White Paper

5th Gen Intel® Xeon® Scalable Processors
Intel® Advanced Matrix Extensions (Intel® AMX)

5th Gen Intel® Xeon® Scalable Processors Empower Winning Health to Expedite the Implementation of Large Language Models in Traditional Medical Scenarios

卫宁健康 WINNING HEALTH

"The innovation and widespread application of LLMs represent an important trend in the development of smart hospitals. However, the lean operations in hospitals underscores the pressing need to better unleash the potential of applying LLMs in smart healthcare services with lower deployment costs. Through our collaboration with Intel, we have found a CPU-based LLM inference solution that not only meets the performance requirements but also offers cost advantages, helping accelerate the deployment of LLMs in hospitals, while providing intelligent knowledge services across various hospital scenarios."

—Zhao Daping
Vice President and CTO,
Winning Health

"The combination of LLMs + healthcare opens up endless possibilities for the healthcare industry. Yet, the obstacles standing between aspirations and applications aren't just technical, but also the steep cost of deploying LLMs. As the latest-generation processors tailored for the AI era, the 5th Gen Intel® Xeon® Scalable processors offer more than powerful AI performance, but also cost-effectiveness and exceptional flexibility in deployment, which means they can better meet the demands of LLMs applied in medical scenarios and expedite the development of smart hospitals."

—Eric Tang
General Manager, Software Technology Solution Group, Intel China

Overview

In the current landscape of smart hospital advancement, it is widely recognized that large language models (LLMs), as a groundbreaking technology, have significant potential to be applied in medical settings. Applications powered by LLMs, such as medical literature analysis, healthcare Q&A, medical report generation, AI-assisted imaging diagnosis, pathology analysis, chronic disease monitoring and management, and medical record sorting, all contribute to leveling up the efficiency and quality of medical services, reducing costs for medical institutions in manpower and other resources, while improving the overall experience for patients. One major obstacle in the wider use of LLMs in healthcare institutions, however, is the lack of high-performance and cost-effective computing platforms. Take model inference: the sheer complexity and scale of LLMs far exceed those of common AI applications, posing a challenge for traditional computing platforms to adequately meet their demands.

Building on its leading medical LLM WiNGPT, Winning Health has introduced the WiNGPT solution based on 5th Gen Intel® Xeon® Scalable processors. The solution effectively leverages the built-in accelerators including Intel® Advanced Matrix Extensions (Intel® AMX) in these processors for model inference. Through collaboration with Intel in areas like graph optimization and weight-only quantization, the inference performance has been increased by over 3 times compared with the platform based on the 3rd Gen Intel® Xeon® Scalable processors[1]. The enhancement meets the performance demand for scenarios like automated medical report generation, accelerating the adoption of LLM applications in healthcare institutions.

[1] Data from Winning Health's internal test results as of November 2023. Test configurations—Baseline: 2S Intel® Xeon® Platinum 8380 processor @ 2.30 GHz, 1024 GB total memory (16x64 GB DDR4 3200 MT/s), 745.2 GB SSD, Ubuntu 22.04.3 LTS; New: 2S Intel® Xeon® Platinum 8592+ processor @ 1.90 GHz, 512 GB total memory (16x32 GB DDR5 5600 MT/s), 1.1 TB SSD, Ubuntu 22.04.3 LTS. Intel does not control or audit third-party data. You should consult other sources to evaluate accuracy. Performance varies by use, configuration and other factors. Learn more at www.Intel.com/PerformanceIndex.

Challenge: The compute conundrum in medical LLM inference

The extensive use of LLMs in various verticals such as healthcare is considered a milestone for the real-world application of this technology. Healthcare institutions are stepping up their investments and have made considerable progress in LLMs for medical diagnostics, services, and management. Research forecasts that 2023 to 2027 will witness a surge in the adoption of LLMs in the healthcare industry, with the market size expected to exceed 7 billion yuan by 2027[2].

LLMs are typical compute-intensive applications, and their training, fine-tuning, and inference all rely on substantial computing resources, resulting in huge computing costs. Among these, model inference stands out as a crucial stage in LLM deployment. When creating model inference solutions, healthcare institutions are commonly confronted with the following challenges:

- The scenarios are complex, with a high demand for real-time accuracy. This requires the computing platform to be powerful enough in inference. Additionally, given the stringent security requirements for medical data, healthcare institutions usually prefer the platform to be deployed locally rather than on the cloud.
- Hardware upgrade does not happen frequently, while LLM upgrades may require GPUs to be upgraded accordingly. As a result, updated models may not be able to work on legacy hardware.
- The hardware requirements for the inference of Transformer-based LLMs have seen a substantial rise than in the past. Both memory and time complexity scale exponentially with the length of the input sequence, making it difficult for previous computing resources to be fully utilized. Consequently, hardware utilization has yet to reach its optimal level.
- From a cost perspective, deploying servers dedicated to model inference would incur higher costs and such servers would be limited in usage. Given this, many healthcare institutions prefer to use CPU-based server platforms for inference to cut hardware expenses with the flexibility to support various workloads.

Solution: WiNGPT based on 5th Gen Intel® Xeon® Scalable processors

WiNGPT by Winning Health is a LLM specifically designed for the healthcare sector. Built on the general-purpose LLM, WiNGPT integrates high-quality medical data, and is optimized and customized for medical scenarios, allowing it to provide intelligent knowledge services across different healthcare scenarios. WiNGPT is characterized by the following three distinctive aspects:

Fine-tuned and specialized

WiNGPT is trained and fine-tuned for medical scenarios and on high-quality data, delivering exceptional data accuracy that meets diverse business requirements.

Low cost

Via algorithm optimization, the deployment based on CPU is already tested to have gained the generation efficiency close to that of GPU.

Support customized private deployment

Private deployment ensures that medical data stays within healthcare institutions, preventing data leaks while offering better system stability and reliability. Moreover, it allows for customized options for organizations of varying needs to accommodate different budget plans.

To accelerate WiNGPT's inference speed, Winning Health has partnered with Intel by opting for the 5th Gen Intel Xeon Scalable processors. These processors offer enhanced reliability and energy efficiency, delivering significant performance gains per watt across various workloads and exceptional performance in AI, data center, network and HPC, all-while maintaining a lower total cost of ownership (TCO). Compared with the previous generation, the 5th Gen Intel Xeon Scalable processors offer increased computing power and faster memory within the same range of power consumption. Additionally, they are compatible with last generation's software and platforms, significantly saving testing and validation efforts when deploying new systems.

The 5th Gen Intel Xeon Scalable processors are built-in with several AI-optimized features including Intel AMX, taking AI performance to the next level. Intel AMX adopts a new instruction set and circuit design, significantly boosting the instructions per cycle (IPC) for AI applications by enabling matrix operations. The advancement leads to a notable performance improvement for both training and inference in AI workloads.

[2] EO Intelligence, Large Language Models in Healthcare Industry Research Report 2023: https://www.iyiou.com/research/202312151293

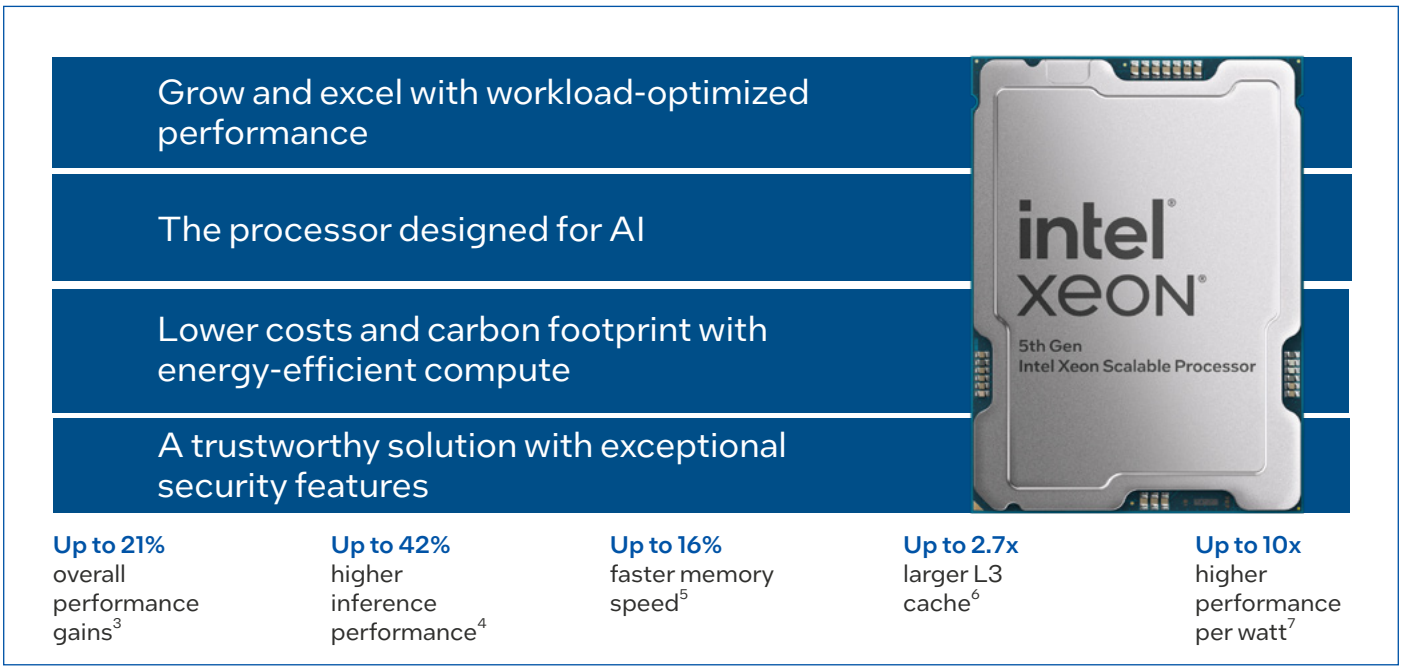

Figure 1. 5th Gen Intel® Xeon® Scalable processors have better performance

In addition to the 5th Gen Intel Xeon Scalable processors, Winning Health and Intel are also exploring ways to address the memory access bottleneck in LLM inference on the current hardware platform. LLMs are usually considered memory-bound due to their extensive parameter size, which often requires billions or even tens of billions of model weights to be loaded into memory for computing. When computing is underway, vast data needs to be stored in memory temporarily and read for subsequent computing. The speed of memory access—instead of the computing power—has thus become the primary hindrance dragging down inference speed.

Winning Health and Intel have taken the following measures to optimize memory access and beyond:

- **Graph optimization**

  Graph optimization refers to the process of merging multiple operators to reduce the overhead of operator/core calls. Combining several operators into a single operation saves the consumption of memory resources once required for the read-ins and read-outs of different operators, thus improving the performance. In these processes, Winning Health has used Intel® Extension for PyTorch to optimize the algorithms, resulting in effective performance boost. With Intel® Extension for PyTorch, Intel uses acceleration libraries such as oneDNN and oneCCL in the form of intel-extension-for-pytorch as a plug-in to improve PyTorch performance on servers based on Intel Xeon Scalable processors and Intel® Iris® X^{e} graphics.

[3] Average performance gain as measured by the geomean of SPEC CPU rate, STREAM Triad, and LINPACK compared to 4th Gen Intel® Xeon® processor. See [G1] at intel.com/processorclaims: 5th Gen Intel® Xeon® Scalable processors. Results may vary.

[4] 1.19x to 1.42x performance gains for ResNet50v1.5, BERT-Large, SSD-ResNet34, RNN-T (BF16 only), Resnext101 32x16d, MaskRCNN (BF16 only), DistilBERT compared to 4th Gen Intel® Xeon® processor. See [A15-A16] at intel.com/processorclaims: 5th Gen Intel® Xeon® Scalable processors. Results may vary.

[5] See [G12] at intel.com/processorclaims: 5th Gen Intel® Xeon® Scalable processors. Results may vary.

[6] See [G11] at intel.com/processorclaims: 5th Gen Intel® Xeon® Scalable processors. Results may vary.

[7] 1.46x to 10.6x performance per watt gains as measured by AI, data, and network workloads using built-in accelerators. See [A19-A25], [D1], [D2], [D5] and [N16] at intel.com/processorclaims: 5th Gen Intel® Xeon® Scalable processors. Results may vary.

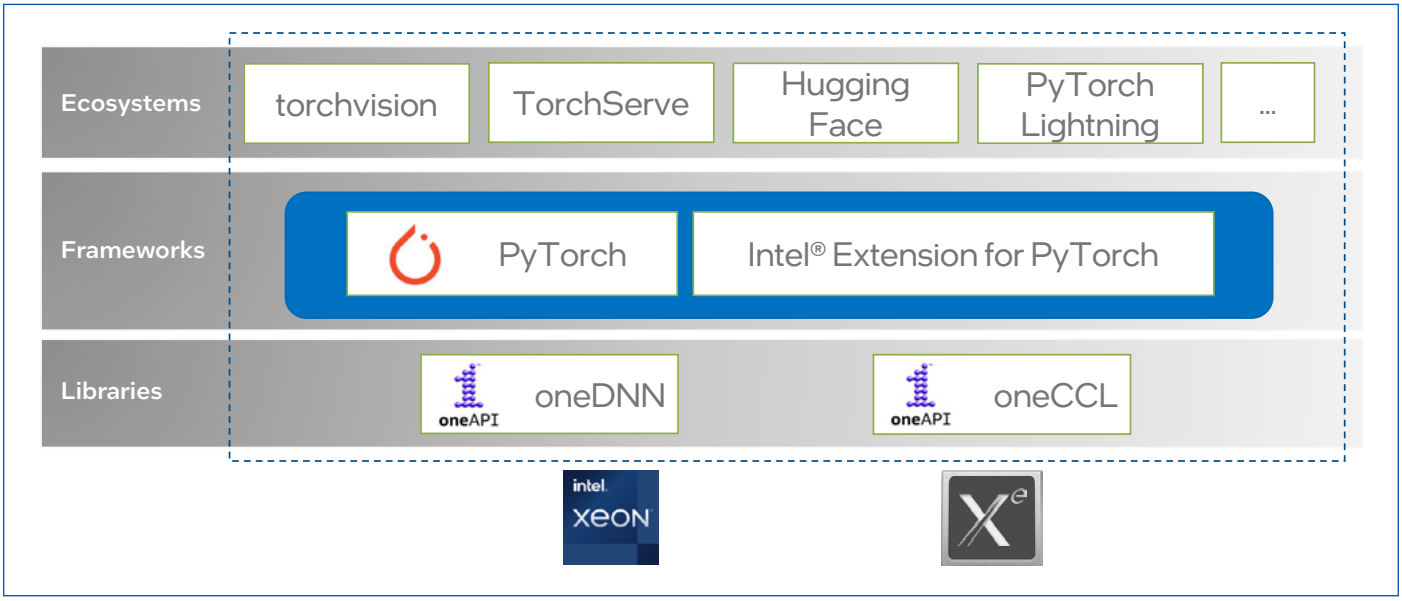

Figure 2. Intel ® Optimizations for PyTorch

- **Weight-only quantization**

  Weight-only quantization is a type of optimization for LLMs. As long as the computing accuracy is guaranteed, the parameter weights are converted to INT8 data type, but restored to half-precision during computing, which helps to reduce the memory space occupied by model inference, speeding up the overall computing process.

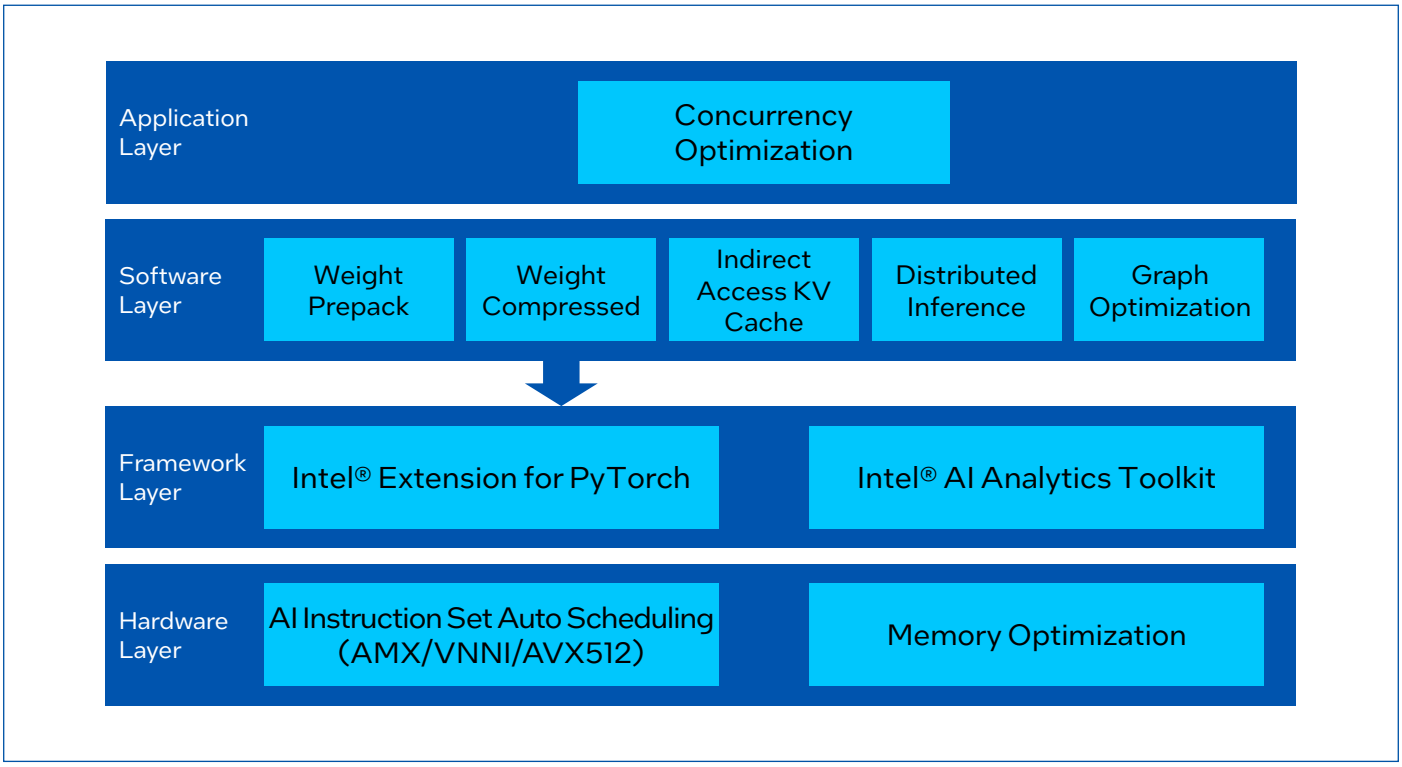

Figure 3. Optimized architecture for WiNGPT

Winning Health and Intel have jointly optimized WiNGPT's inference performance by improving memory utilization. The two has also collaborated to fine-tune the key operator algorithms for PyTorch on CPU platforms, delivering further inference acceleration for the deep learning framework.

In a test-based validation environment, the inference performance of the LLaMA2 model reached 52ms/token. With automated medical report generation, a single output takes less than 3s[8].

During the test, Winning Health also compared the performance of the 5th Gen Intel® Xeon® Scalable processor-based solution with that of the 3rd Gen. The result shows the latest generation processors deliver over 3x performance boost over the 3rd generation[9].

As the business scenarios in which WiNGPT is used are relatively tolerant of LLM latency, the robust performance of the 5th Gen Intel Xeon Scalable processor is sufficient enough to meet user needs. Meanwhile, the CPU-based solution is also easily scalable for inference instances and can be adapted to perform inference on a variety of platforms.

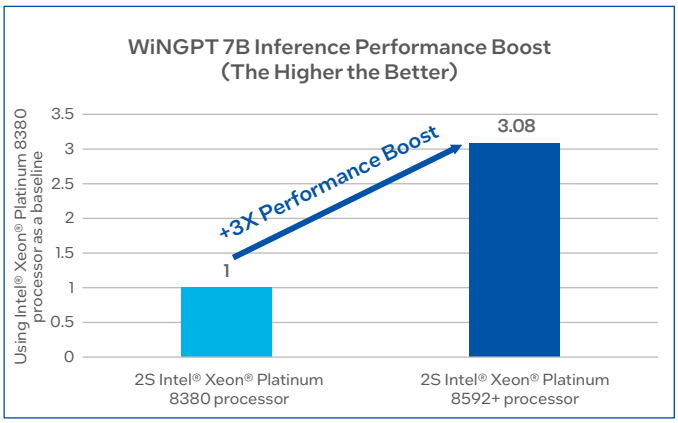

Figure 4. Performance results of WiNGPT on different generations of Intel® Xeon® processors

Benefits

WiNGPT solution based on the 5th Gen Intel Xeon Scalable processors has delivered the following benefits to healthcare institutions:

- **Optimized LLM performance with enhanced application experience:** With technical optimizations by both sides, the solution has fully leveraged the AI performance advantages with the 5th Gen Intel Xeon Scalable processors. It can meet the performance requirements for model inference in scenarios such as medical report generation, resulting in shortened generation time with guaranteed user experience.
- **Improved cost-effectiveness with platform building cost kept under control:** The solution can utilize the general-purpose servers already in use in healthcare institutions for inference, eliminating the need to add dedicated inference servers, which helps to reduce costs of procurement, deployment, operation, maintenance, and energy consumption.
- **Well-balanced between LLMs and other IT applications:** The fact that the solution manages to use CPU for inference means healthcare institutions can flexibly allocate CPU's computing power between LLM inference and other IT applications as needed, which improves the agility and flexibility of computing power allocation.

[8,9] Data from Winning Health's internal test results as of November 2023. Test configurations—Baseline: 2S Intel® Xeon® Platinum 8380 processor @ 2.30 GHz, 1024 GB total memory (16x64 GB DDR4 3200 MT/s), 745.2 GB SSD, Ubuntu 22.04.3 LTS; New: 2S Intel® Xeon® Platinum 8592+ processor @ 1.90 GHz, 512 GB total memory (16x32 GB DDR5 5600 MT/s), 1.1 TB SSD, Ubuntu 22.04.3 LTS. Intel does not control or audit third-party data. You should consult other sources to evaluate accuracy. Performance varies by use, configuration and other factors. Learn more at www.Intel.com/PerformanceIndex.

Looking ahead

The 5th Gen Intel® Xeon® Scalable CPUs provide excellent inferencing performance, especially when used in conjunction with WiNGPT, making the application of the LLM easier and more cost-effective. Both sides will continue to refine their work on LLMs to make Winning Health's latest AI technologies accessible and beneficial to more users.