SLC Flash & ReRAM Heterogeneous Memory System with Multi-Tier 5G Network & Device Co-Design for Smart Manufacturing

Chihiro Matsui and Ken Takeuchi
Univ. of Tokyo
7-3-1 Hongo, Bunkyo-ku, Tokyo 113-8656, Japan
Phone: +81-3-5841-1264 E-mail: matsui@co-design.t.u-tokyo.ac. jp

Abstract

This paper proposes heterogeneous-integrated non-volatile memory system, configured with 3D-SLC flash and ReRAM, in multi-tier 5G network & device co-design (NDCD) for smart manufacturing. Memory system solutions are proposed to meet 5G network latency of 0.25 ms. 1) Distributed multiple write to 3D-SLC flash reduces memory system latency to equal 0.25 ms latency requirement of smart factory. 2) Inter-deck hierarchical ReRAM system achieves 0.1 ms memory system latency with low manufacturing cost.

1. 5G Network and Non-Volatile Memory Co-Design

In the 5G network era, factories become “smart” by machine learning (ML) with generated data in edge machines [1, 2]. The proposed multi-tier smart factory (such as semiconductors) for high-yield manufacturing (Fig. 1) utilizes heterogeneous non-volatile memories, i.e., 3D-TLC flash, 3D-SLC flash, and ReRAM. From the bottom to the top, shorter-latency and smaller-capacity non-volatile memories are utilized because edge machine, edge server, and cloud centralized server handle different data types and amount. However, in 5G smart manufacturing, the network becomes so fast that non-volatile memories become the bottleneck. From 4G to 5G, the network delay, transferring 16 KByte (1 page) data of 3D-SLC flash decreases from 122 us to 6.1 us which is shorter than memory system delay time when reading/writing 3D-SLC flash (Fig. 2). Thus, this paper proposes network and device co-design (NDCD) by adding network to system, circuit, and device [5]. To achieve low memory latency at system level, this paper proposes 2 solutions for edge servers and centralized severs, respectively.

2. Distributed Multiple SLC Flash Write in Edge Servers

In the edge servers, 3D-SLC flash manages both sensor/image data (sequential) from edge machine and model/weights (random) of ML from cloud centralized servers. prxy_1 (random) and src2_2 (sequential) workloads [6] listed in Table I are considered as weight data and sensor/image data, respectively. Fig. 3 shows ECC architecture and decoding operation. Compared with BCH ECC, soft-decoding LDPC has longer ECC latency because of 7 V_{REF} operations to obtain log-likelihood ratio (LLR) for 3D-SLC flash.

This work assumes that the memory system latency of 0.25 ms is acceptable when the network latency of smart factory is 0.25 ms [7]. Fig. 4 shows average latency of 1 chip 3D-SLC flash with different types of ECC. Both prxy_1 and src2_2 workloads exceed the target memory latency of 0.25 ms. Proposed distributed multiple 3D-SLC flash writing, shown in Fig. 5, write sequential data to multiple chips. Thus, writing to Chip #1 and garbage collection (GC) to Chip #2 can be operated simultaneously. As a result, the average memory system latency reduces because latency of GC in one chip of 3D-TLC flash is concealed by read/write operation in the other chips. The average latency of prxy_1 becomes shorter than 0.25 ms of network latency, while that of src2_2 is longer than the network latency. Therefore, prxy_1 and src2_2 is called network bottleneck application and memory bottleneck application, respectively.

3. Inter-Deck Hierarchical ReRAM System in Cloud Centralized Servers

The cloud centralized severs communicate with other centralized servers to update shared model and distribute the model to the edge servers. To meet the latency requirement of 0.25 ms of 5G, inter-deck hierarchical ReRAM system is proposed (Fig. 6) to increase ReRAM capacity with less manufacturing cost. The bottom ReRAM decks have short latency, but the upper decks have large process variation. Upper decks face large line/space variation by processing such as lithography and CMP. Due to variation of capacitance and resistance of bit-lines (BLs) and word-lines (WLs), the estimated read/write latency becomes long. However, process variation has little impact on reliability because conductive filament size of ReRAM is almost the same, irrespective of the feature size. In the proposed inter-deck hierarchical ReRAM system, 128 bit and higher bandwidth I/Os (interconnections) can be adapted because all decks and sense amplifier circuits are fabricated in the same chip. Deck-0, 1, … in the bottom act as non-volatile (NV-) cache, and Deck-*N* acts as large-capacity storage. Assuming 30% process variation, hierarchical double-deck ReRAM system has longer average memory latency than conventional single-deck ReRAM system. Proposed inter-deck hierarchical ReRAM system with single chip solution have cost benefits compared with multi-chip organization of single-deck ReRAM. Therefore, to realize the same memory capacity, one chip of double-deck ReRAM is recommended compared with 2 chips of single-deck ReRAM.

4. Conclusions

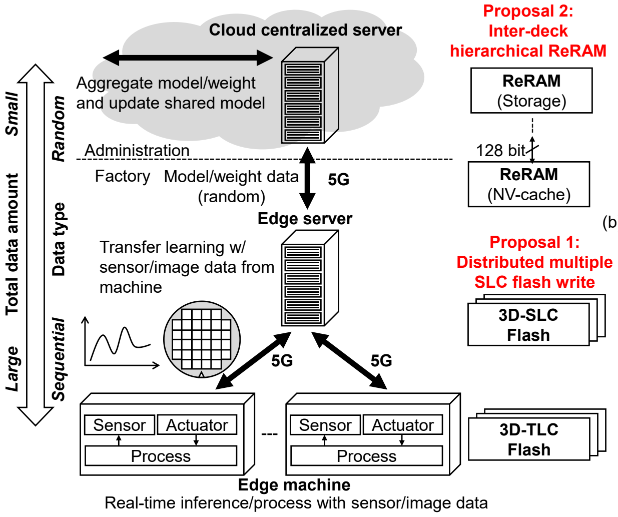
This paper proposed heterogeneous non-volatile memory system with multi-tier 5G network and device co-design. In each tier, less than 0.25 ms 3D-SLC flash system latency, and 0.1 ms inter-deck ReRAM system latency is achieved with less manufacturing cost, respectively. The proposed non-volatile memory system can be applied to smart factory as well as self-driving cars to update/manage dynamic map by multi-tier 5G network.

Acknowledgements

The authors thank R. Yasuhara and T. Mikawa for their ReRAM support. This work is based on results obtained from a project commissioned by the New Energy and Industrial Technology Development Organization (NEDO) and JSPS KAKENHI Grant Number 19K15051.

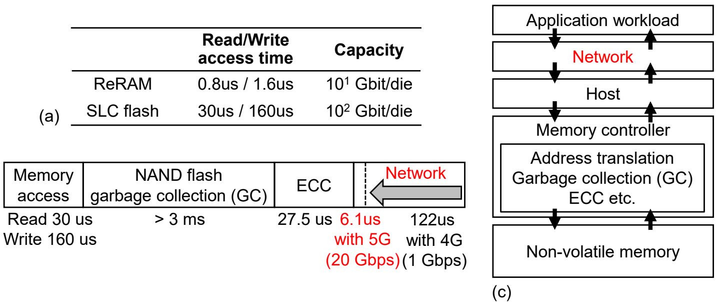
References

[1] N.N. Dao et al., *ICTC*, 2017, pp. 1280-1282. [2] A. Mueller, *ITU Workshop*, 2019. [3] A. Kawahara et al., *ISSCC*, 2012, pp. 432-433. [4] D. Nobunaga et al., *ISSCC*, 2008, pp. 426-427. [5] C. Matsui et al., *VLSI Tech.*, 2019, pp. 234-235. [6] MSR Cambridge Traces, http://iotta.snia.org/traces/388. [7] I. Parvez et al., *COMST*, vol. 20, no. 4, pp. 3098-3130, 2018. [8] T. Sakurai, *TED*, vol. 40, no. 1, pp. 118-124, 1993.


Cloud centralized server
Aggregate model/weight and update shared model
Small
Total data amount
Large
Random
Data type
Sequential
Administration
Factory
Model/weight data (random)
5G
Edge server
Transfer learning w/ sensor/image data from machine
5G
5G
Sensor
Actuator
Process

Sensor
Actuator
Process
Edge machine
Real-time inference/process with sensor/image data
Proposal 2: Inter-deck hierarchical ReRAM
ReRAM (Storage)
128 bit
ReRAM (NV-cache)
Proposal 1: Distributed multiple SLC flash write
3D-SLC Flash
3D-TLC Flash

Fig. 1 Proposed heterogeneous SLC flash & ReRAM system with multi-tier 5G network & device co-design for smart manufacturing.



	Read/Write access time	Capacity
ReRAM	0.8us / 1.6us	10^1 Gbit/die
SLC flash	30us / 160us	10^2 Gbit/die

Fig. 2 (a) Memory characteristics [3, 4] used in heterogeneous non-volatile memory system. (b) Cause of long latency in 3D-SLC flash of 16 KByte (1 page) access with 20 Gbps of 5G. (c) Proposed NDCD evaluation platform including network delay.

Table I Characteristics of workloads [6] for evaluation

	Total write data [GByte]	Total read data [GByte]	Average write request size [KByte]	Average read request size [KByte]	Write request ratio [%]
prxy_1	75.59	129.62	13.53	12.33	36.8
src2_2	59.82	22.79	29.15	68.08	72.4

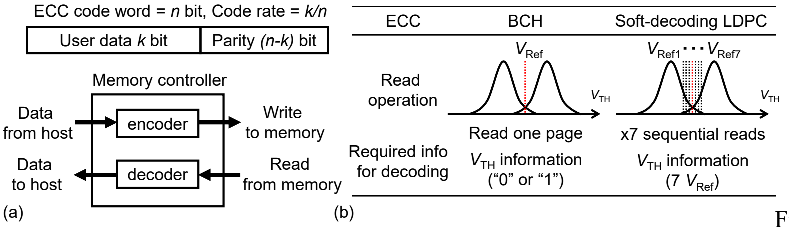


Fig. 3 (a) ECC encoding and decoding. (b) Read operation of BCH and soft-decoding LDPC ECC.

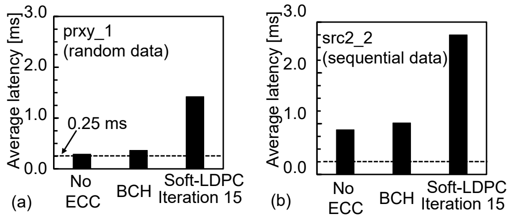


Fig. 4 Average latency of 1 chip 3D-SLC flash memory system with (a) prxy_1 and (b) src2_2. The average latency of 3D-SLC flash memory system exceeds the target memory latency of 0.25 ms.

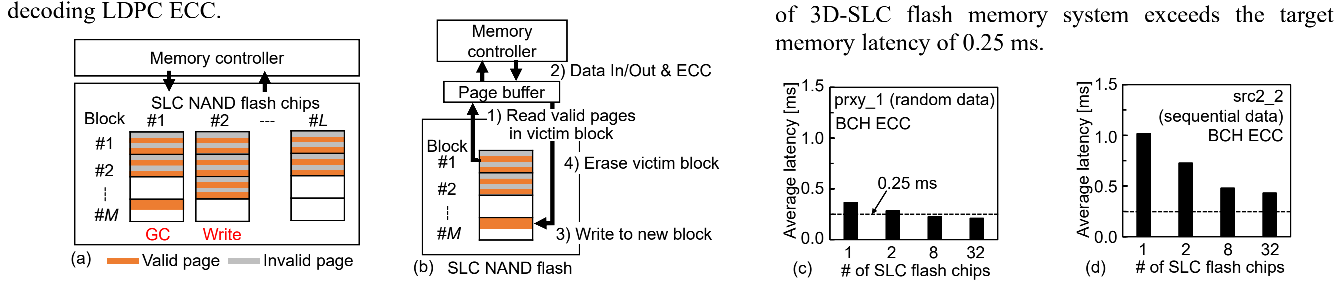
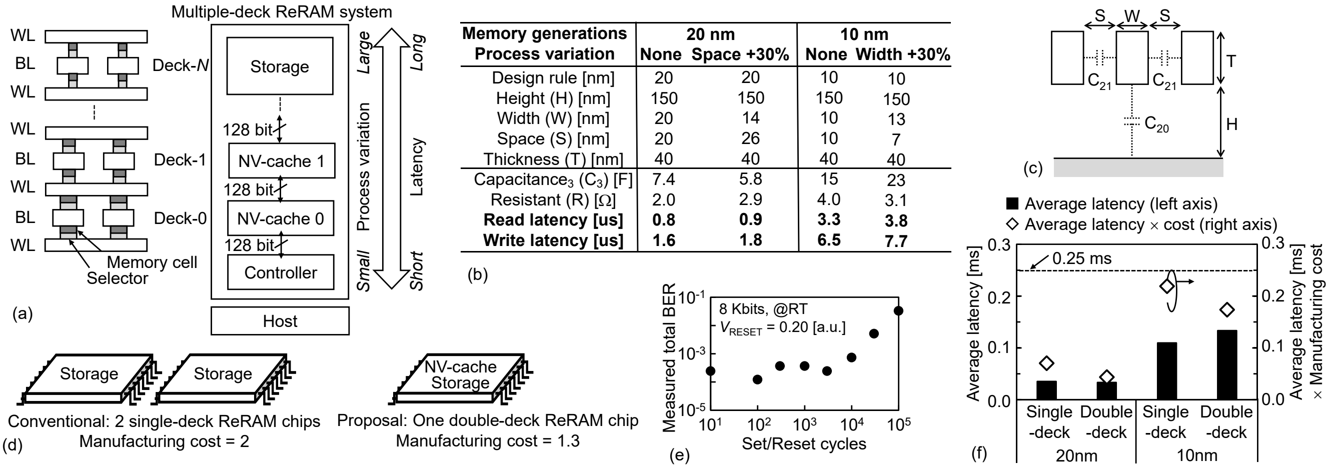


Fig. 5 (a) Proposed distributed multiple write to 3D-SLC flash for edge server. In this example, Chip #1 operates garbage collection (GC) while Chip #2 operates writing. (b) GC operation. Evaluation results of average latency with (c) prxy_1 and (d) src2_2 applications. Multiple chips reduce the average latency, and the average latency of prxy_1 achieves the target memory latency of 0.25 ms.



Memory generations	20 nm		10 nm	
Process variation	None	Space +30%	None	Width +30%
Design rule [nm]	20	20	10	10
Height (H) [nm]	150	150	150	150
Width (W) [nm]	20	14	10	13
Space (S) [nm]	20	26	10	7
Thickness (T) [nm]	40	40	40	40
Capacitance$_3$ (C$_3$) [F]	7.4	5.8	15	23
Resistant (R) [Ω]	2.0	2.9	4.0	3.1
Read latency [us]	**0.8**	**0.9**	**3.3**	**3.8**
Write latency [us]	**1.6**	**1.8**	**6.5**	**7.7**

Fig. 6 (a) Multiple-deck ReRAM architecture and proposed inter-deck hierarchical ReRAM system for cloud centralized server. (b) ReRAM latency estimation for 20 nm and 10 nm generations with process variations [3, 8]. (c) Sectional view of bit-lines (BLs) and word-lines (WLs) used for latency estimation. (d) Chip cost comparison to realize same memory capacity. (e) Measured BER of ReRAM [5]. (f) Average latency comparison of proposed inter-deck hierarchical ReRAM and conventional single-deck ReRAM. One chip of double-deck ReRAM has lower average latency × manufacturing cost compared with 2 chips of single-deck ReRAM.