Volume 6, Issue 9, September 2016 ISSN: 2277 128X

International Journal of Advanced Research in Computer Science and Software Engineering

Research Paper

Available online at: www.ijarcsse.com

3D XPoint and GP-GPUs: Solution to Problem of Big Data

Aayush Sinha, Utkarsh Sharma
CSE Department, Northern India Engineering College, GGSIPU,
Delhi, India

***Abstract:** Memory is an integral component of a computer, without which we cannot initiate it. With the emergence of big data, we have encountered the problems related to the storage and processing of humongous amount of data. In this paper, we will lay emphasis on the revolutions in memory till date, which will also tackle these problems of big data. Our main focus will be to elaborate on the emerging technologies of 3D XPoint and GPGPUs. The main aim of this paper is to update ourselves regarding the biggest game changer in the computer industry for about 30 years.*

***Keywords:** 3DX Point, GPU, GPGPU, DRAM, NAND, Flash Memory, CPU*

I. INTRODUCTION

A. 3D XPoint: The Next Generation Non Volatile Memory

We're finally on the cusp of Solid State storage breaking into mainstream adoption (e.g. becoming more affordable and more relevant), it has been a whopping 26 years since the last completely new type of memory – NAND Flash, was introduced. But now, Intel engineers have realized that breakthrough with 3D XPOINT.

3D XPoint is a fast, cost effective and non-volatile memory technology announced by Intel and Micron in July 2015. Intel refers to future storage devices using the technology under the name Optane while Micron uses the name QuantX. This memory will soon enter mass production and promises to revolutionize the computer industry [1].

B. GP-GPUs

By a rapid development of Graphics Processing Unit (GPU) in recent years, the programmability and highly parallel processing feature of GPU create a chance to allow the general purpose computation to be conducted on GPU, conventionally called GPGPU (General Purpose computation on GPU).

A GPGPU uses parallel processing on multiple GPUs to handle operations in a pipeline manner which allows it to handle computation on applications which are traditionally handled by the CPU [2].

II. THE MEMORIES IN USE TODAY: DRAM, EPROM AND FLASH MEMORY

A. DRAM (1966)

Dynamic random-access memory (DRAM) is a type of random-access memory that stores each bit of data in a separate transistor within an integrated circuit. The capacitor can be either charged or discharged; these two states are taken to represent the two values of a bit, conventionally called 0 and 1. Unlike flash memory, DRAM is volatile memory (vs. non-volatile memory), since it loses its data quickly when power is removed [1].

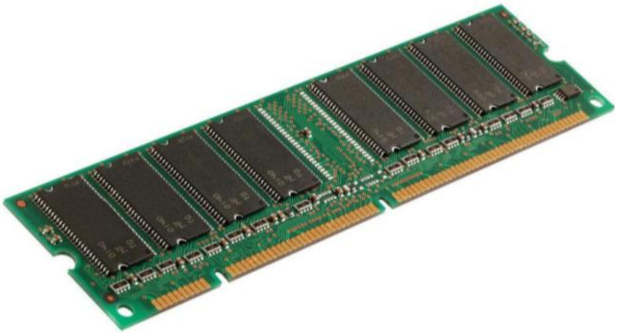


Fig. 1- DIMM module [1]

B. EPROM (1971)

EPROM (erasable programmable read-only memory) is programmable read-only memory (programmable ROM) that can be erased and re-used. Erasure is caused by shining an intense ultraviolet light through a window that is designed into the memory chip. EPROMs are easily recognizable by the transparent fused quartz window in the top of the package, through which the silicon chip is visible, and which permits exposure to ultraviolet light during erasing. This is a non-volatile form of memory [1].

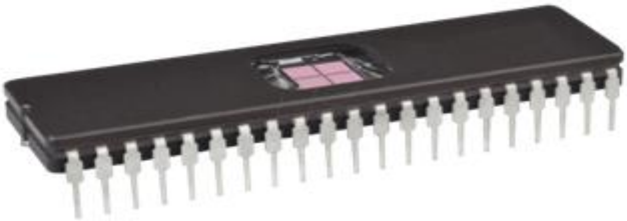


Fig. 2- EPROM Chip [1]

C. Flash memory: NAND (1989)

Flash memory is a type of non-volatile memory that erases data in units called blocks. A block stored on a flash memory chip must be erased before data can be written, or programmed, to the microchip. Flash memory retains data for an extended period of time whether a flash-equipped device is powered on or off. NAND flash memory stores data in an array of memory cells made from floating gate transistors. Insulated by an oxide layer, there are two gates, the control gate and the floating gate [1].

Fig. 3-Flash memory module [1]

III. CURRENT CPU STRUCTURE

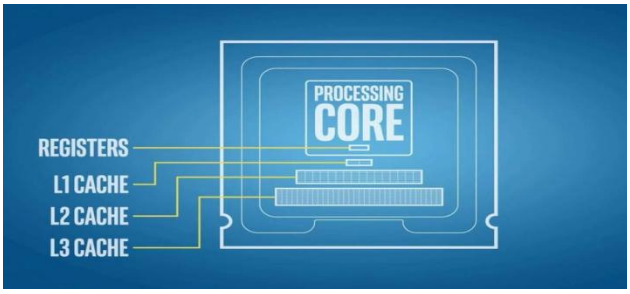


Fig. 4- CPU internal memory hierarchy [4]

In a computer, a register is one of a small set of data holding places that are part of a computer processor. A register may hold a computer instruction, a storage address, or any kind of data (such as a bit sequence or individual characters). Some instructions specify registers as part of the instruction.

L1, L2 and L3 caches are different memory pools similar to the RAM in a computer. They were built in to decrease the time taken to access data by the processor. The time taken is called as latency time.

From registers in processing core, there are 3 levels of cache in a CPU die itself (i.e. L1 Cache, L2 Cache and L3 Cache). These 3 caches are present in an orderly manner such that the cache that is closer to the registers is more expensive and smaller in size. As we move away from the register in a CPU, the cache gets bigger and comparatively slower [3].

Comparison Scale

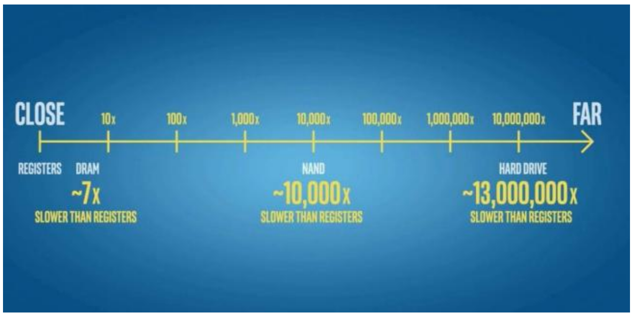


Fig. 5- Size and speed comparison of various kinds of memory [4]

Looking at the comparison scale, we can see that no memory works as fast as registers. When we get off the chip, we have DRAM that is 7 times slower than registers (but volatile memory). Then we have NAND SSDs which are 10,000 times slower than registers, and finally there are hard drives that are nowhere near fast enough being 13,000,000 times slower.

Our main objective is to get the data as close to the CPU as we possibly can, and if we can get that data to be faster, then the CPU waits less for that data. The perfect memory would be inexpensive, fast and non-volatile i.e. it will be able to retain data without power [3].

IV. PAVING THE WAY FOR 3D XPOINT

A. Big Data and its need

What is Big Data? Big data is a term that describes the large volume of data – both structured and unstructured – that inundates a business on a day-to-day basis. Big data helps us make decisions in real time businesses by collecting the humongous amount of data and then analysing it.

Why Is Big Data Important? The importance of big data doesn't revolve around how much data you have, but what you do with it. You can take data from any source and analyze it to find answers that enable:

- Cost reductions
- Time reductions
- New product development and optimized offerings
- Smart decision making. [6]

B. Big Data: Upcoming problems

1) The Storage Problem:

Intel and Micron emphasized that in terms of data, time is not necessarily on our side, thus driving the need for 3D XPoint. By 2020, we're going to generate another 44 Zettabytes of data. A Zettabyte comprises 1000 exabytes, and a single exabyte can hold 36,000 years' worth of HD video

Hence, we need a memory that helps us store all of this data [5].

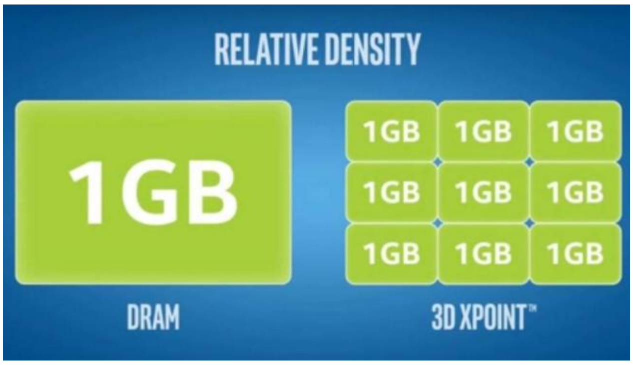


Fig. 6- Deployment size of 3D XPoint in comparison to DRAM [3]

3D XPOINT has a simple, stackable and transistor-less design which enables it to pack up to 10 times more capacity than DRAM in the same space, greatly reducing costs. When it comes to memory performance, low latency is the key. NAND latency is measured in 10s of microseconds whereas 3D XPOINT technology's latency is measured in 10s of nanoseconds. [

2) The processing problem:

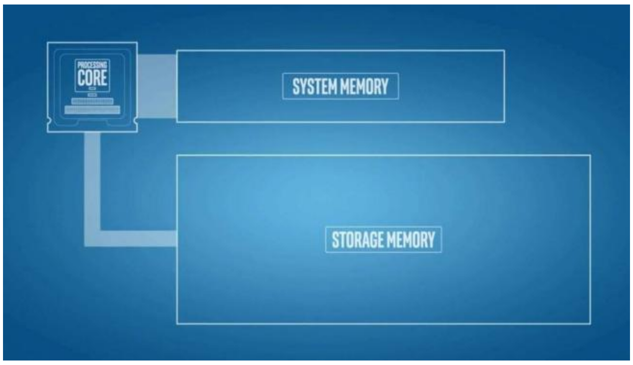


Fig. 7- Current CPU Processor Deployment [3]

When large amounts of data had to be processed, a traditional processing unit had separate segments for system memory and storage memory which comparably slow when interacting with the registers of the processor. This also increased the wait time for CPU and hence made all the processes slower especially when large amounts of data were to be processed.

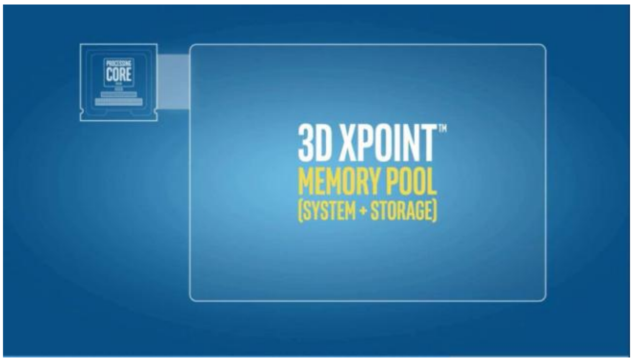


Fig. 8- 3D XPoint Processor Deployment [3]

3D XPOINT serves as a single high speed, high capacity pool of system and storage memory. It takes the best characteristics of DRAM (i.e. speed) and solid state drives (i.e. cost effectiveness and non-volatile) and concatenates the multiple layers of the memory hierarchy into essentially just one layer which allows the computer to process Big Data in a very small amount of time i.e. 10s of nanoseconds. We have something that's big enough and cheap enough to use as storage, and something that's fast enough to use as memory [3].

GPGPU- GPUs, for many years, were just used to accelerate some stages of the graphics rendering pipeline, which helps display the triangles onto screen. However, after the programmability available on this chip, GPU opens a door to developers to take advantage of its ALUs besides graphics processing. Compared with CPU in their architectures, GPU is more suitable for stream computations, it can process data elements with parallel processing. And in many cases, people gain great performance improvement on GPU over CPU. So a need arises to use GPGPUs when we are dealing with humongous amounts of data [2].

V. 3D XPOINT-STRUCTURE AND ARCHITECTURE

1. 3DX POINT is constructed by packing lots of capacity into a tiny footprint. We started by slicing sub-microscopic layers of materials into columns each containing a memory cell and a selector.

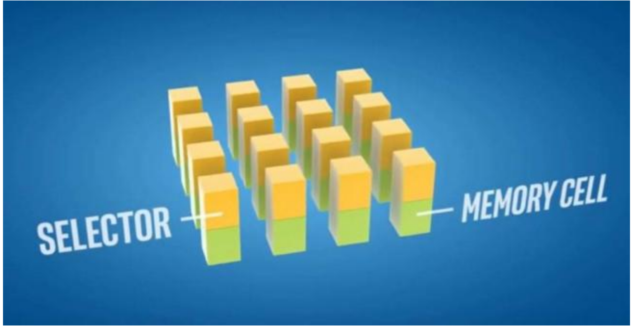


Fig. 9- Selectors and memory cell of 3D XPoint [3]

2. Then we connected those columns using an innovative cross point structure consisting of perpendicular wires that enables memory cells to be individually addressed by selecting one wire on top and another at the bottom.

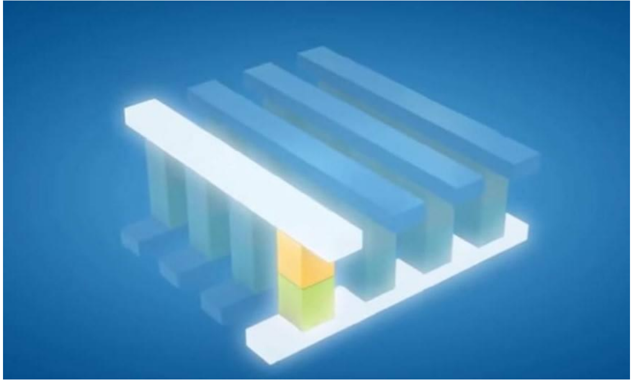


Fig. 10- Perpendicular Wires used in 3D XPoint [3]

3. We can stack these memory bridges three-dimensionally to maximise density and whereas DRAM requires a transistor at each memory cell to access or modify the cell making DRAM big and expensive. Each 3D XPOINT memory cell can be written to or read by simply varying the voltage sent to its selector completely eliminating the need for transistors.

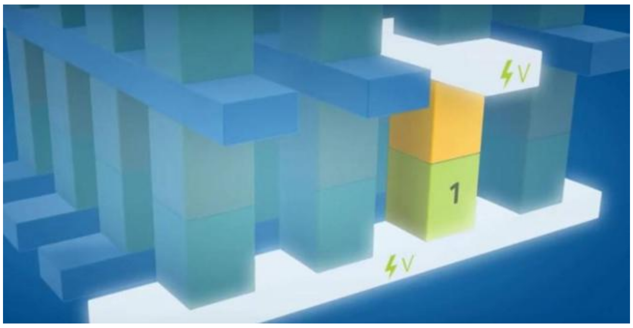


Fig. 11- Voltage transition in 3D XPoint [3]

Final cross point structure

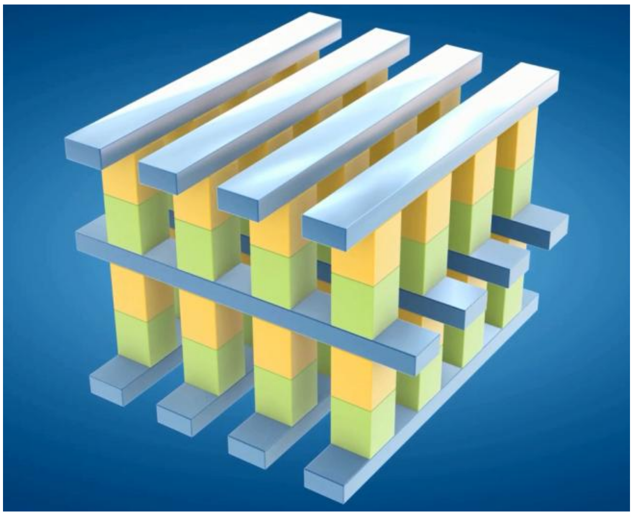
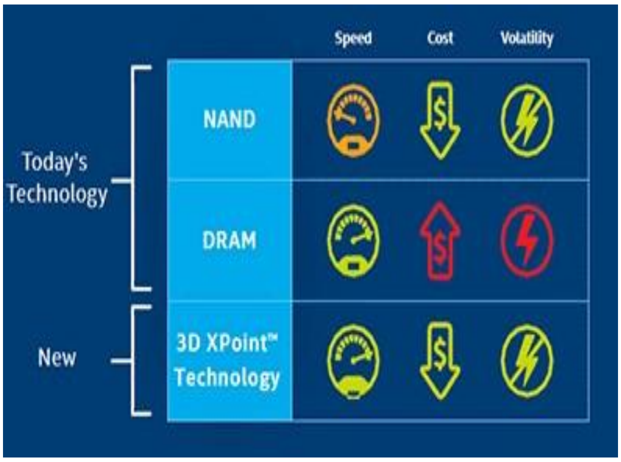


Fig. 12- Final 3D XPoint Structure [3]

Features

1. Non-volatile: The data isn't lost when the power is turned off making it a great choice of storage.
2. High endurance: It is not significantly impacted by number of write cycles it can endure, making it more durable.
3. Stackable: These thin layers of memory can be stacked to further boost density.
4. Low latency: The latency time is significantly reduced due to its structure [3], [4].

VI. CONCLUSION



		Speed	Cost	Volatility
Today's Technology	NAND	⏲	↓$	⚡
	DRAM	⏲	↑$	⚡
New	3D XPoint™ Technology	⏲	↓$	⚡

Fig.13- Comparison table for NAND, DRAM and 3D XPoint [5]

Today's existing memory technologies of NAND flash memory and DRAM have their various drawbacks. While NAND memory is inexpensive and volatile, it is not viable for processing of Big Data whereas DRAM offers incredible speed but it will cost an unreasonable amount of money apart from being volatile.

The attributes of having a low-cost memory comparable to NAND with performance attributes of DRAM, being non-volatile at the same time, sums up to being the holy grail of the memory industry, looking for the universal memory.

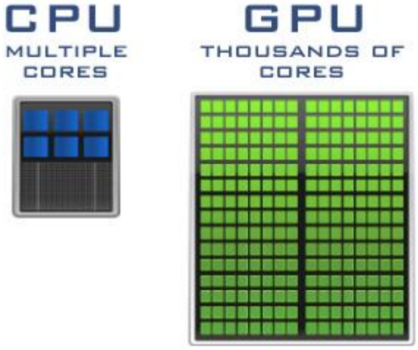


Fig. 14- Comparison diagram for cores of a CPU and GPU [2]

The GPU has thousands of cores in it and has a potential for processing data much faster than a CPU. The only problem is the absence of a programming language that can be used to process all types of commands. A GPGPU uses pipeline flow of commands and parallel processing with other GPUs to carry out a special set of operation, that too at a much faster rate than the current CPU.

Through our Review Paper, we have tried to put light on the effectiveness of 3D XPOINT as a future memory technology and a possible game changer in the industry. We have also elaborated on the use of high performance GPGPUs for cutting through huge clusters of data and processing it.

We believe if these two technologies handle the problem of Big Data so well individually, if they are simultaneously implemented with the invention of a programming language, this could well be the denouement of the complication of Big Data, that is so necessary in today's world.

ACKNOWLEDGEMENTS

This work has been supported by Northern India Engineering College, GGSIPU specially the entire CSE Deaprtment. We would like to extend our special thanks to Professor Saurabh Gupta for supporting the work and for his able guidance.

REFERENCES

[1] Journey of memory from RAM (1947) to 3D XPoint (2015)- An Overview: http://www.ijarcsse.com/docs/papers/Volume_5/9_September2015/V5I9-0172.pdf

[2] Emerging Technology about GPGPU (2015)- University of Macau- https://pdfs.semanticscholar.org/b44e/213be29c9c4bfaea500a5f2a5ff2a555f2bd.pdf

[3] "3D XPoint™ Technology Revolutionizes Storage Memory", www.youtube.com (video, infomercial), Intel

[4] Official Intel unveil of 3D XPoint- http://www.intel.in/content/www/in/en/architecture-and-technology/3d-xpoint-unveiled-video.html

[5] Intel and Micron jointly unveil disruptive, game changing 3D XPoint Memory- Jason Evangelho- http://hothardware.com/news/intel-and-micron-jointly-drop-disruptive-game-changing-3d-xpoint-cross-point-memory-1000x-faster-than-nand

[6] Big Data: What it is and why it matters?- Thomas H. Davenport: http://hothardware.com/news/intel-and-micron-jointly-drop-disruptive-game-changing-3d-xpoint-cross-point-memory-1000x-faster-than-nand