intel

Intel® oneAPI Math Kernel Library (oneMKL)

The Precision Spectrum: Performance and Reproducible Computing

Pavel Dyakov
pavel.dyakov@intel.com

Notices & Disclaimers

Performance varies by use, configuration and other factors. Learn more at www.Intel.com/PerformanceIndex.

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

Your costs and results may vary.

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

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

© Intel Corporation. Intel, the Intel logo, Xeon, Core, VTune, OpenVINO, and other Intel marks are trademarks of Intel Corporation or its subsidiaries.
Other names and brands may be claimed as the property of others.

Agenda

01 Biography

02 oneMKL Overview

03 oneMKL: BLAS Performance for different data types

04 oneMKL: Conditional Numerical Reproducibility

01

Biography

Pavel Dyakov

pavel.dyakov@intel.com



- Math Algorithm Engineering Manager at Intel.
- 10+ years of experience in high-performance computing and mathematical algorithm development.
- Engineering manager and product owner for:
  - Intel® Integrated Performance Primitives
  - Intel® Cryptography Primitives Library
  - Intel® oneAPI Math Kernel Library (oneMKL) Vector Statistics / Data Fitting
  - Intel® Query Processing Library
- The author of multiple patents and co-author contributions to C++ standards related to random number generation.

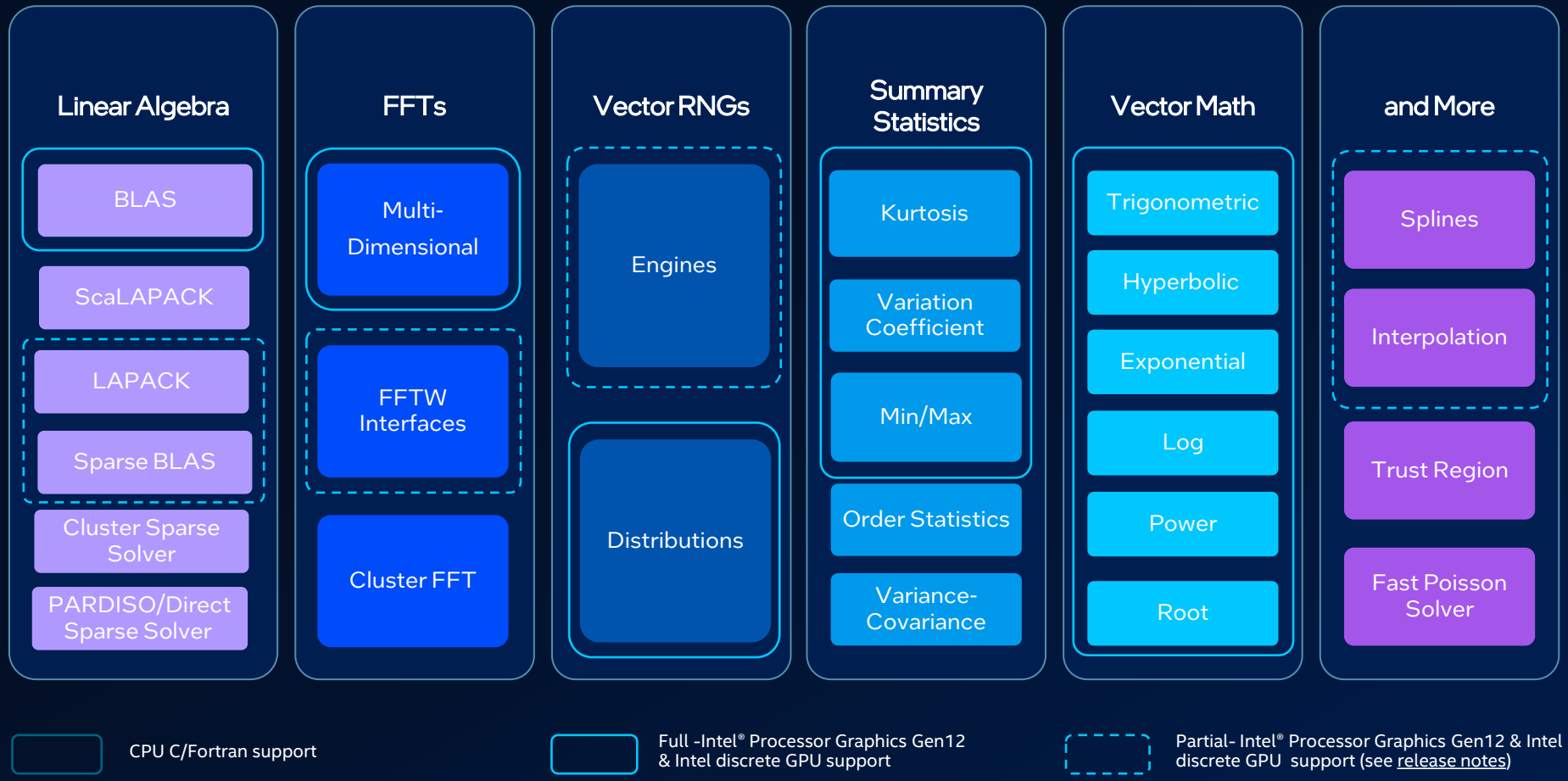
02

oneMKL Overview

oneMKL Overview

Useful links:

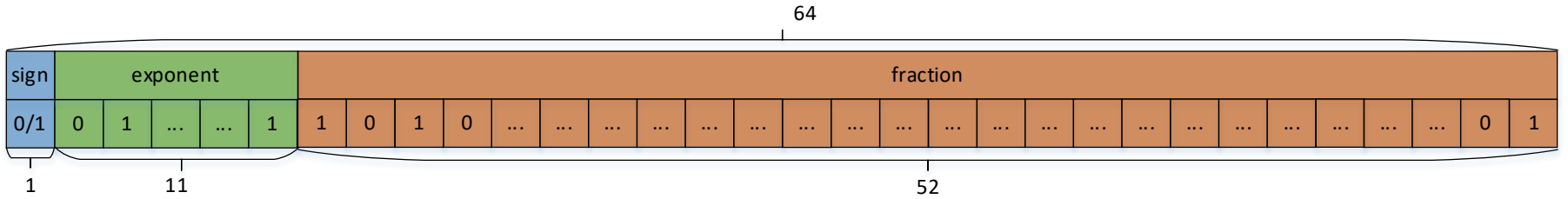
- Download
- C, Fortran, SYCL Developer Reference
- Link Line Advisor



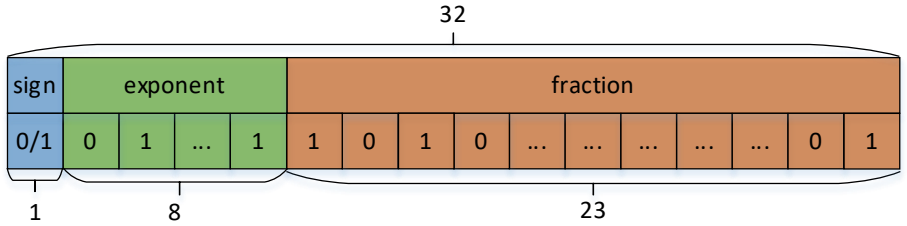
03

oneMKL: BLAS Performance

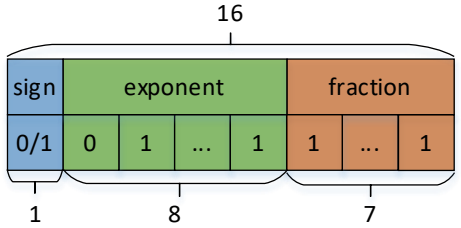
Floating Point Data Types



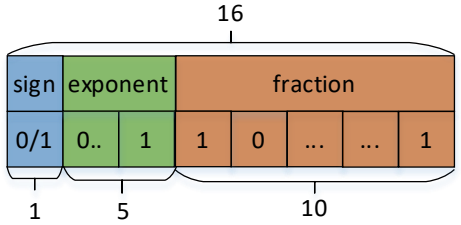
Double



Float



BFloat16

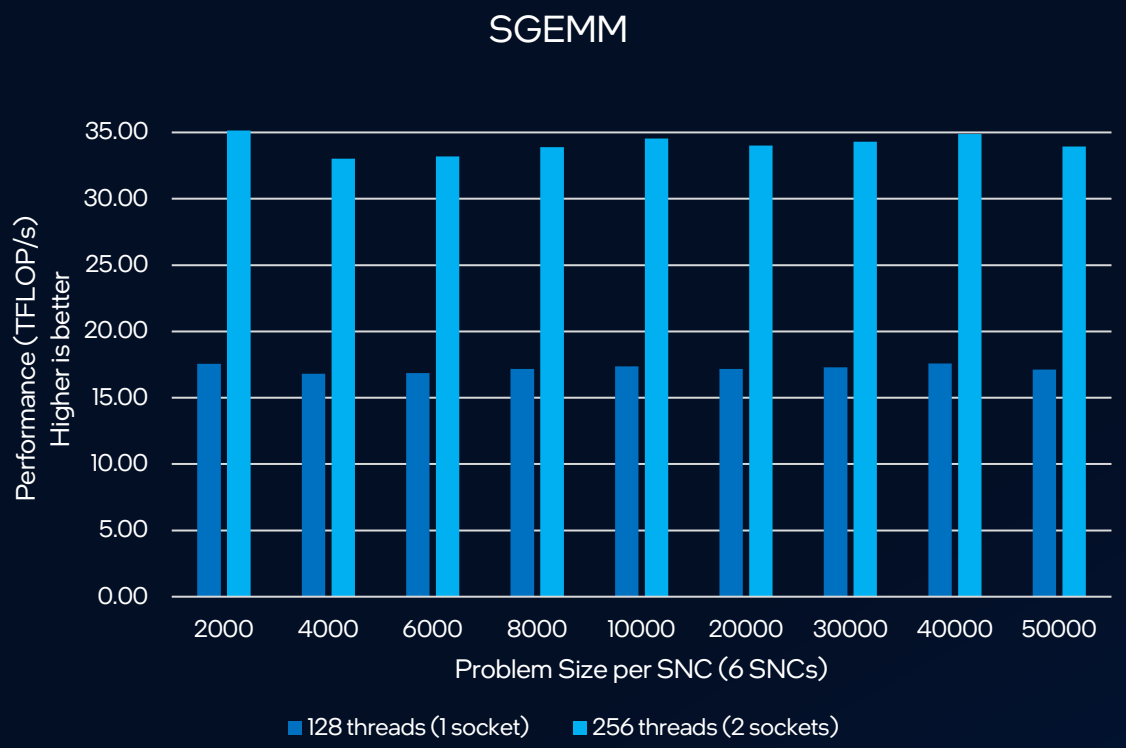


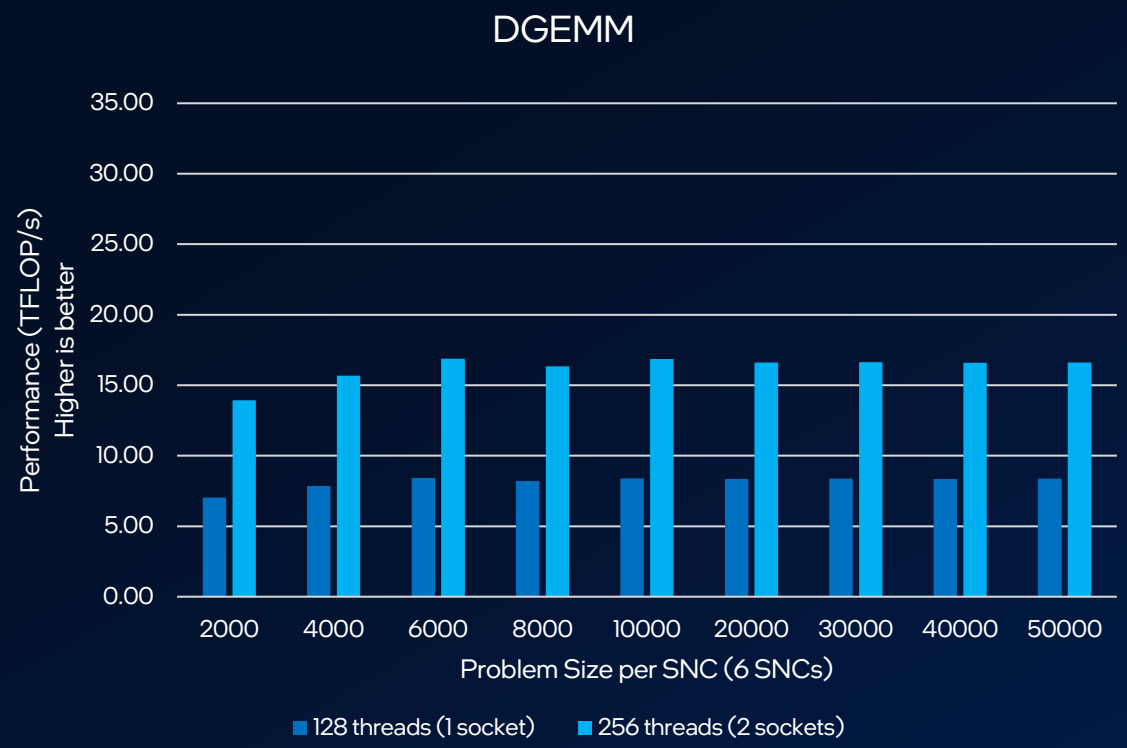
Float16

	Mantissa	Exponent	Range	Application
Double	52 bits	11 bits	$\pm 1.8 \times 10^{308}$	Scientific computing, finance
Float	23 bits	8 bits	$\pm 3.4 \times 10^{38}$	General computing
BFloat16	7 bits	8 bits	$\pm 3.4 \times 10^{38}$	ML training, AI inference
Float16	10 bits	5 bits	$\pm 6.5 \times 10^{4}$	ML inference, mobile devices

oneMKL General Matrix Multiplication (GEMM) Performance

Intel® oneAPI Math Kernel Library (oneMKL) 2026.0.0 GEMM on Intel® Xeon® 6 Processor with P-Cores





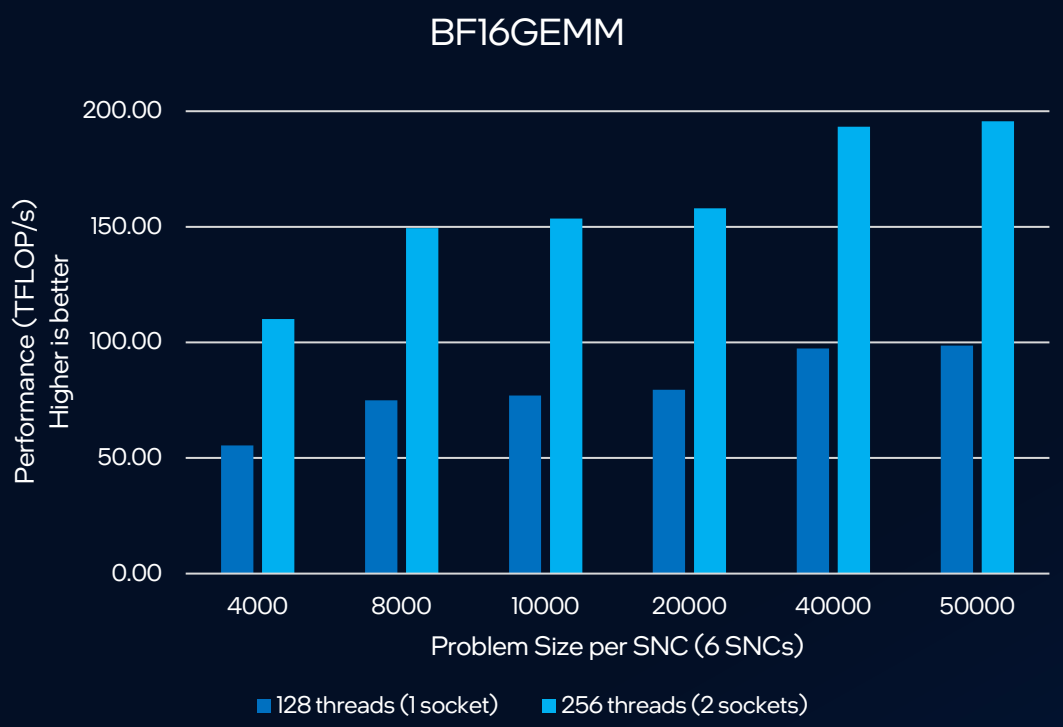
Testing Date: Performance results are based on testing by Intel as of April 24, 2026 and may not reflect all publicly available security updates.

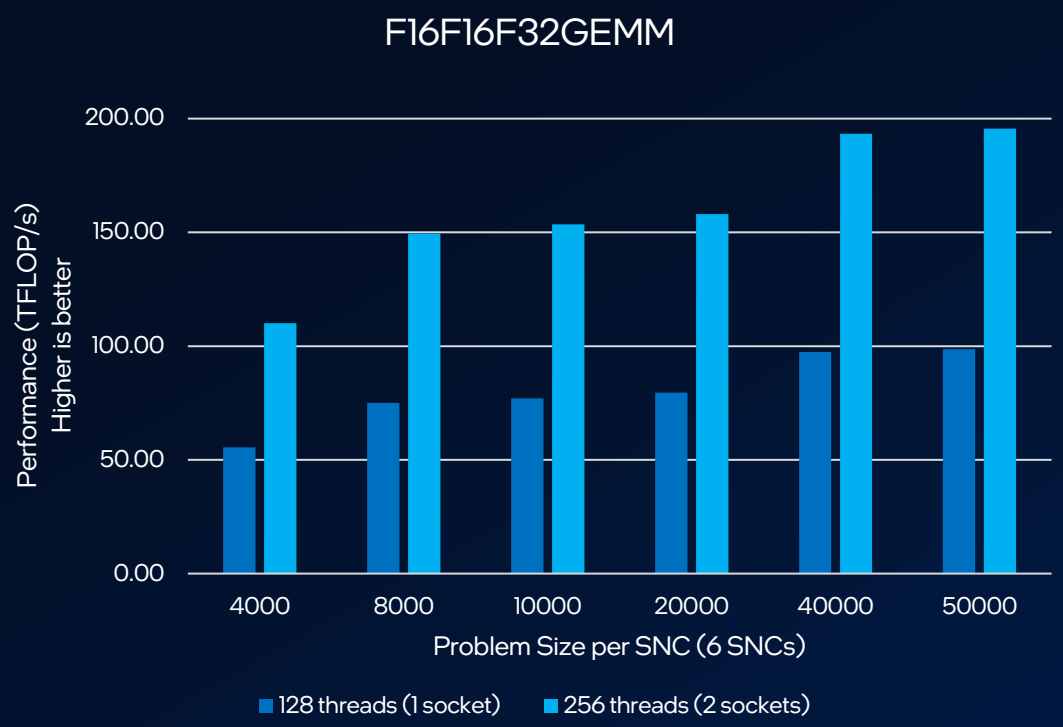
Configuration: 1-node, 2x Intel® Xeon® 6980P processor with 1536 GB (24 slots/ 64GB/ 8800) total DDR5 memory, ucode 0x1000405, HT on (used physical cores only), Turbo on, Ubuntu 24.04 LTS, 6.8.0-79-generic, 1x Micron_7450_MTFDKBG960TFR 960 GB; Intel® oneAPI Math Kernel Library (oneMKL) 2026.0.0 SGEMM & DGEMM performance: square matrices with multiple concurrent problems (1 per SNC).

Performance varies by use, configuration and other factors. Learn more on the Performance Index site.
Performance results are based on testing as of dates shown in configurations and may not reflect all publicly available updates. See backup for configuration details. No product or component can be absolutely secure.
Your costs and results may vary.
Intel technologies may require enabled hardware, software or service activation.
© Intel Corporation. Intel, the Intel logo, and other Intel marks are trademarks of Intel Corporation or its subsidiaries. Other names and brands may be claimed as the property of others.

oneMKL General Matrix Multiplication (GEMM) Performance

Intel® oneAPI Math Kernel Library (oneMKL) 2026.0.0 GEMM on Intel® Xeon® 6 Processor with P-Cores





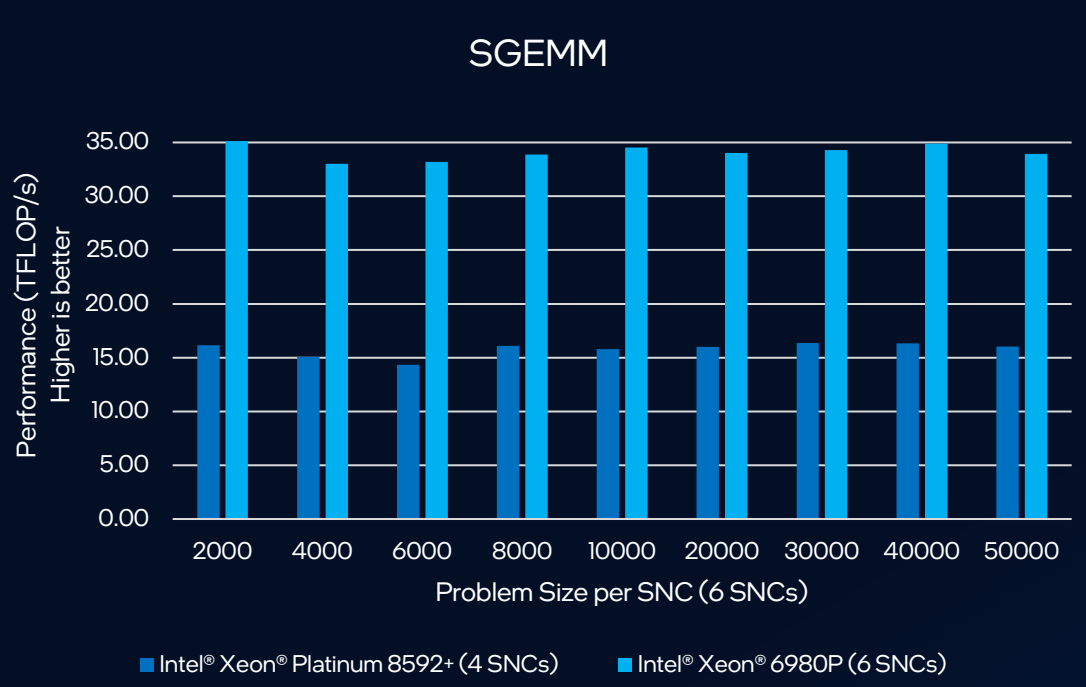
Testing Date: Performance results are based on testing by Intel as of April 24, 2026 and may not reflect all publicly available security updates.

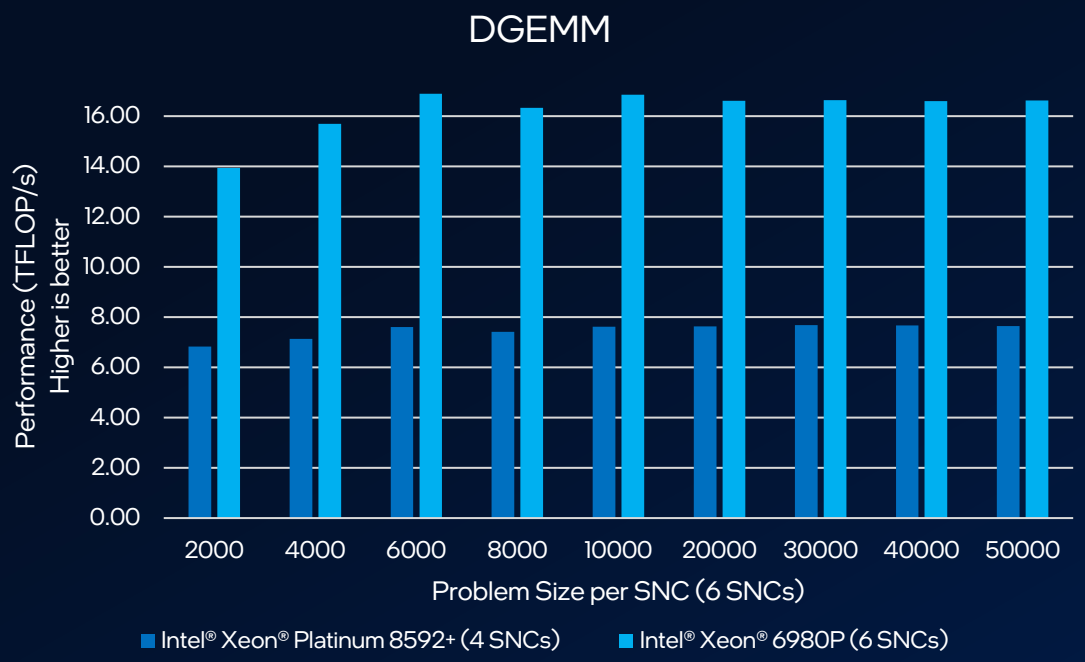
Configuration: : 1-node, 2x Intel® Xeon® 6980P processor with 1536 GB (24 slots/ 64GB/ 8800) total DDR5 memory, ucode 0x1000405, HT on (used physical cores only), Turbo on, Ubuntu 24.04 LTS, 6.8.0-79-generic, 1x Micron_7450_MTFDKBG960TFR 960 GB; Intel® oneAPI Math Kernel Library (oneMKL) 2026.0.0. BF16GEMM & F16F16F32GEMM performance: square matrices with multiple concurrent problems (1 per SNC).

Performance varies by use, configuration and other factors. Learn more on the Performance Index site.
Performance results are based on testing as of dates shown in configurations and may not reflect all publicly available updates. See backup for configuration details. No product or component can be absolutely secure.
Your costs and results may vary.
Intel technologies may require enabled hardware, software or service activation.
© Intel Corporation. Intel, the Intel logo, and other Intel marks are trademarks of Intel Corporation or its subsidiaries. Other names and brands may be claimed as the property of others.

oneMKL General Matrix Multiplication (GEMM) Performance

Intel® oneAPI Math Kernel Library (oneMKL) 2026.0.0 GEMM on Intel® Xeon® 6980P vs. Intel® Xeon® Platinum 8592+





Testing Date: Performance results are based on testing by Intel as of April 30, 2026 and may not reflect all publicly available security updates.

Configuration:

1-node, 2x Intel® Xeon® 6980P processor with 1536 GB (24 slots/ 64GB/ 8800) total DDR5 memory, ucode 0x1000405, HT on (used physical cores only), Turbo on, Ubuntu 24.04 LTS, 6.8.0-79-generic, 1x Micron_7450_MTFDKBG960TFR 960 GB; Intel® oneAPI Math Kernel Library (oneMKL) 2026.0.0.
1-node, 2x Intel® Xeon® Platinum® 8592+ processor with 512GB (16 slots/ 32GB/ 5600) total DDR5 memory, ucode 0x210002a9, HT on (used physical cores only), Turbo on, Ubuntu 24.04 LTS, 6.8.0-85-generic, 1x INTEL SSDSC2KB96 894.3G, Intel® oneAPI Math Kernel Library (oneMKL) 2026.0.0.
SGEMM & DGEMM performance: square matrices with multiple concurrent problems (1 per SNC).

Performance varies by use, configuration and other factors. Learn more on the Performance Index site.
Performance results are based on testing as of dates shown in configurations and may not reflect all publicly available updates. See backup for configuration details. No product or component can be absolutely secure.
Your costs and results may vary.
Intel technologies may require enabled hardware, software or service activation.
© Intel Corporation. Intel, the Intel logo, and other Intel marks are trademarks of Intel Corporation or its subsidiaries. Other names and brands may be claimed as the property of others.

04

oneMKL: Conditional Numerical Reproducibility

Conditional Numerical Reproducibility (CNR)

What causes a variation in results?

With floating-point numbers, the order of computation matters!
Associativity does not always hold $(a + b) + c \neq a + (b + c)$

The CNR mode ensures bitwise reproducible results from run-to-run on a fixed number of threads for a specific Intel instruction set architecture (ISA).

$2^{-63} + 1$	+	(-1)	=	2^{-63}	Infinitely precise result
$(2^{-63}+1)$	+	(-1)	=	0	Correct IEEE double precision result
$2^{-63} + (1$	+	$(-1))$	=	2^{-63}	Correct IEEE double precision result

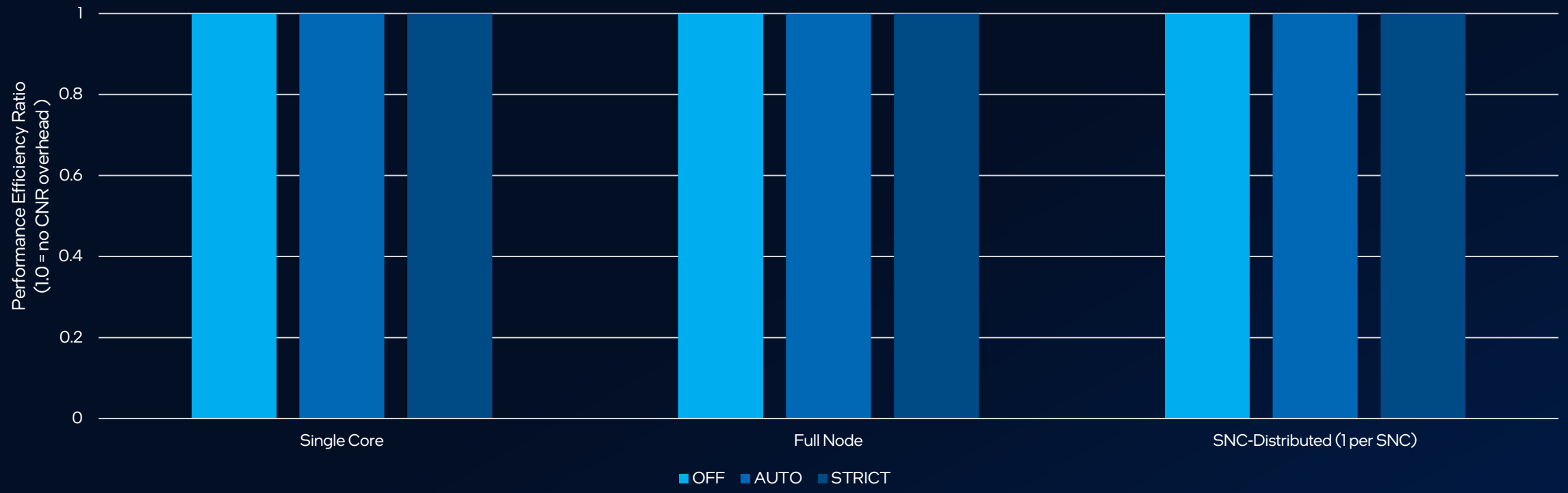
Maximum Compatibility	For consistent results …	Function Call mkl_cbwr_set(…)	Environment Variable MKL_CBWR =
	On Intel and Intel-compatible CPUs supporting SSE2 ISA	MKL_CBWR_COMPATIBLE	COMPATIBLE
	On Intel CPUs supporting SSE2 ISA	MKL_CBWR_SSE2	SSE2
	…	…	…
	On Intel CPUs supporting Intel® AVX-512 with VNNI ISA	MKL_CBWR_AVX512_E1	AVX512_E1
	From run-to-run (but not processor-to-processor)	MKL_CBWR_AUTO	AUTO

The STRICT CNR mode doesn't request the number of threads to be fixed but has more restrictions.

mkl_cbwr_set(<branch>, STRICT)or MKL_CBWR="<branch>,STRICT"

Impact of Conditional Bitwise Reproducibility on oneMKL GEMM Throughput

Intel® oneAPI Math Kernel Library (oneMKL) 2026.0.0GEMM on Intel® Xeon® 6 Processor with P-Cores



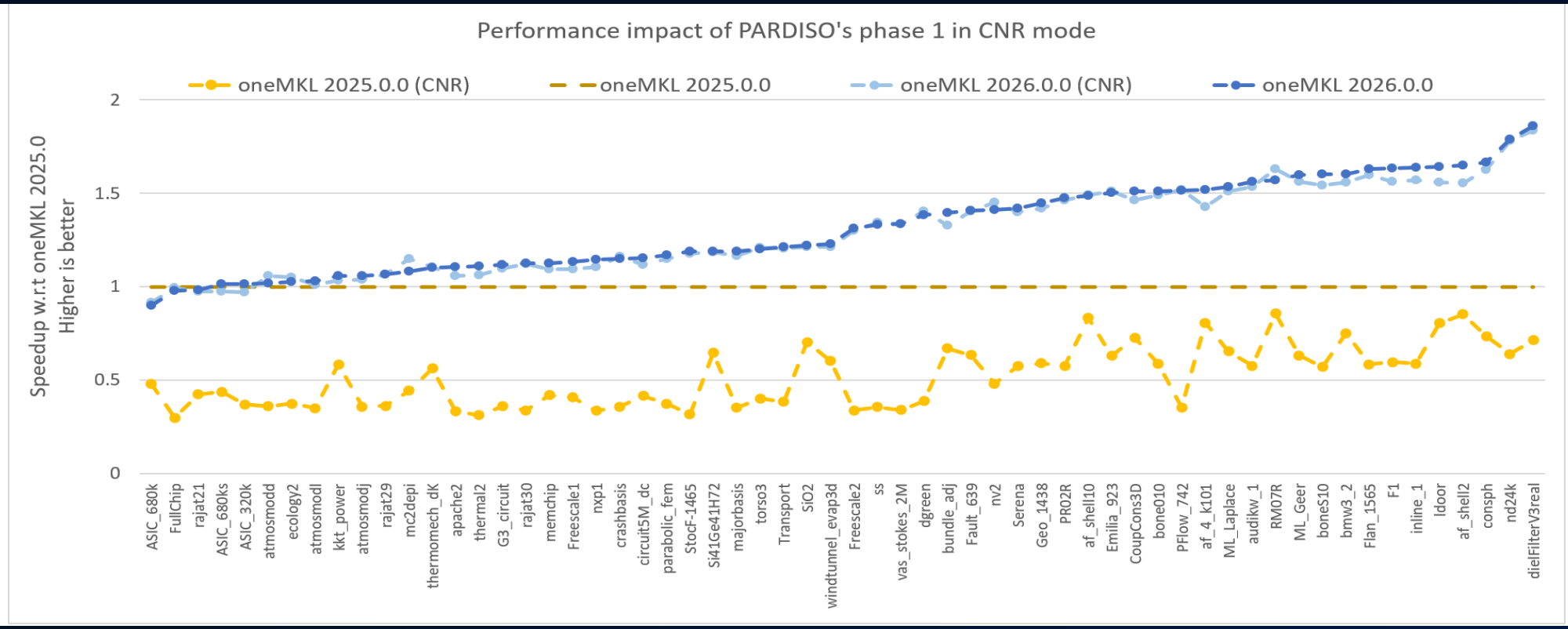
Testing Date: Performance results are based on testing by Intel as of May 10, 2026 and may not reflect all publicly available security updates.

Configuration: : 1-node, 2x Intel® Xeon® 6980P processor with 1536 GB (24 slots/ 64GB/ 8800) total DDR5 memory, ucode 0x1000405, HT on (used physical cores only), Turbo on, Ubuntu 24.04 LTS, 6.8.0-79-generic, 1x Micron_7450_MTFDKBG960TFR 960 GB; Intel® oneAPI Math Kernel Library (oneMKL) 2026.0.0. Normalized to the OFF baseline, this chart compares GEMM efficiency across Single Core, Full Socket, and SNC-Partitioned (1 per SNC) scales, with results representing average values across different problem sizes. AUTO and STRICT modes maintain peak performance parity within 5% noise.

Performance varies by use, configuration and other factors. Learn more on the Performance Index site.
Performance results are based on testing as of dates shown in configurations and may not reflect all publicly available updates. See backup for configuration details. No product or component can be absolutely secure.
Your costs and results may vary.
Intel technologies may require enabled hardware, software or service activation.
© Intel Corporation. Intel, the Intel logo, and other Intel marks are trademarks of Intel Corporation or its subsidiaries. Other names and brands may be claimed as the property of others.

Phase 1 Performance of oneMKL PARDISO

Intel® oneAPI Math Kernel Library (oneMKL) PARDISO 2026.0.0 vs 2025.0.0 on Intel® Xeon® 6 Processor with P-Cores



Testing Date: Performance results are based on testing by Intel as of July 1, 2026 and may not reflect all publicly available security updates.

Configuration: : 1-node, 2x Intel® Xeon® 6980P processor with 1536 GB (24 slots/ 64GB/ 8800) total MRDIMM memory, ucode 0x1000405, HT on (used physical cores only), Turbo on, Ubuntu 24.04.4 LTS, 6.8.0-124-generic; Intel® oneAPI Math Kernel Library (oneMKL) 2026.0.0 and 2025.0.0.

Performance varies by use, configuration and other factors. Learn more on the Performance Index site.
Performance results are based on testing as of dates shown in configurations and may not reflect all publicly available updates. See backup for configuration details. No product or component can be absolutely secure.
Your costs and results may vary.
Intel technologies may require enabled hardware, software or service activation.
© Intel Corporation. Intel, the Intel logo, and other Intel marks are trademarks of Intel Corporation or its subsidiaries. Other names and brands may be claimed as the property of others.

Summary

Intel® oneMKL delivers highly optimized math routines (Linear Algebra, FFT, RNG) engineered to maximize efficiency on Intel architectures.

Comprehensive precision support across FP64, FP32, BF16, and FP16 enables direct throughput gains tailored to specific workload needs.

Conditional Numerical Reproducibility (CNR) ensures bitwise run-to-run consistency, which is vital for validation and compliance with minimal runtime overhead.

Continued engineering investments are actively expanding CNR coverage across more Intel® oneMKL routines while driving ongoing performance improvements.

intel

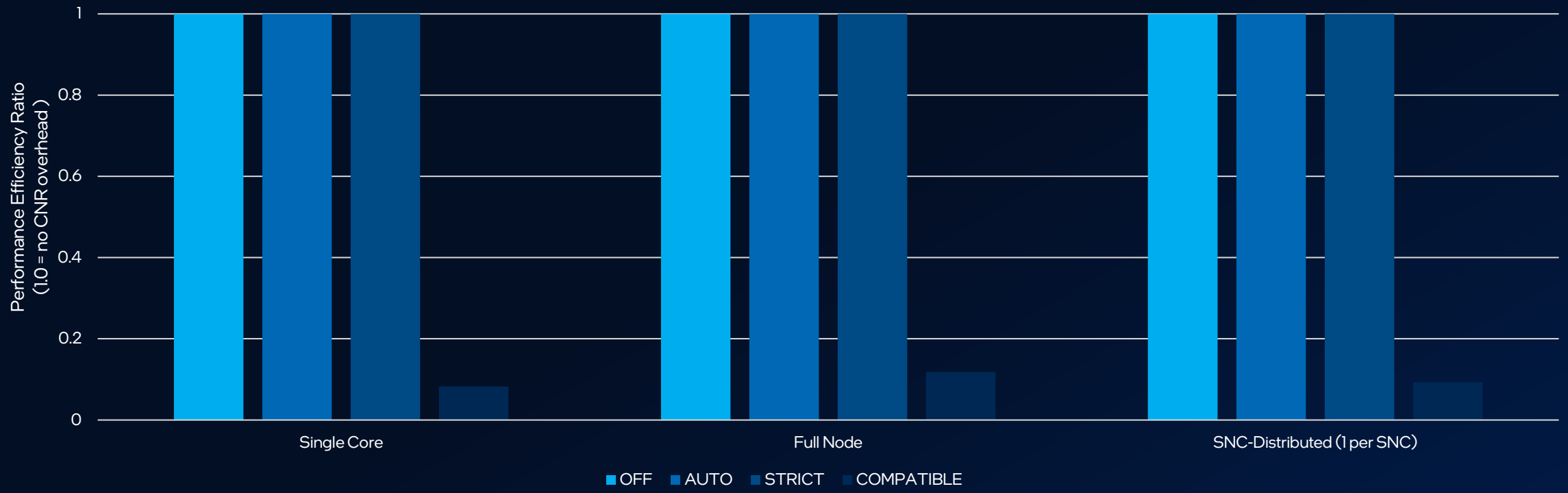
Thank you!

Pavel Dyakov
pavel.dyakov@intel.com


intel

Impact of Conditional Bitwise Reproducibility on oneMKL GEMM Throughput

Intel® oneAPI Math Kernel Library (oneMKL) 2026.0.0GEMM on Intel® Xeon® 6 Processor with P-Cores



Testing Date: Performance results are based on testing by Intel as of May 10, 2026 and may not reflect all publicly available security updates.

Configuration: : 1-node, 2x Intel® Xeon® 6980P processor with 1536 GB (24 slots/ 64GB/ 8800) total DDR5 memory, ucode 0x1000405, HT on (used physical cores only), Turbo on, Ubuntu 24.04 LTS, 6.8.0-79-generic, 1x Micron_7450_MTFDKBG960TFR 960 GB; Intel® oneAPI Math Kernel Library (oneMKL) 2026.0.0. Normalized to the OFF baseline, this chart compares GEMM efficiency across Single Core, Full Socket, and SNC-Partitioned (1 per SNC) scales, with results representing average values across different problem sizes. AUTO and STRICT modes maintain peak performance parity within 5% noise, while COMPATIBLE trades throughput for cross-generation reproducibility.

Performance varies by use, configuration and other factors. Learn more on the Performance Index site.
Performance results are based on testing as of dates shown in configurations and may not reflect all publicly available updates. See backup for configuration details. No product or component can be absolutely secure.
Your costs and results may vary.
Intel technologies may require enabled hardware, software or service activation.
© Intel Corporation. Intel, the Intel logo, and other Intel marks are trademarks of Intel Corporation or its subsidiaries. Other names and brands may be claimed as the property of others.