The New Non-Volatile Based On STT-MRAM Lies in The Design and Implementation of High-Speed Cache of Magnetic Random Memory Material

Tian Liu, Wei Zhang, Tao Xu and Guan Wang

Institution of Information Science and Engineering, Zaozhuang University, Zaozhuang, China;

258062058@qq.com

Abstract. With the development of semiconductor technology, the buffer of processor integration is getting larger and larger. The high-speed design of storage capacity of traditional memory devices has an important connection with cache. How to design new non-volatile magnetic random memory has become an important step. To solve these problems, researchers at home and abroad have discussed a large number of new non-volatile memory technologies, which have excellent characteristics such as non-volatility, low power consumption and high memory density. The key optimization techniques for the disadvantages of high write power, limited write life and long write latency of the new non-volatile memory are analyzed. Finally, the possible research directions of new non-volatile memory devices in future cache optimization are discussed.

1. Introduction

With the rapid development of hardware technology, the power consumption problem causes the processor's main frequency to not increase further, and instead moves toward multi-core processor design. The multi-core technology is becoming more mature and the multi-threading technology is widely used, and the processor performance is greatly improved; however, the access speed of the main memory is slow, which seriously hinders the improvement of the overall performance of the computer system. The performance gap between the processor and the main memory is also gradually increasing, and the cache can alleviate the problem of the access speed mismatch to a certain extent.

The emergence of new non-volatile memories has received widespread attention in academia and industry, and provides new solutions for computer storage technology. The new non-volatile magnetic random-access memory material is promising to replace traditional memory devices because of its high integration, low leakage power, fast access, and non-volatility. Due to the different manufacturing processes and design principles of new memory devices, their size, read and write speed, read and write power consumption, and storage life are slightly different [1].

As computing devices become more demanding on memory, traditionally based volatile memories cannot meet modern demands in terms of density and power consumption. With the development of microelectronics and materials engineering, storage media composed of new storage materials have emerged, which has received extensive attention and research from academia and industry. Since these new storage media are mostly non-volatile, these new storage media can be collectively referred to as
new non-volatile memories [2]. Non-volatile memory is characterized by high density, non-volatility, and low power consumption compared to conventional memories [3].

In order to alleviate the performance bottleneck caused by the difference between the main memory and the external memory speed, the modern operating system will open up a cache area in the main memory as the cache between the main memory and the external memory. When the data needs to be read from the external memory, the system will first check whether the data already exists in the cache [4]. If the data is already in the cache, the data can be read directly from the cache, so that the access speed of the external storage is close to the speed of the main memory. If the data is not in the cache, the system will first load the data from the external memory into the cache, and then read the data from the cache to the user process address space [5]. Although the access speed is the speed of external storage, the subsequent querying of the data can be directly performed in the cache. The same operation is used for write operations, that is, data is first written to the cache from the user process address space, and then written to the external memory by the cache. Therefore, the existence of the cache eases the difference in performance between the external memory and the main memory, greatly improving the system's I/O performance [6].

![Fig. 1 High-speed cache structure analysis of STT-MRAM](image)

By mixing the cache to compensate for the lack of write operations, releasing the volatility of STT-MRAM to improve performance, solving the problem of read and write asymmetry, using multi-level cells to increase storage density and applications in the GPU to reduce power consumption. The methods involved in the cache optimization standard include sensing at the bit level and avoiding unnecessary write operations, reducing write operation conflicts through data prefetching techniques, balancing access between groups in the cache group through data distribution methods, utilizing compilation guide data allocation methods, and modifying Storage architecture, predictive cache dead write operations, number of cache flushes controlled by counters, data migration awareness methods, and more. These methods have greatly promoted the development of STT-MRAM [7].

2. Measures to improve the high-speed cache design of new non-volatile magnetic random memory materials

2.1. Reduce the number of write operations of STT-MRAM

When writing data to the cache, many of the same data bits are already in the cache. In order to judge whether the bit to be written is repeated, the strategy of reading first and then writing is adopted,
because the cost of reading is smaller than that of writing, the reading operation perceives the state of the MTJ by current, and then compares with the bit to be written, if the data is the same, the writing is not required. This data bit is entered, which avoids redundant write operations. This approach can effectively reduce write power consumption, but introduces hardware overhead and additional read operation overhead. Data is kept in the cache, which avoids prematurely writing frequently accessed data to the cache. At the same time, write bias and write cache optimization techniques are proposed to reduce write operations. The technique is to keep the data in the cache as long as possible through the cache replacement algorithm. The technique is to add a small cache to the cache, which is used to store the data replaced from the cache, which can alleviate the data written from the cache to the STT. The number of times in MRAM. This architectural approach reduces dynamic power consumption but increases the hardware overhead of small cache blocks and the overhead of cache management strategies. The behavior of the operation to reduce unnecessary write operations. This method can sense the write operation behavior of the bit unit. It is considered necessary when the bit sequence is written for the first time. It is considered unnecessary when the second sequence is written again. This process can be detected by the value of the bit unit. 1 part sends a current to read the MTJ status bit, so that you can know the value of the currently stored bit unit; the second part compares the sequence to be written with the value of the value, if different, write data, otherwise avoid This write operation, which is done by the write circuit. The method also adopts the method of reading first and then writing to avoid unnecessary write power consumption, but increases the hardware overhead, and needs to design and modify the corresponding bit circuit.

![Fig. 2 STT-MRAM logic design](image)

A flag is added to the cache, and the data value with this flag is zero. When performing a write operation, the cache first checks whether the data to be written is zero. If it is detected, the flag is set to write only non-zero data to the STT-MRAM; when the read operation is performed, only the cache is read. Non-zero flag data. In this way, a large number of write operations can be reduced, and power consumption can be reduced, but each data line is incremented by one flag bit, which increases hardware overhead, and the detection also has a certain price. A low-current probabilistic write operation that reduces power consumption by lowering the write current pulse. However, if the current is too low, the write operation will fail. To solve this solution, we designed the Structure shown in figure. 2.

2.2. Designing an Iterative Framework for STT-MRAM
Step 1 first performs a write operation with a low current; Step 2 reads the data just written; Step 3 compares with the initially written data to determine whether the write operation was successful, and if the comparison is unsuccessful, then 1 step to start execution. All write operations are performed
according to this iterative framework, and the data is written accurately; at the same time, in order to further ensure the correct rate of the write process, the method introduces a Bernoulli probability equation to accurately predict each write operation. The above scheme reduces the write power consumption by dynamically adjusting the strength of the write current pulse, which is relatively novel, but the frequent execution of the iterative frame can degrade the performance of the system. A STT-MRAM cache architecture method for writing predictions. A large amount of data is written to the last level of the cache and has never been used.

![Non-volatile circuit design of STT-MRAM](image)

Fig. 3 Non-volatile circuit design of STT-MRAM

Such a write operation is called dead write. If these dead write operations can be reduced, the cache power consumption can be greatly reduced. In order to achieve this goal, the method first divides the dead write into three categories: then predicts the cache access behavior by sampling and updates the prediction table in real time. The prediction table determines whether each write operation of the cache is a dead write, if it is a write the write operation bypasses the last level cache and writes directly to the main memory or low-level cache. Because the precision of the sampling prediction is very high, this method can reduce the dead write operation of most of the last level cache, thus reducing the write power consumption, the design operation is transferred to the cache, and the cache space and resources of the cache are very effective. Can improve the performance of STT-MRAM system.

2.3. Reduce the write latency of STT-MRAM

A method of managing and optimizing the STT-MRAM cache by increasing data prefetch efficiency. With the basic priority allocation and priority startup methods, the basic priority allocation method is to assign priorities according to the importance of different types of access requests (read, write, prefetch, write back, etc.), and important requests are executed first, which can be reduced. Access violations between requests reduce latency; methods are used to differentiate between application access requests and higher priority for applications with fewer access requests, which speeds up non-intensive program requests and improves overall System performance.

| Workload File | Size  | Number of Files | Average Size | Proportion |
|---------------|-------|-----------------|--------------|------------|
| varmail       | 16K   | 1000            | 4K           | 51%        |
| webproxy      | 16K   | 10000           | 4K           | 84.36%     |
| fileserver    | 128K  | 10000           | 64K          | 35.62%     |
| webserver     | 16K   | 1000            | 16K          | 90.85%     |
Prioritize and mix local global processing control methods. Due to the data processing mechanism, the number of write requests increases rapidly, which increases the waiting time of other requests, and extends the overall access period. The request priority method considers that the read request has the highest priority and should be processed immediately, the write request is followed, and the processing is written. The priority of the back request is lower because the data processed and written back is not used immediately. The hybrid local global processing control method includes local processing control and global processing control, local processing control is a processing strategy for each core; and global processing control is a processing strategy for all cores and shared caches, which periodically extracts each core. The processing frequency and the global access frequency of the LLC are used to evaluate the processing accuracy, which will guide the next processing operation. By improving the efficiency of data processing, it can effectively reduce the conflict of access requests, reduce the overall waiting period, and improve the performance of the cache. This method depends on the accuracy of the request classification and the access characteristics of the program.

![Fig. 4 STT-MRAM of High-speed Cache physical design](image)

2.4. Balance STT-MRAM to increase the lifetime of the cache.
Through the cache management strategy to reduce write fluctuations between groups and groups, the strategy has the ability to write data to the buffer without writing to the cache line. Two characteristics are used to reduce write fluctuations between groups and write fluctuations within groups. Each 2 cache groups are mapped into a pair, each time exchanged between corresponding groups, and if a group writes too many times, it is exchanged with the next adjacent group. This method can make the write between the groups equalized, but cannot solve the problem of write imbalance in the group, because according to the cache management strategy, frequently accessed data will always be written to the beginning of the line. Data that is frequently accessed with a certain probability records frequently accessed data through a global counter. The more frequently the data is accessed, the greater the probability of being accessed. This reduces the frequent access of the data to access certain bits and cause bit corruption. The write operation can be balanced as a whole, but sufficient buffers are needed for write-intensive programs to successfully save frequently accessed data. The hybrid architecture approach of STT-MRAM then proposes a partition-level wear leveling method and an access-aware strategy to reduce unbalanced wear between different partitions. The access awareness strategy is mainly for write management of STT-MRAM and read management of SRAM.
Fig. 5 High speed cache logic analysis

STT-MRAM write management is used to reduce the number of rights and evenly distribute the write operation. The management process is to redirect the write request directly to the SRAM if the write request arrives and there is a writable line in the SRAM, if the SRAM There are no writable lines in the line to redirect write requests to other STT-MRAM lines. The SRAM read management process is similar to the STT-MRAM write management process, copying the data in the SRAM to the STT-MRAM and then performing the read operation in the STT-MRAM. The wear leveling method divides the hybrid buffer into 4 large blocks, each block contains 1 SRAM block, 2 STT-MRAM blocks and 1 STT-MRAM block, and dynamically adjusts the partition path according to the number of SRAM and STT-MRAM Corresponding structure of the computer.

Because the hybrid cache is designed to solve the energy consumption problem of traditional SRAM caches, the analysis of energy consumption is particularly important. The comparison of the energy consumption of the cache scheme shows that the hybrid cache has no static power consumption due to non-volatileness, which is about 40% lower than that of the pure SRAM cache. Although the program with more write instructions has more write operations to the STT-MRAM due to migration, causing more dynamic write power consumption, it can be ignored compared to the overall power consumption. Of course, it can be seen from the figure that a good data migration strategy can further reduce the cache by reducing the number of rights to the STT-MRAM.

Fig. 6 Comparison of high-speed cache of STT-MRAM
The STT-MRAM cache uses SRAM storage technology to support the processor to quickly acquire program instructions and data. In recent years, with the rapid development of computer technology, the cache capacity has increased greatly, and the number of semiconductors integrated on the chip has increased, which makes the traditional SRAM-based cache expansion difficult, and the static power consumption is too high, and has become a constraint cache technology. The obstacles to development, computer researchers have to explore new storage solutions to solve.

3. STT-MRAM Lies in the Design and Implementation of High-speed Cache Analysis

STT-MRAM is favored for its high integration, low static power consumption and long write life, and is considered promising for replacing SRAM for caching. However, due to its slow writing speed and high write power consumption, the use of pure STT-MRAM cache for computer systems can not only improve performance, but may lead to higher energy consumption. Based on this, this paper proposes the research of hybrid cache high-speed Cache based on STT-MRAM and SRAM. This paper analyzes the characteristics of STT-MRAM and other aspects, and finds that there are big differences in read/write latency, storage density and static power consumption, and the advantages and disadvantages are complementary. Therefore, STT-MRAM and Cache are combined to design Parallel hybrid caches may be better at performance and power consumption than they are used for caching alone, which reflects the feasibility and research implications of hybrid cache research. In order to make good use of the advantages of Cache and STT-MRAM, hybrid cache architecture, this paper chooses the parallel partition cache scheme. This design can effectively combine the advantages of STT-MRAM's large capacity, low static power consumption and fast SRAM read/write speed. For the shortcomings of STT-MRAM write latency and write power consumption, the high-speed Cache of immediate migration, delayed migration and adaptive migration is proposed step by step. It is judged whether the block is a frequent block of reading and writing by the previous access condition of the data block. And as much as possible, the write frequent cache block is placed in the SRAM area to reduce the write operation to the STT-MRAM area, thereby achieving an improvement in cache performance and a reduction in power consumption.

4. Conclusion

New storage devices provide high storage density for next-generation high-performance computing, low-power and fast-access storage for big data, non-volatile, low-power storage for the Internet of Things, and high-capacity, high-speed mobile devices Storage devices, these features present tremendous opportunities for computer system architecture and software design. As the manufacturing process of new non-volatile magnetic random-access devices matures, performance is more stable, write power is lower, and usage time is longer, new memory applications will become popular, new cache architectures and system software optimization It will also get unprecedented development. These are all future in-depth research and exploration and, if possible, STT-MRAM as the main memory to implement the cache, which provides more possibilities for the development of future computer systems, especially processors.

References

[1] XU Yonggang, et al. Embedded image recording technology based on NAND flash. Beijing: University of Chinese Academy of Sciences, Vol. 4 (2013) No. 21, p.165-169.
[2] LI Qing, et al. The design and realization of a high speed and large capacity NAND FLASH storage system. Beijing: Beijing Institute of Technology, Vol. 8 (2015) No. 23, p.145-148.
[3] XU Yujie, et al. Design and implementation of high-speed and huge-capacity storage system. Xi'an: Xidian University, Vol. 9 (2014) No.24, p.75-79.
[4] Zhan Wang, Kun Sun, Jiwu Jing, Sushil Jajodia, et al. Verification of data redundancy in cloud storage. Cloud Computing. Proceedings of the 2013 international workshop on Security in cloud computing, Vol. 3 (2013) No. 31, p.56-58.
[5] Mohamme Abdul Azeem, Mohammed Sharfuddin, et al. Ragunathan. Support-based replication algorithm for cloud storage systems. COMPUTE '14: Proceedings of the 7th ACM India Computing Conference, Vol. 12 (2014) No. 32, p.216-219.

[6] Fatemeh Azmandian, David R. Kaeli, Jennifer G. Dy, et al. Securing cloud storage systems through a virtual machine monitor. SRAS 12: Proceedings of the First International Workshop on Secure and Resilient Architectures and Systems, Vol. 10 (2012) No. 35, p.183-185.

[7] Ricardo Batista Rodrigues, Carlo M. R. da Silva, Frederico A. Durao., et al. A File Recommendation Model For Cloud Storage Systems. SBSI 2015: Proceedings of the annual conference on Brazilian Symposium on Information Systems: A Computer Socio-Technical Perspective, Vol. 11 (2015) No. 40, p.355-357