Brief Talk about RAM Technology and Development Prospect

Weike Wu

G33281147

Abstract—In order to explore the development bottleneck and prospect of RAM technology, this paper consults and summarize the history and working principle of RAM as theoretical foundation. After exploring, the main development bottleneck of RAM is memory wall, so RAMCloud could be a possible way to solve this problem.

Index Terms — RAM history, memory wall, RAMCloud

I. INTRODUCTION

RAM (Random-access memory) is an internal storage used to exchange data with CPU directly. It can be read and written by bus addressing at any time with high speed. RAM is the bridge between CPU and external storage such as drives and used as the temporary data storage component for operation system and other programs, which means all programs are running in RAM. Therefore, the performance of RAM will influence the computer greatly. The data in RAM storage cell can be loaded and saved at any time when necessary and the data will lose after power off. According to different work principle of storage cell, RAM has two forms, static RAM and Dynamic RAM.

The industrial pattern that CPU makers and RAM makers are separated leads the result that the development of RAM cannot follow the development of CPU. In the past 20 years and over, the performance of CPU promotes about 55% every year. However, the performance of RAM only promotes about 10% every year. The imbalance development speed leads result that the read and written speed of RAM is far lower than the computing speed of CPU. This makes the high performance CPU cannot be fully used when working and has greatly limited the increasing of high performance computing (HPC). Because RAM is the only component which can exchange data with CPU directly, the low performance of RAM is just like a wall limits the performance of CPU. Therefore, people call the bottleneck of RAM as “Memory Wall”.

This passage reviews the development history and summarizes working principle and bottleneck of RAM technology. Then explores some possible ways to break the bottleneck based on newest research and looks forwards to the prospect of RAM.

II. Development History of RAM

A. The Original of Memory

At the beginning, before 80286 mainboard published, the memory used in PC is just many ICs. ICs are soldered mainboard and their capacity is only 64 ~ 256 KB. Therefore, maintaining memory is a really trouble. After 80286 mainboard published, the demand of memory’s performance become higher for program and hardware. In order to promote the speed and capacity of RAM, designers invented the modularized memory stick and each stick integrated many ICs, which also makes the maintaining of memory become easier.

B. The Initial Memory—SIMM

After 80286 mainboard published, SIMM (Single In-line Memory Modules) interface become be widely used. SIMM's capacity is 30pin, 256KB, and each bank must be composed of 8 data bits and 1 check bit. So in generally, we should use four sticks at same time.

Soon afterwards, during 1988 ~ 1990, PC technology came to a peak of development, which is the time of 386 and 486. At this time, CPU has developed to 16bit, so the low bandwidth of memory becomes the bottleneck of computer's performance. Therefore, 72pin SIMM memory occurs. 72pin SIMM can support 32bit FP DRAM, so the bandwidth of memory promotes greatly. DRAM need constant electric to save data and the data will lose after power off. The refresh rate of FP DRAM can reach hundreds of times per second. However, for FP DRAM use same circuit to load and save data, there will be interval in DRAM's saving and loading time, which leads the speed of saving and loading is not very fast. Moreover, in DRAM, the saving addresses spaces are arranged by page, so changing pages will occupy extra clock cycle of CPU.

C. Remain Stagnant—EDO DRAM

EDO DRAM (Extend Date Out RAM) was popular during 1991 ~1995, which is similar with FP DRAM. Its bandwidth is 32bit and it speed is over 40ns. For EDO DRAM could read next page when sending data to CPU, it eliminated the interval between two periods and the speed promoted 15% ~30% compared with normal DRAM.

During 1991 ~ 1995, the development of RAM almost remained stagnant. At that time, EDO RAM had two forms, 72pin and 168pin. Actually, EDO RAM falls under the category of 72pin SIMM, but it used a new addressing method. According to the improvement of manufacturing process, EDO RAM promoted greatly on cost and capacity.

D. A Classic—SDRAM

After Intel Celeron and AMD K6 and related mainboard published, EDO DRAM again became the limitation of computer performance, so RAM came into the period of SDRAM. The first generation of SDRAM conforms to the specification of PC66, then replaced by PC100 for CPU's external frequency increased to 100MHz. Then PC100 was also replaced by PC133 because PIII and K7 with 133MHz were published. Its bandwidth was 64bit and the read and written speed of SDRAM promoted to above 1GB/sec. For its input and output signal is synchronous with system's external frequency, its speed is far higher than EDO RAM. [1]

E. Highbrow Songs Find Few Singers—Rambus DRAM

For Intel began to develop Pentium 4, SDRAM PC133 again cannot meet the demand of CPU, even its speed is above 1GB/sec. Therefore, Intel combined with Rambus published Rambus DRAM with a brand new RAM architecture. The new architecture was based on RISC (Reduced instruction Set Computing) theory, which made architecture fast and simple and reduced the complexity of data. So it promoted the performance of whole system. It speed can reach about 4.2GB/sec. However, for Rambus DRAM's cost was so high that it was replaced by DDR RAM soon.

F. Classic Again—DDR RAM

DDR SDRAM (Dual Date Rate SDRAM) can be regarded as the improvement version of SDRAM. DDR can transmit data at rising edge of clock and falling edge of clock, which made the DDR’s data transmission speed is double of transmission speed of SDRAM. For it only made another use of falling edge of clock, the power consumption will not change.

G. Star of Nowadays—DDR2 & DDR3

DDR2 SDRAM is the new generation of RAM technology developed by JEDEC. DDR2 also use the method of transmitting data at rising edge of clock and falling edge of clock, but its ability of preload is double of preload ability of DDR (4bit data preload, DDR is 2bit data preload). Which means in DDR2, every clock can read and write data with the speed as four times as external bus speed and running with the speed as four times as internal controlling bus.

DDR3 is the most popular RAM in nowadays. For DDR3 adds another 4bit burst chop based on DDR2, its preload is 8bit, which means its speed is as twice as DDR2’s speed and this is also the most important improvement compared with DDR2.

Fig.1. DDR3 RAM

III. Working Principle of RAM

A. RAM Addressing

RAM will get the data finding instruction from CPU. When finding data, RAM will first find the abscissa (column address), and then find the ordinate (row address), just like drawing a cross on the map to find the position. For computer system, it will determine whether the address is correct, so it will read the signal of the address; [2] both abscissa and ordinate have their own signals, which are RAS (Row Address Strobe) and CAS (Column Address Strobe) and then will read or write in this address.

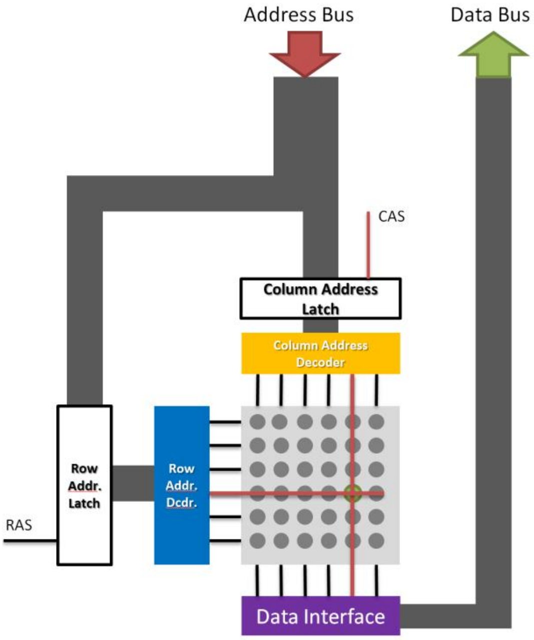

Fig.2. RAM Addressing

B. RAM Transmission

In order to save and load data, CPU will give a particular address when saving and loading data. Then CPU will send the address to RAM by address bus and data bus will send the corresponding data to CPU.

C. Saving and Loading Time

Saving and loading time means the processing time of CPU saving or loading data from RAM, which is also called bus cycle. For instance of loading, when CPU send instruction to RAM, it will ask for a particular data in a particular address, RAM will response and send the data to CPU. Until CPU gets the data, a process of loading is finished. Therefore, the whole process is the CPU send the instruction then RAM responses the instruction and send data to CPU. The whole time will be finished in few nanoseconds. [3] For example, if the cycle finished in 6ns, the RAM’s frequency is 1/6ns=166MHz. For DDR is 2-bit preload, its frequency will be regarded as 333MHz. Similarly, if the RAM is DDR2, its frequency will be regarded as 667MHz. [4]

D. RAM Latency

The RAM latency time (as called incubation period) is the synthesis of following items: latency time between FSB and mainboard chipset (±1 clock cycle), latency time between mainboard chipset and DRAM (±1 clock cycle), latency time form RAS to CAS: RAS latency time (2 ~ 3 clock cycle, depending on row address), CAS latency time (2 ~ 3 clock cycle, depending on column address), one clock cycle use to transmit data and the latency time of data transmitting from DRAM output cache through mainboard shipset to CPU (±2 clock cycle). [5] In generally, the latency of RAM relates to four parameters: CAS latency, RAS-to-CAS latency, RAS precharge latency and act-to-precharge latency. [6] CAS latency is more important in four parameters for it reflects the latency of process from RAM receiving data to RAM finishing data transmission. For example, if the latency of one RAM is 3-3-3-6, the first parameter is CAS latency (CL=3). [7]

IV. The Bottleneck of RAM Development—Memory Wall

Memory wall is the increasing distance between CPU and memory’s speed. The main factor leads memory wall is the limitation of bandwidth. From 1986 ~ 2000, the speed of CPU increase about 55% every year. However, the speed of memory increases only about 10% every year. Therefore, it was expected that the latency of memory will become the overwhelming bottleneck of computer’s performance. Now the increasing rate CPU’s speed has already lower than before because the physical limitation and in some degree, current CPU design has already hit the memory wall.

One possible way to break the memory wall and promote performance of computer is multicore parallel computing technology for computers will have higher ability of multitask processing with more cores. However, some researches show that except the problem of the efficiency of assigning tasks to multicore, multicore parallel computing even brings more serious memory wall. This is because under the processing mode of multicore parallel computing, many cores will share the limit bandwidth and this will cause larger latency. Just like there are four lanes in a high way, but there are more than four cars want to travel parallel, which will cause road congestion and the speed of cars will be slower.

According to SNL's (Sandia National Laboratories) research of simulation experiment of multicore CPU' performance, researchers show that in information science field, more cores for CPU doesn't mean the higher performance. The result shows that because the existing of memory wall, when the number of core exceeds eight, the performance of CPU almost has no promotion. When the number of cores exceeds 16, the performance of CPU will decrease. [8]

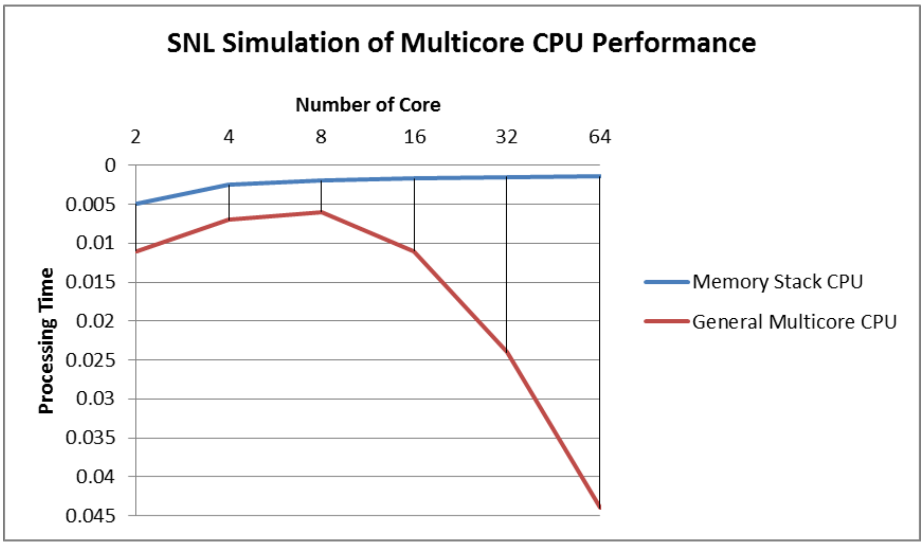

Fig.3. SNL Simulation of Multicore CPU Performance

V. Two Possible Ways to Decrease the Effect of Memory Wall

The main parameters represents the performance of RAM are bandwidth and latency, so we should focus on these two aspects to find the way to break memory wall.

A. RAM's Bandwidth and Promotion Technology

RAM's bandwidth is the amount of data that RAM can transmit through bus in unit time, which can be calculated by the formula "bandwidth= RAM core frequency*RAM bus bit*multiplication ratio/8” and its unit is Byte/s. Bus bit means the number of bit in RAM data bus, core frequency means the clock frequency in RAM and multiplication ratio means the times of data transmission in one clock pulse cycle of every RAM data line. Obviously, the basic way to improve the bandwidth of RAM is to improve the three factors determine the bandwidth of RAM in formula, which are RAM core frequency, RAM bus bit and multiplication ratio.

Improve memory bus bit: Under the existing architecture of independent memory chips, the further increasing of memory bus bit is limited by memory chip data cable pin number, so increasing the memory bus bit to promote bandwidth need to use new memory architecture which can eliminate this limitation. For example of widespread concerned "Memory and processor integrated" technology, it has characteristics of using the method of increasing memory bit to significantly increase memory bandwidth.

Improve memory working frequency: relying solely on improving the working frequency to improve memory bandwidth method will be restricted by memory chip heat and other aspects, so use this method to further improve memory bandwidth is limited.

Improve memory multiplication rate: the method by increasing the transmission rate to increase memory bandwidth is familiar to us. Such as DDR memory is double data rate (Double Data Rate), which each data line can preload data from the storage for 2bit, and 1 data were rising and falling edges of the clock pulse for each transmission, that is, the transmission rate is 2 in one clock cycles. So in the same data transfer frequency, the speed of DDR is as twice as the speed of memory. Similarly, DDR2 memory and DDR3’s memory transfer rate were 4 and 8, and Rambus's "terabyte bandwidth" technology transfer rate can be increased to 32, which greatly increases memory bandwidth.

B. RAM Latency and Its Hiding Technology

RAM latency, which is the waiting time from CPU sending request to RAM to RAM sending data to CPU, in generally uses “ns” as its unit. Compared with the high processing performance of CPU, RAM latency is too much longer. In nowadays, RAM technology cannot decrease its latency greatly and fundamentally, so using cache and parallel computing technology is still the efficient way to decrease the effect of memory wall.

Memory latency hiding technology: data and program code may be accessed by the processor before saved to the cache, the processor will minimize direct access to memory, but to get data from the high-speed cache, which is a typical memory latency hiding technology. This method is based on conventional technology caching mechanism and has greatly reduced "memory wall".

Hardware support for parallel processing technology: Although supported by the hardware multithreading, order execution and other parallel processing technology does not directly address the "memory wall" problem, but multi-threading and parallel processing mechanism out of order execution, can be more effective in reducing the processing tasks during situations processor resources are idle, when the processor is handling a large number of tasks that "resources are not idle" cumulative effect produced, can make a significant increase throughput, so the overall processing efficiency correspondingly somewhat lifting, which to some extent on the impact shield "memory wall".

VI. How to Decrease the RAM Latency in Reality

For it is difficult to promote the RAM working frequency, we focus on how to optimize the use of RAM to decrease the RAM latency. According to researches, there are several aspects which should be focused to decrease RAM latency: cache, preload, multithreading and out-of-order execution. Moreover, CPU integrated RAM controller will also decrease the RAM latency.

A. Multi-level Cache and Preloading Technology

When CPU needs to load data, it will try to find the data in cache first. If CPU can find the data, it will load the data to core to process, which is called “cache hit”. The ration of the time of hit and the whole time of loading data is called hit rate. [9] Therefore, the key of using cache to decrease RAM latency is to increase the hit rate. In nowadays, the trend of multicore CPU cache development is using higher capacity and more level architecture of cache and cache management and data preloading technology with higher efficiency. [10]

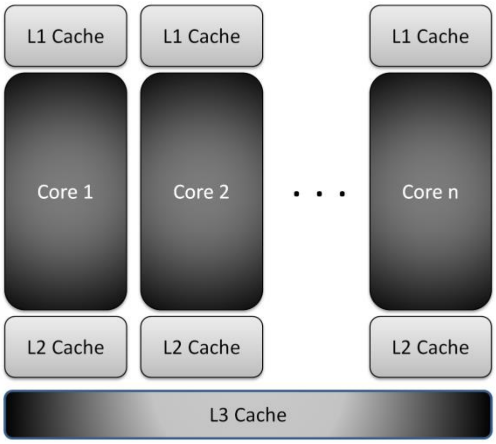

Fig.4. Three-level Cache in Multicore CPU

Enhanced high-capacity multi-level cache design: Increasing cache size can obviously improve the cache hit rate. And increasing the cache hierarchy is another way to improve the cache hit ratio, such a two-stage cache L1, L 2 theory hits rate is about 80% , so the total level cache hit rate is about (80 + 0.2 × 80)% = 96%. According to this, more levels means higher hit rate. So Corei7 and other next-generation quad-core Phenom II processors have adopted large-capacity three-level cache architecture, each processor core has a separate L1, L2 level cache and a shared L3 cache capacity. PhenomⅡ L3 cache capacity is 6MB, Corei7 L3 cache capacity even reached 8MB. [11]

More efficient cache management technique: For multi-core processors cache system, a more efficient cache management technique is also very important. E.g. Corei7 using ASC (Advanced Smart Cache) technology can be dynamically allocated shared L3 cache in accordance with various cores processing load, thereby increasing the respective core read data from a shared L3 in efficiency. ASC of shared L3 cache using the inclusive mechanism, which means L1, L2 of the data are included in the L3.

When a core access L3 did not hit, it will immediately turn to access the memory read data, thus effectively reducing traffic continue to find listeners L1, L2 and the time latency. ASC also uses a mechanism based on L3 inclusive SF (Snoop Filter) technology. When processor hit in L3, it will not read data from L3, but read data from L1 and L2 which decrease the time of CPU getting the data.

More effective data prefetching techniques: Which means the mechanism based on certain prediction, most likely to use the processor reads data from memory to pre-cache to improve the cache hit rate. Data prefetching can be built into the processor prefetcher unit to achieve hardware prefetch mode.

B. Multithread and Out-of-Order Execution Technology

HT (hyper-threading), which is also called SMT (simultaneous multi-threading). HT is begining from Pentium 4 CPU. For the register and many other components has double in CPU core, we can add a thread control unit in CPU which can assign two threads' instruction sequence into double components. This equals staring two threads in same time. When one thread sleeps for waiting data, another thread can start immediately; this avoids the waste of CPU resources and increases the efficiency of CPU.

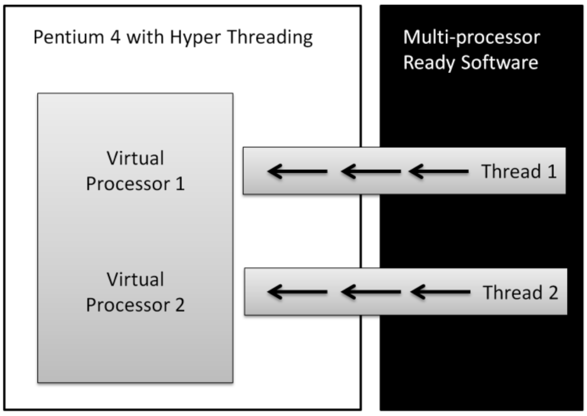
Pentium 4 with Hyper Threading
Multi-processor Ready Software
Virtual Processor 1
Virtual Processor 2
Thread 1
Thread 2

Fig.5. The Essential of HT is Reduce RAM Latency

OOOE (Out-of-order Execution) is a design of instruction set parallel computing of CPU. CPU supports OOOE can execute the instruction without the original order given by program. It will monitor and analysis which processor unit will be unused and which instruction can be executed without order by OOOE, then assign these instructions to unused processor unit to execute and order and return the result in the original order. OOOE uses parallel processing avoids the waste of CPU's resources and increase the efficiency of CPU with HT.

C. Integrated Memory Controller

IMC (Integrated Memory Controller) is firstly used in AMD Athlon in commercial CPU. It integrates the memory controller originally integrated in north bridge chipsets in CPU chip. Therefore, it shortens the physical route when CPU visiting RAM and reduce the latency of RAM loading. It can also make memory controller and CPU running in same frequency, which can reduce the waiting time greatly when CPU visiting RAM's data.

In Phenom II CPU, it integrates the memory controller of dual channel DDR2 1066 and DDR3 1333. Later, Core i7 firstly integrates the controller of triple channel DDR3 1333. Combined with AMD HyperTransport 3.0 and Intel QuickPath Interconnect (QPI) bus technology, IMC can greatly promote cores' efficiency to save and load data from RAM.

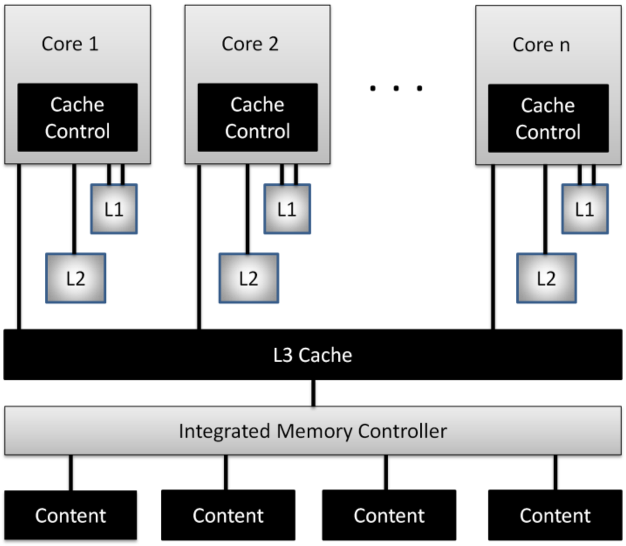

Fig.6. CPU with Integrated Memory Controller

VII. Ways to Promote Bandwidth

Besides using multi-level cache and parallel processing technology to reduce RAM latency, another way to reduce the influence of memory wall is to promote RAM bandwidth to decrease the distance between performance of RAM and CPU. When RAM meet the bottleneck of heating and craftsmanship when promoting RAM's frequency, we should consider other ways to promote RAM's bandwidth, such as increasing RAM transmission efficiency, transmission bit and transmission multiplication ratio.

A. Technology of Using Buffer to Promote RAM's Performance

FDB (Fully-Buffered DIMM) is an efficient way to expand the bandwidth and capacity of DDR2 and DDR3 and its key component is the AMB (Advanced Memory Buffer) invented by Intel. Under FDB RAM architecture, RAM doesn't exchange data with CPU directly; it will be buffered processed in AMB. AMB general memory interfaces can be used to connect DDR2/DDR3. The method used to connect AMB and memory controller is P2P high-speed serial connection replacing the general shared-parallel connection, which means every AMB of RAM inserted in DIMM slot will be cascaded and uses P2P transmission method, so the data will be transmitted by buffer one by one. The advantage of cascading connection is it has stable P2P connection impedance which improves the stability and reliability of signal transmission. [12]

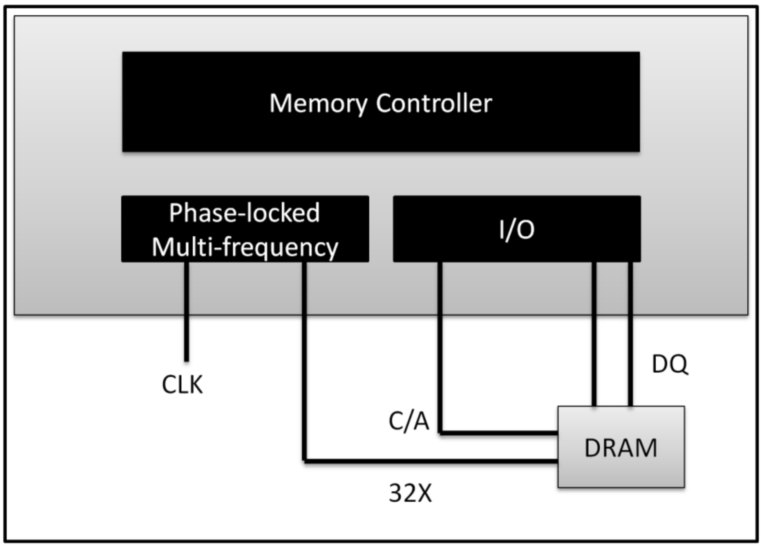

Fig.7. General Parallel Connection and FDB Serial Connection

Each way of AMB serial connection uses SC (Self-Clockin) transmission technology, so the receiver clock frequency can be adjusted automatically according to the size of the data stream, which effectively increases the speed of data transmission and increases transmission bandwidth by 4 times. Because using FBD serial transmission mode, memory can use fewer pins to build more memory channels, and each channel can be connected to a larger capacity memory module. FBD total capacity can be expanded to 384GB. However, due to the higher power consumption and cost of FBD memory, they are mainly used in servers, workstations and high performance computers.

The development trend of FBD memory buffer technology is to abolish the current combination of AMB and memory, but to place AMB chip directly on the motherboard, or use the form of inserting card, which is called BOB (Buffer On Board) technology. In addition, AMD has also developed G3MX (G3 Memory Extender) technology and IMB (Isolation Memory Buffer, IMB) of Inphi of is also a memory buffer technology.

B. Increase Transmission Rate to Promote Memory Performance

By increasing the memory transfer rate to improve memory bandwidth is a major trend in technology development. For example, the technology to enhance the memory transfer rate from the DDR, DDR2 2,4 times the transfer rate to 8 times the transfer rate of DDR3. Many other technologies can also promote the transfer rate, for example, QBM (Quad Band Memory) technology is the use of field-effect tube so that DDR memory can transmit four times in one clock cycle, which will increase memory bandwidth 4 times. Rambus launched the plan named TBI (Terabyte Bandwidth Initiative) in recent years is being highly concered. TBI's goal is to develop one million mega-bandwidth single-chip memory control system. Rambus has developed the 32X Data Rate, FDMA (Fully Differential Memory Architecture), Flex Link scalable links and many other key technologies. The new memory-based and TBI technology systems for high-performance multi-core processors and graphics, games and other applications to provide a complete and effective solution. According to Rambus's plan, based on first-generation technology TBI XDR2 SoC chip, enabling chip DRAM memory bandwidth of 38.4 GB /s, 2nd generation XDR2 SoC chip-chip DRAM bandwidth can reach 51.2GB / s, further goal is to achieve a memory controller connected to 16 pieces DRAM memory chips, so that the total bandwidth will reach to one million megabytes per second, which is 1024GB /s or 1TB /s.

The 32-time data transmission rate technology uses high accurate phase locked loop frequency multiplier circuit, which will convert clock signal to an internal clock signal, so that the memory data channels in an input clock signal period within the 32-bit data transfer to the memory buffer, is a new 32-bit prefetch buffer architecture, with 500 MHz clock frequency can reach 16 GB/s transfer rate, and therefore 32 times the data rate technology will greatly enhance the memory bandwidth. FDMA technology utilizes a strong anti-jamming performance of the differential signal transmission, data, command and address signals to provide a more reliable and more efficient transmission channel between the memory controller and memory in. Flex Link is an innovative high-speed scalable point-style command and address link architecture, two-way link connection speed can reach 16GB/s, which simplifies the connection of memory and memory controller, and easy to expand the memory capacity.

V.III Advanced Memory and Processor Integration Technology

Compared to the technology simply lower memory latency and improve the memory bandwidth, we should consider about another way than is working on more advanced memory and processor integration technology. The most representative of such technology is PIM (Processing in Memory). Its basic idea is to integrate the memory and multi-core processors with a chip. In addition, IRAM (Intelligent Random Access Memory) and EDRAM (Embedded Dynamic Random Access Memory) technology are also based on the principle of integrated memory and processor.

The advantages of PIM technology is mainly represented in two aspects: First, it can effectively reduce memory latency, due to the physical distance between each core of processor and memory is significantly shorter and core memory access latency also effectively reduces, which means it change the latency between chips into chip internal latency. Second, it has the potential to promote the memory bandwidth. The traditional stand-alone memory chip memory architecture is difficult to increase memory bit by using more data line pins due to restrictions on the number of pins. However, PIM technology can make the core and memory establish wider data transmission channel in the same chip and there is no limitation of pin. Therefore, it is more easily to improve memory bit to improve memory bandwidth.

Theoretical calculations shows that PIM technology with these characteristics is possible to reduce the reaction time of the memory 5 to 10 times and upgrade the bandwidth 50-100 times; reduce energy consumption by 50% ~ 75%. At present, many manufacturers are developing 3D stacking chip packaging technology to eventually create PIM-based 3D stacked chips. Such as X-Caliber processor developed by Sandia National Laboratories (SNL) can make DRAM memory stacked on the PIM chip of multicore processor logic layer, its performance will increase with the increase of the number of cores. However, the major problem 3D stacked chip packaging technology is the heat problem; it also needs further breakthroughs in way of stack, heat, power and thermal management technology.

IX. The Prospect of Breaking Memory Wall

More Cores≠Higher Performance: Because of the serious impediment due to "memory wall" and other problems on multi-core processor performance, AMD believes simply increasing processor core is not feasible, Intel also said that under existing development environment the performance will not significantly improve when the number of cores exceeds 16. To enhance the performance of the processor, Intel is more concerned about the improvement of processor floating point capability, but it seems AMD is more concerned about GPU-based stream processing technology. Seen in this light, under the situation that we cannot break the memory wall efficiently, it is difficult to promote computer performance by increasing the number of core, so we need to find another way to meet the increasing demand of high performance computing.

"Make everything be multiple": It is reported that the US expert Joseph Ashwood designed a new architecture of the storage system, its most advanced feature is the ability to achieve parallel access and the access speed is significantly higher than the current serial access memory. Because of this new storage system with parallel processing mechanism is fit for multi-core processors; the memory is called as "multi-core memory." Although the "multi-core memory" now only completes the design work and there is still a long way to go, the idea is similar with some researchers' theory of Multi – Everything. Maybe multithread, multicore processor and multicore memory will be the possible ways to solve the problem of memory wall.

Development of next-generation storage technology: developing new architectures and new devices based on the memory and narrowing the gap between memory and processor performance could be an effective way to break the memory wall. For example, some technologies which are under developing such as "phase-change memory" (PCM), "programmable metallization cell memory" (PMCM), "Magnetic Random Access Memory (MRAM)", "Ferroelectric Random Access Memory" (FRAM), "nanotubes Random Access Memory "(NRAM), Memristor and other new non-lose storage technologies have hope to replace the current flash memory. With the development of technology, some new high-speed memory will be possible to eventually replace the current DRAM memory.

Researches on various new computer architectures: Until now, all computer system architectures are based on von Neumann's "stored program principle," so fundamental solution is probably not using "von Neumann" architecture in new computers. For example, some architectures such as "data flow machine" (DFM) and "artificial neural network" (ANN) has abandoned the principle of stored procedures. Therefore, there is no longer "memory wall" problem.

X. The Prospect of RAM—RAMCloud

The fastest storage method in computer systems is RAM, which traditionally mainly used as memory. Because performance of drives is deteriorating over the years and lower cost of RAM, in recent years, many researchers are exploring how to replace the hard drive with RAM.

In the past four decades, the main storage of the computer system is hard disk, file systems and relational databases are developed according to hard disk. However, although the hard drive capacity increase rapidly (more than 1000 times since the mid-1980s), the improvement of hard drive's performance is slow; transfer rate increases only 50 times, the latency only increases 2 times. If you use capacity / bandwidth (Jim Gray's Rule) to measure the disk access latency, it actually deteriorated sharply. [13]

	Mid-1980s	2009	Improvement
Disk capacity	30MB	500GB	16,667x
Maximum transfer rate	2MB/sec	100MB/sec	50x
Latency (seek+rotate)	20ms	10ms	2x
Capacity/bandwidth (large blocks)	15s	5,000s	333x worse
Capacity/bandwidth (1KB blocks)	600s	58days	8,333 worse
Jim Gray’s Rule[12] (1KB blocks)	5 min	30 hours	360x worse

Fig.8. Development of Storage from 1980s to 2009

In order to solve the problem of data access latency, developers and researchers have proposed various solutions: using cache such as memcached, database partition, more use of flash memory to replace hard drives, SSD, using MapReduce, Hadoop and other asynchronous work scheduling , NoSQL, distributed file systems and so on.

Ousterhout team proposes a new solution --RAMCloud, through large-scale cluster ordinary server's memory, the main online data storage center migrates from hard disk to DRAM, and hard disk is only use for backup and archive. This kind of RAM can achieve large-scale cloud and low latency at the same time.

A. RAMCloud Overview

RAMCloud is most suitable for the data center has been classified into application server (mainly used to generate web pages and execute application logic such as business rules) and storage server (long-term shared storage for application server). [14] These data centers generally support many applications; some are small which may only use a part of resources of one server, and some are big which may use thousands of servers.

RAMCloud put all data in DRAM, so the performance can achieve 100 to 1,000 times or even higher than the current highest-performance disk storage system. In terms of access latency, in RAMCloud program, it only need 5~10μs for a process running in application server to read hundreds bytes of data from the same center storage server through network, but current system generally takes 0.5 ~ 10ms, depending on the data whether in server memory cache or in hard disk. Moreover, a multi-core storage server can process at least 1 million small read requests per second. The same machine in hard disk system can only process 1,000 to 10,000 requests. [16]

B. Challenges of RAMCloud

DRAMs are volatile memory, so it is difficult to realize the keeping of data permanently. Therefore, we need to keep data and service when there is server failure and power off. The easiest method is to save multiple copies in different servers, but the cost is too high and the data will still lose after whole data center power off. [17] So how about back up on servers? There will be large latency if we need to refresh drives when write operations, so it will lose the advantage of RAMCloud. Therefore, Ousterhout put forward buffered logging theory. [18]

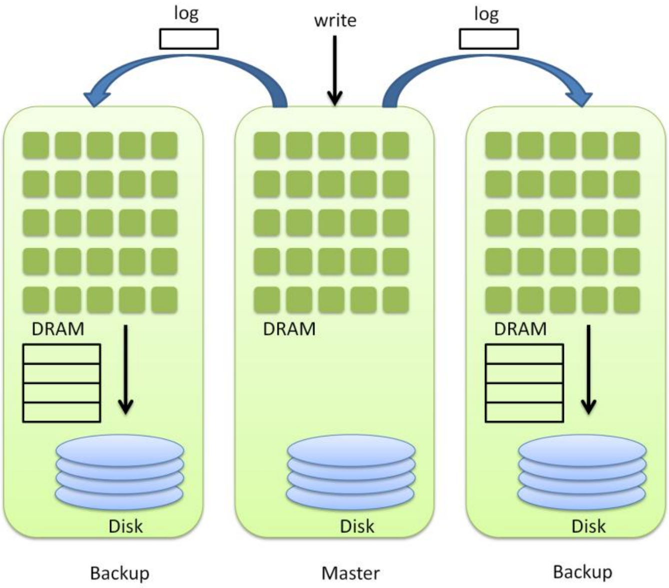

Fig.9. Buffered Logging

C. Comments on RAMCloud

With the cloud computing coming into the mainstream, computing, networking and storage will be provided by few cloud computing services and a serials of highly specialized technologies will also appear. Today, Google, Facebook and other large Internet companies' special technology can be regarded as the beginning of this trend. So, for RAMCloud, nothing is impossible.

Reference

[1] M. Renovell, J. M. Portal, J. Figueras, Y. Zorian, "SRAM-Based FPGAs: Testing the Embedded RAM Modules", *Journal of Electronic Testing: Theory and*

Applications, vol. 14, April 1999, pp. 159-167.

[2] Steve Lu, Rafail Ostrovsky, “Distributed oblivious RAM for secure two-party computation”, *TCC'13: Proceedings of the 10th theory of cryptography conference on Theory of Cryptography*, March 2013.

[3] An-I Andy Wang, Geoff Kuenning, Peter Reiher, Gerald Popek, “The Conquest file system: Better performance through a disk/persistent-RAM hybrid design”, *Transactions on Storage (TOS)*, vol. 2, August 2006, pp. 309-348.

[4] Komal, S. P. Sharma, Dinesh Kumar, “RAM analysis of repairable industrial systems utilizing uncertain data”, *Applied Soft Computing*, vol. 10, September 2010.

[5] Michael T. Goodrich, Michael Mitzenmacher, Olga Ohrimenko, Roberto Tamassia, “Oblivious RAM simulation with efficient worst-case access overhead”, *CCSW '11: Proceedings of the 3rd ACM workshop on Cloud computing security workshop*, October 2011, pp. 95-100.

[6] Benny Pinkas, Tzachy Reinman, “Oblivious RAM revisited”, *CRYPTO'10: Proceedings of the 30th annual conference on Advances in cryptology*, August 2010.

[7] Miklós Ajtai, “Oblivious RAMs without cryptogrpahic assumptions”, *STOC '10 Proceedings of the forty-second ACM symposium on Theory of computing*, June 2010, pp. 181-190.

[8] S. K. Moore, “Multicore is bad news for supercomputers”, *IEEE Spectrum*, vol. 45, November 2008, pp.15-15.

[9] Yiran Chen, Weng-Fai Wong, Hai Li, Cheng-Kok Koh, Yaojun Zhang, Wujie Wen, "On-chip caches built on multilevel spin-transfer torque RAM cells and its optimizations", *Journal on Emerging Technologies in Computing Systems (JETC),* vol. 9, no. 16, May 2013.

[10] Yiran Chen, Weng-Fai Wong, Hai Li, Cheng-Kok Koh, "Processor caches with multi-level spin-transfer torque ram cells", *ISLPED '11: Proceedings of the 17th IEEE/ACM international symposium on Low-power electronics and design*, August 2011, pp. 73-78.

[11] Kyu Hyung Lee, In Hwan Doh, Jongmoo Choi, Donghee Lee, Sam H. Noh, "Write-aware buffer cache management scheme for nonvolatile RAM", *ACST'07: Proceedings of the third conference on IASTED International Conference: Advances in Computer Science and Technology,* April 2007, pp. 29-35.

[12] Yiming Zhang, Rui Chu, Dongsheng Li, "A peer-to-peer IO buffering service based on RAM-grid", *International Journal of Autonomous and Adaptive Communications Systems,* Vol. 2, November 2009, pp. 382-396.

[13] J. Ousterhout, P. Agrawal and D. Erickon, "The case for RAMCloud", *Communications of the ACM,* vol. 54, no. 7, July 2011, pp. 121-130.

[14] John Ousterhout, Parag Agrawal, David Erickson and Christos Kozyrakis, "The case for RAMClouds: scalable high-performance storage entirely in DRAM", *SIGOPS Operating Systems Review,* vol. 43, January 2010, pp. 92-105.

[15] Stephen M. Rumble, Diego Ongaro, Ryan Stutsman, Mendel Rosenblum and John K. Ousterhout, “It's time for low latency”, *HotOS'13: Proceedings of the 13th USENIX conference on Hot topics in operating systems,* May 2011.

[16] Stephen M. Rumble, Ankita Kejriwal, and John Ousterhout, “Log-structured Memory for DRAM-based Storage”, *12th USENIX Conference on File and Storage Technologies (FAST ’14),* Santa Clara, CA USA, February 17–20, 2014.

[17] Diego Ongaro, Stephen M. Rumble, Ryan Stutsman, John Ousterhout and Mendel Rosenblum, “Fast crash recovery in RAMCloud”, *SOSP '11: Proceedings of the Twenty-Third ACM Symposium on Operating Systems Principles,* New York, NY, USA, October, 2011.