13.1 A 1Gb 2GHz Embedded DRAM in 22nm Tri-Gate CMOS Technology

Fatih Hamzaoglu, Umut Arslan, Nabhendra Bisnik, Swaroop Ghosh, Manoj B. Lal, Nick Lindert, Mesut Meterelliyoz, Randy B. Osborne, Joodong Park, Shigeki Tomishima, Yih Wang, Kevin Zhang

Intel, Hillsboro, OR

CMOS technology scaling continues to drive higher levels of integration in VLSI design, which adds more compute engines on a die. To meet the overall performance-scaling needs, high-speed and high-bandwidth memory is becoming increasingly important. Conventional VLSI systems often rely on on-die SRAMs to address the performance gap between CPU and main memory, DRAM. However, with the rapid growth in capacity needs for high-performance memory, SRAM is not always sufficient to meet the demands of bandwidth-intense applications. Embedded DRAM (eDRAM) has been explored as an alternative to satisfy the high-performance and density needs in memory [1-3]. In this paper, a high-performance eDRAM based on a 22nm tri-gate CMOS technology is introduced. This eDRAM technology enables the integration of an eDRAM cell into the logic technology platform [4]. The design features a well-balanced configuration to achieve both optimal array efficiency and bandwidth. By leveraging the high-performance and low-voltage tri-gate transistor at 22nm generation, the eDRAM achieves a wide range in operating voltage, from 1.1V down to 0.7V, which is essential for low-power logic applications.

Figure 13.1.1 shows a 4th-generation Core™ processor, where the CPU is connected to a 1Gb eDRAM die through on-package-IO (OPIO) [5]. This multi-chip-package (MCP) product, the Iris Graphics Pro™, uses eDRAM as L4 cache and provides low-power high-bandwidth memory access to meet high-performance graphics segment needs. The eDRAM bitcell features a low-leakage access transistor in high-k tri-gate bulk technology and a MIM storage capacitor with capacitance of greater than 13fF [4,6]. The eDRAM cell area is 0.029μm^2, less than one-third of the high-density 6T-SRAM bitcell offered in the same 22nm technology [7], enabling design of high-density memory. The bitcell adopts the capacitor-over-bitline (COB) architecture to maximize the surface area of the capacitor. To support high-performance logic in the eDRAM design for GHz operation, the COB is embedded into the high-performance Cu-metallization interconnect layers. Negative word-line voltage (VSS_WL) with wide programmable range is employed to reduce access-transistor leakage (Fig. 13.1.2). To achieve high data-retention time, VSS_WL and threshold voltage of the access transistor are co-optimized to balance subthreshold leakage and gate-induced drain leakage at the storage node. The access transistor is turned on with the wordline overdriven to VCC_WL by an on-die charge pump (CP) to allow the storage capacitor to be fully charged to VCC and to achieve fast sensing and data write-back operations. The range of wordline swing during off and on states is largely determined by the reliability requirement of the access transistor. A NOR-based wordline driver is shown in Fig. 13.1.2, where the input signal voltage swings are designed to limit the gate bias to VCC_WL level and below to meet reliability needs. The final WL-driver voltage for the groups of drivers that are not accessed is kept at VCC in order to minimize the CP loading and leakage power. This design also avoids dual level-shifting circuitry in the same gate, which further reduces the design complexity.

The 256Kb-subarray architecture is shown in Fig. 13.1.3. The array has an open-bitline architecture with 128+ cells on each side, including redundant rows. Similarly, each wordline has a total of 1024+ columns, including redundant columns. The subarray achieves 65% area efficiency. The subarray reads or writes 128+ bits and each bit-slice contains its own set of half-VCC local bitline precharge circuitry, sense amplifier, and 8:1 column mux, as described in Fig. 13.1.3. Subarrays also contain local half-VCC generators, which are programmable for optimal sensing margin. Four bitcell operations are also shown in Fig. 13.1.3, including sense, write-back, wordline-turn-off and local bitline precharge.

Figure 13.1.4 shows the 1Gb array configuration and data-path for read, write and refresh operations. The chip contains 128 independent banks for read and write and 64 bank-groups for refresh, where bank random cycle time (RCT) is equal to six array clock cycles. By providing large number of banks and short RCT, we minimize bank conflict for high-bandwidth random accesses and maximize performance. Four vertical 256Mb quarters are activated simultaneously during each operation, where each bank reads out 64×2 bits in two consecutive cycles after column and row repairs to get 512b-wide word size. The OPIO is clocked at twice the array frequency and double data-rate to meet area and bandwidth requirements. The array has separate data buses for read and write operations but shares a common address bus, hence, it supports read and write operations in alternating array clock cycles to different banks. The refresh operation can occur during a read or write since it has a separate refresh bank-group address. There are two copies of CPs and regulators, each supporting the top or bottom 512Mb, which occupy less than 2% of the die area. The chip also contains fuses, programmable built-in self test (PBIST), test access port (TAP) and a digital thermal sensor (DTS).

Figure 13.1.5 describes the CP circuits that support wordline over- and under-drive voltages. A portion of the CPs always run to support leakage and the remaining portions are activated only when there is array access (read, write or refresh) to compensate the wordline-activation charge. Although positive and negative CPs can generate up to 2*VCC and –(VCC/2) output voltages, respectively, output voltages are regulated to a programmable value through on-die-generated reference voltages. The cumulative distribution of both VCC_WL and VSS_WL measurement data at hot (95°C) and cold (-10°C) are also shown in Fig. 13.1.5 for unregulated and two different regulator setting cases. Silicon data show that regulated voltages have less die-to-die variation and better control across temperatures. By introducing programming control over the CP output voltages, the design is able to provide a large window to compensate process and temperature variation while achieving a balance between performance and reliability.

Figure 13.1.6 shows the voltage-frequency shmoo of the 1Gb eDRAM array tested at 95°C and 100μs retention time. The design achieves 2GHz operation frequency at a supply voltage of 1.05V, hence 3ns of RCT. The array also supports a wide range in power supply, down to 0.7V at 1GHz frequency.

Figure 13.1.7 shows die micrograph of the 1Gb eDRAM and the feature summary table. The chip is fabricated in a 22nm high-performance tri-gate CMOS technology. The die size is 77mm^2 with 0.029μm^2 eDRAM bitcell area and an array density of 17.5Mb/mm^2 at the 128Mb macro level.

Acknowledgements:
The authors gratefully acknowledge many members of PTD and IDG technical staffs for their contributions to this work.

References:

[1] J. Barth et al., "A 45nm SOI Embedded DRAM Macro for POWER7™ 32MB On-Chip L3 Cache", *ISSCC Dig. Tech. Papers*, pp. 342-344, Feb. 2010.
[2] K. Hijioka et al., "A Novel Cylinder-Type MIM Capacitor in Porous Low-k Film (CAPL) for Embedded DRAM with Advanced CMOS Logics," *IEDM Technical Digest*, pp. 756-759, Dec. 2010.
[3] S. Romanovsky et al., "A 500MHz Random-Access Embedded 1Mb DRAM Macro in Bulk CMOS", *ISSCC Dig. Tech. Papers*, pp. 270-271, Feb. 2008.
[4] R. Brain et al., "A 22nm High Performance Embedded DRAM SoC Technology Featuring Tri-Gate Transistors and MIMCAP COB", *VLSI Tech. Symp.,* June 2013.
[5] N. Kurd et al., "Haswell: A Family of IA 22nm Processors," *ISSCC Dig. Tech. Papers*, Feb. 2014
[6] Y. Wang et al., "Retention Time Optimization for eDRAM in 22nm Tri-Gate CMOS Technology," *IEDM Technical Digest*, Dec. 2013.
[7] E. Karl et al., "A 4.6GHz, 162Mb SRAM Design in 22nm Tri-Gate CMOS Technology with Integrated Active Vmin-Enhancing Assist Circuitry," *ISSCC Dig. Tech. Papers*, pp. 230-232, Feb. 2012.

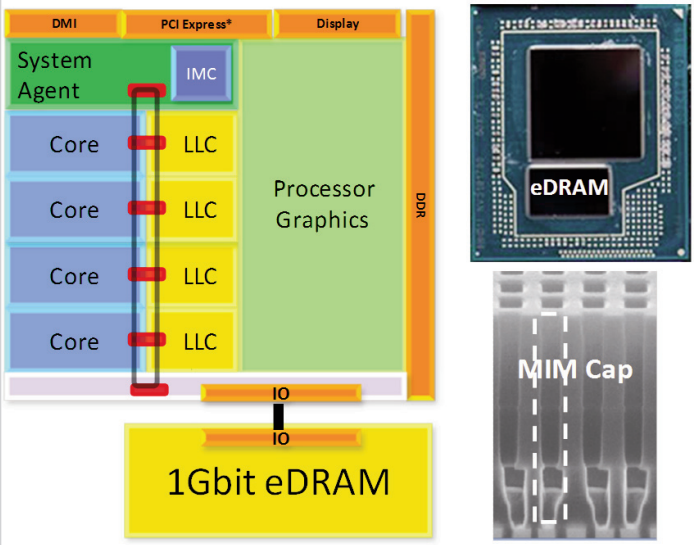


Figure 13.1.1: Intel Iris Pro™ with 1Gb eDRAM in 22nm tri-gate technology.

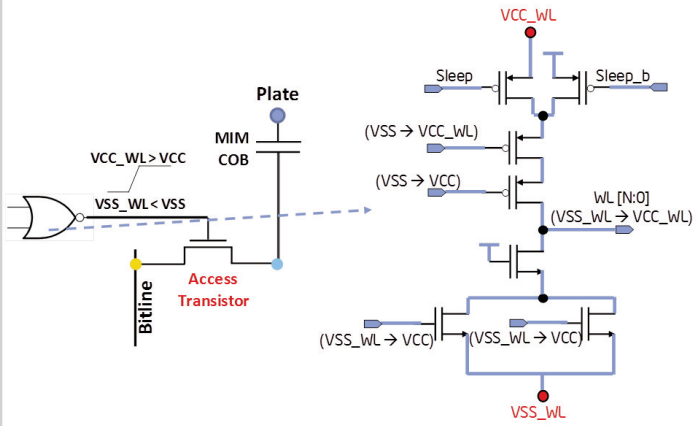


Figure 13.1.2: NOR WL driver with V_{max} protection and leakage reduction.

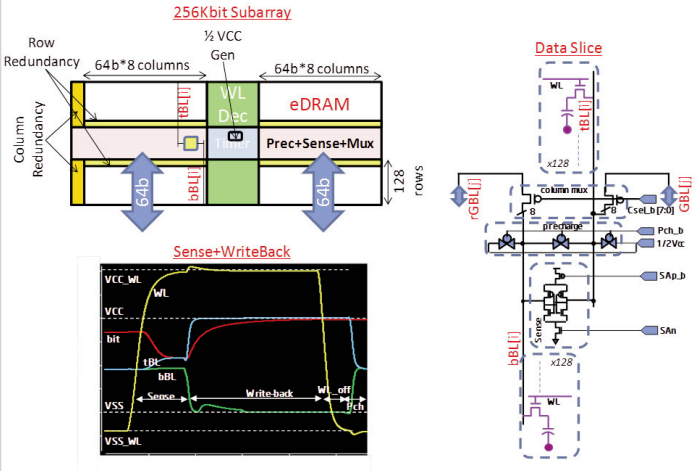


Figure 13.1.3: Floorplan of 256Kb eDRAM subarray and bit-slice architecture.

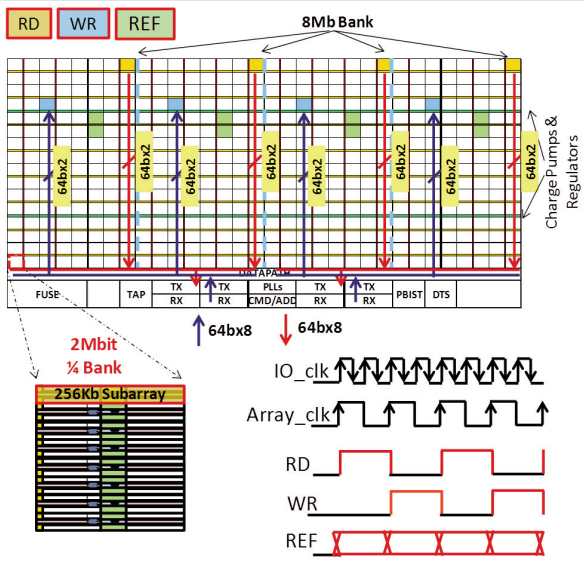


Figure 13.1.4: 1Gb eDRAM architecture and RD/WR/REF operations.

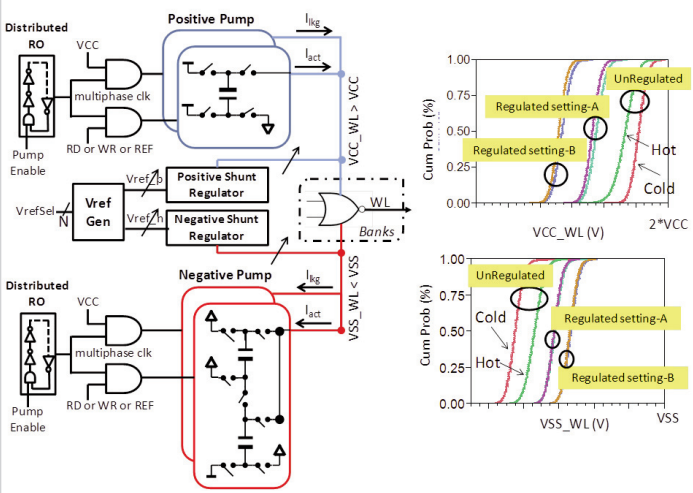


Figure 13.1.5: Charge pumps for WL voltages and measured silicon data.

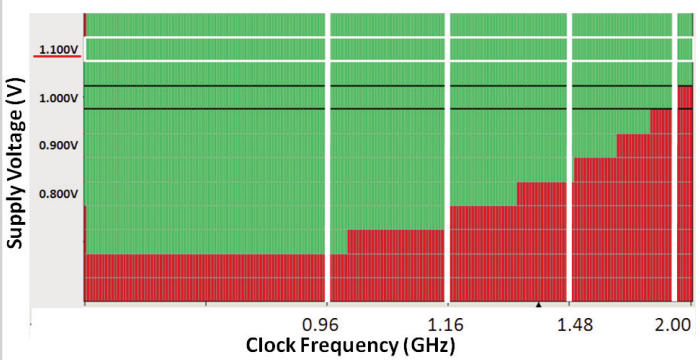
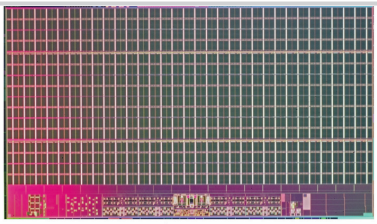


Figure 13.1.6: Frequency shmoo of 1Gb eDRAM array at 95°C and 100µs retention time.



Technology	22nm Tri-gate CMOS
Cell Size	0.029μm^2
Macro Area	17.5 Mb/mm^2 @ 128Mbit Macro
Chip Organization	¼ Bank: 8 Subarrays (2Mb) Bank: 4 Quarter Banks (8Mb) Chip: 128 Banks (1Gb)
Subarray Configuration & Array Efficiency	256 Word-line x 1024 bit-lines & 65%
Chip Size	77mm^2
Supply	1.05V
Clock, Random Cycle Time	2GHz, 3ns
Retention Time	100μs @95C

Figure 13.1.7: Die micrograph and feature table of 1Gb eDRAM die.

978-1-4799-0920-9/14/$31.00