Applications Emerging to Employ Embedded Non-Volatile Memory

May 2011

By: Linh Hong
Vice President of Marketing
Kilopass Technology Inc.
www.kilopass.com

To learn more about Kilopass, stop by the Chip Estimate Booth (#1731) at DAC.

Abstract: The increasing complexity of system on chip (SoC) designs being implemented in process geometries 40nm and below is creating demand for multi-time programmable (MTP) embedded non-volatile memory (eNVM) to replace embedded read-only memory (ROM) and external flash memory. NVM based on floating-gate technology, flash memory and some forms of EEPROM (MTP) cannot be implemented in process geometries below 90nm. The ROM supplied by foundries is undesirable because of security concerns, inability to migrate to other foundries, and large area for memories density above four kilobits (4kb). A new MTP NVM solution that leverages anti-fuse technology addresses both shortcomings and provides solutions to a number of high-volume consumer and mobile applications.

Other benefits include cost reduction, improved performance, enabling secure storage and configurability. This paper detail embedded NVM technologies and applications that benefit most from using embedded NVM.

SoC designers requiring NVM storage in their designs have a number of options to choose from, including the following:

• Embedded flash solutions that use either a split-gate architecture or a floating-gate architecture, depending on area, power and performance requirements;
• Via or diffusion ROM that produce “1”s and “0”s through a double metal via or a change in Vt behavior from the implant in the diffusion;
• Electrical fuse solutions—eFuses, typically provided by the foundry, that blow the silicide on the poly line, creating a change in resistance;
• Floating gate or charge trapping solutions that use hot-carrier injection as the programming mechanism; and
• antifuse solutions that produce “1”s from a hard oxide breakdown of the gate, causing a resistive change;

The summary in Table 1 below contrasts the benefits and shortcomings of each of these memory types, including: (1) cell structure, (2) logic process compatibility, (3) bit cell and macro areas, (4) endurance, (5) standby and active current, (6) random access time, (7) security, (8) scalability, (9) additional manufacturing steps, (10) manufacturing backend turnaround time (TAT) and (11) voltage/temperature tolerance.
Table 1: Embedded NVM Benchmark

<table>
<thead>
<tr>
<th>Cell Structure</th>
<th>Embedded Flash</th>
<th>ROM</th>
<th>Electrical Fuse (OTP)</th>
<th>CMOS Floating Gate (MTP)</th>
<th>CMOS Floating Gate (OTP)</th>
<th>Antifuse (OTP)</th>
<th>Antifuse (MTP)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Standard CMOS Compatible</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Bitcell Area (Normalized)</td>
<td>3</td>
<td>&lt;1</td>
<td>300</td>
<td>20</td>
<td>10</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>128Kb Area in 65nm (Normalized)</td>
<td>n/a</td>
<td>0.5</td>
<td>8</td>
<td>24</td>
<td>n/a</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Endurance</td>
<td>100-1000k</td>
<td>No</td>
<td>No</td>
<td>10K</td>
<td>&lt;5</td>
<td>1</td>
<td>Up to 1000</td>
</tr>
<tr>
<td>NStandby &amp; Active Current</td>
<td>Med</td>
<td>Low</td>
<td>High</td>
<td>Med</td>
<td>Med</td>
<td>Low</td>
<td>Low</td>
</tr>
<tr>
<td>Security</td>
<td>Med</td>
<td>Low</td>
<td>Low</td>
<td>Med</td>
<td>Med</td>
<td>High</td>
<td>High</td>
</tr>
<tr>
<td>Scalability</td>
<td>Med (up to 90nm)</td>
<td>High</td>
<td>High</td>
<td>Med</td>
<td>Low (mainstream up to 180nm)</td>
<td>High</td>
<td>High</td>
</tr>
<tr>
<td>Additional Steps</td>
<td>+10 Mask</td>
<td>None</td>
<td>None</td>
<td>Wafer Bake</td>
<td>UV Erase, Wafer Bake</td>
<td>None</td>
<td>None</td>
</tr>
<tr>
<td>Manufacturing Backend TAT</td>
<td>None</td>
<td>+30 days to respin</td>
<td>None</td>
<td>None</td>
<td>None</td>
<td>None</td>
<td>None</td>
</tr>
<tr>
<td>Low Temperature &amp; Voltage Tolerance</td>
<td>Low</td>
<td>High</td>
<td>Med</td>
<td>Low</td>
<td>Low</td>
<td>High</td>
<td>High</td>
</tr>
</tbody>
</table>

1.0 Embedded Flash

Embedded flash is the memory of choice for code storage that changes often. It is the most expensive of the embedded NVM technologies by far requiring at least 10 additional mask steps. Although there is a higher upfront cost of the technology, it is the most flexible with high endurance thus code can be programmed post fabrication and can be changed often. Common in microcontrollers (MCU), embedded flash provides flexibility to end applications by configuring a variety of different products from a single silicon image. For example, an MCU vendor may sell a low end 8-bit MCU to 12 washing machine makers. With embedded flash, the MCU vendor need only develop one product to fit the needs of all the washing machine makers. The washing machine makers load the embedded flash with the unique code for each make and model using the same MCU. Both the vendor and the customer benefit from the flexibility to control inventory and provide differentiation.

The other major drawback of embedded flash for SoC designers building devices in leading-
edge process technologies is a big one. Embedded flash lags the advanced logic process
technology by at least three generations. Furthermore, with each new leading-edge logic
process generation, the endurance of flash memory decreases and uncorrectable bit error
rates increase. Technical literature reports that embedded flash technology has a hard stop
at 20nm.

2. 0 Read-Only Memory (ROM)

ROM is the exact opposite of flash; it cannot be programmed post fabrication. It is used for
code storage that is fixed. This code does not change often, such as audio recordings of
“Happy Birthday” in a musical greeting card or fonts in an ink jet printer. It is the least
expensive of the embedded NVM technologies but also the least flexible; if a designer needs
to change the content post silicon, a minimum of 30 days is required for respin. ROM
programming is done during fabrication. Once the IC is fabricated, the content cannot
change without re-spinning mask layers. There is also added cost of inventory with ROM
solutions; choosing the wrong content can result in bad parts or extra inventory for
products that aren’t sold quickly.

The two types of foundry-supplied ROM are Via or diffusion ROM. The former produces a
“1” or “0” by connecting or disconnecting a double metal via in the CMOS ROM transistor
comprising the ROM memory array. The latter changes the Vt behavior of the implant in
the diffusion in a CMOS ROM transistor comprising the ROM memory array thus turning the
transistor permanently on or off to store a “1” or “0”.

3.0 Standard CMOS technologies

Unlike ROM, the other standard CMOS technologies, electrical fuses, floating gate and
antifuse can be programmed post fabrication. One pervasive use of these technologies is to
recover parametric yield loss due to process variation or design marginality. For example,
Ethernet requires +/- 100 ppm clock frequency accuracy that may not be achieved with
silicon design, unless there is fine tuning during automated test equipment (ATE) test.
Likewise, Ethernet has a digital-to-analog convertor (DAC) that needs calibration for the
target voltage range. Calibrated values for the DAC or trim values for the clock can be
stored using NVM technologies.

3.1 Electrical Fuse (eFuse)

Electrical fuse (eFuse) technology is a one-time programmable memory programmed by
forcing a high-current density through a conductor link to rupture it or make its resistance
significantly higher. The modern eFuse is built on polysilicon with Cobalt or Nickel silicide
on top. The fuse is programmed by a well-known reliability mechanism called
electromigration in which electron momentum pushes the silicide atoms out of the
conductor link.

Still, most fuses can only be programmed during wafer processing and have stringent power
requirements for programming, making programming in packaged parts difficult. The bitcell
is the largest of the standard CMOS NVM technologies. For higher bit density memory applications, e.g. greater than four kilobits, the size of the fuse quickly begins to take up too much area of an SoC. eFuse is usually custom-designed and provided by the foundry as macros. As a result, it cannot be legally ported to another foundry without the consent of the foundry.

Traditional eFuses as shown in Figure 1 are made of polysilicon or metals and are programmed by rupturing the conductor links. In addition to strict requirements for surrounding passivation and metals, these Fuses have unacceptable reliability because debris and shards produced during programming can cause healing over time, causing the bit cell to reverse its state.

![Debris and Shards](image)

**Figure 1. Poly/Metal Fuse**

### 3.2 CMOS Floating Gate

One of the most widely used cell types is the stacked-gate 1T cell, a MOS transistor with one floating gate and one contacted gate overlapping each other, shown in Figure 2. This is the technology found in flash and EEPROM devices. A high-quality oxide above and below insulates the floating gate. The floating gate is typically programmed by channel hot electron (CHE) injection from the drain of the transistor. It is erased by Fowler-Nordheim (FN) electron tunneling through the bottom tunnel oxide. The presence or absence of charges (electrons or holes) in the floating gate determines if the transistor is on or off during read operations. Thus, a “1” or “0” is detected according to the channel conductance of the cell transistor. Floating-gate devices can be erased and electrically programmed multiple times, up to 100K times.
CMOS floating gate technology require additional masks and processing steps in addition to standard CMOS logic process, adding manufacturing cost and making high-performance CMOS devices more difficult to deliver. MTPs that are CMOS compatible use at least two single poly transistors with gate oxide available for I/O devices. For gate oxide less than 70Å (3.3V I/O), both program/erase endurance and data retention become unacceptable. Therefore, the lack of migration path below 0.13um is a limitation for this kind of embedded NVMs. Foundry porting is also a formidable challenge because memory performance and reliability are sensitive to minor differences in junction doping profiles, side-wall oxidation, spacer composition and back-end of line (BEOL) dielectrics.

4.0 One-Time Programmable (OTP) Antifuse

An antifuse is the opposite of an eFuse. The circuit is open (high resistance) in its unprogrammed state and is programmed by applying electrical stress that creates a low-resistance conductive path. Once programmed, the memory cannot be changed, hence the description one-time programmable. Antifuse NVM has been implemented for many decades using additional processing steps. Kilopass was the first to use antifuse in a standard CMOS process with no additional processing steps and holds patents for several flavors of bitcells, including the 1T and 2T.

4.1 Two Transistor (2T) Antifuse

A hard-gate oxide breakdown is used as the one-time programmable NVM mechanism and is achieved by applying a high voltage on the program gate word line program (WLP) shown in Figure 3. Before the breakdown, the region between the gate and the source of the program transistor is isolated and behaves like a capacitor. After the breakdown, the region between the gate and the source behaves like a resistor. The program transistor is isolated from the select (read) transistor WLR. Both the program and read transistors are implemented using standard foundry devices. As the technology scales, the bitcell scales.
The advantages of 2T include:
- Design for manufacturing (DFM) design rules are followed
- No extra requirements for both lithography and metrology
- Channel width and length of select and antifuse (program) transistors are controlled by active and poly CD alone
- Gate oxide reliability is no different from standard MOSFETs
- Uniform breakdown under the entire gate; the current distribution is predictable and Gaussian

4.2 One Transistor (1T) Antifuse

The 1T bitcell, often called the split-gate bitcell, is shown in Figure 4. The programming mechanism is the same, but the program and select transistors are combined into one, becoming a new device structure that employs the manufacturing step used to build the I/O transistor. If implemented in standard CMOS, the 1T bitcell will be larger than the 2T structure because the thick-gate device (I/O transistor) uses an OD2 (oxide diffusion second layer) mask with non-critical low resolution. Given that the OD2 layer is low grade, minimum geometry rules cannot be utilized. As a result, the bitcell will be larger than the 2T.
In the 1T solution, one word line is shared by the antifuse and select transistor, resulting in area savings in decoder logic. Although the breakdown is in the thin oxide channel region, several factors are compromised to achieve this. The thin oxide area is determined by the triple overlay of active, OD2, and poly. The interface between thin and thick oxide is defect prone. Likewise, the select transistor channel length cannot be well controlled by standard fabrication tools, i.e., it is sensitive to critical dimension control and OD2 layer alignment to the OD layer. To ensure high yield, a special mask step may be required for alignment or tight process control. Given the potential added cost of a special mask layer and the yield loss due to defects at the thin gate/thick gate interface, the 2T bitcell is more reliable and cost effective.

4.0 Multi-Time Programmable (OTP) Antifuse

With 2T antifuse technology scaling 25X from 180nm to 20nm, the evolution of OTP to MTP was inevitable. Redundancy in the memory to enable re-programmability consumes only a small percentage of the overall area due to the scaling of the 2T antifuse bit cell in advanced process nodes. As a result, Kilopass Technology Inc. developed a multi-time programmable (MTP) NVM called Itera. Integrated in a SoC, the MTP NVM provides a total system cost savings that is 70% less than an external flash or EEPROM MTP device.

5.0 Applications

For code that is fixed, ROM is the best embedded NVM technology. For code that changes frequently, embedded flash is the best embedded NVM technology. In between, there are valued applications where standard CMOS NVM fills a need. To determine which solution is best for an application, review the tradeoffs (Table 1) between each technology. For example, if it is a high-volume product that will utilize a second-source foundry, the eFuse solution should not be chosen. A designer cannot legally port a foundry solution without consent. To minimize design re-work, it would be beneficial to look for a foundry-agnostic solution.
Pervasive uses of embedded NVM technologies are in Table 2.
<table>
<thead>
<tr>
<th>NVM Technology</th>
<th>Pervasive Market</th>
<th>Density</th>
<th>Application</th>
<th>Key Usage</th>
<th>Benefit</th>
</tr>
</thead>
<tbody>
<tr>
<td>Embedded Flash</td>
<td>MCU</td>
<td>64Kb – 8Mb</td>
<td>Code Storage</td>
<td>Program code, Boot Code/Firmware</td>
<td>•Low Current •Field Update Flexibility •Product Customization</td>
</tr>
<tr>
<td>ROM</td>
<td>Greeting Cards, Toys, Electronic Dictionary</td>
<td>4Kb – 1Mb</td>
<td>Code Storage</td>
<td>Boot Code/Firmware</td>
<td>•Small Area</td>
</tr>
<tr>
<td>eFuse</td>
<td>Analog and Mixed Signal ICs, Consumer SOCs</td>
<td>256bit – 4Kb</td>
<td>Manufacturing &amp; Usability, Configuration</td>
<td>Trim and Calibrate, IDs</td>
<td>•Product Customization</td>
</tr>
<tr>
<td>Logic NVM Floating Gate (OTP)</td>
<td>Analog and Mixed Signal ICs</td>
<td>1Kb – 16Kb</td>
<td>Manufacturing &amp; Usability, Configuration</td>
<td>Trim and Calibrate, IDs</td>
<td>•Low Current •Field Update Flexibility •Product Customization</td>
</tr>
<tr>
<td>Logic NVM Floating Gate (MTP)</td>
<td>Analog and Mixed Signal ICs</td>
<td>1Kb – 16Kb</td>
<td>Manufacturing &amp; Usability, Configuration</td>
<td>Trim and Calibrate, IDs</td>
<td>•Low Current •Field Update Flexibility •Product Customization</td>
</tr>
<tr>
<td>Antifuse</td>
<td>Conditional Access, DTCP, HDCP, ePayment, Mobile Phone</td>
<td>16bit – 8Kb</td>
<td>Security</td>
<td>Chip ID, Encryption Key, Anti-piracy codes, Certifications, IMEI/DRM</td>
<td>•High Security Level •Small Area</td>
</tr>
<tr>
<td>Itera MTP NVM</td>
<td>RF/Analog IC, CMOS Image Sensor, MEMS, Audio, LCD Display, RFID, embedded SOC, Mobile Handset chipsets</td>
<td>128Kb – 4Mb</td>
<td>Manufacturing &amp; Usability, Configuration</td>
<td>Trim &amp; Calibration, Pixel repair</td>
<td>•Small Area •On-board Programmable •Low Current •Field Update Flexibility •Product Customization</td>
</tr>
<tr>
<td>Itera MTP NVM</td>
<td>MCU, Fixed and Wireless Broadband (RF, Wi-Fi, WiMax, LTE, etc...), Set-top Box, DTV, DVD</td>
<td>128Kb – 4Mb</td>
<td>Code Storage</td>
<td>Boot Code/Firmware Storage, Patch ROM code</td>
<td>•Small Area •Field Update Flexibility •Minimize Re-spin Cost •Reliability •No Extra Board Space</td>
</tr>
</tbody>
</table>

**Table 2. Embedded NVM Applications**

**6.0 Conclusion**

Five embedded NVM technologies are on the market: embedded flash, ROM, eFuse, CMOS floating gate, and antifuse. One benefit of the wide offering of technologies is that SoC designers and managers can choose the best solution for their end application.
When choosing between the different technologies, an SoC designer or manager needs to consider the application usage and tradeoffs between each embedded NVM solution. Each has its advantages and disadvantages. For code that changes frequently, SoC designers can decide between MTP solutions such as embedded flash and CMOS floating gate. For code that may change infrequently and requires field programmability, OTP solutions are viable. Although Antifuse is an OTP, it can emulate an MTP for up to a 1000 cycles of endurance. If 128Kb of code may need to be updated up to 8 times in the field, a 1Mb memory can be used.

###

Author Bio: Linh Hong is vice president of marketing at Kilopass, responsible for marketing Kilopass’ solutions globally. With 13 years of solid semiconductor industry experience, primarily focused on logic NVM IP, high-speed SERDES IP and broadband communication ASICs, Hong served for three years in various director and management positions in field applications engineering and applications marketing at Kilopass before assuming her current role in 2009. Prior to joining Kilopass, she was a design consultant and design manager at LSI Logic, where she also served in various design and marketing engineering functions. She began her career as a component engineer at Sun Microsystems. Hong holds a Bachelor of Science degree with honors in physics, and a Master of Science degree in Electrical Engineering, both from the University of California, Davis.