Report

I7 Raptor Lake

Machine Learning on CPUs

Discover how the latest CPUs enhance machine learning workloads and optimize performance and cost efficiency,

Student: Chan Duong Nguy

Student id: 63408

April 21 2023

Contents

1. Introduction

The field of machine learning (ML) has been evolving rapidly over the years[6], and most of the models are run on edge devices. However, as the size of the models grows, the need for a central processing unit (CPU) that can run ML tasks efficiently also grows. While most of the popular ML models we see today are trained on clusters of graphics processing units (GPUs) or tensor processing units (TPUs) then converted and compressed[8] to run on CPUs.
The 13th generation Intel Core i7 processor, code-named Raptor Lake, was released in October 2022, and in this report, we will delve into the technologies that support these ML models

Figure 1: 13th Gen Intel Core Processors

1.1 Technical Specifications

Processor	Cores	P-cores	E-cores	Threads	Frequency
i7-13700KF	16	8	8	24	5.40 GHz
i7-13700	16	8	8	24	5.20 GHz
i7-13700F	16	8	8	24	5.20 GHz
i7-13700T	16	8	8	24	4.90 GHz
i7-13700K	16	8	8	24	5.40 GHz
i7-13650HX	14	6	8	20	4.90 GHz
i7-13700H	14	6	8	20	5.00 GHz
i7-13620H	10	6	4	16	4.90 GHz
i7-13700TE	16	8	8	24	4.80 GHz
i7-13705H	14	6	8	20	5.00 GHz
i7-1365U	10	2	8	12	5.20 GHz
i7-1370P	14	6	8	20	5.20 GHz
i7-13800HE	14	6	8	20	5.00 GHz
i7-1360P	12	4	8	16	5.00 GHz
i7-1355U	10	2	8	12	5.00 GHz
i7-13850HX	20	8	12	28	5.30 GHz
i7-13800H	14	6	8	20	5.20 GHz
i7-13700HX	16	8	8	24	5.00 GHz
i7-13700E	16	8	8	24	5.10 GHz
i7-1370PE	14	6	8	20	4.80 GHz
i7-1365UE	10	2	8	12	4.90 GHz

Having more cores is generally better for any tasks including Machine Learning, as it allows for highly parallel vectors processing. However, this can lead to power consumption issues, particularly for laptop devices. To address this challenge, Intel has developed new technologies such as Deep Learning Boost, Gaussian and Network Accelerator, and Smart Sound Technology,... [2] Which are designed to improve the efficiency of running model inference. These technologies aim to make the most out of core counts while minimizing power consumption. With these features, processors equipped with these technologies are expected to be a competitive choice for machine learning workloads, particularly for high-performance computing needs.

One of the key technologies developed by Intel to improve machine learning performance is Deep Learning Boost (DL Boost or DLB). DL Boost is a set of instructions that accelerates deep learning inference workloads, allowing for faster and more efficient processing. DL Boost works by optimizing the use of available hardware resources, including the CPU, GPU, and memory, to provide faster and more efficient execution of deep learning workloads[1]. This technology is particularly useful for tasks that require high levels of computation, such as image recognition and natural language processing[1]. In the next section, we will explore DL Boost in more detail, including its features and performance benefits.

2. Deep Learning Boost

Intel's Deep Learning Boost [3] is a technology developed for the Xeon lineup. It extends the AVX512 instruction set with Vector Neural Network Instructions (VNNI) to significantly accelerate inference performance for deep learning workloads.

2.1 Advanced Vector Extensions 512

Intel AVX-512 is a powerful set of instructions designed to increase the processing capabilities of CPUs. Unlike scalar instructions, which can only process one piece of data at a time, AVX-512 is a SIMD instruction that allows CPUs to process multiple pieces of data simultaneously. This is achieved by increasing the register width to 512 bits, which supports 32 double-precision and 64 single-precision floating-point numbers. Additionally, AVX-512 provides up to two 512-bit fused-multiply add (FMA) units, which doubles the number of registers and computational capacity.[4]

One of the main advantages of Intel AVX-512 is that it can be used without requiring significant modifications to existing applications. Moreover, the Intel compilers can automatically optimize the use of vector instructions for AVX-512, making it easier for developers to take advantage of its capabilities. Some of the applications that benefit the most from AVX-512 include high-performance computing (HPC) simulations, DNA sequencing, 3D modeling, and financial analytics. It is also useful for image and audio/video processing, cryptography and data compression, and AI and deep learning applications that leverage Intel Deep Learning Boost (Intel DL Boost).

In summary, Intel AVX-512 is a powerful instruction set that can significantly increase the processing capabilities of CPUs. It achieves this by allowing CPUs to process multiple pieces of data simultaneously, thereby improving computational efficiency. Its compatibility with existing applications and automatic optimization capabilities make it a valuable tool for a wide range of applications, including HPC, cryptography, and AI.

2.2 Vector Neural Network Instructions

The Deeplearning boost is a combination of two important features in modern computing, AVX512 and Vector Neural Network Instructions (VNNI). VNNI is a powerful tool that combines three execution instructions into one, which increases the inference performance of INT8 models. VNNI is particularly beneficial for quantized models, which can sacrifice some accuracy to achieve faster inference speed and lower power consumption. This is especially important for mobile and edge devices, where processing power and battery life are limited. By using VNNI, the execution of INT8 models can be completed more efficiently with a lower computational cost, making them more suitable for deployment on mobile and edge devices. Although quantized models may have lower accuracy compared to their full-precision counterparts, the reduction in computational cost and power consumption can lead to faster and more practical AI solutions that are better suited for real-world applications on these devices.[1]

When combined with AVX512, VNNI becomes a powerful tool for deep learning

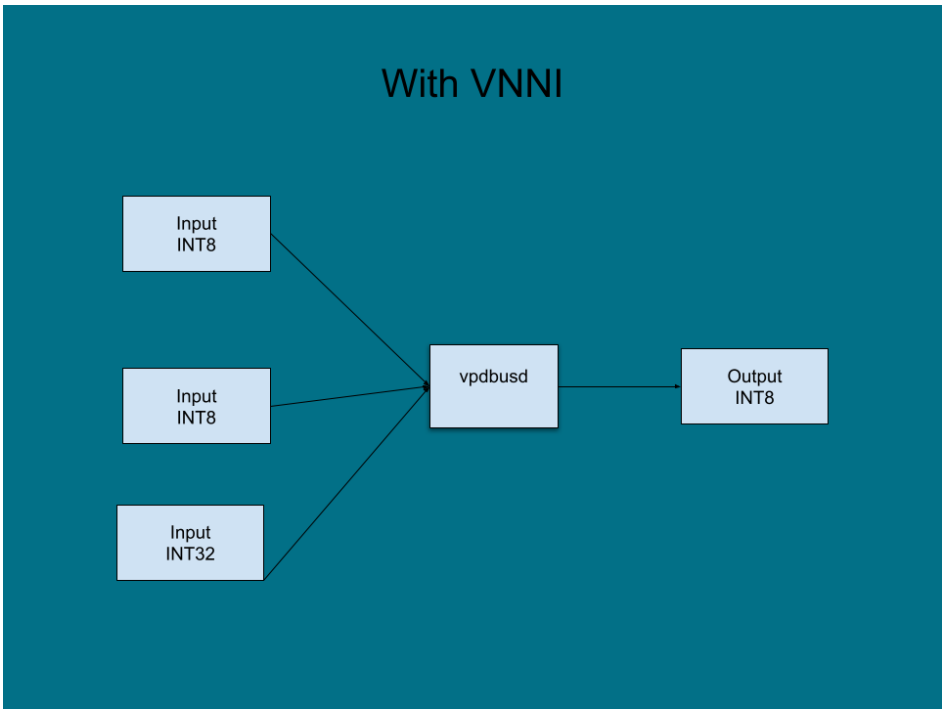

Figure 2: Platforms using VNNI require only one instruction, vpdpbusd, to complete the INT8 convolution operation

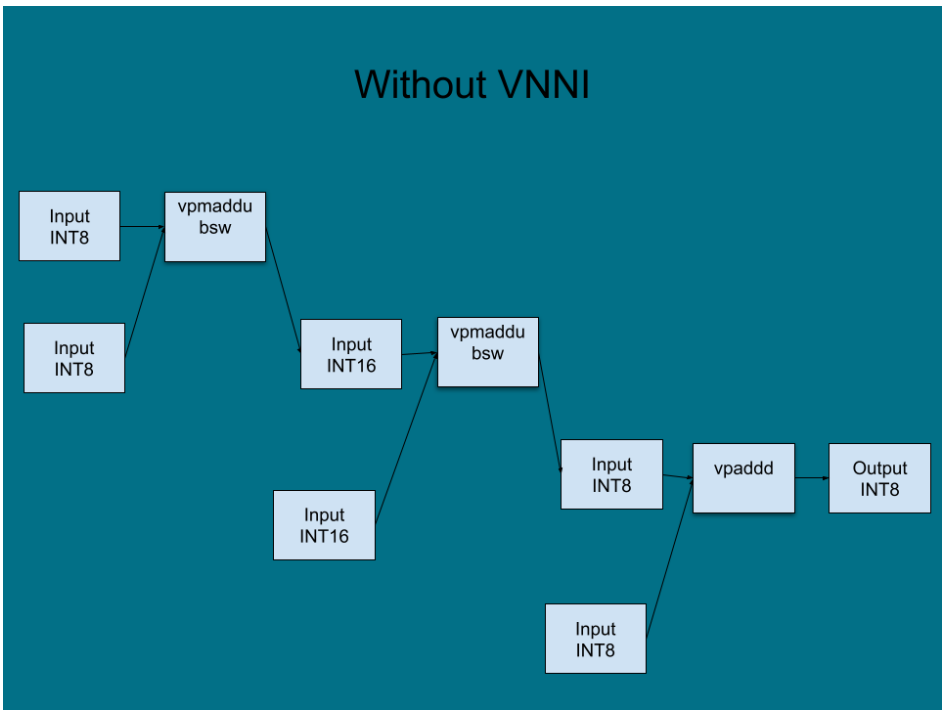

Figure 3: Without VNNI, three instructions need to be run separately, which can add latency to AI processing tasks

applications. The integration of VNNI with AVX512 allows for significant improvements in the performance of AI tasks. By using these technologies together, the latency of AI processing tasks is reduced, resulting in faster, more accurate results. The combination of AVX512 and VNNI is a significant advancement in the field of AI and computing, and it is expected to play an important role in the development of more efficient and advanced AI applications in the future.

3. Gaussian Neural Accelerator

3.1 Understanding GNA

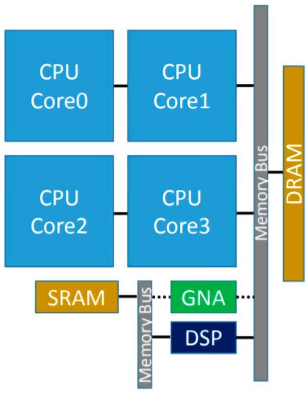

Figure 4: Diagram of GNA [5]

The 13th generation Intel Core i7 includes a specialized component called the GNA (Gaussian and Neural Accelerator)[7]. It is designed to accelerate neural network-based applications, particularly those related to audio tasks such as denoising, as well as image denoising. Although it is part of the CPU package, the GNA operates independently of the CPU's execution graph. To use it, users need to access a dedicated Intel API. The GNA is an efficient and low-power accelerator that enhances performance for AI-related tasks on compatible devices.

3.2 GNA Usage

To use GNA, a dedicated Intel API such as the Intel Deep Learning SDK tool is required. The process typically involves training a neural network, ideally using quantization training awareness to take into account the noise created by low precision numbers when compressing the model in the next step.

The next step involves using the Intel Deep Learning SDK Deployment to import the newly trained model. The API works with popular DL frameworks like the Intel DLSDK inference Engine or the native GNA libraries.

When the model is run, it will be executed solely on GNA instead of the CPU. This approach offers significant performance improvements and energy efficiency compared to running the model on the CPU alone

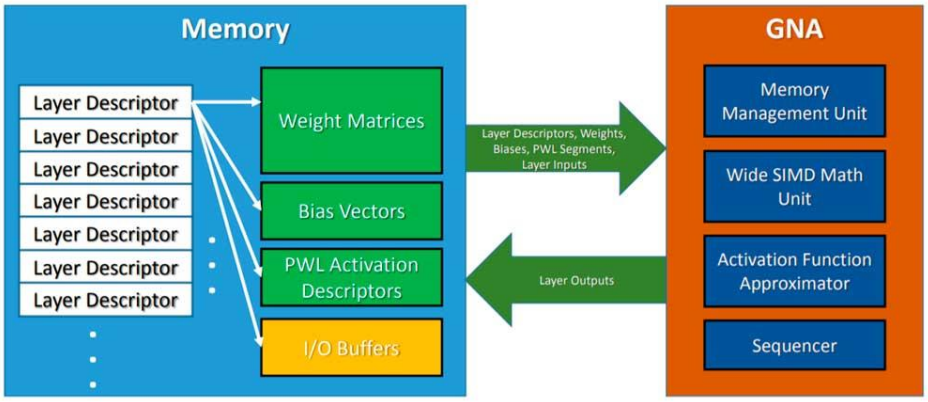

Figure 5: Architecture of GNA[5]

4. Price

The best processor for a particular task will depend on a variety of factors, such as the specific ML algorithm being used, the size of the dataset being processed, and the available hardware resources. Additionally, ML workloads often benefit from the use of specialized hardware, such as GPUs or TPUs, rather than relying solely on CPUs. Therefore, when assessing the cost-effectiveness of processors for ML, it's important to consider the entire system architecture and workload requirements.

Processor	**Price**	**Total Cores**
i7-13700KF	$384.00	16
i7-13700	$384.00	16
i7-13700F	$359.00	16
i7-13700T	$384.00	8
i7-13700K	$409.00	16
i7-13650HX	$485.00	8
i7-13700H	$502.00	8
i7-13620H	$502.00	4
i7-13700TE	$390.00	8
i7-13705H	$502.00	8
i7-1365U	$426.00	8
i7-1370P	$438.00	8
i7-13800HE	$460.00	8
i7-1360P	$480.00	8
i7-1355U	$469.00	8
i7-13850HX	$428.00	12
i7-13800H	$457.00	8
i7-13700HX	$485.00	8
i7-13700E	$390.00	8
i7-1370PE	$441.00	8

For students studying ML, it may be more practical to choose a mid-range processor that provides a good balance between performance and cost, such as the i7-13700 or i7-13700KF. These processors offer 16 cores and are priced at around $384.00, making

them a good choice for students who need to run moderately-sized ML models and algorithms.

5. Conclusion

The development of AI products has been accelerating rapidly in recent years, driving the need for more efficient AI accelerators. However, the inclusion of new ML technologies such as DLB embedded in new cores series chips may cause battery usage problems, while GNA may increase unnecessary complexity and drive up the cost of the chip if not adapted by developers. This raises a critical question: do the benefits of implementing these technologies outweigh the risks? While these technologies have the potential to significantly improve the performance of AI products, manufacturers and developers must weigh the potential risks and consider the trade-offs involved in incorporating them into their products. Ultimately, it is important to carefully evaluate the impact of these new technologies on the overall efficiency, cost, and functionality of AI products before incorporating them.

Bibliography

[1] Intel. Enhance artificial intelligence (ai) workloads with built-in accelerators.

[2] Intel. Intelligent pcs with integrated ai.

[3] Intel. Enhance artificial intelligence (ai) workloads with built-in accelerators, 2022.

[4] Intel. Get outstanding computational performance without a specialized accelerator, 2022.

[5] Matt Mills. Intel gausian & neuro accelerator: The low power coprocessor for intel inference, 2020.

[6] Jaime Sevilla, Lennart Heim, Anson Ho, Tamay Besiroglu, Marius Hobbhahn, and Pablo Villalobos. Compute trends across three eras of machine learning, 2022.

[7] Hayashi Kan Shiori Sugawara. Noise cancellation for simple neural network models, 2020.

[8] Andre Ye. Three model compression methods you need to know in 2021, 2021.