Application Brief

Ansys® Fluent®
Intel® Xeon® CPU Max Series

Faster Simulations at Scale: Ansys® Fluent® and the Intel® Xeon® CPU Max Series

Innovations of the Intel Xeon CPU Max Series, including on-package High Bandwidth Memory (HBM), increase Ansys Fluent performance across Ansys benchmarks.

Each new generation of CPUs enables HPC clusters across industry, government and academia to run more sophisticated simulations on larger data sets while accelerating results. At the same time, increased compute resources must be balanced by advances in the memory subsystem to keep redesigned, high-performance cores supplied with sufficient data. The performance of computational fluid dynamics (CFD) software such as Ansys® Fluent® tends to be memory-bandwidth-bound, showcasing requirements for high memory bandwidth to deliver optimal performance.

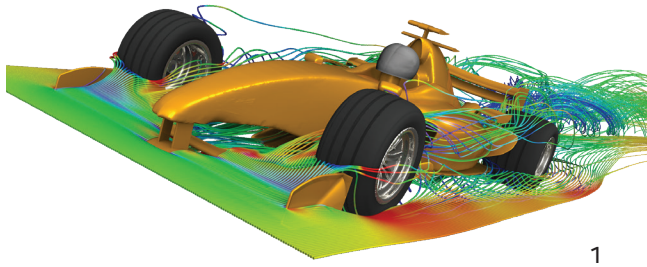
The hours or days required to complete complex simulations may restrict the scope of analysis possible, as well as extend cost and time requirements. Overcoming such limitations is the subject of ongoing joint work by Ansys and Intel. Enablement on Intel® architecture includes code optimization of Fluent binaries and testing on new platforms. The Intel Xeon® CPU Max Series introduces significant advances in platform memory architecture that directly benefit Fluent workloads.

This application brief introduces the co-engineered capabilities of Ansys Fluent on the Intel Xeon CPU Max Series and presents test results that demonstrate their combined unique value proposition. It shows that, together with other platform innovations, the CPU's very high memory bandwidth provided by 64GB of in-package memory substantially improves Fluent performance.

Hardware Platform Selection for Ansys Fluent

Ansys Fluent is computational fluid dynamics (CFD) software, widely used to model phenomena such as fluid mechanics, heat flow and mass transfer. Commonly run on computing clusters, Fluent is a memory-bandwidth-bound workload, making the speed of the memory subsystem critical to overall Fluent performance. Improved performance per node enables customers to use more sophisticated models based on richer datasets, to run more iterations in a given amount of time or to complete simulations in less time.

In creating the most cost-effective cluster environment to run Fluent, architects must consider that Ansys licensing is based on the number of cores in use. Higher performance per core is important in making the most efficient use of Ansys licenses. High core-count processors are advantageous in building out greater compute density, but that higher per-node density also creates a higher memory bandwidth demand. Intel Xeon CPU Max series leverages HBM to meet that higher memory bandwidth demand and enable efficient compute density.



Likewise, using too few cores per socket or NUMA domain causes memory bandwidth to be underutilized, reducing per-node efficiency. While there can be no substitute for testing within individual circumstances, industry best practices suggest that the sweet spot for Fluent cluster-node selection exists generally around moderate core counts of approximately 32 cores per socket. As a starting point for obtaining the best possible total cost of ownership, that moderate core count must be supported by high memory bandwidth.

Memory Innovation in the Intel Xeon CPU Max Series

The Intel Xeon CPU Max Series is the only x86-based platform with HBM built into the processor package. This innovation specifically targets large memory-bound workloads such as Fluent and other HPC applications, as well as deep learning and analytics. The novel memory subsystem features 64GB of built-in HBM2e memory, up to 112.5MB of shared last-level cache and eight DDR5-4800 memory channels per socket. With core counts up to 56 per CPU, the platform affords more than 1GB per core of HBM capacity.

The CPU cores are arranged on four tiles, as illustrated in Figure 1, which are interconnected using Intel's embedded multi-die interconnect bridge (EMIB) technology. A new core architecture provides a significant boost in instructions per clock (IPC) with high power efficiency. Platform enhancements and hardware optimizations built into the Intel Xeon CPU Max Series help maximize performance of the HBM subsystem, including the following:

- Refactored hardware prefetching algorithms
- Enhanced uncore frequency scaling
- Direct-to-core response on all local memory requests
- Enhanced snoop filter for cross-socket coherency

To help provide a balanced system, the Intel Xeon CPU Max Series incorporates enhanced I/O and cluster support. Throughput between the processor and devices is accelerated with up to 80 lanes of PCIe 5.0 per socket. The platform also supports Compute Express Link (CXL) 1.1 for high-speed fabric interconnect.

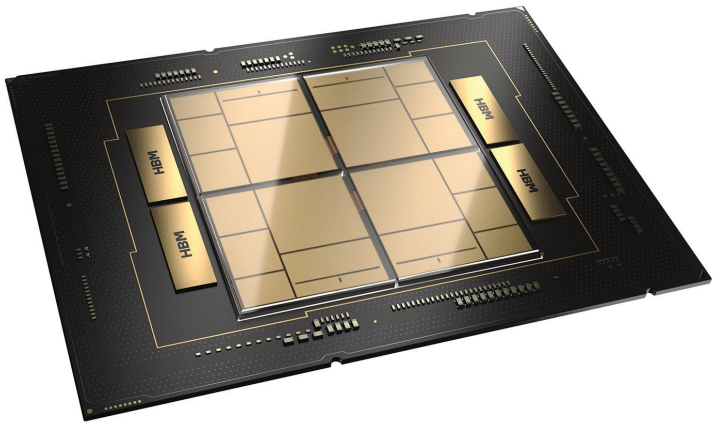


Figure 1. The Intel® Xeon® CPU Max Series.

The memory subsystem of the Intel Xeon CPU Max Series allows for flexible roles of in-package HBM and DRAM. Accordingly, three memory modes are available:

- **HBM-Only Mode** uses in-package HBM exclusively (and not DRAM), suited to workloads that fit within 64GB capacity and scale at 1-2GB per core.
- **HBM Flat Mode** provides separate memory regions of HBM and DDR5 memory for the flexibility to scale workloads beyond 64GB. Some run-time tuning or even re-coding is required to take full advantage of the separate memory regions.
- **HBM Cache Mode** configures the HBM as a memory-side cache, rather like a super-sized last level cache, using HBM as a cache for DRAM. This mode provides a seamless out-of-the-box experience, requiring no tuning or re-coding to take advantage of both HBM memory and large DDR5 memory capacity.

The testing reported in this brief is based on HBM Cache Mode.

Math Acceleration with Intel oneAPI Math Kernel Library

Longstanding collaboration between Ansys and Intel has optimized performance of Fluent for successive generations of Intel platforms. Intel oneAPI Math Kernel Library (oneMKL), part of the Intel oneAPI Base Toolkit, is a key enhancement to Fluent performance. Built specifically to accelerate math operations involved in solving large computational problems, oneMKL provides a range of highly optimized parallel routines for both sparse and dense linear algebra. In addition to the traditional performance advantages, oneMKL adds a set of SYCL interfaces to enable code for heterogeneous platforms. Key functional areas of oneMKL include BLAS and LAPACK linear algebra routines, random number generators, vector math and Fast Fourier Transforms (FFTs).

Optimizations to Fluent include implementation of the oneMKL sparse LDU smoother, which provides acceleration up to 15% compared to the Ansys Fluent native incomplete lower/upper (ILU) smoother.[1] When solving equations based on the sparsely populated matrices that are common in CFD domains, smoothers algorithmically provide approximate solutions to portions of the equations set involved. Those approximated solutions enable important patterns to stand out that provide insights to improve the overall efficiency of calculations.

OneMKL Sparse LDU Smoother:

1 oneAPI

- 15% acceleration over Ansys native smoother.[1]
- Implements Intel® AVX-512 with two FMA units per core.
- Future-ready optimization for upcoming Intel platforms

The oneMKL sparse LDU smoother is based on Intel AVX-512 instructions, which accelerate mathematically intensive operations to provide higher performance per core for better return on investment. Intel AVX-512 technology can pack 32 double-precision and 64 single-precision floating-point operations per clock cycle within the 512-bit vectors, as well as eight 64-bit and sixteen 32-bit integers. The technology doubles the width of data registers, the number of registers available and the width of fused add-multiply (FMA) units relative to its predecessor, Intel AVX 2.

FMA units combine addition and multiplication into a single operation to reduce the number of steps in computations, driving up throughput. Current implementations on Intel architecture provide two FMA units per core, which increases capacity. Another optimization of note is that oneMKL is enabled by default for Fluent, which delivers acceleration on the latest Intel platforms, including the Intel Xeon CPU Max Series and future platforms.

Quantifying Per-Core Performance and Scalability

Testing by Intel reveals that Fluent scales well on the Intel Xeon CPU Max Series up to 32 nodes, as illustrated in Figure 2. At this scale, the benefit of HBM provided by the hardware platform is readily apparent by comparing results on the Intel Xeon CPU Max Series and the 4th Gen Intel Xeon Scalable processor. In addition, both current-generation Intel processors are shown to outperform predecessor platforms.

This testing is based on the Ansys Fluent Benchmarking Suite, which consists of a suite of fluid-flow problems selected to represent typical usages. The test cases are explicitly developed for the purpose of benchmarking hardware performance for comparison across multiple system architectures and generations. Their development by Ansys helps ensure objectivity and consistency in measurements among hardware providers. The following set of test cases is reported on in these results:

- **F1_Racecar_140M:** Aerodynamic flow around a Formula-1 racecar, 140 million cells.
- **Open_Racecar_280M:** External aerodynamic flow around an open wheel racecar, 280 million cells.
- **Combustor_71m:** Flow through a combustor, species transport and combustion, 71 million cells.

Results are reported using the Ansys Solver Rating as a metric, defined as the number of times that a given benchmark can be run sequentially by a given machine within a 24-hour period. It is computed by dividing the number of seconds in a day by the number of seconds to run the benchmark, with a higher rating corresponding to higher expected performance on corresponding production workloads. Testing against a range of use cases such as those presented here is a best practice to provide a robust basis for performance comparison across hardware platforms.

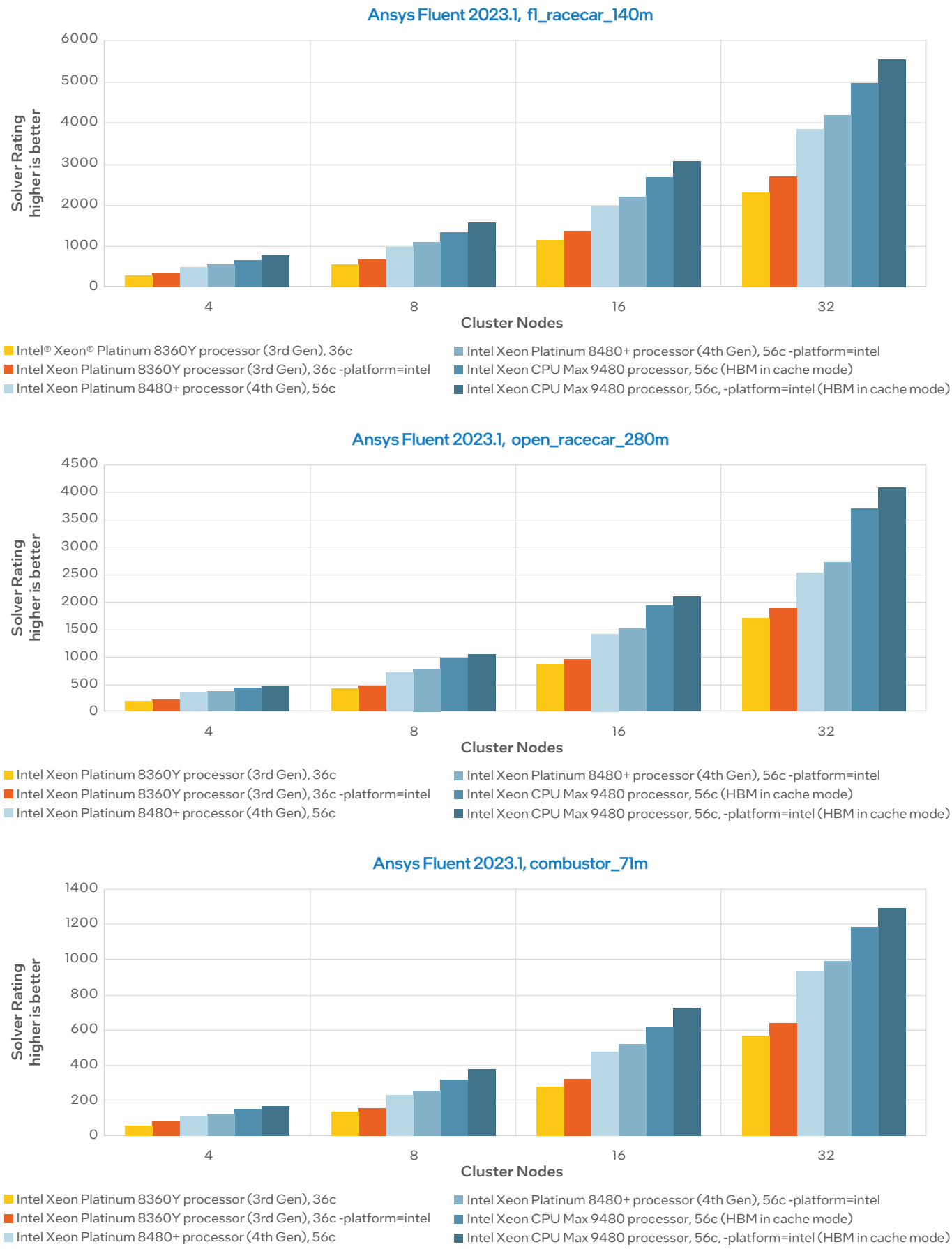


Figure 2. Performance comparisons across Ansys Fluent test cases, hardware and cluster sizes.[2]

Conclusion

The Intel Xeon CPU Max Series delivers significant performance upside for Fluent workloads compared with competing and predecessor platforms. Performance advantages across Ansys benchmarks demonstrate the opportunity for Fluent implementations. Platform advances including HBM built into the processor package as well as platform-specific tuning and optimizations in the software enable faster, more cost-effective simulations.

Learn More:

Ansys® Fluent®
Intel® Xeon® CPU Max Series

Solution provided by:

intel + Ansys

[1] Ansys, March 28, 2022. 2021 Annual Report. https://investors.ansys.com/static-files/d3e39ada-4e58-4dba-87f7-93e41eff55e4.

[2] 3rd Gen Intel Xeon Scalable processor: Intel Xeon Platinum 8360Y processor (2.40 GHz, 36 cores per socket), Intel Hyper-Threading Technology enabled, Intel Turbo Boost Technology enabled, 16x 16GB DDR4-3200, dual OPA fabric, BIOS SE5C6200.86B.0020.P23.2103261309, Microcode 0xd000270, Rocky Linux release 8.7 (Green Obsidian), kernel: 4.18.0-372.32.1.el8_6.crt3.x86_64, ib driver provided by base OS, single rail OmniPath 100.

4th Gen Intel Xeon Scalable processor: Intel Xeon Platinum 8480+ processor (2.00 GHz, 56 cores per socket), Intel Hyper-Threading Technology enabled, Intel Turbo Boost Technology enabled, SNC4, 16x 32GB DDR5-4800, Samsung M321R4GA3BB0-CQKVG, BIOS SE5C7411.86B.9525.D13.2302071332, Microcode 0x2b000190, Rocky Linux release 8.7 (Green Obsidian), kernel: 4.18.0-372.32.1.el8_6.crt3.x86_64, Mellanox OFED 5.9-0.5.6.0, single rail Mellanox HDR200.

Intel Xeon CPU Max Series: Intel Xeon CPU Max 9480 processor (1.90 GHz, 56 cores per socket), Intel Hyper-Threading Technology enabled, Intel Turbo Boost Technology enabled, SNC4, fake NUMA enabled, 8x 16 GB @ 3200 MHz HBM cache mode, 16x 32 GB DDR5-4800, Intel HMCG88MEBRA174N, BIOS SE5C7411.86B.9525.D13.2302071332, Microcode 0x2c000170, Rocky Linux release 8.7 (Green Obsidian), kernel: 4.18.0-372.32.1.el8_6.crt3.x86_64, Mellanox OFED 5.9-0.5.6.0, single rail Mellanox HDR200.

Performance varies by use, configuration and other factors.

Performance results are based on testing as of dates shown in configurations and may not reflect all publicly available updates. See configuration disclosure for configuration details. No product or component can be absolutely secure.

Intel does not control or audit third-party data. You should consult other sources to evaluate accuracy.

Your costs and results may vary.

Intel technologies may require enabled hardware, software, or service activation.

You may not use or facilitate the use of this document in connection with any infringement or other legal analysis concerning Intel products described herein. You agree to grant Intel a nonexclusive, royalty-free license to any patent claim thereafter drafted which includes subject matter disclosed herein.

The products described may contain design defects or errors known as errata which may cause the product to deviate from published specifications. Current characterized errata are available on request.

0623/JW/MESH/353912-001US