Extended Abstracts of the 2013 International Conference on Solid State Devices and Materials, Fukuoka, 2013, pp894-895

TLC/MLC NAND Flash Mix-and-Match Design with Exchangeable Storage Array

Shogo Hachiya[1], Koh Johguchi[1], Kousuke Miyaji[1,2] and Ken Takeuchi[1]

[1] Chuo Univ. 1-13-27 Kasuga, Bunkyo-ku, Tokyo 112-8551, Japan, [2] Shinshu Univ. E-mail: johguchi@takeuchi-lab.org

Abstract

This paper proposes TLC/MLC NAND flash mix-and-match design method for exchangeable storage array. The proposed Round-Robin frozen data collection achieves 56% higher write performance and 29% write energy reduction compared with the conventional MLC only SSD. SSD card exchange method is also presented to realize sustainable and flexible storage arrays.

1. Introduction

Recently, NAND flash based solid-state-drives (SSDs) are widely used in high speed and low power storage array at data centers storing the big data. To save bit-cost, triple-level-cell, TLC, is now available for SSD instead of multi-level-cell (MLC) [1]. However, it is complicated and slow to write TLC due to block unit writing [2, 3]. Alternatively, although a ReRAM/MLC hybrid SSD has been proposed in [4,5], ReRAM is still high cost. Hence in the proposal, as a low cost storage array application, we propose a TLC/MLC hybrid NAND flash SSD storage array with wireless SSD cards [6] (Fig. 1). With Round-Robin frozen data collecting algorithm (RR-FDCA) (Fig. 2), MLC NAND flash not only works as a simple buffer for TLC but aggressively screens extremely cold data, which is rarely accessed from the host. To reduce TLC access, only the screened static frozen data are stored in TLC. With the proposed RR-FDCA, SSD write performance is enhanced up to 56% compared with the conventional all MLC SSD. Additionally, a SSD card exchange method without storage system suspension is proposed. With the proposed method, system administrators can also choose the optimum TLC/MLC capacity ratio and the round number, corresponding to the screening period, that realizes the best mix-and-match design configuration of the storage array.

2. TLC NAND Flash Memories and Proposed Round-Robin Frozen-Data Collection Algorithm

Fig. 3 gives the measurement results of TLC and MLC NAND flash memories. Compared with MLC, the program disturb and data retention errors of TLC are significantly increased. Fig. 4 shows block and page definitions of TLC NAND. The page and block are the write and erase unit of NAND flash, respectively. The page size depends on word-line (WL) length. In TLC, one word-line corresponds to three pages. Fig. 5 illustrates write order and V_{TH} distribution of TLC. Because of small V_{TH} margin and many reference voltages for read in TLC, the read latency becomes long and the write verify iteration are increased by 2.7-times. Moreover, due to complicated write order [2], TLC requires block unit writing. This causes the write performance degradation because all valid pages in the TLC block have to be read and written to another block (block-modify-write) if an overwrite operation is occurred in TLC. Table I summarizes the specification of MLC and TLC NAND flash memories [7]. The read/write latencies of TLC are longer than that of MLC and TLC write unit is a block. Therefore, to deal with TLC's characteristics, TLC/MLC storage array and the suitable control are proposed as shown in Figs. 1 and 2.

Fig. 6 explains Round-Robin page/block selection algorithm (RR algorithm) used in this paper. The target block is selected for write or erase in the cyclic order and Blk#0 becomes the next target if Blk#127 is used up. The conventional garbage collection (GC) is shown in Fig. 7. If the blank block number is not enough, GC operation is triggered and the erase target block is selected according to RR algorithm. All valid pages in the selected block are moved to other blocks. However, this conventional RR algorithm does not include special hot/cold data screening. Hot/cold data are defined as data accessed frequently/rarely, respectively. If hot data are moved to TLC, system performance and reliability can be degraded due to TLC's write-latency and W/E cycle limitation. Furthermore, hot data induces a block-modify-write, which makes the performance worse.

Thus, RR-FDCA is proposed as shown in Fig. 8. When blank MLC block number is not enough, RR-FDCA starts and selects two victim blocks for GC. If the number of valid pages in the selected MLC blocks is less than the page number per block (PNPB) of TLC ($N_{PNPB,TLC}$), RR-FDCA collects more valid pages until the collected pages reach two MLC blocks ($2\times N_{PNPB,MLC}$) to accumulate valid pages to be $N_{PNPB,TLC}$. After block copy, the corresponding blocks are erased and the round number for cold data screening, N_{RDS} is set to 0 for the erased blocks and set to 1 for the copied blocks, respectively (Step A1-A4). In case when the accumulated valid page number is large enough, RR-FDCA migrates to cold data screening process. First, RR-FDCA screens cold data by skipping copy (Step B1-B2). During the rounding period, corresponding to N_{RDS}, cold data is screened since the data accessed during the period is moved to another MLC block. Then, the screened cold data, or static frozen data is moved to TLC and the block is erased (Step C1-C4). If there are residual pages after frozen data eviction to TLC, they are written back to another MLC blocks (Step C5). As a result, only the static frozen data are stored in TLC and the number of TLC access can be drastically reduced, resulting in accelerated system performance.

3. Results of Proposed RR-FDCA

The write performance of the proposed storage array is evaluated by the transaction level modeling simulator [4]. Fig. 9 shows the write frequency distribution of *tcc-mysql* [8], used in the simulation as a benchmark. Fig. 10 depicts hot/cold data distributions. In the conventional MLC only array, all data are stored ramblingly. Meanwhile, the proposed storage array with RR-FDCA can effectively collect cold data and move the static frozen data to TLC. Fig. 11 indicates the performance improvement as a function of the threshold of N_{RDS}, N_{RDSMAX}. In this simulation the chip number is fixed for the evaluation under the same cost constraint. When $N_{RDS} = 1$, the proposed RR-FDCA improves the throughput and energy by 56% and -29%, respectively. In addition, $N_{W/E,MLC}$ also decreases by 36% compared with MLC only NAND. Fig. 12 also demonstrates the TLC/MLC capacity dependence. The results provide the optimum design point, which is 50% of TLC/MLC capacity ratio in tccc-mysql case. In the conventional case, moving cold data by GC, the throughput and energy are aggravated. On the other hand in the proposal, MLC free space is not occupied with cold data. A slow speed and small $N_{W/E,TLC}$ do not affect the system performance and reliability due to less TLC access. Therefore, the proposed TLC/MLC storage array has higher system performance compared with MLC only storage despite of moderate TLC characteristics.

4. Chip Exchange Method without System Suspend

If a NAND chip reaches the W/E cycle limit, the degraded chip must be exchanged. Hence, a chip exchange method is proposed (Fig. 13). In the conventional system, the system must be suspended due to enforce data copy when replacing the chip. Assuming 8GB valid pages in the degraded chip, the continuous dead time, which corresponds to the system suspend time, reaches 960 s (Fig. 14), that is unacceptably long. In the proposed method, the degraded chip is continuously used as read only mode, the valid pages are gradually evicted from the degraded chip by normal overwrite or GC operation (Step 2-3). At the timing after finishing the eviction of all valid pages, the degraded chip becomes ready for exchanging (Step 4). As a result, the continuous dead time of the proposal determines ~200 ms of the required time for GC. With the read only mode and proposed method, NAND chips can be removed and replaced. Thus, the proposed exchange method enables to choose the best NAND configuration to enhance the performance or save the cost.

5. Conclusions

Table II indicates the summary of this work. In *tccc-mysql* trace case, RR-FDCA realizes 56% the write throughput enhancement and 29% energy saving with 37% W/E cycle decrease. In case of *prxy_1* trace [9], the optimum TLC capacity percentage is changed to 87.5%. Therefore, the system cost can be decreased by 71%. Basically, the optimum organization can change according to usage or applications. Besides, the W/E cycle limitation and chip price of NAND depends on generations. Since the chip organization of the proposed storage array and N_{RDS} can change by the proposed methods, system administrators can select the best storage parameter on the performance or cost. Table II also demonstrates the W/E cycles ratio of MLC and TLC NAND W/E cycles. The W/E cycle of TLC NAND is acceptably lower than the requirement ($N_{W/E,TLC}/N_{W/E,MLC} < 1/100$).

References [1] J. da Silva, *presented in Samsung SSD Global Summit*, 2012, [2] S.H. Shin, *Dig. Tech. Papers, Symp. VLSI Circuits 2012*, pp. 132-133, 2012, [3] I.J. Chang, *IEICE ELEX*, 9, 23, pp. 1775-1779, 2012. [4] H. Fujii, *Dig. Tech. Papers, Symp. VLSI Circuits 2012*, pp. 132-133, 2012 [5] C. Sun, *Abst. NVMTS2012*, pp. 87-88, 2012, [6] W.J. Yun, *Dig. Tech. Papers, ISSCC2012*, pp. 52-53, 2012, [7] T. Vali, *presented in 1st Micro and Nano Electronics Workshop*, 2010, [8] *tpcc-mysql*, https://code.launchpad.net/~percona-dev/perconatools/tpcc-mysql, [9] MSR Cambridge Traces. http://iotta.snia.org/traces/388.

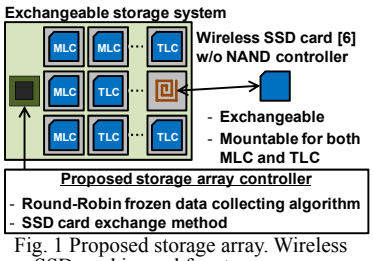

Fig. 1 Proposed storage array. Wireless SSD card is used for storage.

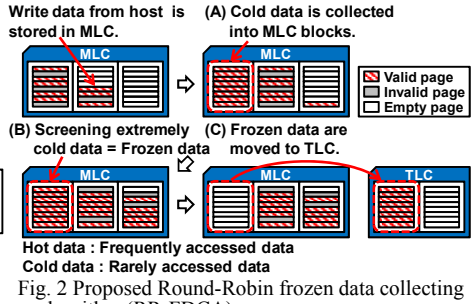

Fig. 2 Proposed Round-Robin frozen data collecting algorithm (RR-FDCA).

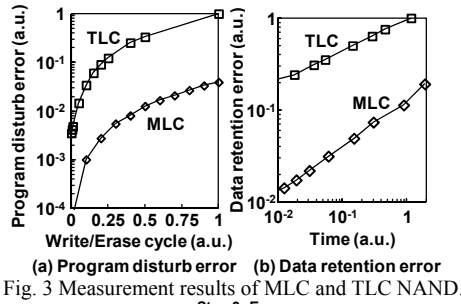

Fig. 3 Measurement results of MLC and TLC NAND.

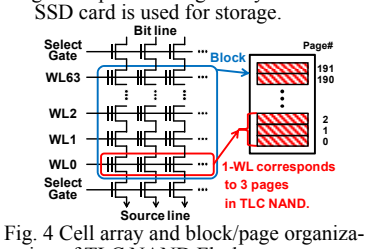

Fig. 4 Cell array and block/page organization of TLC NAND Flash memory.

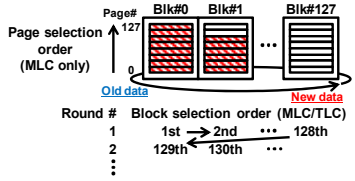

Fig. 6 RR block selection approach.

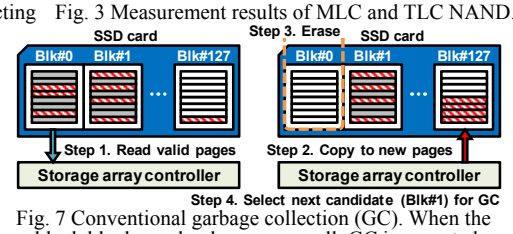

Fig. 7 Conventional garbage collection (GC). When the blank block number becomes small, GC is executed.

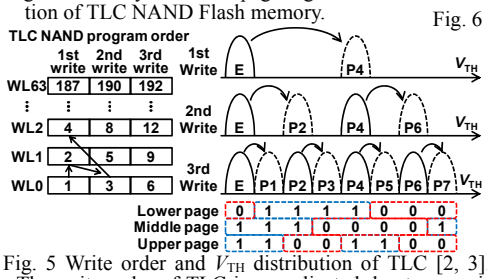

Fig. 5 Write order and V_{TH} distribution of TLC [2, 3]. The write order of TLC is complicated due to capacitive interference, resulting in block unit write and long write latency.

Table I MLC/TLC NAND specifications [7].

	MLC NAND	TLC NAND
Read latency (Typ.)	50 µs/page	100 µs/page
Write latency (Avg.)	900 µs/page	×2.7 → 2400 µs/page
Erase latency (Typ.)	2000 µs/block	3000 µs/block
Read unit	Page	Page
Write unit	Page	Block
Page number per block (PNPB), N_{PNPB}	128	×1.5 → 192
Bit cost (Normalized)	1	0.66
W/E cycle limitation	10000 cycles	×0.1 → 1000 cycles

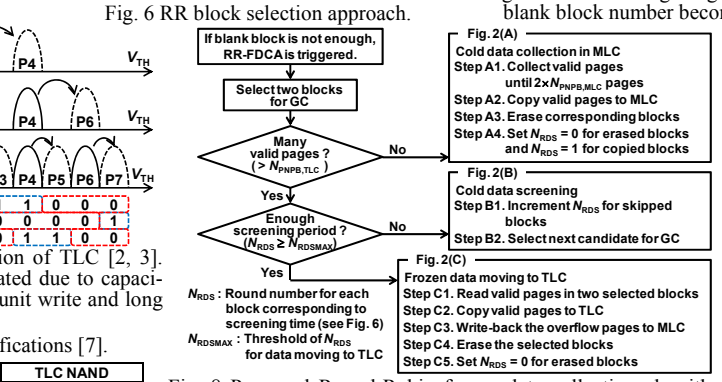

Fig. 8 Proposed Round-Robin frozen data collecting algorithm (RR-FDCA).

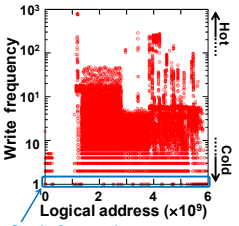

Fig. 9 Write frequency distribution of tpcc-mysql [8].

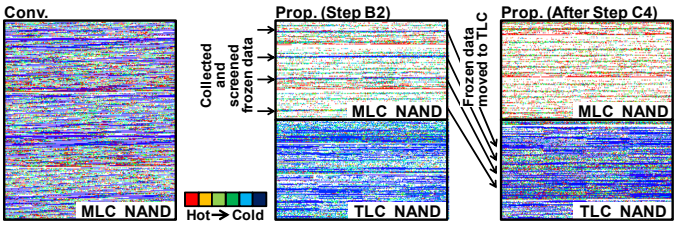

Fig. 10 Hot/cold data distribution in conventional and proposed storage systems.

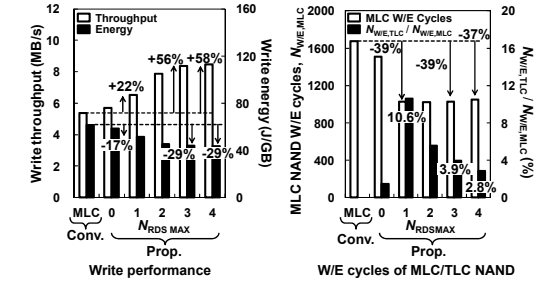

Fig. 11 (a) Write performance and (b) average W/E cycles of MLC/TLC NAND for conventional and proposed storage arrays. N_{RDSMAX} corresponds to screening period.

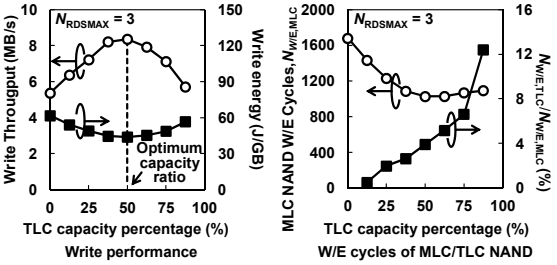

Fig. 12 TLC/MLC capacity ratio dependence of write performance (a) and MLC/TLC NAND W/E cycles (b) for proposed storage system.

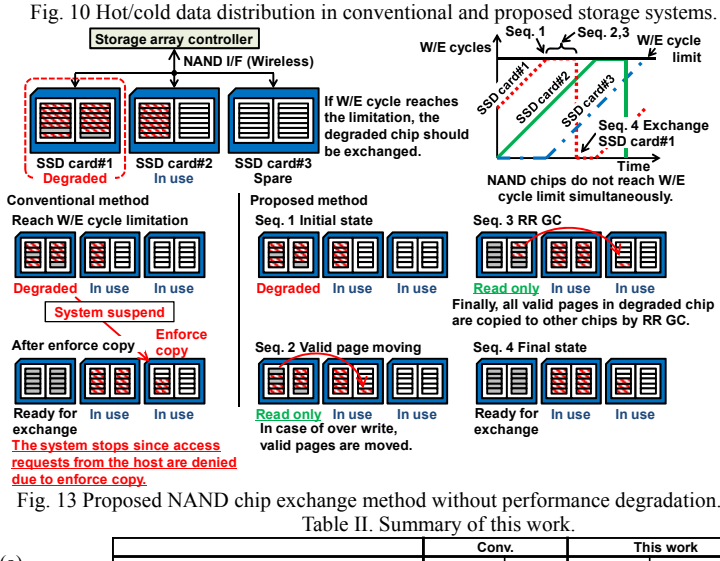

Fig. 13 Proposed NAND chip exchange method without performance degradation.

Fig. 14 Estimated continuous dead time for conventional and proposed exchange methods. In the conventional method, to move all valid pages in degraded chip to other chip, storage system must be suspended. The proposed storage array can be accessed without system suspend because of read only mode.

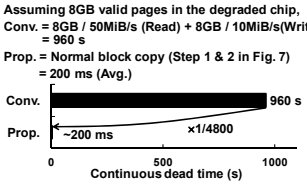

Table II. Summary of this work.

	Conv.		This work	
Benchmark	*tpcc-mysql*	*prxy_1*	*tpcc-mysql*	*prxy_1*
Total capacity (a.u.)	1	1	1	1
Optimum TLC NAND capacity (%)	N/A	N/A	50	87.5
Cost	1	1	0.83	0.71
N_{RDS}	N/A	N/A	3	3
Write throughput (MB/s) (Normalized by conv.)	5.36 (×1)	7.52 (×1)	8.36 (×1.56)	10.18 (×1.35)
Write energy (J/GB) (Normalized by conv.)	61.61 (×1)	(×1)	43.8 (×0.71)	42.5 (×0.82)
MLC NAND W/E cycles (Cycle/GB)	1678	1285	1026	1009
TLC NAND W/E cycles (Cycle/GB)	N/A	N/A	40.3	0.82
W/E cycles ratio (TLC/MLC)	N/A	N/A	3.92%	0.08%