Application Performance Analysis: A Report on the Impact of Memory Bandwidth

Yinzhi Wang(✉), John D. McCalpin, Junjie Li, Matthew Cawood, John Cazes, Hanning Chen, Lars Koesterke, Hang Liu, Chun-Yaung Lu, Robert McLay, Kent Milfield, Amit Ruhela, Dave Semeraro, and Wenyang Zhang

Texas Advanced Computing Center, The University of Texas at Austin,
Austin, TX 78758, USA
iwang@tacc.utexas.edu

Abstract. As High-Performance Computing (HPC) applications involving massive data sets, including large-scale simulations, data analytics, and machine learning, continue to grow in importance, memory bandwidth has emerged as a critical performance factor in contemporary HPC systems. The rapidly escalating memory performance requirements, which traditional DRAM memories often fail to satisfy, necessitate the use of High-Bandwidth Memory (HBM), which offers high bandwidth, low power consumption, and high integration capacity, making it a promising solution for next-generation platforms. However, despite the notable increase in memory bandwidth on modern systems, no prior work has comprehensively assessed the memory bandwidth requirements of a diverse set of HPC applications and provided sufficient justification for the cost of HBM with potential performance gain. This work presents a performance analysis of a diverse range of scientific applications as well as standard benchmarks on platforms with varying memory bandwidth. The study shows that while the performance improvement of scientific applications varies quite a bit, some applications in CFD, Earth Science, and Physics show significant performance gains with HBM. Furthermore, a cost-effectiveness analysis suggests that the applications exhibiting at least a 30% speedup on the HBM platform would justify the additional cost of the HBM.

Keywords: Benchmarking · Performance analysis · Memory bandwidth

1 Introduction

In recent decades, high-performance computing (HPC) has become an indispensable part of numerous scientific and engineering domains, such as molecular

This work is supported by the National Science Foundation through the Frontera award (OAC-1854828) and the CSA award (OAC-2139536).

A. Bienz et al. (Eds.): ISC High Performance 2023 Workshops, LNCS 13999, pp. 339–352, 2023.
https://doi.org/10.1007/978-3-031-40843-4_25

dynamics, computational fluid dynamics, climate modeling, and others. Consequently, the performance of modern computing systems has emerged as a crucial factor in advancing research in these fields. In response to the rising demand for data-intensive applications, memory bandwidth has become a critical consideration in the procurement of large-scale HPC systems, alongside the augmentation of computing capabilities to enhance overall system performance.

The significance of memory bandwidth in HPC systems has resulted in the creation of technologies aimed at enhancing memory bandwidth. Apart from conventional measures such as improving memory channels, clock speed, bus width, etc., novel architectures like High-Bandwidth Memory (HBM) have been developed for parallel computing with efficient power usage. Despite the notable increase in memory bandwidth in contemporary systems, no research has comprehensively assessed the impact of memory bandwidth on a diverse set of HPC applications on HBM systems and justified the extra cost of HBM.

In this study, we aim to explore the impact of memory bandwidth on the performance of a diverse range of HPC applications as part of the planning process for the Leadership-Class Computing Facility (LCCF). Specifically, we investigate the performance improvement of these applications when executed on three different architectures, a compute node on Frontera with two Intel Xeon Platinum 8280 processors (Cascade Lake), a test node with two Intel Xeon Max 9480 processors (Sapphire Rapids) with HBM disabled, and another identical test node with DDR5 disabled but HBM enabled. We then compare the results from the applications with those obtained from standard benchmarks like STREAM, SPEC CPU 2017, and SPEChpc 2021. The objective of this research is to gain insights into the design and evaluation of the memory bandwidth requirement of next-generation HPC systems based on real-world application workloads.

2 Background

2.1 Sapphire Rapids and HBM

Sapphire Rapids is the latest generation of Intel Xeon Scalable processors designed for high-performance computing. It features up to 8-channel DDR5 memory interface with 4 memory controllers, providing up to 300 GB/s of memory bandwidth per socket. HBM is a type of memory technology that offers significantly higher bandwidth than traditional DDR memory [11]. It achieves this by stacking multiple memory dies vertically, and connecting them to the CPU via a high-speed interface. This allows for faster transfer of data between the CPU and memory, which can result in improved performance for memory-intensive applications. The Sapphire Rapids Max processor we are testing supports up to 4 HBM2 stacks onboard to provide a total of 64 GB of memory per socket. It can provide up to 1 TB/s of memory bandwidth per socket.

2.2 Characteristic Science Application

The Characteristic Science Applications (CSAs) are a set of computer codes and challenge problems selected to represent a diverse range of scientific domains and computational approaches. The primary goal of the project is to transform these CSAs to enable next-generation science on the NSF LCCF. The first phase of the CSA has a total of 20 projects covering scientific domains including Astronomy and Astrophysics, Biophysics and Biology, Computational Fluid Dynamics, Earth Science, Materials Science, and Physics. Here, we selected 15 applications from these projects that can well-utilize the performance of the CPU and the WRF benchmark from TACC's internal benchmark applications as our application benchmark to understand how memory bandwidth may further improve the performance on the next-generation CPU platforms.

2.3 Benchmarks and Memory Bandwidth

The STREAM benchmark is the industry standard benchmark designed to measure the sustainable memory bandwidth of a system by testing four basic vector operations: copy, scale, add, and triad [13]. The benchmark reports the bandwidth achieved by each of these operations in MB/s, and it is considered the maximum achievable memory bandwidth of the system. Because the memory bandwidth measurements from STREAM may not be representative of the performance of real-world applications, it is often used in conjunction with other benchmarks, such as the SPEC CPU 2017 [3,15], and SPEChpc 2021 [2,12] benchmarks, to provide a more comprehensive view of the performance of a system.

SPEC CPU 2017 is a set of standardized benchmarks that measure the performance of processors on a variety of tasks, such as integer, floating-point, and memory-intensive workloads. SPEChpc 2021 is a set of benchmarks designed to measure the performance of HPC systems on scientific applications. One limitation of both SPEC CPU 2017 and SPEChpc 2021 is that the selected workloads and applications are mostly memory-bound, especially for SPEChpc 2021, this is partially due to the fact that SPEC benchmarks don't allow strong dependence of external libraries, therefore no math libraries like BLAS or LAPACK are involved [2,15]. Therefore, they will be biased when used alone to evaluate the impact of memory bandwidth on scientific applications.

There are two commonly used application benchmarks: the CORAL-2 and SPP benchmarks. The CORAL-2 (Collaboration of Oak Ridge, Argonne, and Livermore) benchmark was developed to assess the performance of HPC systems for scientific simulations and data analytics commonly supported by the Department of Energy. The SPP (Scalable Parallel Programming) benchmark focuses on the scientific community and includes parallel kernels and full applications that represent typical HPC workloads in academia. Although some applications overlap with the ones in our benchmark, such as AWP-ODC, MILC, and NAMD, the SPP benchmark was last updated in 2017 [1] and may not reflect the latest developments in scientific research. The application benchmark presented in our work serves as an update and expansion, incorporating the most recent scientific test cases and new applications developed by the community.

3 Implementation

3.1 Resources

Three different platforms are used in this work: 1) a compute node on Frontera with two Intel Xeon Platinum 8280 processors (CLX) with 6-channel DDR4 (2933 MT/s) memory per socket, 2) a test node with two Intel Xeon Max 9480 processors (SPR) with 8-channel DDR5 (4800 MT/s) memory per socket and its HBM disabled, and 3) another SPR test node with 4 HBM2 stacks onboard per socket and its DDR5 memory disabled. The CLX node has 28 cores per socket and the SPR node has 56 cores per sockets (Table 1). Both of the SPR nodes are configured to the flat mode in their sub-NUMA clustering configurations.

3.2 Standard Benchmarks

The STREAM benchmark can run with many different configurations. The selected configurations here are aimed to produce consistent and robust results that lie within the 80% to 90% percentile range of tests. It uses all the cores while ensuring that the size of arrays is at least four times larger than the combined sizes of all the L2 and L3 caches. We also tested compiling the benchmark with streaming stores or allocating stores and use the best configuration.

The SPEC CPU 2017 and SPEChpc 2021 runs use binaries compiled from scratch with the icpx, icx, and ifx compilers from Intel oneAPI 2022.1. Compilers flags are chosen from the most recent vendor submissions in order to be comparable. All SPEC runs utilize all physical cores available on the nodes.

3.3 Application Benchmarks

The 16 selected applications include ChaNGa, Enzo-E, Athena++, NAMD, Amber, PSDNS, CHyPS and Plascom, ISSM, SeisSol, CESM, WRF, AWP-ODC, EPW, Parsec, MuST, and MILC. Below, we include a short description of each application as well as its single-node benchmark performed. All the benchmarks were run on all the cores available on the node unless otherwise is documented below. Also, most of the benchmarks were measured by time in seconds, and the exceptions are also described in details below.

ChaNGa. (Charm N-body GrAvity solver) [8] is a cosmological code that performs collisionless N-Body simulations, including optional hydrodynamics using the Smooth Particle Hydrodynamics (SPH) method. A Barnes-Hut tree algorithm is used to structure particles and solve gravity equations within a volume with periodic boundary conditions. ChaNGa uses the Charm++ [9] runtime system for parallelism and relies on its dynamic load-balancing scheme to achieve good parallel efficiency on distributed systems. The 50 million particle zoom-in simulation 'dwf1.6144' case was used as a benchmark with a performance metric obtained by summing the elapsed runtime of 3 'BigSteps'.

Enzo-E. Enzo-E is an extension of the Enzo parallel astrophysics and cosmology application. The Enzo-E application is capable of running numerical simulations to address current scientific questions in astrophysics and cosmology. Enzo-E is a parallel adaptive mesh refinement (AMR) hydrocode. The AMR algorithm leverages the Cello framework. Parallel implementation is achieved through the use of Charm++. Enzo-E consists of roughly 75,000 lines of C++ with a bit of FORTRAN thrown in. The Cello framework consists of about the same number of lines of C++. The benchmark problem consists of a root grid that is 128^3 cells in size with a block decomposition of 8^3 resulting in 16^3 cells per block. The test case was run from scratch without checkpointing enabled to eliminate I/O from the run times. Performance was measured in total elapsed time in seconds.

Athena++. Athena++ is a astrophysical radiation magnetohydrodynamics code. A great strength of the code is the broad range of physics it models, and therefore the wide variety of problems to which it can be applied. The fundamental framework of Athena++ is based on a block-based AMR mesh organized in an oct-tree. A dynamic execution model that implements task-based parallelism is used to improve parallel performance by overlapping communication and computation and simplify the inclusion of a diverse range of physics.

NAMD. NAMD is a massively scaled molecular dynamics simulation package to model the physical movements of atoms and molecules in biological systems and functional materials. [16] Its remarkable parallel performance greatly benefits from Charm++ [9], an adaptive load balancing framework for efficient inter-process communication. In addition, accelerated computing through GPU offloading is now available at NAMD, which can take advantage of the highly optimized NVIDIA cuFFT library for fast Fourier transform when treating the time-consuming electrostatic interactions in an MD simulation. The benchmark case for NAMD is a satellite tobacco mosaic virus (STMV) dissolved in water that entails a total of 1.06 million atoms. Its performance is measured in seconds/step, i.e., seconds of simulation time per MD step.

Amber. Amber is a software suite of molecular simulation programs for biomolecular systems and is most often used with its namesake (amber) force field. Its primary molecular dynamics (MD) engine, PMEMD, is designed for large-scale parallel CPU and GPU computing systems (most of its features support GPU acceleration). The benchmark is the STMV_production_NPT_4fs case from the Amber20 Benchmark Suite (available at https://ambermd.org). In the simulation, the dynamics of a 1.067 million-atom solvated satellite tobacco mosaic virus (STMV) system is propagated under the isothermal-isobaric (NpT-ensemble) condition for 10,000 steps (with 4fs/step). The benchmark ran with 56 (CLX) and 112 (SPR) MPI tasks respectively. For better performance, the I/O was done on the local /tmp directory instead of using a shared filesystem. The performance is evaluated by averaging the timings for all steps and is reported in ns/day.

CHyPS and Plascom. CHyPS and Plascom are multiphysics simulation codes that work together to solve of a full hypersonic vehicle simulation. This including shocks, chemical reaction, radiation, laminar-to-turbulent boundary layer transition, gas-surface interaction, and surface material modeling (degradation, ablation, and oxidization). The benchmark case can only run with 20 MPI tasks, so only 20 cores were used in all the tests.

PSDNS. PSDNS [5] is a software to model large-scale turbulent flow under constant or nearly constant density conditions through the Fourier pseudo-spectral method, which is particularly powerful for investigating nonlinear scale interactions often exhibited by the classical energy cascade. In the case of substantial density fluctuations, PSDNS can still capture the significant departures from classical cascade idealizations in incompressible flows along with the dynamical effect of strong compression and expansion using higher-order compact finite differences for discretization in space. As a result, the parallel performance of PSDNS heavily relies on the efficiency of the large-scale fast Fourier transform library it interfaces with. Our chosen benchmark case for PSDNS consists of $12888 \times 12888 \times 12888$ grid points and 339,738,624 particles, placing an extremely high demand on an HPC system's memory bandwidth. The performance is measured in seconds/step, i.e., seconds of simulation time per propagation step.

ISSM. The Ice-sheet and Sea-level System Model (ISSM) is an open-source software package designed to simulate ice sheet and sea level behavior [10]. It uses a finite element approach to model ice flow and the interactions between ice sheets and the ocean. ISSM can be used to model past, present, and future ice sheet behavior under different climate scenarios, making it a useful tool for studying the impacts of climate change on sea level rise. The code is written in C++, and uses PETSc as the numerical solver. The benchmark case is a medium-sized mesh with 4.7×10^6 elements, and we run it with 5 timesteps. The performance is measured using the total core solution elapsed time reported by the code.

SeisSol. SeisSol is an earthquake simulation software that solves seismic wave propagation in viscoelastic media and dynamic rupture problems on geometrically complex, heterogeneous 3D models using clustered local time stepping on unstructured statically-adaptive tetrahedral meshes [6]. It is a hybrid MPI + OpenMP code. The computational kernels for many CPU architectures are generated via the Yet Another Tensor Toolbox, which uses small-BLAS back-ends such as LIBXSMM or Eigen. The benchmark case is a spontaneous rupture on a vertical strike-slip fault in a homogeneous halfspace. It has $2,051,112$ cells and $346,222$ vertices, and we set the simulation end time to be 2 s. The performance is measured with the elapsed time reported by the code.

CESM. The Community Earth System Model (CESM) [7] is a fully coupled, global climate model that provides state-of-the-art simulations of the Earth's past, present, and future climate states. The EarthWorks Modeling System leverages the CESM, but is especially focused on high-resolution ESM research at Global Storm Resolving (GSR) resolutions. It differs from standard CESM model configurations primarily in the use of the Model for Prediction Across Scales (MPAS) infrastructure for ocean, sea-ice and atmosphere components. The test case is a low-resolution Aqua Planet case (called a QPC6 component set), in which the atmospheric component is run with full physics. The test case makes the simplifying assumption that the entire planetary surface is covered by water. A data ocean component supplies the SST (Sea Surface Temperature) as a lower atmospheric boundary condition. To fit on one node, the resolution is set on a quasi-uniform grid at 120 km (40962 cells) with 32 vertical levels. The model ran for five simulation days. Performance was measured by total elapsed time of the model run.

WRF. The Weather Research and Forecasting(WRF) Model [18] is a widely used numerical weather prediction system used for both research and operational forecasts. WRF has been used as a standard benchmark for HPC procurements for many years. It is primarily a Fortran code implemented using MPI and OpenMP for distributed computing. The benchmark presented here is the standard CONUS 2.5 KM case used to compare the performance of WRF across a variety of architectures. Specifically, it simulates the weather across the continental United States at a horizontal resolution of 2.5 km. The performance is measured from the total elapsed time taken during the domain 1 execution phase with the exception of the first time step. This removes overhead from the initialization steps and the initial I/O.

AWP-ODC. AWP-ODC simulates the propagation of seismic waves. The equations are formulated using a finite difference scheme and a stencil update is used to advance the simulation in time and space (3D). The current production versions of the code are written in C and C/CUDA, respectively. The code base has been under active development for 20+ years. The number of lines of the C code is approximately 6,000 lines. The benchmark case is a dynamic wave propagation study in a homogeneous halfspace. The dimension of the 3D solid is $179.2 \times 102.4 \times 51.2$ (km), with a uniform mesh size of 200 m. The simulation duration and time step are 19.99 sec and 0.01 sec, respectively. The performance is measured with the elapsed time reported by the code.

EPW. The EPW code (https://epw-code.org) is an open-source code released under the GNU GPL consisting of approximately 67K lines of Fortran with MPI/OpenMP. EPW is the most popular code for first-principles calculations of electron-phonon interactions and finite-temperature properties. EPW is highly optimized to compute efficiently and accurately an array of properties and phenomena related to the electron-phonon interaction, e.g. electrical transport, phonon-assisted optical properties, and superconductivity [17]. The benchmark case MgB2 (magnesium diboride) is the phonon-mediated superconductor with the highest superconducting critical temperature (39K) at ambient pressure. This system provides an ideal testbench for developing more accurate and more predictive first-principles theories and algorithms for superconducting materials. The dimensions are nk1=nk2=10, nk3=6, nq1=nq2=10, nq3=6, nkf1=nkf2=nkf3=nqf1=nqf2=nqf3=16. Due to memory size, the benchmark runs with 56 tasks on the SPR nodes.

PARSEC. PARSEC is a versatile Density Functional Theory (DFT) code that solves the Kohn-Sham equations by expressing electron wave-functions directly in real space, without the use of explicit basis sets. It is capable of handling both periodical boundary conditions and confined-system boundary conditions. A finite-difference approach is used for the calculation of spatial derivatives. Pseudopotentials are used to describe the interaction between valence electrons and ionic cores. The code is comprised of approximately 50k lines of Fortran code. Additionally, PARSEC is highly scalable to thousands of nodes, and makes efficient use of AVX512 through ScaLAPACK and BLAS math libraries. The benchmark case is a single-point energy calculation of $Si_{1947}H_{604}$. The grid is set at 0.9Å grid spacing and boundary sphere radius of 50 bohr. Total number of calculated states is 4800.

MuST. MuST is an open-source package designed to perform ab initio electronic structure calculations for the study of quantum phenomena in disordered materials. The code has approximately 250,000 lines, mostly written in FORTRAN-90, and has been under active development for around 25 years. The MuST package is developed based on full-potential multiple scattering theory, also known as the KKR method, with Green's function approach to the Kohn-Sham equation in density functional theory (DFT). It is capable of performing KKR, KKR-CPA, and linear scaling LSMS calculations for materials with complex structures. It also allows for electronic conductivity calculation based on Kubo-Greenwood formula. For details of the LSMS method used in the benchmark case, see [19]. The benchmarking case is a Cantor alloy, CrMnFeCoNi, one of the best-known examples of high entropy alloys (HEAs) with excellent mechanical properties. System size is 56 atoms and 32 energy points. The benchmark runs with 112 tasks on the SPR nodes.

MILC. Lattice QCD is an approach to studying Nature's strong interaction, also called the nuclear force. This force is responsible for holding atomic nuclei together and for binding quarks into the protons and neutrons that comprise the atomic nuclei. Lattice QCD is a nonperturbative technique in which the quantum fluctuations of the quarks and gluons are treated somewhat analogously as in a statistical mechanical system. The MILC collaboration code is one of the community codes that can be used to produce the ensemble of configurations. The code is typically characterized as being memory-bound. It is written in C and contains about 350k lines of code. Libraries being used are QPhiX, QUDA, Grid and FFTW. In this benchmark, we set the lattice grid to be $32 \times 32 \times 32 \times 32$, and the lattice spacing is approximately 0.03 fm. The strange and charm quark masses are set to their physical values, and average values are used for the up and down quarks. The code is run in single precision. The step size is 0.0125 and the total number of steps is 40.

4 Results

The results of the STREAM benchmark are presented in Table 1. The CLX node exhibits a maximum bandwidth of approximately 220 GB/s, while the SPR nodes with DDR5 and HBM demonstrate maximum bandwidths of approximately 399 GB/s and 1400 GB/s, respectively. The outcomes derived from testing the CLX node exhibit negligible fluctuations from one execution to the next. Similarly, the test results for the SPR node with DDR5 indicate minimal variations in performance, although it yields slightly better results without streaming stores. In contrast, the SPR node with HBM demonstrates a significant degree of performance variability, approximately 10%, in relation to the problem size and array alignment. The selected subset in Table 1 is a considerably high Triad result, while the other results are at least 5% lower than the highest across the entire tests.

Table 1. STREAM Benchmark Results

Platform	Sockets	Cores	Copy	Scale	Add	Triad	Size (M)
CLX	2	56	204,396	204,391	220,498	220,219	1600
SPR w/DDR5	2	112	378,852	375,917	397,918	398,578	640
SPR w/HBM	2	112	1,371,992	1,370,889	1,343,482	1,400,131	3200

Table 2 presents the scores obtained from the SPEC benchmark, while Fig. 1 illustrates the speedup achieved by the two SPR nodes over the CLX node. Specifically, the SPECspeed2017_fp_base benchmark demonstrates that the SPR node with DDR5 provides a speedup of 1.66, with an additional 10% improvement achieved through the use of HBM. On the other hand, the SPECrate2017_fp_base metric measures the system's throughput, making higher memory bandwidth more desirable. The SPR node with DDR5 demonstrated a speedup of 2.10, and HBM gave an additional 29% improvement resulting in a speedup of 2.71. The SPEChpc 2021_tny_base benchmark demonstrates an even greater sensitivity to memory bandwidth, with the SPR nodes exhibiting speedups of 2.19 and 3.30, respectively.

Table 3 lists the results from the applications benchmarks, and Fig. 2a displays the speedup of the two SPR nodes relative to the CLX nodes for these

Table 2. SPEC Benchmark Scores

Benchmark	CLX	SPR w/DDR5	SPR w/HBM
SPECspeed2017_fp_base	169	280	308
SPECrate2017_fp_base	350	736	950
SPEChpc 2021_tny_base	3.15	6.89	10.40

Higher score is better.

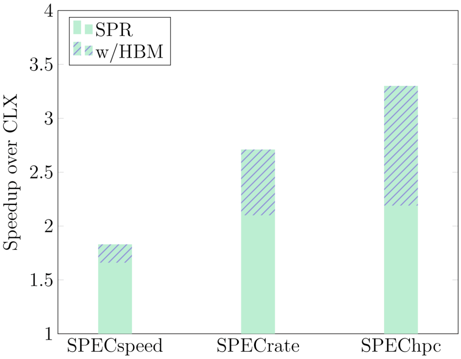
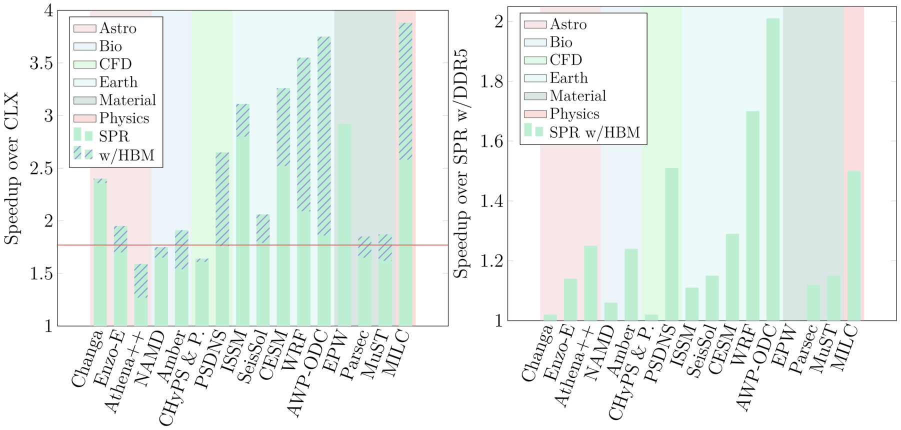


Fig. 1. The performance improvement of the SPEC benchmarks on the two Sapphire Rapids (SPR) nodes compared to the Cascade Lake (CLX) node. The additional performance gained from HBM is shaded in blue in the bar plot. (Color figure online)

Table 3. Application Benchmark Results

Application	Area	CLX	SPR w/DDR5	SPR w/HBM
ChaNGa	Astro	143.70	60.90	60.00
Enzo-E	Astro	763.41	447.76	391.14
Athena++	Astro	243.20	191.30	152.80
NAMD	Bio	0.34	0.21	0.19
Amber	Bio	2.05	3.16	3.92
CHyPS & P	CFD	32.70	20.35	20.02
PSDNS	CFD	916.66	520.83	345.91
ISSM	Earth	162.02	57.79	52.11
SeisSol	Earth	2075.76	1159.84	1006.01
CESM	Earth	1326.00	527.00	407.00
WRF	Earth	1810.74	865.40	509.25
AWP-ODC	Earth	328.00	176.03	87.36
EPW	Materials	137.71	47.17	48.11
Parsec	Materials	574.36	348.25	310.46
MuST	Materials	1631.10	1007.11	872.91
MILC	Physics	2018.30	783.40	520.90



(a) Speedups of two SPRs over CLX

(b) Speedup of HBM over DDR5

Fig. 2. The performance improvement of the application benchmark (a) on the two Sapphire Rapids (SPR) nodes compared to the Cascade Lake (CLX) node, and (b) on the Sapphire Rapids (SPR) node with High Bandwidth Memory (HBM) compared to the one with DDR5. The additional performance gained from HBM is shaded in blue in the bar plot. The applications are classified based on six scientific fields and arranged in increasing order according to the speedup gained from HBM. The red line marks the median of the speedups on the SPR node with DDR5 only. (Color figure online)

benchmarks. The SPR node with DDR5 achieved a speedup ranging from 1.27 to 2.92 with a median of 1.77, while the HBM one achieved a speedup ranging from 1.59 to 3.87 with a median of 2.23. The speedup achieved by utilizing HBM is depicted in Fig. 2b, exhibiting a range of 1 to 2.01, with a median of 1.15.

5 Discussion

The result from the STREAM benchmark indicates that the SPR w/DDR5 has about double the memory bandwidth than the CLX. The improvement is a lot more when moving to the SPR w/HBM node, with an additional 3.5× increase in memory bandwidth. These numbers represent the maximum performance gains that a memory-bound application can achieve.

The results from the SPEC benchmarks require further analysis as the performance improvement is influenced by changes in core count, clock speed, and memory bandwidth. To evaluate a purely compute-bound problem, we used the High-Performance Linpack (HPL) Benchmark [4]. The best HPL results obtained from the CLX, SPR w/DDR5, and SPR w/HBM nodes were 3.25, 5.39, and 5.73 TFLOPS, respectively. These numbers were achieved without fine-tuning but provide insight into the behavior of a compute-bound code. The slightly higher performance of the SPR w/HBM node can be attributed to the higher clock speed of the CPU when using HBM with lower power consumption [14].

It suggests that if the performance improvement of the SPR w/HBM node over the SPR w/DDR5 node is within 6%, it may not be solely attributed to the increased memory bandwidth.

The relative speedup from the SPECspeed2017_fp_base results is very close to that from the HPL benchmark, suggesting that the improvement of the SPECspeed2017_fp_base metric on the two SPR nodes represents the performance gain when the application is not memory-bound. The SPECrate2017_fp_base and SPEChpc 2021_tny_base, on the other hand, reflect the behavior of a more memory-bound application. It is worth noting that none of the benchmarks achieved the same level of performance gain on the HBM platform as the STREAM benchmark, indicating that the bandwidth provided by the HBM is more than what our applications require with given core count on the SPR platform.

Regarding the application benchmarks, the performance improvement on the two SPR nodes varies depending on the specific application. On average, the speedups of the SPR w/DDR5 and SPR w/HBM nodes are 1.98 and 2.51, respectively. Both of these speedups are lower than the corresponding metrics obtained from the SPEChpc benchmark. These results suggest that the applications selected for our analysis are less memory-bound.

We did not observe any general trends by categorizing the applications into different science domains. However, we did observe that PSDNS, WRF, AWP-ODC, and MILC achieved more than 40% performance gain when moving from DDR5 to HBM. We may also evaluate the cost-effectiveness of HBM based on the application benchmark results. The current "Recommended Customer Price" of the Intel Xeon CPU Max 9480 Processor, which was used in this work, is listed on Intel's website as $12,980. The closest comparable processor without HBM is the Intel Xeon Platinum 8480+ Processor, which is listed as $10,710. Therefore, we may conclude that it is appropriate to acquire the HBM processor if the SPR w/HBM is 21% faster than the SPR w/DDR5. However, this calculation does not take into account other associated costs in a system procurement, and we may need to offset the 6% performance gain from the power consumption. Overall, a more reasonable criteria for cost-effectiveness evaluation would be 30%, and only PSDNS, CESM, WRF, AWP-ODC, and MILC meet this requirement.

6 Conclusion

We compared the performance of three different architectures, the CLX, SPR w/DDR5, and SPR w/HBM, using a suite of benchmarks and application tests. We found that HBM has significantly improved memory bandwidth over DDR5 and results in higher performance gains for memory-bound applications. However, the performance improvement varies depending on the specific application, and on average in our application benchmarks, the speedup of HBM over DDR5 on the SPR nodes is 27%. We observed that PSDNS, WRF, AWP-ODC, and MILC achieved significant performance gain when moving from DDR5 to HBM, and only these applications plus CESM met our criteria for cost-effectiveness evaluation. This suggests that domains such as CFD, Earth Science, and Physics may benefit more from the high memory bandwidth offered by HBM.

References

1. Bauer, G., et al.: Updating the SPP benchmark suite for extreme-scale systems. In: Proceedings of Cry User Group Meeting (CUG-2017) (2017)
2. Brunst, H., et al.: First experiences in performance benchmarking with the new SPEChpc 2021 suites. In: 2022 22nd IEEE International Symposium on Cluster, Cloud and Internet Computing (CCGrid), pp. 675–684 (2022). https://doi.org/10.1109/CCGrid54584.2022.00077
3. Bucek, J., Lange, K.D., v. Kistowski, J.: SPEC CPU2017: next-generation compute benchmark. In: Companion of the 2018 ACM/SPEC International Conference on Performance Engineering, ICPE 2018, pp. 41–42. Association for Computing Machinery, New York (2018). https://doi.org/10.1145/3185768.3185771
4. Dongarra, J.J., Luszczek, P., Petitet, A.: The LINPACK benchmark: past, present and future. Concurr. Comput.: Pract. Experience **15**(9), 803–820 (2003). https://doi.org/10.1002/cpe.728
5. Donzis, D.A., Yeung, P.K., Sreenivasan, K.R.: Dissipation and enstrophy in isotropic turbulence: resolution effects and scaling in direct numerical simulations. Phys. Fluids **20**(4), 045108 (2008). https://doi.org/10.1063/1.2907227
6. Heinecke, A., et al.: Petascale high order dynamic rupture earthquake simulations on heterogeneous supercomputers. In: Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis, SC 2014, pp. 3–14 (2014). https://doi.org/10.1109/SC.2014.6. ISSN 2167-4337
7. Hurrell, J.W., et al.: The community earth system model version 2 (CESM2). J. Adv. Model. Earth Syst. **11**(12), 3761–3802 (2019). https://doi.org/10.1029/2019MS001916
8. Jetley, P., Gioachin, F., Mendes, C., Kale, L.V., Quinn, T.: Massively parallel cosmological simulations with ChaNGa. In: 2008 IEEE International Symposium on Parallel and Distributed Processing, pp. 1–12. IEEE (2008). https://doi.org/10.1109/IPDPS.2008.4536319
9. Kale, L.V., Krishnan, S.: CHARM++: a portable concurrent object oriented system based on C++. In: Proceedings of the Eighth Annual Conference on Object-Oriented Programming Systems, Languages, and Applications, OOPSLA 1993, pp. 91–108. Association for Computing Machinery, New York (1993). https://doi.org/10.1145/165854.165874
10. Larour, E., Seroussi, H., Morlighem, M., Rignot, E.: Continental scale, high order, high spatial resolution, ice sheet modeling using the ice sheet system model (ISSM). J. Geophys. Res.: Earth Surface **117** (2012). https://doi.org/10.1029/2011JF002140
11. Lee, D.U., et al.: 25.2 A 1.2V 8Gb 8-channel 128GB/s high-bandwidth memory (HBM) stacked DRAM with effective microbump I/O test methods using 29 nm process and TSV. In: 2014 IEEE International Solid-State Circuits Conference Digest of Technical Papers (ISSCC), pp. 432–433 (2014). https://doi.org/10.1109/ISSCC.2014.6757501. ISSN 2376-8606
12. Li, J., et al.: SPEChpc 2021 benchmark suites for modern HPC systems. In: Companion of the 2022 ACM/SPEC International Conference on Performance Engineering, ICPE 2022, pp. 15–16. Association for Computing Machinery, New York (2022). https://doi.org/10.1145/3491204.3527498

13. McCalpin, J.D.: Memory bandwidth and machine balance in current high performance computers. IEEE Comput. Soc. Tech. Committee Comput. Archit. (TCCA) Newsl. **2**, 19–25 (1995)
14. Nabavi Larimi, S.S., Salami, B., Unsal, O.S., Kestelman, A.C., Sarbazi-Azad, H., Mutlu, O.: Understanding power consumption and reliability of high-bandwidth memory with voltage underscaling. In: 2021 Design, Automation & Test in Europe Conference & Exhibition (DATE), Grenoble, France, pp. 517–522. IEEE (2021). https://doi.org/10.23919/DATE51398.2021.9474024
15. Panda, R., Song, S., Dean, J., John, L.K.: Wait of a decade: did SPEC CPU 2017 broaden the performance horizon? In: 2018 IEEE International Symposium on High Performance Computer Architecture (HPCA), pp. 271–282 (2018). https://doi.org/10.1109/HPCA.2018.00032
16. Phillips, J.C., et al.: Scalable molecular dynamics on CPU and GPU architectures with NAMD. J. Chem. Phys. **153**(4), 044130 (2020). https://doi.org/10.1063/5.0014475
17. Poncé, S., Margine, E.R., Verdi, C., Giustino, F.: EPW: electron-phonon coupling, transport and superconducting properties using maximally localized Wannier functions. Comput. Phys. Commun. **209**, 116–133 (2016). https://doi.org/10.1016/j.cpc.2016.07.028
18. Skamarock, W., et al.: A description of the advanced research WRF version 3. Technical report, University Corporation for Atmospheric Research (2008). https://doi.org/10.5065/D68S4MVH
19. Wang, Y., Stocks, G.M., Shelton, W.A., Nicholson, D.M.C., Szotek, Z., Temmerman, W.M.: Order-N multiple scattering approach to electronic structure calculations. Phys. Rev. Lett. **75**, 2867–2870 (1995). https://doi.org/10.1103/PhysRevLett.75.2867