博士論文 (要約)

High Performance Solid-State Storage Systems
with Memory-Aware Software Design

(メモリと制御ソフトウェアの統合設計によるソリッド・
ステート・ストレージシステムの高性能化)

孫　超

The market of the solid-state drives (SSDs) is expanding due to SSDs' higher performance and endurance while lower power consumption, compared with the hard disk drives (HDDs). The continuously decreasing cost of the NAND flash memory, thanks to the scaling and multi-bit technologies, is another key driving force. Since there is no seek time for the data access in SSDs, the read performance of the SSDs is higher than HDDs. Moreover, SSDs also have a significant sequential write performance advantage over HDDs. However, due to the "erase-before-write" characteristics and endurance problems of the NAND flash memory, the SSDs have little or even no advantage in random write performance than the HDDs. Therefore, the SSD write performance should be improved.

In this dissertation, research is carried out to improve the SSD performance and endurance, reduce its energy consumption and cost. The solution is the memory-aware design in the storage stacks, including the memory device layer, flash translation layer (FTL), operating system (OS, kernel) layer and the application layer.

In the memory device layer, the design guidelines of the storage class memory (SCM) and NAND flash memory are provided for the proposed three-dimensional through-silicon-via (TSV) SCM/NAND flash hybrid SSD (Hybrid SSD). From the experimental results, the required SCM capacity will be dependent on the application speed requirement, workload characteristics, data management algorithms like SCM wear leveling and SCM latency parameters. Strategies to configure the required SCM and NAND flash capacities are presented to achieve the low cost. In addition, the SCM chip design examples are provided, which meets the system speed requirement with a low SCM chip cost. Lastly, the sensitivity of the NAND flash organization, block and page sizes, on the system performance is discussed.

In the FTL, a write cache buffer replacement algorithm, cold data eviction (CDE), is proposed for the Hybrid SSD. According to a 50 ns HfO2 ReRAM measurement results, the NAND like interface (I/F) is proposed for ReRAM. Frequently accessed (hot) data and random data are stored in ReRAM, which reduces the data traffic to the NAND flash and decreases the SSD data fragmentation issue. When ReRAM is almost full, the

CDE algorithm aggregates and evicts the cold and less fragmented data to the NAND flash memory. Then the ReRAM space can be reclaimed for the new writes. Combined with the page-level address mapping scheme, 12.6-times performance improvement, 92.8% energy consumption reduction and 9.7-times endurance enhancement are achieved for the financial application.

In the OS layer, a NAND flash aware middleware, LBA scrambler, is proposed to improve the SSD performance by minimizing the garbage collection (GC) overhead, which is the bottleneck of the SSD random write performance due to many page-copy operations. The concept of the LBA scrambler is to intentionally write data to the fragmented pages in the next erase block. As a result, the page-copy overhead can be minimized. To achieve this, the recommended writing page information is sent from the SSD controller to the LBA scrambler. With these hints, the LBA scrambler maps the logical block address (LBA) from the write requests to the scrambler LBA (SLBA) and send the new write requests with SLBA to the SSD controller. With the LBA scrambler, 17%-400% performance improvement, 23%-60% energy consumption reduction and 9%-55% endurance enhancement are achieved. The SSD issues due to the unaligned writes are eliminated as well.

Moreover, since file system (FS) in the OS manages the file storage and access, the effect of the FS on the SSD performance is evaluated. By analyzing the SSD workload characteristics under different FSs, it is proved that the success of NAND flash aware FSs in improving the SSD performance is realized by grouping the random writes into sequential and reducing the metadata updates. The proposed middleware LBA scrambler presents the effectiveness to improve the SSD performance even for the NAND flash aware FS.

In the application layer, a storage engine assisted SSD (SEA-SSD) is proposed for the database applications. It is based on the idea that upper layer in the host contains more rich information of the data activity, which is useful for reducing the GC overhead. By passing several hints from the SE to the SSD controller, data with similar activity are identified, clustered and aggregated the in the same block of the NAND flash. As a

result, maximum 24% SSD performance improvement, 16% energy consumption reduction and 19% lifetime extension are achieved.