

Literary Survey of Intel® Xeon Family Processors

Abishek Arumugam Thiruselvi
Department of Electrical and Computer
Engineering
Concordia University
Montreal, Canada
abishekarumugam@icloud.com

Abstract—This paper describes the information on historical background, Instruction set properties, Evolution, application domain, latest trends and reason for becoming obsolete of Xeon family processors.

Keywords—High Performance Computing, SIMD, Hyper Threading, Integrated Memory Controller, Quick Path Interconnect, Thermal Design Power, Voltage, Current, Instruction Level Parallelism, Instruction Set Architecture, Reduced Instruction Set Compiler, Advanced Vector Extensions, Knights Landing, Vector Neural Network Instructions.

I. INTRODUCTION

The Xeon Processors are the modern processors that make the foundation of data center innovation and high-performance computation. The Xeon's multiprocessor efficiency and performance is in upward trend.[1] Moore's law states that the number of transistors per chip doubles every year. Fig. 1 represents the graphical representation of Moore's law.

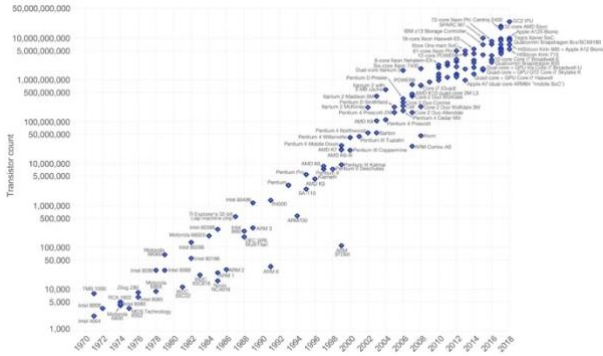


Fig. 1. Transistors count in comparison with the year

HPC requires efficient use of SIMD, and many modern processors utilize the SIMD execution, which is an efficient way to boost performance [8]. The Intel's Xeon multiprocessors processors are designed to station at automotive environment for embedded applications or WSC/Server application. The primary advantage of Xeon processors is the multiple cores, ECC (Error Code Correction) memory and Hyper Threading capability.

The critical characteristics of Xeon server family processors are availability, scalability, desired efficient throughput, overall efficiency, cost-effectiveness, and handling multiple requirements per unit time [1].

[1] Dennard Scaling signifies even with the use of more transistors, the processor gets faster but uses less power. He observed power density is constant for a given area of silicon.

Later in 2004, Dennard Scaling failed as V, and I dropped. This also helped develop multicore processors and introduction of ILP. [3] Multicore technology was developed when engineers increased the speed of single-core chips, which created too much heat. Tests show that, compared to a single processor, the addition of secondary processors can improve the processing speed by 93%. Multicore technology plays a primary role in server and desktop technology for maximum business advantage and IT potential.

II. HISTORICAL BACKGROUND

The first member of the Xeon processor family is Xeon Pentium II, followed by Pentium III, which is an improved server line-up by Xeon microprocessors. Netburst based Xeon processor used RDRAM supported HT technology, later E7 models had DDRAM and gave advancements over Pentium III processors like Gallatin, Prestonia, and Foster. Failure of Itanium due to low demand and competition over AMD's x86-64 processor, Xeon later in 2004 released Xeon Nocona processor. After that, Xeon's Sandy Bridge and Ivy Bridge based processors came into existence, which could work in the multiprocessor system. Finally, Xeon's Nehalem, the successor to Xeon Core microarchitecture, is defined in two vectors, one with explicit Xeon core functionality universal for all the Nehalem family and second on Xeon uncore elements and highly optimized to fit best for the application requirement. Xeon 5500 processors has the first QPI (Quick Path Interconnect) and IMC (Integrated Memory Controller). Currently, we have Intel's 3rd gen. Xeon SPs, which is efficient as showed in TABLE II. [4] Currently, we have 3rd generation Xeon SP's, which record performance and reduce power consumption. However, switching loss and leakage currents in transistors has severe effects on the energy efficiency of the processors (1). [1] In addition, if the frequency can't be easily maintained, thermal management will be an issue. Modern processors use the voltage indexing method to slow the processor down when the peak current increases. Also, failing to provide required cooling increases the junction temperature and, in turn, leads to the failure of the processors. This led to creating TDP [1].

$$E_{dynamic} \propto \frac{1}{2} * Capacitive Load * Voltage^2 \quad (1)$$

III. INSTRUCTION SET PROPERTIES

The X86 architecture incorporated in the Xeon server processors is very sophisticated and used in a wide area of applications. [1] The instruction set architecture helps in the virtual elimination of ALP and standardized vendor-independent OS. This enabled the development of RISC.

[1] Mostly, computer has identical ISA and similar organisation, but they vary in their hardware implementation. For example, Xeon core i7 and Xeon E7 has the same ISA and organization but has different clock and memory system. Making Xeon E7 effective for server computers.[4] TABLE I. compares the specification of 2nd and 3rd Gen. Xeon SP. TABLE II. shows the specification data of Xeon family E3, E5 and E7 processors.

TABLE I. SPECIFICATION TABLE OF 2ND AND 3RD GENERATION XEON SP

Features	2 nd Gen Intel Xeon Scalable Processors	3 rd Gen Intel Xeon Scalable Processors
Cores and Threads Per Processor	[8200] Up to 28 cores and 56 threads [9200] Up to 56 cores and 112 threads	Up to 40 cores and 80 threads
Data Center Cache Hierarchy	Up to 38.5 MB (non-inclusive) Shared L3 Cache	Up to 60MB
PCIe Lanes	Up to 48 Lanes PCIe 3.0 (2.5, 5, 8 GT/s)	Up to 64 Lanes PCIe 4.0
Memory	Up to 6 channels at 2933 MT/s per processor Up to 2 DPC RDIMMs, LRDIMMs, 3DS LRDIMMs, supporting up to 16Gb DDR4 devices Intel® Optane™ DC Persistent Memory Module support for up to 4.5TB system memory per processor	Up to 8 channels at 3200 MHz. Intel Optane™ DC Persistent Memory Module support for up to 6TB system memory per processor and supporting DDR4-3200
Turbo Frequency	Boost across Stack	3.40 GHz

TABLE II. SPECIFICATION TABLE OF XEON E3, E5 AND E7

	Intel® Xeon® Processor E3-1285 v6	Intel® Xeon® Processor E5-2699A v4	Intel® Xeon® Processor E7-8894 v4
Total Cores	4	22	24
Total Threads	8	44	48
Max Turbo Frequency	4.50 GHz	3.60 GHz	3.40 GHz
Cache	8 MB Intel® Smart Cache	55 MB	60 MB
Bus Speed	8 GT/s	9.6 GT/s	9.6 GT/s
Max Memory Size (dependent on memory type)	64 GB	1.5 TB	3 TB
Memory Types	DDR4-2400, DDR3L-1866	DDR4 1600/1866/2133/2400	DDR4-1333/1600/1866 DDR3-1066/1333/1600
Max # of Memory Channels	2	4	4
Max Memory Bandwidth	37.5 GB/s	76.8 GB/s	85 GB/s
ECC Memory Supported ‡	Yes	Yes	Yes
Scalability	15 Only	25	585
PCI Express Revision	3	3	3
PCI Express Configurations ‡	1x16, 2x8, 1x8+2x4	x4, x8, x16	x4, x8, x16
Max # of PCI Express Lanes	16	40	32
Instruction Set	64-bit	64-bit	64-bit

Xeon processors are designed for prolonged usage and low power consumption. To achieve that, Xeon is clocked at low frequency to lower operating temperature. Typically, Xeon runs many concurrent processes in different users. Therefore, memory bandwidth becomes more important, and these processors offer large cache and more aggressive memory systems to boost that bandwidth. [4] As compared to the previous processors in the Xeon line, the scalable processors introduced new key features:

- 512-bit vector registers and AVX-512 instruction set
- A re-architect L2 & L3 cache hierarchy.
- Improved performance and scalability with mesh on-chip interconnect.

Stacked DRAMs are used in the top-end GPUs from AMD and NVIDIA to achieve the highest memory performance. Xeon PHI coprocessor utilises stacked DRAM, which helps in HBM (High Bandwidth Memory).

The vector registers and instruction set allow for maximum SIMD were introduced in the Xeon Phi line processor, particularly in KNL chips installed in a supercomputer at TACC [6]. Xeon SP advanced to Mesh-Interconnect architecture rather than coupled-ring architecture because, mesh interconnect helps connect the cores efficiently inside the processors. [4] Additional features introduced in Cascade Lake processors is AVX512 VNNI instruction set, intended to support computation with AI and Deep Learning Neural Networks. Fig. 2 represents the graph of MAC/Cycle across SIMD Extension.

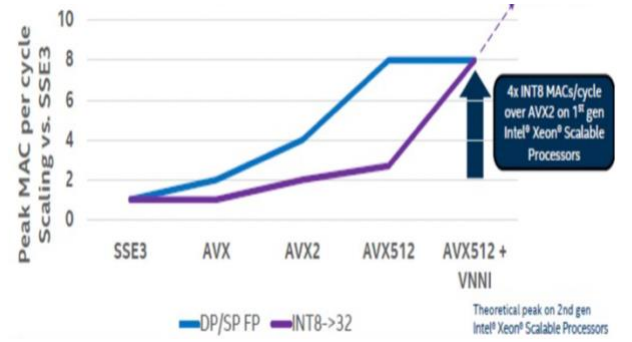


Fig. 2. Comparison of MAC/Cycle across SIMD Extensions.

IV. EVOLUTION/TRANSFORMATIONS IN XEON PROCESSORS

Integration of multi-core in a single chip helped achieve high performance. [4] Xeon launched Xeon PHI, a coprocessor that can work synergistically with Xeon processors and achieve parallel performance up to 1TFLOPS of double-precision peak performance in every chip. Typical Xeon architecture enables it to run on the complete specification in Xeon-based family servers. Simultaneously executing two threads in a single processor rather than changing between lines is called Hyper Threading technology. HT technology allows simultaneous scheduling of two or more threads on the same processor core to utilize the processor's resources better. [4] In addition, all the innovations in systems and microarchitecture in Xeon 5500 processors are designed with Hi-k metal gate Si technology, a power gating technology, SRAM needs power to hold the stored values the only way to cope with leakage is to turn off power in chips (2).

$$P_{static} \propto I_{leakage} \quad (2)$$

This uses a new combination of Hi-k metal gate dielectrics and conductive materials to enhance material properties such as chip size, power consumption, leakage current, and manufacturing cost.

V. INTEL XEON FAMILY PROCESSORS DOMAIN/APPLICATIONS

This processor family is used largely in desktop and WSC/Server systems. Therefore, their primary footprint in an application is on Server, Networking, Big Data and Storage.

[1] Clusters are a combination of desktop or server computers connected with LAN to act as a single large computer, and each node runs its OS. WSC is designed in such a way to use 10,000 servers as one computer. The scalability of WSC is by LAN-connected network and not by integrated hardware. The processors are stationed at diverse fields like,

- Government and Enterprise Markets
- Service Providers in Cloud
- Financial Services in Healthcare
- Retail Markets

VI. LATEST TRENDS AND RESONS FOR BECOMING OBSOLETE

Xeon PHI processor has the newest high throughput architecture targeted at HPC. Intel's two new recent technologies,

- OPTANE
- 3-D NAND

Xeon Optane SSD has high-performance and low latency. Enhancements in PCIe storage in processors, like LED management and error containment functions, these supports are provided by the Xeon Volume Management Device. Introduction of new performance enhancement software from Intel like PCIe-based RAID technology, Rapid Storage Technology Enterprise, and Xeon Cache Acceleration Software helped increase performance. Mesh architecture rendered maximum performance, high throughput, low latency, and optimized data sharing between all CPU cores for ideal memory bandwidth.

Cascade Lake was created by Intel with the goal of improving AI and deep learning performance [7]. Intel has developed an energy efficient DPC technology, which helps coordinate the workload in cores to sleep or slow down while others work. Xeon enables the software community to take full advantage of Xeon scalable processors, and Xeon is actively involved in optimizing the number of open-source AI frameworks. Fig. 3 shows some features of 3rd Generation Xeon Scalable Processor. These processors have opened the opportunity to reduce HPC applications energy consumption and explore new trade-offs between energy and performance [5].

[4] Intel's Turbo-Boost 2.0 technology allows processors to run faster than their rated frequency when they are operating below the specification limit. Intel created many technologies handle junction temperature, to maximize energy efficiency, to manage power and frequency, to

increase speed, improved DL boost, enhanced performance, more Ultra Path connect and more DDR4 memory.

Some additional key attributes of Xeon SP:

- Xeon Stratix 10 NX FPGA
- Xeon Infrastructure Management Technologies (Xeon IMT)
- Support for Xeon Optane persistent memory 200 series (Xeon Optane PMem)
- New Xeon Platform Firmware Resilience (Xeon PFR)
- Integrated Xeon Quick Assist Technology (Xeon QAT)
- Xeon Speed Select Technology

Prior to the launch of Ice Lake processor, Intel stopped its 1st generation Xeon processors due to the shift of demand in market. And, Intel has stopped most of the Skylake processors and moved towards new Cascade Lake processors.

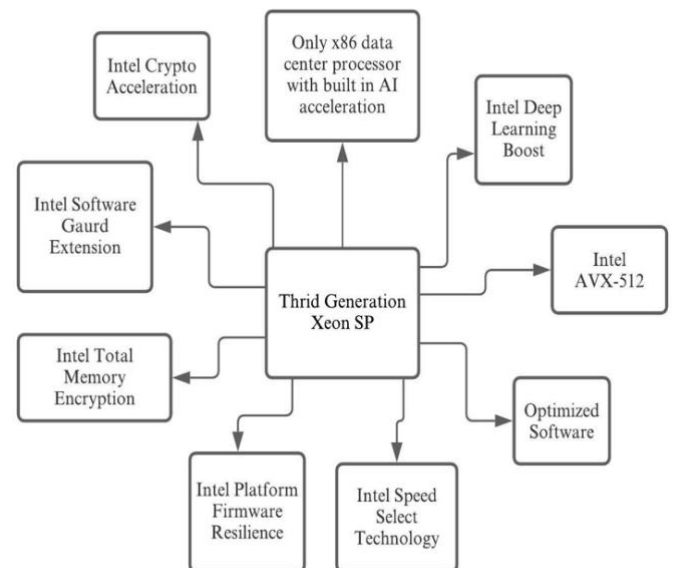


Fig. 3. Block Diagram of 3rd Generation Xeon Scalable Processor.

VII. CONCLUSION

In Xeon processors, there are several advancements made from single core to multicore technology. Incorporating coprocessor helped increase the overall efficiency of the system. [8] Several benchmark tests proved the low-end and increased efficiency of Xeon PHI. All technological advancements are created to improve the efficiency of processors and help reduce bottleneck issues. Currently, the third generation SPs made its footprint in diverse fields. In future, the success of many ISA depends on the workload and user-friendly software.

REFERENCES

- [1] John L. Hennessy, David A. Patterson., vol. 2. Oxford: Clarendon, "Computer Architecture, A Quantitative Approach", November 2017.
- [2] P.Gepner, D.L. Fraser and M.F. Kowalik, "Evaluating Performance of New Quad-Core Intel Xeon 5500 Family Processors for HPC", *R. Wyrzykowski et al. (Eds.): PPAM 2009, Part I, LNCS 6067, pp. 1–10, 2010.*
- [3] T.W. Burger, "Intel® Multi-Core Processors: Quick Reference Guide", August 2005. "<http://www.intel.com>"
- [4] "<https://www.intel.com/content/www/us/en/developer/overview.html>"
- [5] "L. Szustak , R. Wyrzykowski , T. Olas , and V. Mele", "Correlation of Performance Optimizations and Energy Consumption for Stencil-Based Application on Intel Xeon Scalable Processors", *Ieee Transactions on Parallel and Distributed Systems, Vol. 31, No. 11, November 2020.*
- [6] "A. Ramachandran, J. Vienne, R. Wijngaart, L. Koesterke, I. Sharapov, "Performance Evaluation of NAS Parallel Benchmarks on Intel R Xeon Phi", *42nd International Conference on Parallel Processing*, 2013.
- [7] "M. Arafa, B. Fahim, S. Kottapalli, A. Kumar, LP. Looi, S. Mandava, A. Rudoff, I.M. Steiner, B. Valentine, G. Vedaraman, and Sujal Vora, "Cascade Lake: Next Generation Intel Xeon Scalable Processor", *IEEE Micro ,Volume: 39, Issue: 2, March-April 2019.*
- [8] Pennycook, S. J., C. J. Hughes, M. Smelyanskiy, and S. A. Jarvis. "Exploring SIMD for Molecular Dynamics, Intel Xeon Processors and Intel Xeon Phi Coprocessors", "*IEEE 27th International Symposium on Parallel & Distributed Processing*", 2013.

Faculty of Engineering and Computer Science Expectations of Originality

This form sets out the requirements for originality for work submitted by students in the Faculty of Engineering and Computer Science. Submissions such as assignments, lab reports, project reports, computer programs and take-home exams must conform to the requirements stated on this form and to the Academic Code of Conduct. The course outline may stipulate additional requirements for the course.

1. Your submissions must be your own original work. Group submissions must be the original work of the students in the group.
2. Direct quotations must not exceed 5% of the content of a report, must be enclosed in quotation marks, and must be attributed to the source by a numerical reference citation¹. Note that engineering reports rarely contain direct quotations.
3. Material paraphrased or taken from a source must be attributed to the source by a numerical reference citation.
4. Text that is inserted from a web site must be enclosed in quotation marks and attributed to the web site by numerical reference citation.
5. Drawings, diagrams, photos, maps or other visual material taken from a source must be attributed to that source by a numerical reference citation.
6. No part of any assignment, lab report or project report submitted for this course can be submitted for any other course.
7. In preparing your submissions, the work of other past or present students cannot be consulted, used, copied, paraphrased or relied upon in any manner whatsoever.
8. Your submissions must consist entirely of your own or your group's ideas, observations, calculations, information and conclusions, except for statements attributed to sources by numerical citation.
9. Your submissions cannot be edited or revised by any other student.
10. For lab reports, the data must be obtained from your own or your lab group's experimental work.
11. For software, the code must be composed by you or by the group submitting the work, except for code that is attributed to its sources by numerical reference.


You must write one of the following statements on each piece of work that you submit:

For individual work: **"I certify that this submission is my original work and meets the Faculty's Expectations of Originality"**, with your signature, I.D. #, and the date.

For group work: **"We certify that this submission is the original work of members of the group and meets the Faculty's Expectations of Originality"**, with the signatures and I.D. #s of all the team members and the date.

A signed copy of this form must be submitted to the instructor at the beginning of the semester in each course.

I certify that I have read the requirements set out on this form, and that I am aware of these requirements. I certify that all the work I will submit for this course will comply with these requirements and with additional requirements stated in the course outline.

Course Number: COEN6741
Name: ABISHEK ARUMUGAM THIRUSELVI
Signature: 

Instructor: Dr. Sofiène Tahar.
I.D. # 40218896
Date: 31.01.2022

¹ Rules for reference citation can be found in "Form and Style" by Patrich MacDonagh and Jack Bordan, fourth edition, May, 2000, available at <http://www.encs.concordia.ca/scs/Forms/Form&Style.pdf>.