning an exclusive share of cache leaves performance on the table.
- The Temporal Blocking (TB) method improves data locality further by adding another level of blocking along the time dimension via a diamond tiling mechanism. Introduced in GIRIH, the Multicore Wavefront Diamond blocking (MWD) method [1, 2] optimizes practically relevant stencil algorithms by combining the concepts of diamond tiling and multi-core aware wavefront temporal blocking, leading to significant increase in data reuse and locality. The Last Level Cache (LLC) is shared among cores to reduce memory access across successive iterations. A recent study [3] shows the impact of MWD on the performance of seismic applications and highlights its performance superiority versus SB.
- We evaluate the performance of MWD on a variety of new multi-core architectures. Among all of them, the new AMD EPYC™ 7003 processors with AMD 3D V-Cache™ Technology, codenamed Milan-X, provide an unprecedented LLC capacity, achieve significant performance gain over its predecessor Rome and still better than Milan, which has the same cores but less LLC.

State-of-the-art : Spatial Blocking

- Leverage cache reuse
- Support most of stencil-based applications
- Provide simplicity and flexibility



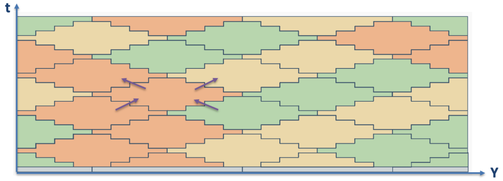
25-point 3D star stencil



3D domain tiled with cubes and computed with SB

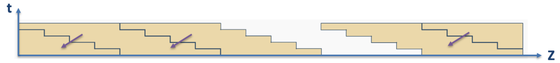
Design of MWD Blocking

- Diamond tiling (outer-level OpenMP) using dynamic scheduling on different groups of threads.



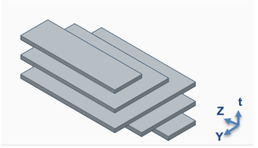
Diamond tiling and the dependency among diamonds

- Wavefront parallelism (inner-level OpenMP) cache reuse among threads and between contiguous wavefront steps.

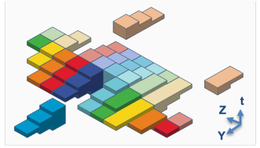


Wavefront and the dependency among different wavefront steps

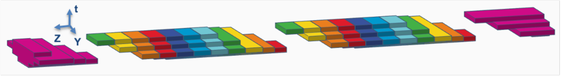
- Multi-core computation of one wavefront step cache reuse inside each wavefront step.



3D view of one wavefront step inside a diamond

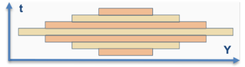


Well-balanced workload among threads



3D view of one diamond with its various wavefront steps (computed using multi-threads)

- Data reusage among various time steps inside the cache.

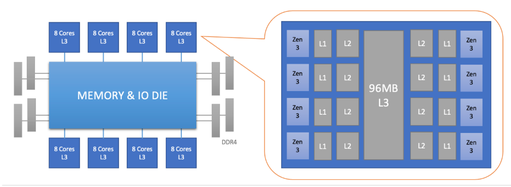


Alternated memory usage between each time step

In-memory status after the computation of the diamond

AMD Milan-X Chip Architecture

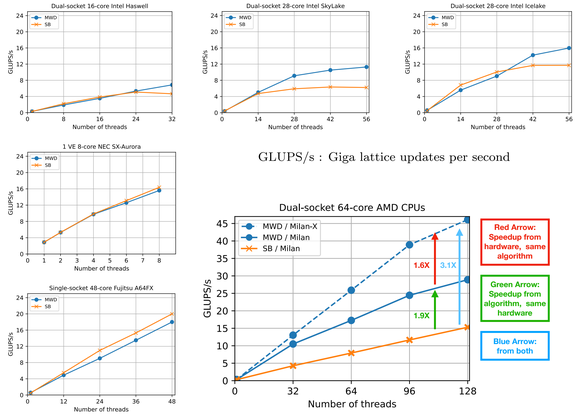
- Two-socket 64-core AMD Milan-X processors have 3 times more L3 cache capacity compared to its predecessors Milan / Rome.
- Each Core Complex Die (CCD) composed of 8 Zen3 cores has 96MB of shared L3 cache.
- Each Zen3 core has 512KB and 64KB of L2/L1 cache, resp.
- Overall 768MB of aggregated L3 cache per socket.



AMD Milan-X Cache Hierarchy.

Performance Results

- MWD technique achieves performance superiority over SB.



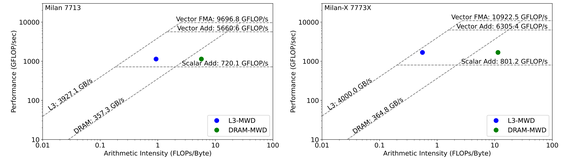
GLUPS/s : Giga lattice updates per second

Results of $1024 \times 1024 \times 512$ domain size with 1000 time steps.

- A plateau is observed for SB approach as the memory bandwidth is utilized, except for multiprocessors with HBM, i.e., NEC and A64FX, while MWD-based stencil computation continues to scale.

Performance Analysis

- Similar sustained memory/L3 bandwidth for Milan / Milan-X.



Roofline Performance Model Analysis

- More cores in sharing the L3 cache.
- With a larger L3 cache capacity, more time steps can fit into the diamond which is computed inside each CCD.
- MWD becomes L3-bound on Milan-X with significant performance gain over Milan.

Summary and Future Work

- MWD improves performance of stencil computations over SB on large grid sizes.
- MWD works best on hardware with a wide bandwidth gap between LLC and main memory.
- MWD achieves high performance in presence of large shared LLC and is agnostic to main memory technology (e.g., HBM).
- Future work: porting on AMD Instinct GPUs and integrating into the reverse time migration [3].

Software Release and References

- GIRIH can be found at https://github.com/ecrc/girih

1 Malas T., Hager G., Ltaief H., Stengel H., Wellein G. and Keyes D. (2015) Multicore Optimized Wavefront Diamond Blocking for Optimizing Stencil Updates. SIAM J. Sci. Comput. 37(4): 439–464.

2 Malas T., Hager G., Ltaief H. and Keyes D. (2017) Multidimensional Intratile Parallelization for Memory-Starved Stencil Computations. ACM Trans. Par. Comput. 4(3): 12:1–12:32.

3 Qu L., Abdelkhalek R., Said I., Ltaief H. and Keyes D. (2022) Exploiting Temporal Data Reuse and Asynchrony in the Reverse Time Migration, Int. J. High Perf. Comput. Appl.