23.2 A 4b/Cell 8Gb NROM Data-Storage Memory with Enhanced Write Performance

Ran Sahar1, Avi Lavan1, Eran Geyari2, Amit Berman1, Itzic Cohen1, Ori Tirosh1, Kobi Donan1, Yair Sofer1, Yoram Bitzer1, Amichai Givanti1, Alexander Kushnarevko1, Yaal Horshi1, Ron Elyahu2, Eduardy Muayap1, Boaz Eitan1, Wang Pei Jen1, Yan Feng2, Lin Ching Yao2, Kwon Yi Jin2, Kwon Sung Woo2, Cai En Jing2, Yi Jing Jing2, Kim Jong Oh2, Yi Guan Jiun2

1Safun Semiconductors, Netanya, Israel, 2SMIC. Shanghai, China

The increasing demand for cost reduction of data storage solutions calls for both cell-size shrink, as well as compressing more bits of data into the storage element. Yet the low cost solution is required to have fast enough write operation to fit density-hungry applications. We present an 8Gb data flash storage device based on 4b/cell NROM technology. A major improvement in write time is accomplished through improvement of programming techniques, erase scheme and sensing methods.

A key issue in multi-bit design is to accurately place the four distributions calls for both cell-size shrink, as well as compressing more bits of data into the storage element. Yet the low cost solution is required to have fast enough write operation to fit density-hungry applications. We present an 8Gb data flash storage device based on 4b/cell NROM technology. A major improvement in write time is accomplished through improvement of programming techniques, erase scheme and sensing methods.

The NROM cell has two physically separate charge-storage areas per cell [3]. The program characteristics of a storage area in a cell depend on the programming level of the other storage area in the same cell. The more a storage area in a cell is programmed, the lower the drain voltage that is required to program the other storage area. This basic nature of operation leads to a further possible improvement in programming speed. Specifically, two different bit line voltages are generated per level, and the selection between the two, when coming to program a storage area in a cell, depends on the programming state of the other storage area in that same cell. The circuit in the right Fig. 23.2.2 shows the final scheme resulting in six different voltage levels per bit slice.

A total of 256 sense amplifiers are used for program verify of all 3 levels in parallel. During the verify stage, the original programmed data pattern is transferred from the SRAM to the bit slice and choose to verify. The data pattern is transferred from the SRAM to the bit slice and choose to verify. The data pattern is transferred from the SRAM to the bit slice and choose to verify.

The NROM four-bit program algorithm includes two phases. The first phase is based on a drain-stepping algorithm, while the second phase is based on a gate-stepping algorithm [1]. The internal data path bus is 256 bits wide, where each of the bits slices has its own control circuitry both for program, erase and read operation (Fig. 23.2.1). The bus may consist of various data combinations to program and verify, resulting in a different program level per control circuitry. In order to program all different levels at different array locations, at least three different bitline (BL) voltage generators are required. For example, programming level 01 may require a BL start voltage of 3.5V, whereas programming level 10 may need to start at 4.5V. Each bit slice consists of a dedicated BL voltage driver as shown in Fig. 23.2.2. The driver is built up from 3 independent source followers shorted to one output via a PMOS switch. The selection from the 3 possible voltage levels depends on the program data in this bit slice. Three different voltage buffers are commonly used to drive all 256 bits slices as shown in Fig. 23.2.2.

The NROM cell has two physically separate charge-storage areas per cell [3]. The program characteristics of a storage area in a cell depend on the programming level of the other storage area in the same cell. The more a storage area in a cell is programmed, the lower the drain voltage that is required to program the other storage area. This basic nature of operation leads to a further possible improvement in programming speed. Specifically, two different bit line voltages are generated per level, and the selection between the two, when coming to program a storage area in a cell, depends on the programming state of the other storage area in that same cell. The circuit in the right Fig. 23.2.2 shows the final scheme resulting in six different voltage levels per bit slice.

A total of 256 sense amplifiers are used for program verify of all 3 levels in parallel. During the verify stage, the original programmed data pattern is transferred from the SRAM to the bit slice and choose to verify. The data pattern is transferred from the SRAM to the bit slice and choose to verify. The data pattern is transferred from the SRAM to the bit slice and choose to verify.

The basic virtual-ground array slice [2] in this design consists of 36 NROM cells located between two isolations. The 4K × 256 byte page is comprised of 256 bit slices, each having 34 NROM cells with 4b data. The edge bits in a slice store internal error-detection (ED) bits [1]. Each main bitline (MBL) connects to the local BL via either 6 or 7 select transistors (Fig. 23.2.3). A total of 2 MBL are used to access the memory-cell drain and source terminals both in read and program operation.

The basic virtual-ground array slice [2] in this design consists of 36 NROM cells located between two isolations. The 4K × 256 byte page is comprised of 256 bit slices, each having 34 NROM cells with 4b data. The edge bits in a slice store internal error-detection (ED) bits [1]. Each main bitline (MBL) connects to the local BL via either 6 or 7 select transistors (Fig. 23.2.3). A total of 2 MBL are used to access the memory-cell drain and source terminals both in read and program operation.

One of the challenges of sensing a virtual-ground array [2] is to eliminate possible drain-side leakage current. The main limiting factor is the fact that drain-side charging not only drives the cell current, but also charges the side pipe capacitance. The severity of this effect varies depending on the actual data pattern programmed in the specific slice. The design uses a read technique that overcomes the drain-side leakage-current problem by means of pipe precharging. Page read operation is a sequential flow—the first step in the scheme is to perform a dummy read to precharge the pipe to a voltage that is close to the final BL drain voltage. Figure 23.2.3 shows the slice structure together with the read order. The sequence is such that the drain terminal is always kept in the same direction, and side leakage is reduced to minimum since the voltage potential across the neighbor cell of the cell being sensed is greatly reduced.

Fig. 23.2.4 depicts the read drain driver circuit. The drain bias is determined by devices T1 and T2, which are biased with Vcs and Vcs respectively. These biases are common to all 256 sense amplifiers, together with the three sense amplifiers used for the read REF cells [1]. The actual sensing current is a function of a constant current driven by T1 NROM devices together with the core cell current. This current charges the integration capacitor, Cin, which is connected as an input to the sense amplifier. The same scheme is used for the 3 REF sense amplifiers required to detect the four levels [1].

System write performance is greatly affected by the total block-erase time. Prior to erasing a memory block, pre-programming is required to avoid over erase phenomena and degradation of program-erase marginality. Pre-programming, however, may consume significant portion out of the total erase time. The approach in this design is to introduce data scrambling during user-page programming command. This allows skipping the pre-program phase before erase since the rotating scrambling patterns assures that each cell suffers only limited over erase condition, and endurance is not be degraded. The data scrambling is shown as an example in Fig. 23.2.5.

Data scrambling is implemented by performing a bitwise XOR between user data and a fixed scrambling pattern (Fig. 23.2.2). During a read command, the page is loaded to the internal SRAM buffer, only after going through a de-scrambling operation.

The NROM cell erase mechanism is based on tunnel enhanced hot hole (TEHH) injection [3]. During erase, cell gates are biased with negative potential. The drain side is biased with high voltage, while the other side is kept floating. Since this operation consumes significant charge-pump power, the erase time is limited by the maximum number of cells being erased in the same pulse. Moreover, the erase current-pulse characteristic is such that at a given fixed drain voltage, a large peak current is drawn from the charge pump, and this peak decays throughout the erase pulse. The solution is to implement a fixed-current driving to the BL, using a feedback loop from the charge pump, the internal micro controller, and the BL erase voltage regulator. As shown in Fig. 23.2.6, a comparator circuit compares the voltage drop on a given resistor (R1) with a scaled-down resistor connected to some of the charge pump phase drivers. During the erase pulse, the micro controller applies an initial voltage step to the BL voltage regulator. The micro controller steps up the voltage as long as the current drawn from the charge pump is below the threshold. Once this threshold is reached, the BL voltage stays at a fixed level until the pulse ends. This scheme better exploits the charge pump current drive and boosts erase performance by handling more cells in parallel.

The 8Gb die micrograph is shown in Fig. 23.2.7 with its major blocks highlighted. The manufacturing process is based on a wordline half-pitch of 90nm, and a BL half pitch of 100nm cell array integrated with CMOS. Divided by the bit number per cell, this device achieves bit size of 0.009µm², and die size of 155mm².

References

Figure 23.2.1: Block diagram of data path.

Figure 23.2.2: Schematics of triple BL driver (right) and hex BL driver (left).

Figure 23.2.3: Slice structure and read order.

Figure 23.2.4: Schematic of drain driver.

Figure 23.2.5: Data scrambling flow.
Figure 23.2.6: Constant-current erase.

Figure 23.2.7: Die micrograph.