AN998
APPLICATION NOTE

FLASH+™ Multiple Memory Technology

EPROM, Flash and EEPROM devices all use the same basic floating-gate mechanism to store data, but they use different techniques for reading and writing. This application note discusses the similarities and differences between these technologies, and introduces ST's new FLASH+ multiple memory technology.

In each technology mentioned above, the memory cell consists of an MOS transistor with two gates:

- a control gate that is connected to the read/write control circuitry
- a floating-gate (located between the control gate and the MOSFET channel) that is completely surrounded by an insulating layer of silicon dioxide.

Because the floating-gate is very close to the MOSFET channel, even a small electrical charge on it has a readily detectable effect on the threshold voltage of the transistor. This mechanism is usually used to store a single bit, which is read by comparing the present threshold voltage with a reference value. (With more sophisticated read/write techniques it is possible to distinguish between more than two charge states, thus allowing two or more bits to be represented on each floating-gate.)

Several techniques are available for moving electrons to and from the electrically isolated floating-gate:

1. Channel Hot Electron injection (CHE)
2. High energy light rays
3. Electron tunnelling

The first method is used in EPROM and flash memory devices, for charging the floating gate. The MOSFET channel is filled with high energy electrons by applying relatively high voltages to the control gate and the drain. Some of these "hot" electrons have sufficient energy to cross the potential barrier between the channel and the floating-gate. When the high voltages are removed, these electrons remain trapped on the floating-gate.

In EPROM devices, the discharging of the floating-gates is achieved in parallel by flooding the entire memory array with ultra-violet light. The high energy light rays penetrate the chip structure and impart enough energy to the trapped electrons to allow them to escape from the floating-gate. As well as being simple and effective, this method is immune to problems of over-erasure: continuing to expose the gate to UV light, after it has already been discharged, does not discharge it further to some non-operative region.

The third method, used in EEPROM devices, relies on a quantum-mechanical effect called *tunnelling.* A voltage is applied to the source that is sufficient to cause electrons to "tunnel" across the insulating oxide layer. This effect can be used both to charge or discharge the floating-gate, according to the polarity of the applied tunnelling voltage.

For an EPROM device, the thickness of the oxide layer that separates the floating-gate from the source is typically 20 to 25 nm (200 to 250 Angstroms). For EEPROM, though, the number of electrons that can tunnel across in a given time depends on the thickness of the layer and the value of the applied voltage. Consequently, in order to meet realistic voltage-level and erase-time constraints, the insulating layer has to be

very thin, typically 10 to 13 nm (100 to 130 Angstroms). The quality of this oxide layer has a profound effect on the performance and reliability of the device.

Table 1. Summary of Methods for Charging and Discharging the Floating-gate

Technology	Charge	Discharge
EPROM	channel hot electron	ultra-violet light
Flash memory	channel hot electron	tunnelling
EEPROM	tunnelling	tunnelling

From the summary in Table 1, it can be seen that flash memory is programmed like an EPROM and erased like an EEPROM. It therefore inherits many of the performance and reliability limitations of both EPROMs and EEPROMs. On the other hand, flash memory avoids the main disadvantages of both: it has the electrical erasability of EEPROM at a packing density close to that of EPROM. The high packing density is made possible by a compromise in which, unlike EEPROM, flash memory is erased a sector at a time.

This, however, brings its own technological problems. A special erase algorithm is implemented, in hardware, that first writes to all the cells in the sector, to ensure that every floating-gate is fully charged, and then applies a short erase pulse to all of the cells simultaneously. The algorithm tests the state of the cells in the sector, and repeats the erase-test cycle until all of the cells have been erased.

A problem may occur, however, if one cell discharges too quickly. The erase pulse will continue to be applied to it, until all the other cells have been erased, so causing the problem cell not only to be discharged further, but to become increasingly charged at the opposite polarity. Charged with a surfeit of holes (over depleted of electrons) the transistor will be placed in its depletion mode of operation. The problem cell will exhibit a “stuck at” fault. The entire chip will thereby be rendered useless.

This problem can be easily prevented in low density EEPROMs by design techniques that are not feasible in multi-megabit flash devices. The solution is to combine optimized circuit design with a meticulously tuned process that ensures a uniform and tightly controlled distribution of cell characteristics. For example, increasing the thickness of the oxidation around the corners of the floating-gates prevents the occurrence of high, local electric fields that can accelerate discharge. Special test modes and structures can also be built into the devices to assist in process monitoring.

THE NEED FOR MULTI-TYPE MEMORY DEVICES

The reason that all three technologies survive, and continue to be supported, is that they each have their own advantages. EEPROM is the most versatile of the non-volatile memory technologies, but is also the most expensive. Since cells can be erased individually, or on a bytewise basis, each cell must include a switch to enable it to be isolated from its neighbours. There must be as many switches to do this as there are cells in the memory device. The most efficient EEPROM device, therefore, needs a two transistor cell.

EPROM, on the other hand, does not need transistors to isolate the individual bytes during the erase cycle, and so can be implemented with a one transistor cell. EPROM is, therefore, the cheapest of the non-volatile memory technologies discussed here. Large areas of storage that do not need to be changed often are more effectively implemented in EPROM than in EEPROM.

Flash memory attempts to achieve the advantages of both, and the disadvantages of neither. With a single transistor cell, it achieves most of the advantages of EPROM, allowing it to be used for large, unchanging storage, such as for holding the microcontroller’s executable program. However, this is made possible by

limiting its erase operation only to whole sectors at a time. Flash memory, therefore, does not quite achieve the flexibility of EEPROM.

For some applications, where the flexibility of EEPROM is required, it is necessary to include two memory chips:

- one for the large executable program
- one for the frequently changing non-volatile data (data that needs to be retained from one power-on period to another, but which can change often during any one power-on period).

To get round this problem, techniques are available that can emulate the flexible behaviour of EEPROM in flash memory, thus allowing a single chip to be used for both types of memory. These techniques are introduced in *AN997*, and analysed in more detail in *AN931*. A glossary for these techniques is available in *AN996*. These application notes, listed in Table 2, are written to highlight the advantages offered by the M39432 memory, implemented in ST's new FLASH+™ technology.

Table 2. Bibliography of the Application Notes on FLASH+ Technology

Application Note	Content	Title
AN997	Introduction	M39432: a FLASH+™ Multiple Memory Device
AN931	Greater detail	On-Chip Hardware EEPROM Emulation versus Flash Memory Software Solutions
AN996	Glossary	Emulating the Unification of Flash and EEPROM: a Glossary
AN998	Technology	FLASH+™ Multiple Memory Technology
AN999	Drivers	Software Drivers for the M39432 FLASH+™ Multiple Memory

FLASH+ TECHNOLOGY

FLASH+ technology uses the same 0.6 micron process that is used for our standard Flash memories. The main difference is the use of a thinner tunnel oxide layer. The result is a "double poly/double metal" process that is a little more complex than the industry standard "double poly/single metal" EEPROM process, but provides significantly greater memory densities. For example, a 1 Mb parallel EEPROM complete with error correction circuitry could be built with a cell size less than that of a 256 Kb serial EEPROM built with the industry standard process.

Because this new technology is so closely related to flash, it has opened up the possibility of integrating both types of memory on a single chip. This, in turn, leads to further efficiency, since much of the support circuitry (such as charge pumps for on-chip generation of the programming voltage, address logic, state machine and I/O buffers) can all be shared.

The M39432 features all of the following on a single chip:

- 4 Mb array of Flash memory, with:
  - sector protection
  - erase suspend/resume features
- 256 Kb array of full-featured EEPROM, with:
  - software data protection
  - 64-byte page-mode
  - enhanced end-of-write detection.

If you have any questions or suggestions concerning the matters raised in this document, please send them to the following electronic mail address:

apps.eeprom@st.com

Please remember to include your name, company, location, telephone number and fax number.