25nm 64Gb 130mm^2 3bpc NAND Flash Memory

M. Goldman and K. Pangal
Intel Corporation

G. Naso and A. Goda
Micron Technology

***Abstract*—25nm NAND technology-based 64Gb 3bpc NAND Flash memory with die size of 130mm^2 (6.15GB/cm^2) is presented. The design including the array architecture is optimized for 25nm process technology while achieving world class performance of 100µs tRD, 2300µs tPROG for write throughput of 6.8MB/s with reliability meeting business requirements.**

Keywords-component; NAND, 25nm, SBL, ONFI, LMU, TWN

I. Introduction

The demand for NAND chips continues to grow [1] with smart phones, tablets and SSDs continuing to consume NAND bits at ever increasing densities. This paper will present Intel-Micron's highest density NAND product yet measuring 130mm^2 at a density 6.15GB/cm^2 using 3bpc NAND manufactured on Intel-Micron's 25nm NAND technology [2]. This chip extends the concepts and techniques developed for Intel-Micron's 3bpc chip at 34nm [3] and delivers a tRD of 100µs, a tPROG of 2300µs, and a write throughput of 6.8MB/s.

II. Array Architecture

Figure 1 shows the chip micrograph while Table 1 summarizes the key device features. The array is divided into two 32Gb planes which can be operated in either single or dual plane mode. Each plane contains 1368 blocks. Within each block, the page length is 8KB and the string length is 66 cells. The extra cells at each edge of the string are used to minimize edge wordline effects [4]. Given this architecture, each block configured for 3bpc operation can address up to 384 pages. The IO interface to the array is fully ONFI2.1 compliant with support for both asynchronous and synchronous timing modes up to a max transfer rate of 166MT/s.

The array architecture was optimized with a wordline air gap in order to reduce the wordline RC (Fig. 2) and with a tungsten bitline with air gap to lower the bitline capacitance, a critical metric for fast sensing (Fig. 3). Both shielded bitline (SBL) and all bitline (ABL) architectures [5] were evaluated. For ABL however, cell placement is susceptible to overshoot caused when the adjacent cell is inhibited which leads to an increase in VT from the floating gate (FG) to adjacent channel coupling. This effect, noted as the kink in Fig. 4, gets worse as the bitline pitch is reduced and is estimated to degrade the state width by 50-100mV at the 25nm node.

In addition, the die cost to support ABL is ~6% greater than that of SBL [6] while the performance cost due to increased pages/block has been mitigated by the controller/firmware companies by subdividing the writes during internal data move operations into smaller, less time consuming, chunks. As a result of these concerns, an SBL architecture was chosen.

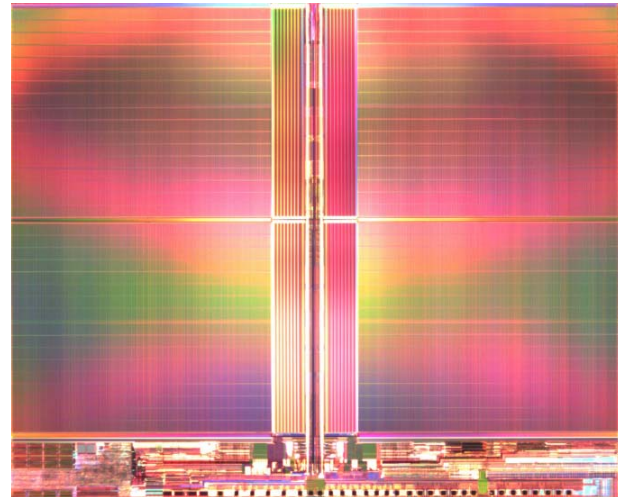

Figure 1: 25nm 3bpc 64Gb Die Micrograph

Table 1: Summary of Features

Technology	25nm, 3 metals
Cell Size	0.0034um^2 (select gates included)
Chip size	130mm^2
Plane Organization	8192 bytes/page x 384 pages x 1368 blocks
Spare Area per page	976 bytes
Read time	100us max
Program time	2.3ms typ
Erase time	10ms typ
Clock cycle time	12ns

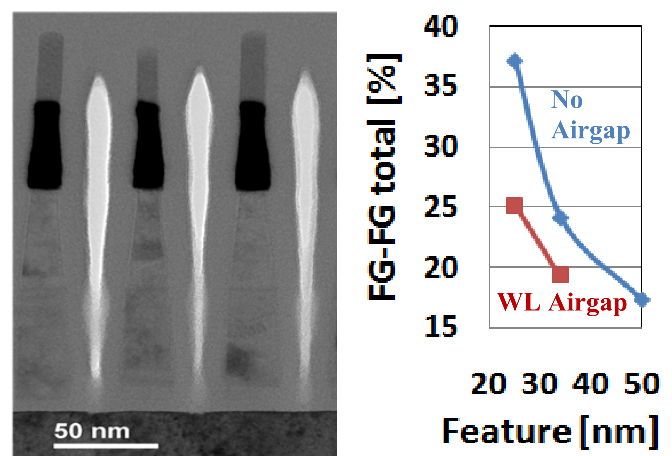

Figure 2: Cell cross section across wordlines showing the air gap for ~25% reduction in FG-FG and WL-WL capacitance

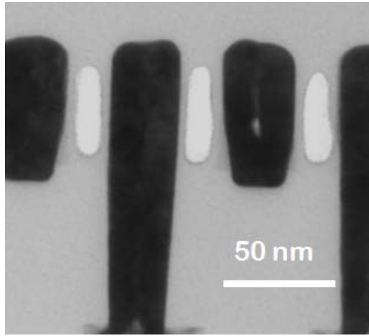

Figure 3: Cell cross section across the bitlines showing the W metallization and air gap for ~30% reduction in BL-BL capacitance

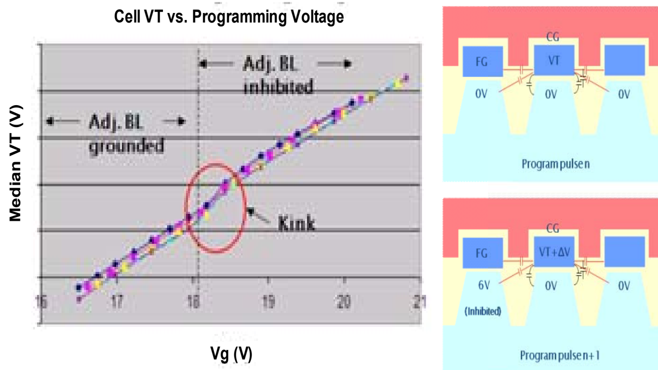

Figure 4: Behavior of the cell threshold highlighting the kink when adjacent bitline transitions from grounded to inhibited state and corresponding coupling diagrams

As is visible in the micrograph (Fig. 1), the floor plan of the die has the arrays rotated such that the data cache is perpendicular to the pads. This choice, made in order to optimize the number of die per reticle field, resulted in a number of design challenges. First and foremost among these, is the additional length of the signals traveling between the array drivers and power supplies located in the periphery of the chip and the array. Secondly, the additional routing of signals from the periphery to the array and data cache creates a routing channel which adds to the overall die size of the part.

To help address both the proximity challenges and the die size inefficiency, the bulk of the string driver, column decoder, redundancy, and data cache pre-driver circuits were moved from the periphery directly into the routing channel. With this, more than 50% of the diffusion in the routing channel is utilized and the impact to the overall die size is reduced to less than 1.5%. In addition, the primary power supplies for the array were relocated to the bottom corners of the array to minimize interconnect distance and the primary power bussing for the array was accomplished using metal directly over the array rather than through the routing channel.

III. PERFORMANCE

To enable 3bpc on 25nm, several key algorithmic features were included. The LMU program algorithm as described in [3] was refined to better mitigate FG-FG coupling. The erase distribution was compacted during programming of the middle page to reduce FG-FG coupling impact when placing cells to the upper page (Fig. 5). The well and source bias voltage were adjusted during verify and read to enable placing states below 0V which reduces the maximum programming voltage required while minimizing the SILC related charge loss [7]. Separate exit criteria for each page in the program flow were created so that a more aggressive exit strategy can be employed for upper page programming when the applied gate voltage is at its maximum. Finally, the program start voltage is adjusted based on the response during lower or middle page programming eliminating the need for guard banding which then speeds up the program algorithm. Fig. 6 shows the final distributions achieved with all the above mentioned techniques.

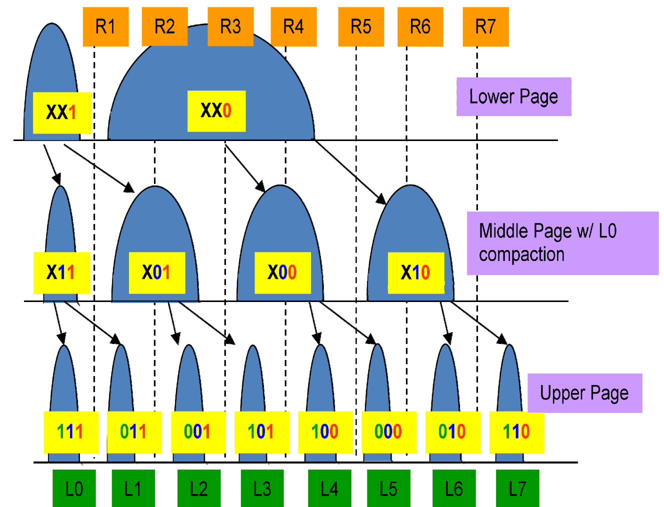

Figure 5: Schematic illustrating the LMU programming algorithm and the erase compaction during middle page programming to minimize impact of FG-FG coupling by reducing the ΔVT of neighboring cells

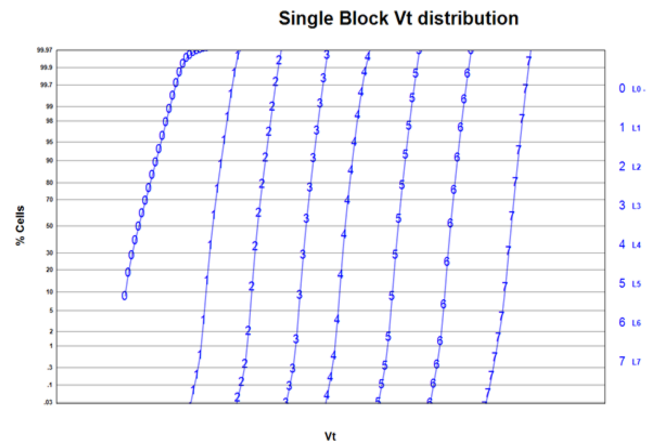

Figure 6: 25nm 3bpc placement for a single block showing 8 levels

In addition to algorithmic enhancements, introduction of an air gap between the wordlines was critical to reduce the FG-FG parasitic coupling thereby increasing the control gate to floating gate capacitance. The higher control gate to floating gate capacitance reduces the maximum program voltage

required and lowers program noise leading to tighter placement [8].

Finally, from an architectural point of view, the synchronous read performance is achieved through the application of a pipeline including a 4N prefetch with a 2:1 multiplexor in front of the sense amplifiers as shown in Fig. 7. Given the need to support asynchronous data transfer at 20ns/byte and synchronous data transfer at 6ns/byte with the extra routing from the rotated arrays, the times for each of the pipeline stages were reallocated and the overall data path re-optimized. Like the array drivers, the redundancy circuits have also been placed directly in the routing channel to maximize die efficiency. The redundancy circuits include bad column address storage, matching circuits, and a replacement mux.

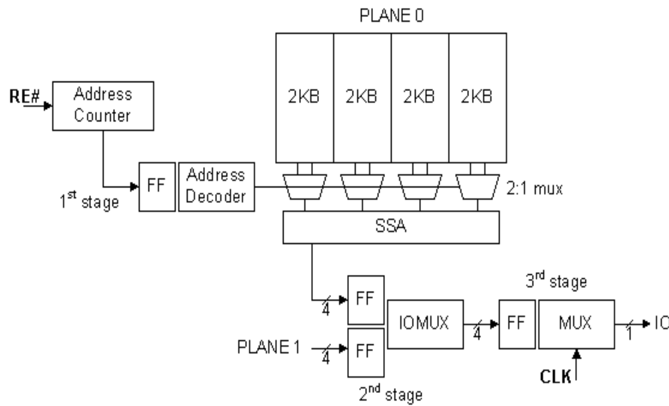

Figure 7: Datapath Architecture

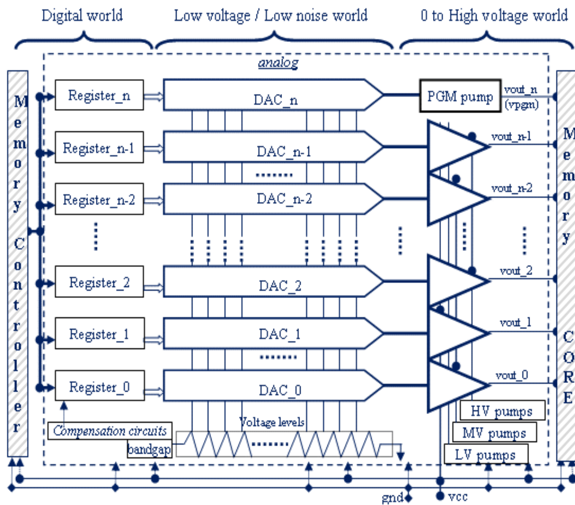

Figure 8: Analog Architecture

IV. PERIPHERY CIRCUITS

The analog architecture of the chip consists of a voltage reference generator, a digital thermometer with compensation circuitry, a bank of digital to analog converters along with linear voltage regulators, and triple well n-device (TWN) charge pumps. The system is built around a 10-bit resistor ladder feeding a bank of digital to analog converters to accurately perform analog operations in precision low-voltage low-noise environment. A digital thermometer is used to monitor system temperature and compensate the DAC outputs as needed for optimal array operation. The DAC outputs are then amplified by linear regulators, which use the TWN pumps as supplies. This entire system is controlled by the on-chip controller via dedicated control registers, as shown in Fig. 8.

The charge pumps are the largest physical analog block on the chip. The newly implemented TWN charge pumps use an isolated well for charge pump transfer devices to overcome the body effect which normally hinders charge transfer time (Fig. 9). This allows for significantly more efficient charge transfer between the pump stages, which when combined with increased pump clocking frequency, provides more output current per array. With more efficient pump arrays, it is possible to reduce the number of arrays required and reduce the overall size of the die. This chip achieved 35% pump area reduction over the non-TWN pump design while doubling the pump clock frequency.

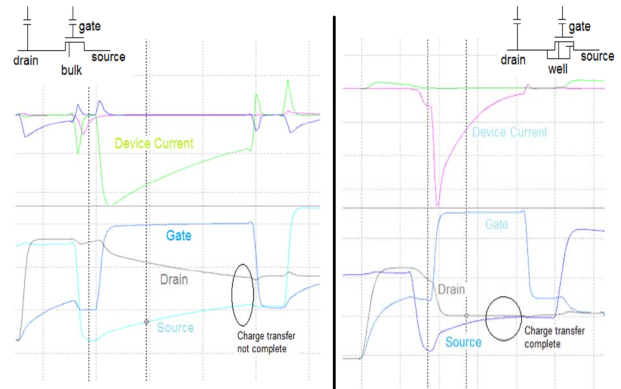

Figure 9: TWN Device Stage Charge Transfer

V. RELIABILITY

To achieve the stated reliability for the part, development focused on both enhancing the tunnel oxide and IPD beyond what was described in [2] to further reduce charge leakage and trapping during cycling. In addition, internal algorithms were optimized to reduce program disturbs and SILC leakage as described in section III. In addition, 3bpc has higher sensitivity to RTS beyond 2bpc [2] leading to wider state-widths and requiring tunnel oxide and channel implant optimization [9, 10]. As shown in Fig. 10, with these improvements, the part is now capable of meeting end customer reliability requirements.

Furthermore, on the algorithmic front, a new customer feature has been introduced that allows the external controller to retry a read operation that exceeds the maximum error recovery capability of the controller using one of several alternate read profiles and/or modified algorithms for more than 2x gain in endurance capability. For each of these retry options (7 alternate trim profiles are provided for this product), the NAND component adjust timings or reference values used

during the read or may even apply a different read algorithm. Fig. 11 shows just how simple it is for the controller to employ this new feature.

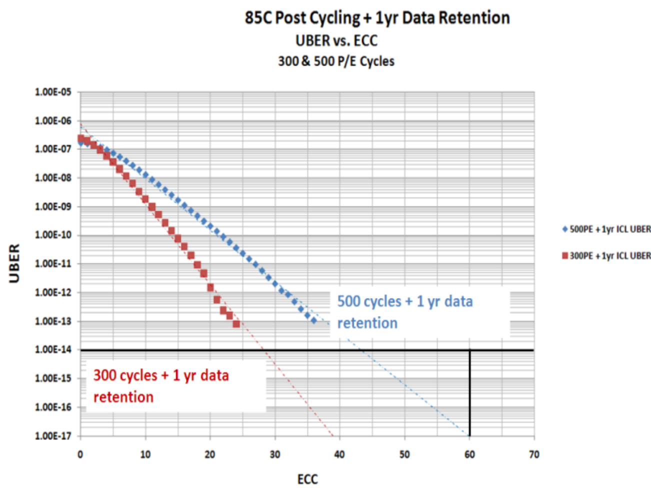

Figure 10: UBER vs. ECC post program/erase cycles and 1 year retention bake showing >500 cycles capability for requirement of 1E-14 at 60bits of ECC

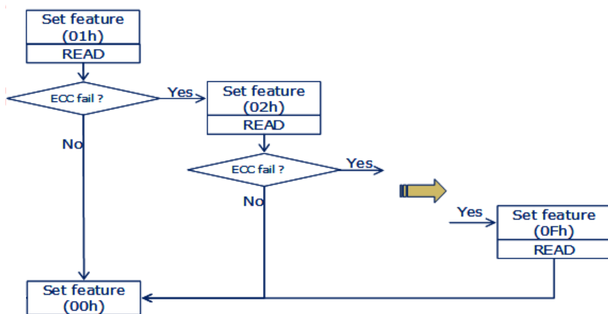

Figure 11: Read Retry Flow Diagram

VI. Conclusion

This paper introduces the first 25nm NAND technology-based 3bpc 64Gb with world-class memory density of 6.15GB/cm^2 while delivering competitive performance of 100µs read and 2300µs programming times with reliability meeting business requirements.

References

[1] Gartner Inc., "Gartner Says Worldwide Semiconductor Revenue to Grow 31.5 Percent in 2010", Sep. 9th, 2010.

[2] K. Prall & K. Parat, "25nm 64Gb MLC NAND Technology and Scaling Challenges", IEEE International Electron Device Meeting (IEDM), Dec. 2010.

[3] G.G. Marotta, et al., "A 3bit/Cell 32Gb NAND Flash memory at 34nm with 6MB/s Program Throughput and with Dynamic 2b/Cell Blocks Configuration Mode for a Program Throughput Increase up to 13MB/s", ISSCC Dig. Tech. Papers, pp. 444-446, Feb. 2010.

[4] R. Zeng, et al., "A 172mm^2 32Gb MLC NAND Flash Memory in 34nm CMOS", ISSCC Dig. Tech. Papers, pp. 236-237, Feb. 2009.

[5] R. Cernea, et al., "A 34MB/s-Program-Throughpu 16Gb MLC NAND with All-Bitline Architecture in 56nm", ISSCC Dig. Tech. Papers, pp. 420-422, Feb. 2008.

[6] G. Wong, Applications for Three and Four bit per cell NAND Flash Memories, Report No. FI-NFL-MBC-0210, Forward Insights, pp. 22-23, Feb. 2010.

[7] H. P. Belgal, et. al., "A new reliability model for post-cycling charge retention of flash memories", Reliability Physics Symposium Proceedings,pp. 7-20, Aug. 2002

[8] C. Compagnoni, et. al., "First evidence for injection statistics accuracy limitations in NAND Flash constant-current Fowler-Nordheim programming", IEDM Tech. Digest, Dec. 2007.

[9] C. Compagnoni, et. al., "Random Telegraph Noise Effect on the Programmed Threshold-Voltage Distribution of Flash Memories", IEEE Electron Device Letters, vol. 30, No. 9, pp. 984-986, Sept. 2009.

[10] A. Ghetti, et. al., "Scaling trends for random telegraph noise in deca-nanometer Flash memories", IEDM Tech. Digest, Dec. 2008.