Solution Brief

Telecom RAN

Tietoevry Improves Intel® FlexRAN™ Software on Advantech Server

Tietoevry R&D team improves Intel FlexRAN software 5G/LTE PHY performance using Advantech SKY-8134S-11 servers featuring 4th Gen Intel® Xeon® Scalable processors with Intel® vRAN Boost[i]

Improving the performance of FlexRAN software, especially in the PHY, is an ongoing priority. Tietoevry, an Intel® Industry Solutions Builders member, has a mission of "creating purposeful technology that reinvents the world for good." As part of that mission, Tietoevry provides FlexRAN software customization consulting services so that its mobile network operator (MNO) and independent software vendor (ISV) customers get the maximum features and benefits.

Intel has created the Intel® FlexRAN™ Reference Architecture to give ISVs and MNOs a head-start on developing 4G / LTE open virtualized radio access networks (O-RANs), specifically distributed units (DUs) / centralized units (CUs). FlexRAN software is comprised of modular, virtualized control functions with well-defined interfaces that enable vendors to virtualize DU and CU functions and host them on servers based on Intel® Xeon® Scalable processors.

In this paper, the Tietoevry R&D team, using vRAN-optimized servers from Advantech, sought to improve the performance of a FlexRAN™ technology matrix inversion algorithm by vectorizing the algorithm using the new Intel® Advanced Vector Extensions 512 (Intel® AVX-512) FP16 instruction set architecture (ISA).

Tietoevry Targets 5G RANs

FlexRAN software / vRAN customization is part of the suite of services provided by the Tietoevry Create division, which is the company's leading digital accelerator for innovation and sustainable value creation.

The company uses its combination of real-time and data-driven software development expertise combined with its worldwide presence and operations to deliver services and software with agility and speed. This know-how enables Tietoevry to help companies develop networks that exceed their RAN performance goals.

Specific benefits of RAN development services include:

- Cost-efficient network operation: Achieve maximum network automation to minimize network operational costs. Tietoevry designs software that takes into account all aspects of a network service, from device to operation to maintenance.
- Shared multi-purpose platform: this helps Tietoevry address a variety of telco workloads, evolving also to other applications. O-RAN uses this platform to enable mix-and-match of hardware and software from multiple vendors.
- Increased flexibility and scalability: By reducing bottlenecks from the delay / CPU load perspective which means fewer cores are needed for L1 processing.

- Server efficiencies: Cloud-native O-RAN deployments enable edge applications to be run on the same COTS server.
- Balanced priorities: Tietoevry's solution enables customers to achieve balance between real-time software performance and the desired energy efficiency for processing PHY and layer 2 workloads, fulfilling KPIs for a given network product.

Tietoevry FlexRAN Software Optimization Project

FlexRAN software uses vectorization, which is the process of converting an algorithm from operating on a single value at a time to operating on a set of values (vector) at one time. This is achieved by utilizing the Single Instruction Multiple Data (SIMD) intrinsics available in the architecture instruction set in the Intel® Intrinsics Guide (see https://www.intel.com/content/www/us/en/docs/intrinsics-guide/index.html). FlexRAN technology uses these SIMD instructions in the 5G PHY algorithms to deliver low latency and optimized implementation.

The 4th Gen Intel® Xeon® Scalable Processor brings a new FP16 ISA for Intel AVX-512. The new ISA supports a wide range of general-purpose numeric operations for 16-bit half-precision IEEE-754 floating-point and complements the existing 32-bit and 64-bit floating-point instructions already available in servers based on Intel® Xeon® Scalable processors.

The Intel AVX-512 ISA also provides complex-valued native hardware support. The new ISA is ideal for numeric operations where reduced precision can be used, such as signal and media processing. For example, wireless signal processing operations such as beamforming, precoding, and minimum mean squared error (MMSE) perform well with this ISA.

This instruction set also works well with traditional signal processing, for example, with real or complex-valued fast Fourier transforms (FFTs). The advantage of using reduced precision in these cases is that because fewer bits are processed for each element, the overall compute throughput can be increased, allowing precision and speed to be traded against each other.

Building on the vectorization of calculations introduced to the Intel FlexRAN software codebase by Intel, the Tietoevry R&D team updated an implementation of a matrix inversion algorithm with the latest Intel AVX-512 FP16 vector instructions and based on its analysis, sought to streamline certain aspects of data handling, including calculation module input and output data shuffling.

Advantech SKY-8134S-11 Compact Telecom Server

The FlexRAN software optimization project was conducted on the Advantech SKY-8134S-11, a carrier-grade (NEBS Level 3), ultra short depth (300 mm) 1U-high server designed to optimize performance, density and total cost of ownership (TCO) for advanced telecom applications such as 5G, vRAN or edge computing.

The server is built with a LAN-on-motherboard (LOM) design that features up to 16 25GbE ports with onboard support for high-precision time synchronization protocols including IEEE 1588 PTP, SyncE and integrated global navigation satellite system (GNSS). The server also provides an extra PCIe 5.0 or OCP 3.0 slot for additional acceleration or I/O.

The unique design of the SKY-8134S-11 provides the highest density in the market[1] for a 1U, 300 mm depth vRAN server which brings greater scalability and TCO benefits to network operators deploying vRAN. Service provider and enterprise customers can also cost-optimize the server further, with configurations starting at 8 25GbE ports, to drive the adoption of private 5G networks.

The performance of the SKY-8134S-11 (see Figure 1) comes from the 4th Gen Intel® Xeon® Scalable processors with Intel® vRAN Boost, which offer energy-efficient high performance by combining up to 32 high-performance processor cores with built-in accelerators.

The SKY-8134S-11 offers environmental hardening for use in a network edge application. This includes a wide operating temperature range (-40 to +65C) and protections for environmental shock, vibration and dust conditions. The SKY-8134S-11 ultra-compact form factor and front accessible network and power interfaces are ideal for network edge sites where limited space is available.

The server can also be deployed in IP65 pole mount, roadside unit or street side cabinets. Its high reliability design, including redundant DC power supplies, the ability to withstand single fan failures, redundant BIOS and firmware images with failsafe remote updates, and hot swappable field-replaceable units (FRU) makes the SKY-8134S-11 the platform of choice for network edge applications that require virtually zero downtime.

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

Figure 1. Advantech SKY-8134S-11 1U edge server for O-RAN and multi-access edge computing (MEC).

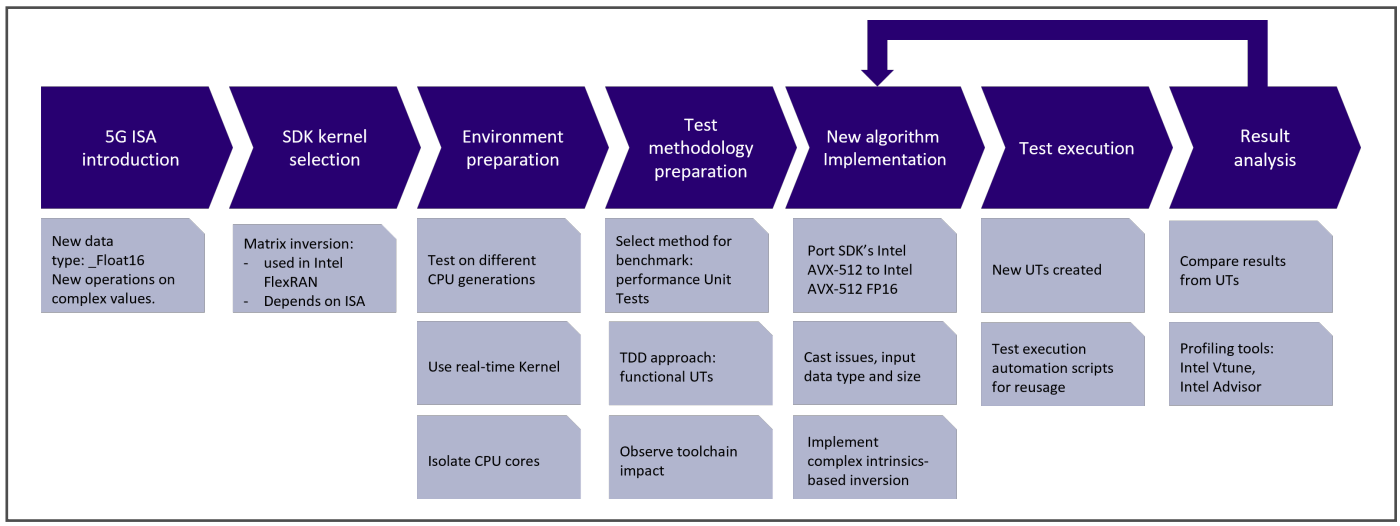

Figure 2. Tietoevry's iterative process used to isolate and optimize PHY bottlenecks.

Optimizing FlexRAN Software Execution Time

The Tietoevry R&D team designed a server configuration that sought to improve the time of execution in the PHY component of the FlexRAN technology by analyzing it, measuring its performance, finding bottlenecks, and then finding a way to optimize these performance barriers in order to squeeze additional performance from the hardware platform on which it is running.

Once developed, the testing procedure was applied across three generations of Intel® architecture CPUs.

Figure 2 shows the R&D team's process for isolating the PHY system for testing:

- First, the team analyzed the 5G ISA included within the 4th Gen Intel Xeon Scalable processor.
- Next, the team selected the PHY software component that both consumed a high number of CPU cycles and that could be vectorized efficiently--in this case, the block matrix inversion.
- The team then prepared several test setups and test procedures that varied only to the extent required to accommodate the hardware capabilities of each processor. As a testing framework, the team selected existing performance unit tests (UTs) within the Intel FlexRAN software codebase.
- The team next prepared a new software implementation utilizing Intel AVX-512 for vRAN, memory layout optimization, and testing automation.
- The team then tested the new and existing implementations and compared their performance, analyzing the results both visually and via specialized profiling software.
- Based on results, the team then fine-tuned the software and performed iterative testing and software adjustments until performance improvements were maximized.

Performance Results and Intel® VTune™ Profiler Analysis

Figures 3 and 4 summarize the results of the 4x4 Hermitian matrix inversion collected by running performance UTs. Note the build environment key located below Figure 4, which summarizes the toolchains used and the Intel FlexRAN software build flags.

Where environment key:

A. Intel® oneAPI Base Toolkit 2022.1.2 with default Intel FlexRAN software compilation flag (Intel AVX-512)

B. Intel oneAPI Base Toolkit 2022.1.2 with platform Intel FlexRAN software compilation flag (SPR for 4th Gen Intel Xeon Scalable processors, SNC for 3rd Gen Intel® Xeon® Scalable processors)

C. Intel oneAPI Base Toolkit 2023.1.0 with default Intel FlexRAN software compilation flag (Intel AVX-512)

D. Intel oneAPI Base Toolkit 2023.1.0 with platform Intel FlexRAN software compilation flag (SPR for 4th Gen Intel Xeon Scalable processors, SNC for 3rd Gen Intel Xeon Scalable processors)

The FlexRAN software compilation flag allowed the code to be compiled in variants that are platform-optimized, such as using the 5G ISA that is available with the 4th Gen Intel Xeon Scalable processors. Implementations based on Intel AVX-512 have a vector length of 512 bits. Results from bare C-code direct calculations were also listed to highlight the performance gain acquired with vectorization.

Intel® VTune™ Profiler

Intel® VTune™ Profiler optimizes application performance, system performance, and system configuration for high-performance computing, cloud, IoT, media, storage, and more. The software can tune an entire application's performance.

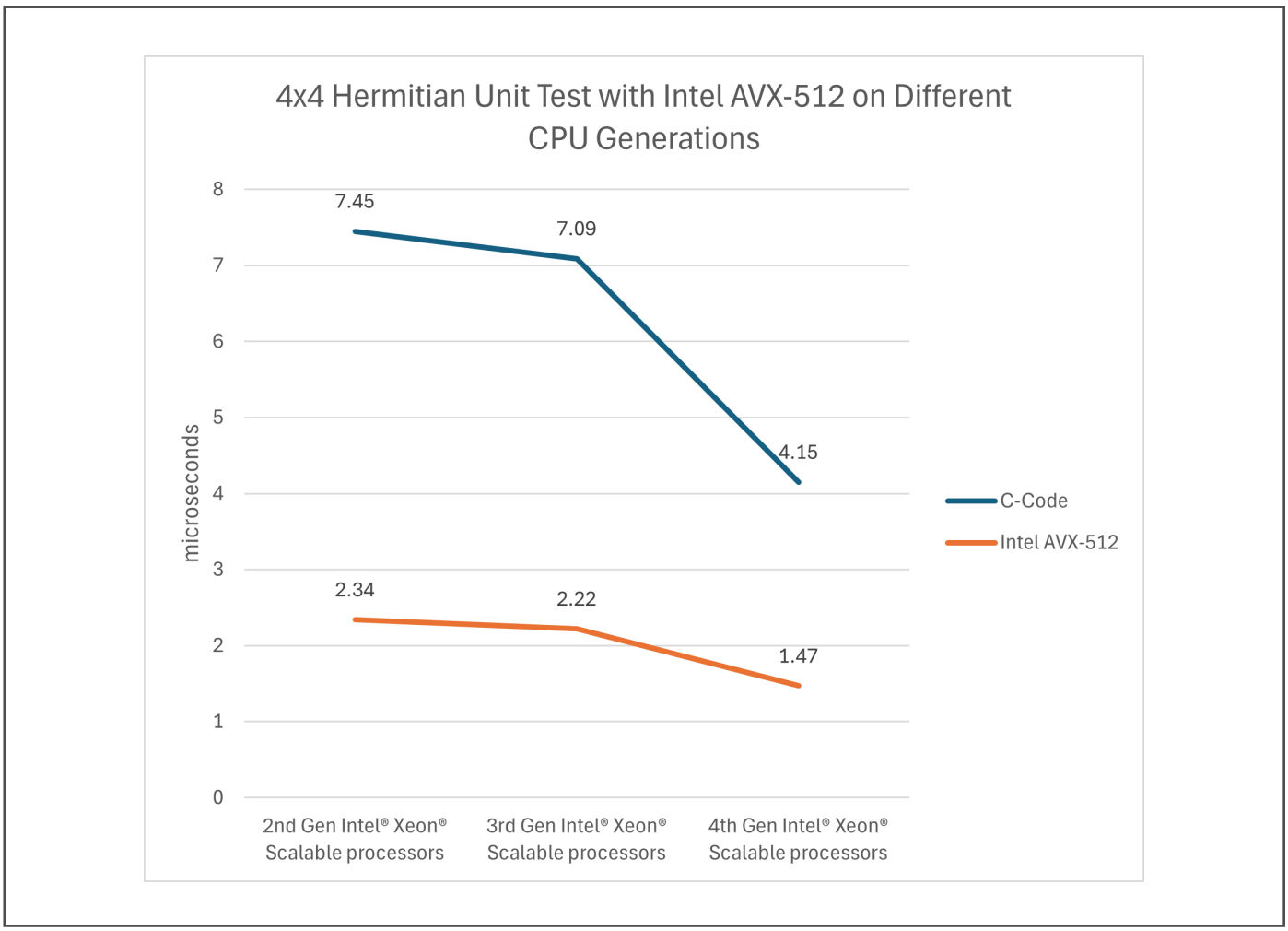

Figure 3. Intel AVX-512 implementation on different generations of Intel architecture-based CPUs. All compiled for Intel AVX-512 complier flag (lower is better).

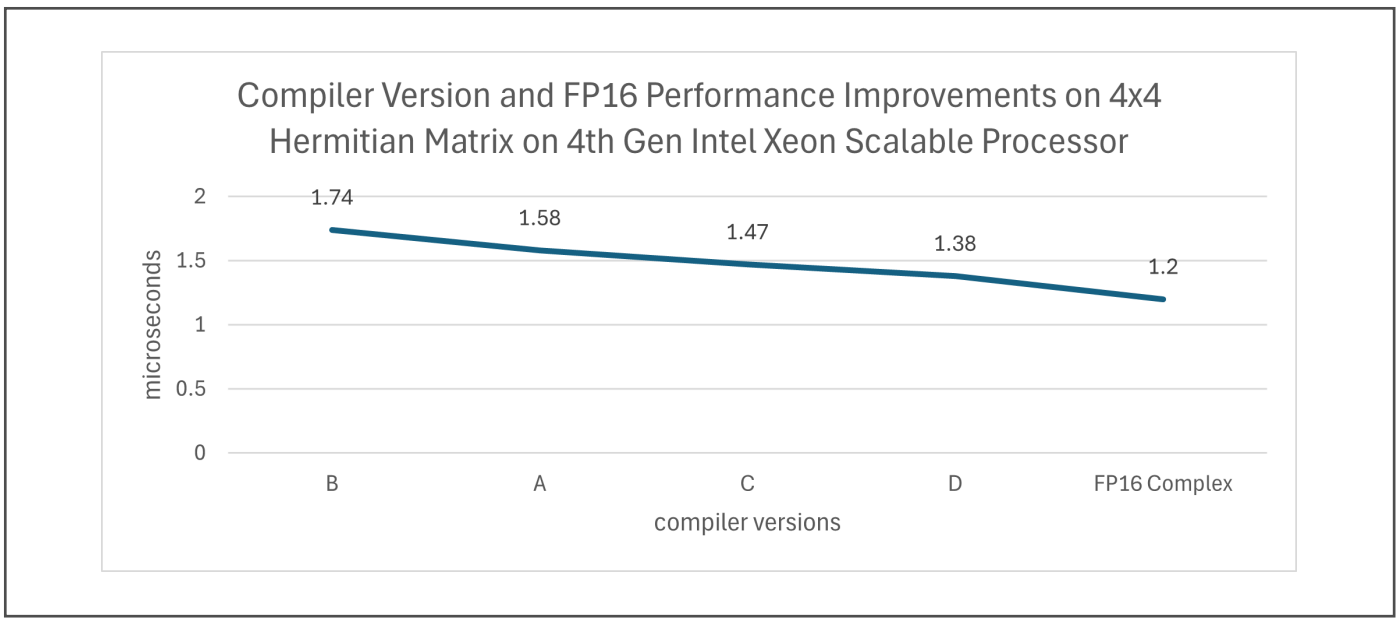

Figure 4. Impact of compiler version unit test performance (lower is better).

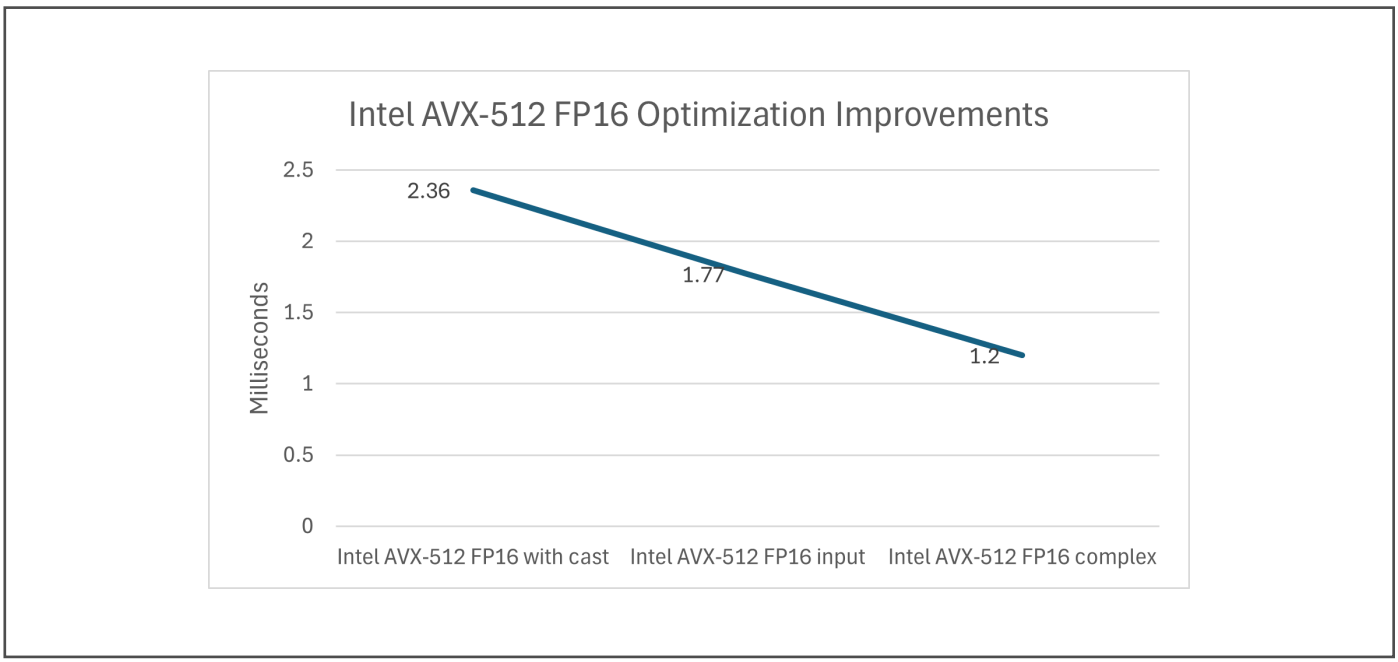

Figure 5. Performance improvements per optimization phase (lower is better).

The Tietoevry R&D team took an iterative and pragmatic approach to optimizing the function for Intel AVX-512 FP16, resulting in three main phases of optimizations as shown in Figure 5.

- Intel AVX-512 FP16 with Cast –input data was 32-bit and was casted to float16 for the duration of calculations, and afterwards it was converted back to 32-bit.
- Intel AVX-512 FP16 – data was fed into calculating function in the form of 16-bit floats, and the result was also accepted as float16.
- Intel AVX-512 FP16 Complex – special 5G ISA instructions were used, and specialized implementation was created. In this scenario, data was fed in as a vector of float16 complex pairs. This was tested only on Intel oneAPI Base Toolkit 2023.1.0 to enable full 5G ISA intrinsics support.

Tietoevry Insights Through Developer Tools

On the 4th Gen Intel Xeon Scalable processor, additional profiling was done using Intel VTune Profiler. The target of the tests was to evaluate each function of the code in terms of performance and check how much time is consumed by each.

For this purpose, the UT was prolonged to allow the loop count to be increased from 30 to 1 million. The code was tested in a build environment that includes Intel oneAPI Base Toolkit 2023.1.0 with platform FlexRAN compilation flag set for SPR for 4th Gen Intel Xeon Scalable processors. Then, the Intel VTune Profiler was run on five different implementations, with the results presented in Figure 6.

Intel VTune was used to obtain data from 4th Gen Intel Xeon Scalable processors. Evaluation was performed on Intel AVX-512 and all FP16-based variants.

Task / ISA	Intel AVX-512	Intel AVX-512 FP16 with cast	Intel AVX-512 FP16 input data	Intel AVX-512 FP16 complex intrinsics
Data load convert_float_to_vector convert_input_and_split	19.10 s (100%)	65.16 s (341%)	39.54 s (207%)	12.99 s (68%)
Calculation matrix_inverse_4x4_hermitian	8.62 s (100%)	4.23 s (49%)	4.67 s (54%)	9.60 s (111%)
Data store convert_vector_to_float convert_matrices_to_output_hermitian	24.62 s (100%)	24.23 s (98%)	24.63 s (100%)	25.38 s (103%)
Total time	52.340 s (100%)	93.62 s (179%)	68.84 s (132%)	47.97 s (92%)

Note:
Below sum of time spent on given task (load, calculate, store) additional percentage metric given. Base for percent is the time that same job took for Intel AVX-512 inversion variant. Note the drop of calculation time when using longer vectors and related significant increase of loading time.

Info:
- Time is expressed in seconds: Intel Vtune Profiler hotspot analysis sums-up the time which evaluated program spent in each of the function or its children (given here is sum of time in function and in its children). I.e.: while performing 40 x 1 000 000 x 64 inversions with Intel AVX-512, total time was 52.34 s.

Figure 6. Intel VTune results using 4x4 Hermitian matrix.

The results should be understood in terms of total time taken by each action: load, store and calculation, showcasing the performance cost of each operation.

Although total execution time was best for Intel AVX-512 FP16 complex intrinsics, as expected, it is worth noting that this is not due to calculation time. Several hypotheses were tested to explain this result, with the most plausible explanation being the lack of exploitation of the Hermitian matrix properties.

Conclusion

The optimizations that Tietoevry performed – including performance unit tests of new code with existing interface and data-oriented design, Intel VTune Profiler and Intel FlexRAN software timer mode tests - led to one important observation that efficient implementation of vectorization relied equally on the design of the software and properly utilized features of the hardware platform. The results also highlight the value of using the latest version of Intel oneAPI toolkit, as compiler improvements can improve performance.

The optimal solution is one that fully utilizes the FlexRAN software capabilities. It requires proper preparation with a constructed benchmarking methodology that not only implements efficient Intel AVX-512 ISA-based solutions, but it also uncovers bottlenecks with memory handling and other properties of the ISA instructions. The tools used in these tests provide a complete picture of the problem, allowing for a better solution.

While the cycle count performance gains from FP16 are clear, care must be taken to ensure the algorithm can be appropriately optimized with FP16. Some algorithms in the Intel FlexRAN software SDK without algorithm adoption will be impacted by loss of precision from FP32 to FP16.

These tests show that memory layout and memory handling are proven to be major issues, impeding performance. Minimizing the requirement of storing/loading data into/from vectors, by maintaining codebase with vector-friendly data containers (e.g. properly organized and indexed arrays) can save a significant portion of overall execution time.

Learn More

Tietoevry Create

Advantech SKY-8134S-11

Intel® Industry Solution Builders

4th Gen Intel® Xeon® Scalable Processors with Intel® vRAN Boost

FlexRAN™ Reference Architecture for Wireless Access

Intel® Advanced Vector Extensions 512 (Intel® AVX-512)

Intel® VTune™ Profiler

Intel® oneAPI Base Toolkit

[i] Config 1 SUT: 1-node, 1x Intel® Xeon® Gold 5423 processor with 20 cores and 40 threads. Total DDR4 memory was 125 GiB; microcode was ISA AVX512, ISA AVX512-FP16. Intel® Hyper-Threading Technology - enabled; and Intel® Turbo Boost Technology - enabled. Software: OS was CentOS; kernel was 3.10.0-1160.83.1.rt56.1228el7.x86_64. Benchmark/workload software: FlexRAN 23.03; Compiler was GCC 9.4.0; Libraries were Intel oneAPI 2022.1.2 and 2023.1.0. Other software: Intel FlexRAN 23.03. Test conducted by Tietoevry on April 04,2023.

Config 2 SUT: 1-node, 1x Intel® Xeon® Gold 6330 processor with 28 cores and 56 threads. Total DDR4 memory was 251 GiB; microcode was ISA AVX512. Intel® Hyper-Threading Technology - enabled; and Intel® Turbo Boost Technology - enabled. Software: OS was CentOS. Benchmark/workload software: FlexRAN 23.03; Compiler was GCC 9.4.0; Libraries were Intel oneAPI 2022.1.2 and 2023.1.0. Other software: Intel FlexRAN 23.03. Test conducted by Tietoevry on April 04,2023.

Notices & Disclaimers

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 details. No product or component can be absolutely secure.

Intel optimizations, for Intel compilers or other products, may not optimize to the same degree for non-Intel products.

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.

See our complete legal Notices and Disclaimers.

Intel is committed to respecting human rights and avoiding causing or contributing to adverse impacts on human rights. See Intel's Global Human Rights Principles. Intel's products and software are intended only to be used in applications that do not cause or contribute to adverse impacts on human rights.

Intel® Turbo Boost Technology requires a PC with a processor with Intel Turbo Boost Technology capability. Intel Turbo Boost Technology performance varies depending on hardware, software and overall system configuration. Check with your PC manufacturer on whether your system delivers Intel Turbo Boost Technology. For more information, see http://www.intel.com/technology/turboboost

0325/CG/DJA/PDF Please Recycle 364690-001US