Our group topic is in-memory processing. In this topic we divide to three parts to explained it. First, let me go through the introduction part.

In the usual way, we usually moving data from the memory/storage unit to the computation unit. Every time, it will move data through the DRAM bus. Plus, DRAM access is taking long time. After the CPU gave the data and the control signal, it should delay long time for data input and output. Moreover, in this time the CPU cannot do other work. Also, the bulk copy and initialize need to use cache to completely the duty. As you can see in this picture, when passing through the data from computing unit to DRAM takes 137 nanojoule. Almost 62.7% of power usage in moving data.

This is a picture from \_\_\_\_\_\_ As you can see, the retiring and front-end bound is necessary processing which takes less than 40% performance cost. And the bad speculation only takes less than 10%. The most important part is back-end bound, this one is more than 50 percent.

So nowadays, basic on Von Neumann structure, It has these way. However, today we want to talk about the in-memory processing. The main idea of in memory processing is \_\_\_\_\_\_. It can be used in the Deep learning, and so on. We also can use GPU to achieve that.ow HHhhh

Actually, in past of years, many people consider about in-memory processing. However, they did not actually achieve that. Till today, this concept has been came up today. Here is two approaches, we want to introduce about in memory processing.

Let’s look at about the current design and compare it to final requirement. In the final requirement, we want high performance, Energy effect and sustainable system. However, currently, the energy wasted.

So here is the Von Neumann basic structure.it contains three part：computing unit, communication unit and memory/storage unit. In the past of years, the designer put more focused on CPU not memory part. The designer raised the speed of internal transistor switching to produce the fast logic. However, they only require the memory/storage unit has huge capability. This caused the speed of accessing memory is lower than the speed of data processing. As the result of this, most of time is the processor is waiting for the data come back to the memory.

Therefore, it has an imbalance system and overly complex processor. they need to tolerate the data from the memory, which is long and slow

Nowadays, it has some improve technology, however, they are nothing different. They all need to wait the memory. It turns to the basic problem which is processing of data is performed far away from the data

Moreover, nowadays, the artificial intelligence appeared giving a chance to in memory processing. The in-memory processing starts to be built. The artificial intelligence needs to process the huge data. If under the Von Neumann structure, moving data from memory union to calculator union frequently will reduce the calculator energy of convolution network. And in training part of the artificial intelligence, it needs a lot

of regular calculator, for example, multiply and adding operation. In usually way, it will add quantity of parallel computing unit. However, in the same time, every unit get themselves capacity decreasing. In here, the memory became the biggest problem of dealing artificial intelligence ‘s data, which can transfer to power usage and the bandwidth. To solve this problem, you can use more GPU or using in memory processing.