13.5 A 512Gb 3-bit/Cell 3D Flash Memory on 128-Wordline-Layer with 132MB/s Write Performance Featuring Circuit-Under-Array Technology

Chang Siau[1], Kwang-Ho Kim[1], Seungpil Lee[1], Katsuaki Isobe[2], Noboru Shibata[2], Kapil Verma[1], Takuya Ariki[1], Jason Li[1], Jong Yuh[1], Anirudh Amarnath[1], Qui Nguyen[1], Ohwon Kwon[1], Stanley Jeong[1], Heguang Li[1], Hua-Ling Hsu[1], Tai-yuan Tseng[1], Steve Choi[1], Siddhesh Darne[1], Pradeep Anantula[1], Alex Yap[1], Hardwell Chibvongodze[1], Hitoshi Miwa[1], Minoru Yamashita[1], Mitsuyuki Watanabe[1], Koichiro Hayashi[1], Yosuke Kato[1], Toru Miwa[1], Jang Yong Kang[1], Masatoshi Okumura[1], Naoki Ookuma[1], Muralikrishna Balaga[1], Venky Ramachandra[1], Aki Matsuda[1], Swaroop Kulkani[1], Raghavendra Rachineni[1], Pai K. Manjunath[1], Masahito Takehara[1], Anil Pai[1], Srinivas Rajendra[1], Toshiki Hisada[2], Ryo Fukuda[2], Naoya Tokiwa[2], Kazuaki Kawaguchi[2], Masashi Yamaoka[2], Hiromitsu Komai[2], Takatoshi Minamoto[2], Masaki Unno[2], Susumu Ozawa[2], Hiroshi Nakamura[2], Tomoo Hishida[2], Yasuyuki Kajitani[2], Lei Lin[1]

[1]Western Digital, Milpitas, CA
[2]Toshiba Memory, Yokohama, Japan

Advancements in 3D-Flash memory-layer-stacking technology has enabled density scaling that circumvents the lithography limitations which have prevented 2D-NAND Flash memory from scaling [1]. Bit densities as high as 5.95Gb/mm² on a single die were recently reported [2], where a 512Gb NAND Flash was built on 96 layers of memory. As memory density increases, with memory layers increasing from 96 layers to 128 layers, higher bit density can be achieved by adopting larger capacity die; however, NAND performance per bit density is reducing with the 2-plane architecture. In this work, we propose a 512Gb 3b/cell 128WL-layer NAND Flash, with a bit density of 7.80Gb/mm²: a 31% improvement over the previously reported. Three key performance improving technologies have been implemented. (1) A 4-plane architecture with circuit under array (CUA) technology to improve performance per bit density. (2) A multi-die peak-power management (PPM) system to manage peak-power consumption in the system, via the ZQ pin. (3) A 4KB-page-read mode to reduce power consumption. Figure 13.5.1(a) summarized the key features and Fig. 13.5.1(b) shows the die photograph and the floorplan for this work. Figure 13.5.2 shows a table comparing this work to previous work.

Previous work [3] has reported a CUA architecture, which divides the array into 32 tiles and requires higher array interconnect overhead. In this work, the array is arranged into 4 planes, thereby minimizing array interconnect. Each plane consists of 683 blocks plus spare blocks, each 24MB in size. Each block has 1536 pages, each 16KB in size. Planes are divided in the BL direction to reduce BL length. The WL length is kept at 16KB length using the even/odd combined decoding structure [4]. Block selection signals are routed on top of array to WL-transfer gates, which reside under WL hookup Area. Connections are made by contacts outside of the array, resulting in no die area impact. In a traditional implementation, four planes of NAND memory will incur an area penalty of 15% due to duplication of the sense amplifier and data latches (SADL). In a CUA implementation, SADL will be hidden under the array, which reduces the area impact to below 1%. With the BL length cut in half, the BL *RC*-component is reduced by 4×, reducing the BL settling time by 31%. In a 3b/cell technology, program-verify operations occupy 60% of the total programming time (t_{PROG}). Figure 13.5.3(a) shows that with a reduced BL settling time, we estimate t_{PROG} to be improved by 16%. At the same time, 4-plane parallel operation will boost performance by another 100%, compared to a 2-plane device. We measured programming performance at 132MB/s with four planes operating in parallel. Figure 13.5.3(b) shows the normalized fail bit count (FBC) vs the average read latency (t_R). This work achieves a 56μs t_R.

The 4-plane architecture increases performance, but it also has a draw back with respect to peak power consumption. The peak current increases by 100% due to the increased WL and sense amplifier loading. For a system using multiple NAND die, the simultaneous peak power consumption by the multiple die can cause system power supply to droop and cause functional failure. A traditional system-level approach is to assume a peak-current consumption for the duration of t_{PROG}, and that a limited number of NAND die are activated for parallel operations. Typically, the peak power consumption is 3.5× higher than the average power consumption, and is only active for less than 10% of the total operation period. Command-based peak-power management was proposed [5] to resolve this issue, but this approach increases system management overhead by requiring the system to monitor the status and to issue stop/resume commands. In this work, we propose a multi-die peak-power management (PPM) scheme utilizing a shared ZQ-pad connection. This allows the NAND die in a system to schedule peak-power operations without involving the system management, as shown in Fig. 13.5.4. This proposal skews the sub-operation of NAND die such that peak-power consumption does not occur at the same time. Once peak-power consumption between different die is skewed, the system will be able to fully utilize the available power without considering the possibility of peak-power collisions among multiple die in the system, which would result in power droops. Figure 13.5.5 shows an example of a 3-die system with skewed peak-power, which allows for higher average power consumption while lowering the peak-power consumption. Based on the operation current consumption profile, we define an index of current consumption for each sub-operation of a program, read, or erase operation. This index is then translated by the state machine into a ZQ pin pull-up strength. Before each sub-operation begins on a NAND die, the state machine will issue a current consumption code to the ZQ pin and increase the ZQ pin pull-up strength. At the same time, a polling operation commences to sense the voltage on the ZQ pin. If the voltage on the ZQ pin, which accumulates the pull-up current from all of the NAND die in the system, is above a pre-set level the sub-operation will be postponed. If the voltage on the ZQ pin is lower, then the total peak current consumed in the system is within limits and the sub-operation of the polling NAND die will commence. In the event that there are multiple die polling at the same time, a randomized delay is used to reschedule polling, as shown in Fig. 13.5.6; indefinite sub-operation postponing can be avoided with this scheme. PPM allows the system to enable all NAND die to operate at the same time without introducing power droops due to peak power consumption from multiple die in the system.

To further reduce power consumption, we introduce 4KB-page read operation. As opposed to [3], where a 4KB read is achieved by enabling 2 tiles in tile array; in this work the array is arranged into planes and additional transistors are added in the sense amplifier to facilitate consecutive 4KB selection, as shown in Fig. 13.5.7. Each 4KB-page will be pre-charged and sensed at the same time using the all-BL (ABL) sensing method. Two transistors have been added to facilitate this operation: the BLC4K transistor for BL selection and the NLO4K transistor for BL discharging. For a selected 4KB BL, BLC4K will be turned on, thus biasing BL to the BL bias voltage. The unselected 12KB BLs, the BLC4K transistor will be turned off while the NLO4K transistor will be turned on to bias the unselected BLs to the cell source level. Average current consumption is reduced by 40% compared to a 16KB read. To prevent a BL on the 4KB boundary from becoming the timing bottleneck in a 4KB read, the control of BLC4K and NLO4K are designed such that there is always 16-BL past the 4KB boundary that are biased to the BL bias voltage, thus ensuring that all 4KB BLs see the same pre-charge conditions.

Acknowledgements:
The authors would like to thank the design, layout, verification CAD, device, test, and process teams, as well as Kazuaki Isobe and Farookh Moogat for their support.

References:
[1] H. Tanaka, et al., "Bit Cost Scalable Technology with Punch and Plug Process for Ultra High Density Flash Memory," *VLSI*, pp.14-15, 2007.
[2] H. Maejima, et al., "A 512Gb 3b/Cell Flash Memory on a 96-Word-Line-Layer Technology," *ISSCC*, pp. 336-337, 2018.
[3] T. Tanaka, et al., "A 768Gb 3b/cell 3D-Floating-Gate NAND Flash Memory," *ISSCC*, pp. 142-144, 2016.
[4] R. Yamashita, et al., "A 512Gb 3b/cell flash memory on 64-word-line-layer BiCS technology," *ISSCC*, pp. 196-197, 2017.
[5] M. Sako, et al., "A Low-Power 64Gb MLC NAND-Flash Memory in 15nm CMOS Technology," *ISSCC*, pp. 128-129, 2015.
[6] S. Lee, et al., "A 1Tb 4b/Cell 64-Stacked-WL 3D NAND Flash Memory with 12MB/s program throughput," *ISSCC*, pp. 340-341, 2018.

 978-1-5386-8531-0/19/$31.00 ©2019 IEEE

Features	
Capacity and die size	512Gb, 66mm2
Bit density	7.80Gb/mm2
Organization	3b/Cell (16KB + ECC) / Page, 1536 Pages / Block, (683 + EXT) Blocks / Plane, 4 Planes
tR Read	56us
Program Performance	132MB/s
I/O	1.066Gbps DDR, X8
Power Supply	VCC: 2.3V to 3.6V VCCQ: 1.2V, 1.8V

Figure 13.5.1: a) Key features (top). b) Chip Floorplan and Die photo (bottom).

ISSCC	This work	[2]	[6]	[4]
Technology	128 WL Layers	96 WL Layers	64 WL Layers	64 WL Layers
Capacity	512Gb	512Gb	1Tb	512Gb
# of bit/cell	3	3	4	3
# of planes	4	2	2	2
Program Performance	132MB/s	57MB/s	12MB/s	55MB/s
Die Size mm^2	66	86	182	132
Bit density Gb/mm^2	7.80	5.95	5.63	3.88

Figure 13.5.2: Performance and bit density comparison with previous works.

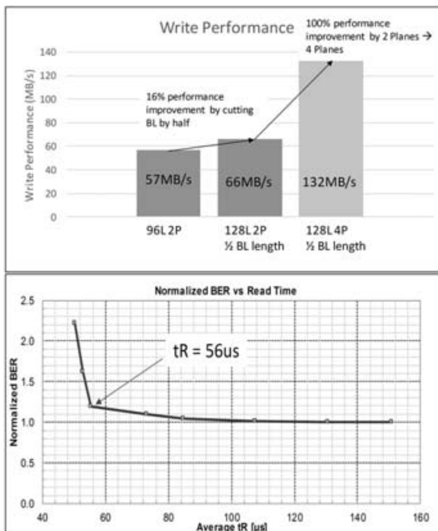

Figure 13.5.3: a) Write performance (top). b) Normalized Fail Bit count vs. t_R (bottom).

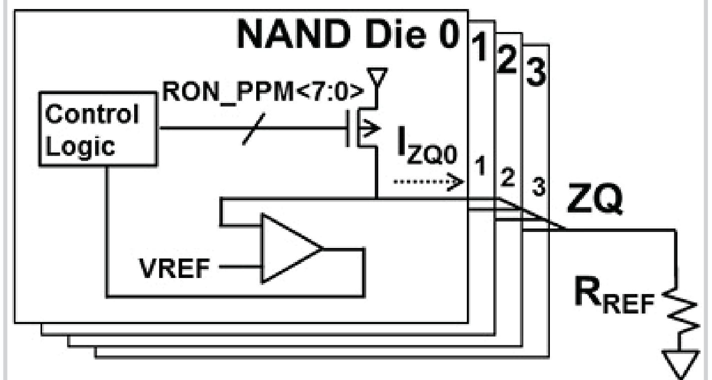

Figure 13.5.4: Multi-Die Peak Power Management scheme utilizing comparator used for ZQ calibration.

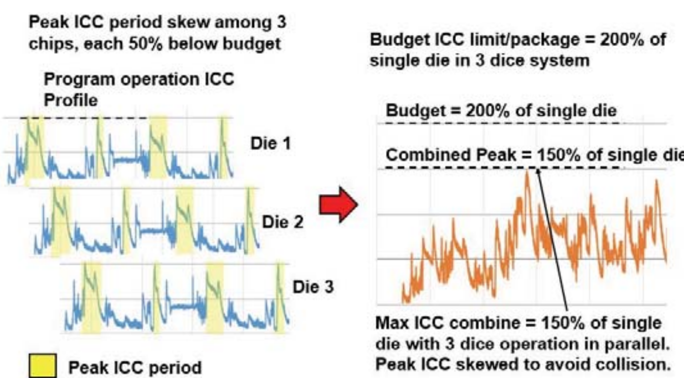

Figure 13.5.5: Peak Power Management scheme skew peak current consumption period to avoid peak current collision. Peak power budget can be fully utilized.

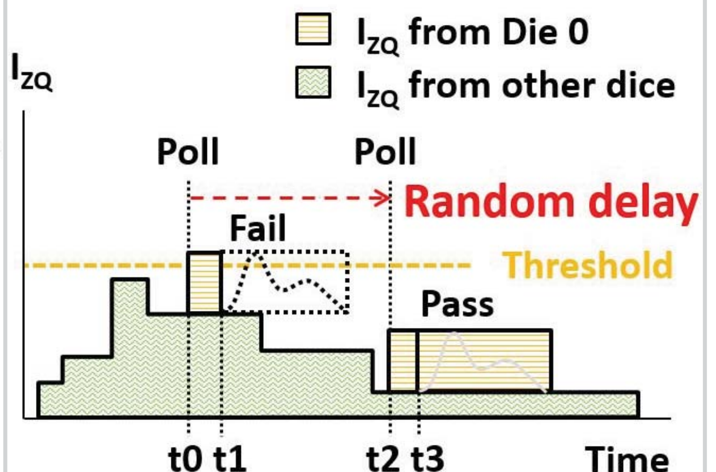

Figure 13.5.6: ICC polling with random delay.

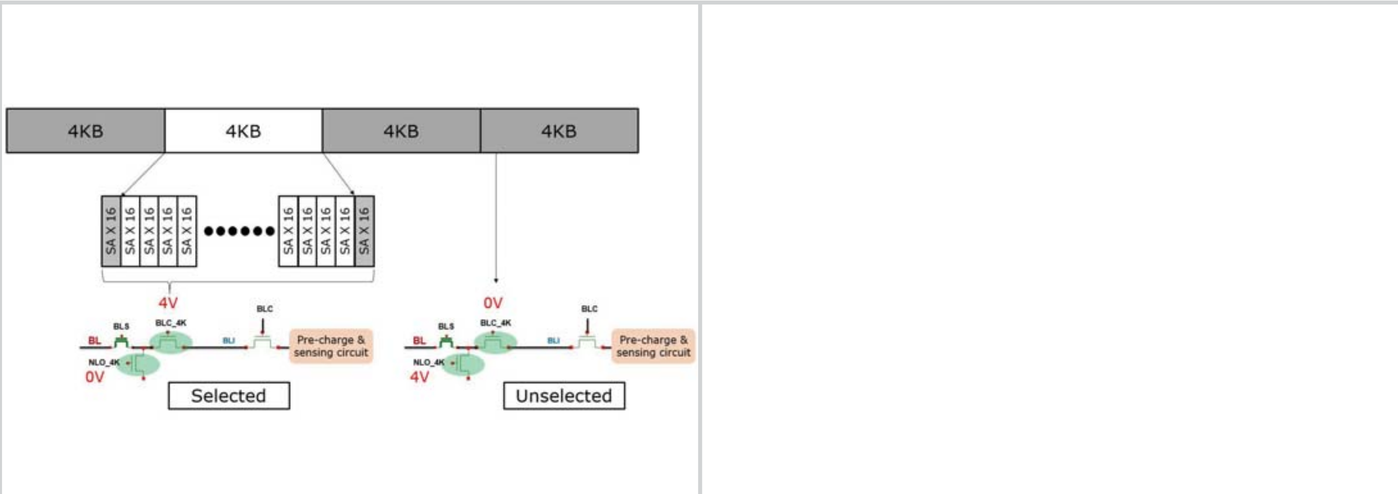

Figure 13.5.7: BL biasing condition for BLC4K and NLO4K for selected and unselected case. Edge words are biased at the same condition as selected word.

978-1-5386-8531-0/19/$31.00 ©2019 IEEE