The Art of Processing

by Rao Saifullah

In the recent past, there was a time when people used to keep personal information like phone numbers, addresses, DOB of their and also their family and friends in their mind. But soon they realized that this information is taking a lot of time when they have had to remember them again. So they thought how about designing a machine. A machine which will not only help them to save this information, but also to solve the complex problems. During the Second World War, they designed a machine, a device which helped them to break the code of their rival armies. at the end of the Second World War, the scientists realized the importance of that machine and they thought that if this machine will be use in the domestic use, it will bring a revolution for our life. So in the 1950's they made the simple and the basic form of a machine and named it Computer.

The computer which we are using these days, it was started back in 1950. So the computer is usually categorized in 5 generations, first, second, third, fourth and fifth. The first and second generation computers were so enormous and huge that they took the entire room. Not only they took the entire room, they also used a lot of electricity, generated a lot of heat. They were expensive and they were very difficult to move from one place to another. In 1960's a very small miniature electronic device was developed, named an “integrated circuit.” The size of the integrated circuit helped the scientist to scaled down the computer size from 1000's to the 10's . Not only scaled down the size, it also helped the scientists to make more efficient and helpful computer. The price was decreased, the performance was increased, good for everyone of us. Then after twentieth century, a new field came, a field which is still helping us, known as “artificial intelligence.” Anything which can be done without the external or manual help from human being is come under the umbrella of artificial intelligence. Like voice recognition, face recognition. every type of security we are doing comes in artificial intelligence. And this field is still expanding a lot.

These all are machines, we all are humans. we are controlled by brain these machines are also have their own brains but we don't call them brains, we call them processors or cores. A processor controls the machine, sends the information and takes back the result. Similarly we get the information from our brain and we do a specific task. So thing is that, if the machine is taking a lot of time, what is the solution? Let's take one example here. An example of a shopping mall. We all love shopping, right? If we go to a shopping mall and there is a long queue and we have to wait to entertain by the cashier or from sales person, we all are feeling annoyed, right? If there is only a counter and we did some shopping and we are at the hundredth place at the last of the queue then it will take 100s seconds until we will be entertained, right? This is called serially, one by one. So the scientists thought about this problem because they faced this problem in their

computer also. With CPU you cannot increase the size, and you can't increase design so what is the solution?

So the scientist thought about that. Is that the end of the road, is this the limitation, the technology is facing today. So they thought about this and they came with a solution. A very easy solution. How about using multiple brains? Like if one brain is taking a lot of time, how about increasing the brain number. Let's go again to the same shopping mall but this time the shopping mall has developed and it has now 10 counters. And each counter has one cashier and you are still at the end of the row but now the row size has shrunk. Why? Because there are multiple counters so the customers can go multiply on different counters, right? It takes only 10 seconds. If there are 10 counters, 100 peoples, it will scale down to 10 seconds. So using multiple brains, but the point here is that can we use multiple brains in the CPU? The easy answer is “No” because there are many limitations, like heat, electricity, and design there are many limitations so the scientist made a new thing and named it GPU, “graphical processing unit.”

Any process which takes a lot of time, the CPU send that information to GPU to do it. Similarly a CPU has 10s of cores or processor where GPU has 1000s of cores, or 100s of cores. So you can see if you have 1000s of brains you can do everything very easily and conformably within few seconds. Similarly CPU has higher clock speed whereas GPU has higher transitory counts. Here is the simplest and basic diagram of GPU which is showing you that GPU has blocks. In GPU we call them blocks and but in CPU we call them processor or core. If you see here there are multiple or 1000s of cores and every core can do information separately and independently from other core.

Let's take one simple example and do it in a CPU way. In a serial way. Let's suppose I have three arithmetic problems and I want to solve them. How it will be done? if I have one resource let say I have one student and I ask him to do this. He will do something like this first this problem, then this problem, and then the last problem, right? If it takes one second to solve a single problem it took three seconds, right? How about doing it the GPU way? I have the same problems and I want to do it in less time. How can I do it? I will use three students. Resources increased, the time decreased. There are many applications where resources don't matter compared to time. We have to decrease the time in some applications. So the same thing is done more easily.

Here is GPU internal structure, GPU doesn't work independently. GPU works side by side the CPU or you can say the CPU controls how and when the GPU will work. Similarly this is CPU and GPU. GPU internal structure. For clear understanding of the GPU internal structure, let’s think about school. A simple school. School has students, GPU has threads. Students combined together makes one class right? Threads combine together make one block. Blocks combine together make a grid. Similarly, in a school all the classes combine together make one school. School is grid, school is grid, block is class and student is thread. It is like a hierarchy. It works like this. How the information is processed? CPU sends a command to GPU memory. Memory

sends it to thread to do the processing and sends the data back to the CPU. Information given and result received. This is how a GPU works.

This time, let's take one more complex problem. Adding two 3 by 8 matrices. We have two matrices and I have to add them element by element. If I will do it in a CPU way, it will be done something like this. First element will be added, a second element will be added. It will go all around the way to come to end and it will take approximately 24 seconds to solve this in a serial way or in a CPU way. Now let's do it in a GPU way or we can call it in a parallel way, right? Same matrices and now I have to add them. How will it be done? It will be done something like this. It didn't have to wait like serially. It will just overlap and do everything at the same time and we will get our result.

So here as you saw in the my presentation and also explained like GPU is very helpful. So the question comes in our mind instantly, “What is the future of the GPU?” In my opinion, the GPU has a very bright future. Why? These days we all are surrounded by electronic gadgets. We are using mobile phone, we are using laptops, we are using computers everywhere and we have to process that information. But as you saw in the presentation, CPU sometimes takes a lot of time. But if we transfer this information to a GPU, time will be saved. Like in our automobile or aerospace engineering. Like the Curiosity rover on Mars. Everywhere they all are doing some processing. But the thing is that sometimes they take time because the distance is large and information is very complex. If we transfer those information to a GPU, time will be saved. We all are used to playing games, we all are playing a lot of games. If you see the graphics on mobile games, it's very good these days but it was not used to be good. How and why? This is done with the help of GPU. So in my opinion, whatever the future is coming and we will face that future, the GPU will be in it.