The Linley Group

MICROPROCESSOR REPORT

www.MPRonline.com

THE INSIDER'S GUIDE TO MICROPROCESSOR HARDWARE

KILOPASS BRINGS GUSTO TO MEMORY

Improved Antifuse Nonvolatile Memory Gives SoC Designers More Options

By Tom R. Halfhill {6/14/10-01}

Remember memory? It's often an afterthought when considering the obstacles to higher SoC integration. Much is written about rising NRE costs, longer verification cycles, growing complexity of system-level design tools, and ever-tighter design rules for deep-submicron fabrication processes, but the shortcomings of embedded-memory technologies are another factor limiting the consolidation of system functions on a single chip.

The old standbys are ROM, one-time-programmable (OTP) fuse memory, antifuse OTP, and various incarnations of EPROM and EEPROM. Flash memory is a relative newcomer that offers the most capacity and flexibility—if the project can stand the higher cost. Also, newer, more exotic technologies, such as Z-RAM, MRAM, and FeRAM, are available.

Now designers have another option. Kilopass, already an established player in nonvolatile memory (NVM), has introduced an improved version of its antifuse OTP. Called Gusto, it's licensed as process-portable intellectual property (IP). It is the industry's first 4Mb OTP, quadrupling the capacity of existing OTP memories. It's large enough to store boot code and system firmware, rather than just code patches, configuration code, and trim settings for analog components. In addition, Kilopass claims Gusto reads memory two to four times faster, cuts active power consumption by an order of magnitude, and slashes current leakage in standby mode by a factor of 40.

Gusto OTP is delivered as a hard macro for a specific fabrication process and foundry. Chips with Gusto memory have already taped out at three foundries, and further development is underway. Compared with existing embedded-memory technologies, Gusto is a higher-density alternative to existing OTP, a lower-cost alternative to flash memory, and a more flexible alternative to ROM.

Patents Focus on Antifuse Technology

Kilopass was founded in 2001 and remains a small private company, with all 40 employees in Santa Clara, California. Originally focused on FPGAs, the company split in two in 2006, spawning SiliconBlue Technologies to pursue the FPGA strategy. (SiliconBlue makes low-power, low-cost FPGAs that use Kilopass OTP memory to securely store on-chip configuration data.) Though small, Kilopass has more than 80 licensees and more than 250 design wins. Kilopass says its memory technology is used in more than two billion chips. Foundry partners include Dongbu, GlobalFoundries, IBM, Samsung, SMIC, Tower, TSMC, and UMC.

For such a tiny outfit, Kilopass is unusually active in research and development. The company has 54 patents granted or pending, mostly for antifuse memory technology. Some patents describe antifuse bit cells with one, two, three, or more transistors, but Kilopass favors two-transistor (2T) cells, as Figure 1 shows.

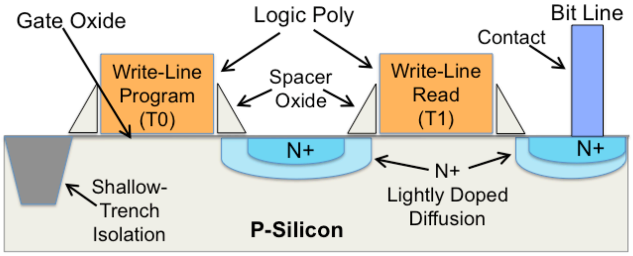

Figure 1. Two-transistor antifuse bit cell. One N-channel transistor is the programmable transistor, and the other is the select or read transistor. When the programmable transistor's gate oxide is intact, the read transistor returns a binary zero. Breaking the gate creates a conductive path, returning a binary one.

Kilopass's 2T cells are actually smaller than its 1T or "split gate" cells, which require an additional oxide-diffusion mask layer to build a thicker gate. One competitor—Ontario-based Sidense—has patented a slightly different technology that uses smaller 1T antifuse cells without an additional mask layer. (Kilopass has filed a lawsuit against Sidense, claiming that the Canadian company's 1T cells infringe on one of Kilopass's patents.)

Several Kilopass patents cover the arcane subject of gate-oxide breakdown, the key technology of antifuse memory. Although most semiconductor companies strive to avoid the erosion of their gate oxide, Kilopass keeps seeking better ways to obliterate it. Breaking down the gate oxide severs the electrical connection that determines whether a bit cell's programmable transistor returns a one or zero.

Unlike some memory cells, a Kilopass bit cell has no capacitors. Omitting the capacitor makes the cell more compatible with the standard CMOS processes used for digital logic, because deep trenches need not be etched in the silicon. Nor does a Kilopass 2T cell need a floating gate. Instead, an N-channel transistor with an unbroken gate behaves like a capacitor and passes a high current, indicating a binary zero. After the gate is broken, the transistor behaves like a resistor, indicating a binary one. Figure 2 illustrates the electrical transformation.

Breaking the gate requires a brief surge of current, such as would blow a fuse (hence the term "antifuse memory"). For a chip manufactured in a 40nm G or 45nm silicon-on-insulator (SOI) process, burning the fuse requires about 5V; in 40nm LP, it's 6.25V. If the memory is preprogrammed during manufacture, the fab's test equipment applies the voltage to each die on the wafer. If the customer prefers in-system programming, the chip doesn't need a 5V or 6.25V power supply for this purpose—just 1.8V is enough, because optional on-chip charge pumps (powered by an on-chip capacitor) raise the voltage to the required threshold. Breaking down the gate oxide of a transistor takes about four microseconds, not counting the time required to ramp the voltage if the charge pump is used. Overall, programming 1Mb of memory takes about one second.

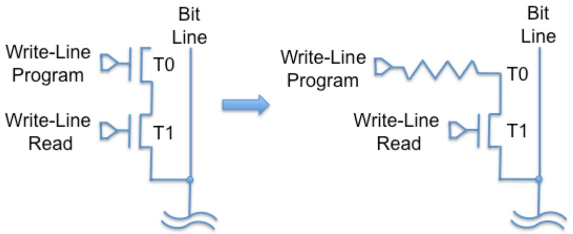

Figure 2. Antifuse bit-cell circuit diagrams. The diagram on the left shows the circuit of an antifuse bit cell before the programmable transistor's gate oxide is broken by a surge of current. The diagram on the right shows the circuit after the breakdown.

The gate-oxide breakdown is, of course, irreversible. Once the fuse is blown, it's blown for good. For this reason, this memory technology is called OTP and isn't reprogrammable like flash memory. The breakdown leaves the gate of the second transistor in the bit cell intact. Because this transistor also has a very thin gate, it could succumb to normal oxide breakdown over time. Kilopass guarantees the transistor for at least ten years, however, even with continuous reads. Internal voltages are regulated to avoid stressing the read transistor beyond its rated time to dielectric breakdown (TDDB), which varies slightly from one fabrication process and foundry to another.

Both transistors in a Kilopass 2T bit cell are standard CMOS elements, so the cell can scale with the fabrication technology and with Moore's law. Kilopass has qualified its previous antifuse technology in about 30 fabrication processes from 180nm to 40nm and is working toward qualification in a 28nm high-*k* metal-gate process. Gusto is designed for processes at 65nm or below and will initially be qualified for 40nm bulk CMOS and 45nm SOI processes at three foundries: IBM, TSMC, and UMC. Kilopass has initial silicon and expects qualification in 4Q10. Figure 3 is a scanning electron micrograph of a Kilopass 2T bit cell fabricated in a 40nm process.

Gusto Quadruples Memory Density

Gusto builds on Kilopass's existing eXtra-Permanent Memory (XPM) 2T-cell technology, which was introduced in 2003 and has a maximum capacity of 1Mb (assuming 40nm fabrication). To increase memory capacity, Kilopass adopted a better error-correction scheme that reduces overhead by a factor of seven for this function.

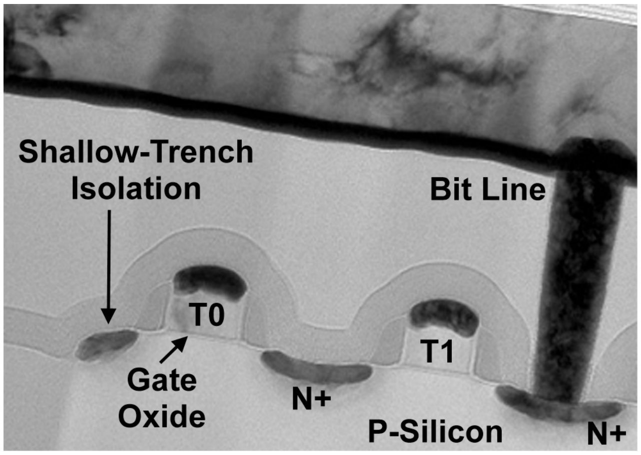

Figure 3. Electron micrograph of a Kilopass 2T antifuse bit cell. This image shows a cross-sectional view of the 40nm cell before gate-oxide breakdown. The transistors (T0 and T1) are the humps near the center of the image; the large vertical black structure on the right is a contact for the bit line. (Photo courtesy of Kilopass)

The old error-correction scheme—full redundancy—was easy for Kilopass to design but was rather primitive. Each 2T cell had another 2T cell watching its back, essentially doubling the size of the array. Gusto substitutes ECC with single-error correction and double-error detection. ECC is commonly used to detect and correct soft errors in other memory types.

By devoting a larger percentage of transistors to actual memory storage and by reducing the error-correction overhead, Gusto crams more cells into the same space. It also improves production yield by simplifying the arrays and reducing the effective number of transistors per bit. Figure 4 compares the size of a Gusto memory array with that of a previous-generation XPM array.

Boosting the maximum size of the array from 1Mb to 4Mb is significant because it makes Gusto suitable for a wider range of embedded-memory applications. Before, 1Mb (128KB) of memory simply wasn't enough to store the boot code or firmware of many systems. Although Gusto still doesn't approach the capacity of flash memory, it's much cheaper to manufacture than flash memory and outdistances other types of OTP memory. Kilopass estimates that 30% of the $5 billion spent last year on serial flash memory and EEPROM was for systems using 4Mb of memory or less.

Expanding the role of NVM can dramatically reduce the number of chips in a system. Kilopass notes that teardowns of Apple's iPhone 3GS have found about 10 serial flash memories and EEPROMs. (Most are die-bonded inside the packages of larger chips and are invisible in system-level teardowns.) As smartphones continue to offer more features, they will need more of these tiny memories. Figure 5 shows Kilopass's vision of a future cell phone that replaces all or most of those separate memories with embedded NVM.

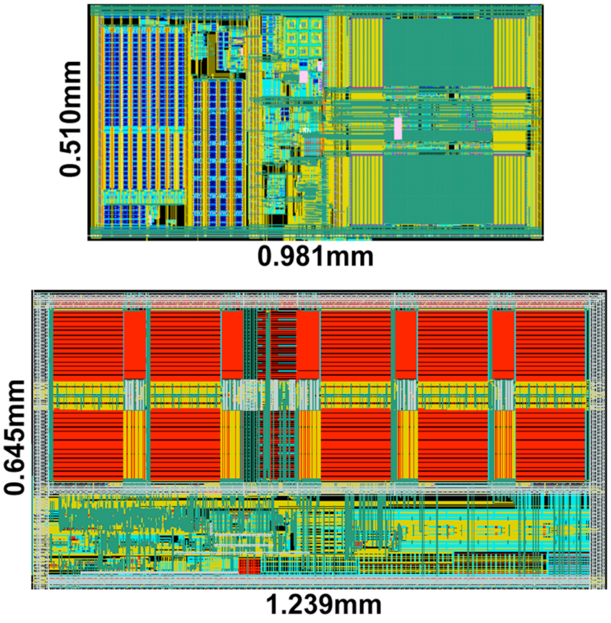

Figure 4. XPM versus Gusto area comparison. At top is a previous-generation eXtra-Permanent Memory (XPM) array from Kilopass; it has 128Kb of memory and is 0.5mm^2. At bottom is a Gusto 1Mb array, which is only 0.8mm^2 despite having eight times the XPM's memory. Both arrays are implemented in a 40nm LP process and include optional charge pumps for in-system programming.

Still Slower Than ROM

Gusto's random-access time is 40ns: almost twice as fast as previous-generation XPM. Page-mode access time (burst read) is 20ns: about four times faster than XPM. (Actually, XPM doesn't support page mode; everything is random access.) When Gusto memory is used to store program code—a likely application, given its larger capacity—page-mode performance is generally more important than random-access performance. Kilopass says it improved Gusto's performance by optimizing critical circuit paths.. Table 1 compares Gusto with XPM.

Sidense quotes random-access times ranging from 50ns at 180nm to 15ns at 40nm for its antifuse OTP memories, but the memory arrays are smaller. (Sidense says it has made 40nm test chips with up to 5.5Mb of antifuse OTP but that no customer has asked for more than 500Kb.) Although Gusto is slower than SRAM, DRAM, or even plain old ROM, it's usually fast enough to allow firmware to execute in place (XIP) without copying the code into DRAM. The Gusto memory array connects to the system over a 32-bit parallel interface that runs synchronously with the on-chip bus, so it's four times faster than quad serial flash memory or serial EEPROM.

Interestingly, Kilopass says Gusto is useful for slow designs (under 100MHz) and fast designs (over 333MHz),

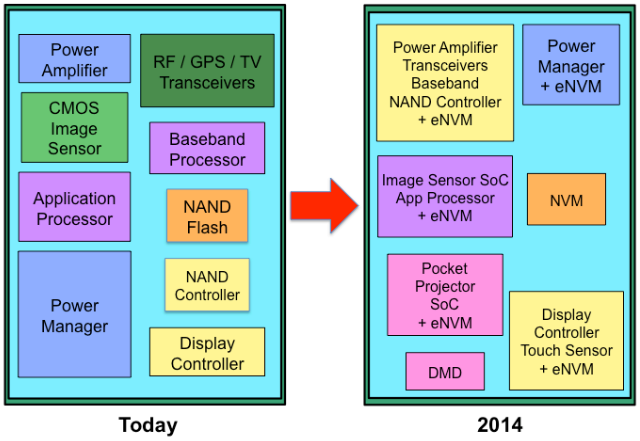
Power Amplifier
RF / GPS / TV Transceivers
CMOS Image Sensor
Baseband Processor
Application Processor
NAND Flash
NAND Controller
Power Manager
Display Controller
Today
Power Amplifier Transceivers Baseband NAND Controller + eNVM
Power Manager + eNVM
Image Sensor SoC App Processor + eNVM
NVM
Pocket Projector SoC + eNVM
Display Controller Touch Sensor + eNVM
DMD
2014

Figure 5. Integration comparison of two hypothetical smartphones. On the left is a block diagram of a typical smartphone today. On the right is a smartphone for 2014 that achieves higher integration by using embedded nonvolatile memory (eNVM) instead of serial flash memory or EEPROM on separate die. (Source: Kilopass)

Price and Availability

Gusto antifuse OTP memory is licensable intellectual property (IP). Deliverables include hard macros ported to specific fabrication processes at independent foundries. Gusto is intended for fabrication at 65nm and smaller. Kilopass has initial silicon and expects qualification in 4Q10 for 40nm bulk CMOS and 45nm silicon-on-insulator (SOI) processes at three foundries: IBM, TSMC, and UMC. Multiple licensing options are available, depending on whether the customer wants a single-project license, a multiple-project license, or an unlimited license. The upfront licensing fee ranges from $300,000 to a few million dollars. In addition, Kilopass collects a royalty per wafer. For more information about Gusto, visit *www.kilopass.com/products/memory-ip/gusto/*.

but it sometimes is unsuitable for designs in between. When the processor runs at 100MHz or less, antifuse OTP memory like Gusto can easily satisfy the system's throughput requirements. Examples of designs in this category are low-speed baseband processors, demodulators, power amplifiers, and physical-layer (PHY) interfaces.

Processors running at 333MHz or faster usually have an L2 cache to buffer I/O and an on-chip bus or crossbar switch with relatively low latency but high bandwidth. In these designs, the OTP array attaches to the system over its 32-bit interface and again provides enough throughput.

Processors running between these performance extremes, however, often rely on single-cycle ROM and omit the L2 cache. Longer-latency OTP memory, buffered only by a small L1 cache, may not satisfy the system's throughput requirements. Despite its marked improvements over the previous OTP generation, Gusto doesn't quite close the performance gap.

In recent years, new types of memory based on exotic technologies have appeared, but none has become popular. In 2005, Innovative Silicon introduced Z-RAM, which uses 1T DRAM cells without capacitors, relying instead on the floating-gate effect of SOI transistors. (See *MPR 10/25/05-03*, "Z-RAM Shrinks Embedded Memory.") In 2006, Freescale Semiconductor shipped the world's first magnetic random-access memory (MRAM) based on spintronics technology. (See *MPR 9/11/06-01*, "MRAM: A New Spin on Memory.")

Specification	Kilopass Gusto	Kilopass XPM
Bus Width (read)	32 bits	8 bits
Bus Width (write)	32 bits	1 or 8 bits
Random Access	40ns	70ns
Page-Mode Access	20ns	—
Write Time Per Mbit	1s	26s
Power Consumption (read, typical)	0.3mW/MHz per 32Mb	3.2mW/MHz per 32Mb
Power Consumption (standby, typical)	1 microwatt	41 microwatts

Table 1. Gusto versus XPM. Gusto offers several performance improvements over Kilopass's previous-generation NVM. Foremost among the improvements are a new page mode for burst reads, a wider I/O interface, and much lower power consumption. (Source: Kilopass)

The higher cost of SOI appears to have stalled Z-RAM as an embedded-memory alternative—most SoCs are fabricated in bulk CMOS. Also, Z-RAM isn't a direct replacement for embedded ROM or OTP memory, because it's volatile. MRAM is nonvolatile but requires extra processing steps for integration with digital logic and isn't as dense as other types of memory. So far, MRAM is found only in a few discrete memory chips from Freescale.

Faster Turnaround Than ROM

Mask ROM is the golden oldie that other NVM types struggle to replace. ROM is fast, dense, cheap, low power, field proven, well understood, and easy to use, and it requires no additional manufacturing steps. If it didn't exist, we'd have to invent it again. To lure customers away from such a successful NVM technology, Kilopass and other OTP vendors aim at ROM's biggest weakness: immutability. Programming ROM is like carving the code in stone.

Actually, ROM is worse than stone. Even a stonecutter engraving ones and zeroes on granite tablets could deliver faster turnaround times than today's fabs do when respinning a chip. In older fabrication processes, remasking a ROM commonly takes 30 to 40 days. For metal-programmable ROM, it's about the same. But at 40nm, remasking a diffusion ROM can stretch to 85 days. Kilopass says the average cellular chip set needs about three mask revisions before it's ready for prime time, so the respins can delay a project by several months. And don't forget the rising cost of masks, which at 40nm can penalize the project many thousands of dollars per spin.

Of course, OTP memory is immutable after it's programmed, but the entire array needn't be programmed all at once. Designers can program different memory blocks at different times, applying incremental patches and updates instead of remasking a ROM. Revisions can be applied at the fab during manufacture or in the system after customers take delivery of the chips. In effect, Gusto's larger 4Mb capacity will allow some developers to use it like few-times-programmable (FTP) memory instead of OTP—new code can override the old.

Some chip designers use OTP for early versions of a chip, then switch to ROM if they judge the firmware to be stable or if sales volumes soar. Kilopass offers a low-cost option for converting Gusto OTP memory to ROM. Another strength of OTP is security, although ROM is good for that, too. Kilopass says customers are using the company's existing OTP memory to securely store code for mobile banking and digital-rights management, among other applications. Smaller-capacity OTP memories are

	Antifuse OTP	Fuse OTP	Floating-Gate OTP	Floating-Gate FTP	Floating-Gate MTP	Mask ROM	eFlash NAND/NOR
Memory Technology	Oxide breakdown	PolyFuse or eFuse	Floating-gate EPROM (UV)	Floating-gate EEPROM	Floating-gate EPROM	Fixed bit cells	Floating-gate or split-gate EEPROM
Fabrication Technology	Standard CMOS to 40nm and below	Standard CMOS to 40nm and below	Standard CMOS to 130nm	Standard CMOS to 65nm	Standard CMOS to 65nm	Standard CMOS to 40nm and below	Memory CMOS for 90nm
Extra Mask Steps	0	0–5	0	0	0	0, but >30 days to respin mask	>10 masks
Bit-Cell Structure	1T, 2T, 3.5T (Gusto: 2T)	Polycide	2T	2T-4T	3T-6T	1T	1T/2T
Bit-Cell Size (Normalized)	1	30–300	10	15	20	< 1	3–10
Maximum Capacity	1Mb (Gusto: 4Mb)	4Kb	512Kb	16Kb	16Kb	2Mb	8Mb
Random Access	Fast	Slow	Medium	Medium	Medium	Fast	Medium
Write Endurance	At fab or in system	Usually at fab only	At fab only	At fab or in system	At fab or in system	At mask only	In system
Process Scalability	Yes	Yes	Up to 130nm	Up to 130nm	Up to 65nm	Yes	Up to 90nm
Design Characteristics	Some flexibility	Some flexibility	Some flexibility	Much flexibility	Much flexibility	Cheap but inflexible	Costly but most flexible
Relative Power	1×3.0Gbps	1×3.0Gbps	None	1×3.0Gbps	Medium	Low	Medium
Environmental Tolerance	None	480Mbps	480Mbps	480Mbps	Low	High	Low
Primary System Functions	Code, config, analog trim, patches, crypto	Config, analog trim	Config, analog trim	Config, patches, prefs	Config, data, prefs, analog trim	Code	Code, data, patches, config

Table 2. Comparison of nonvolatile-memory (NVM) technologies. Gusto is two-transistor antifuse OTP that expands the maximum capacity of an array to 4Mb. Embedded flash memory is the most flexible NVM because it's easily reprogrammable, but it's also the most expensive solution. Mask ROM wins in most categories but is difficult to revise. Other NVMs—such as few-times-programmable (FTP) and many-times-programmable (MTP) technologies—offer various compromises between the reprogrammability of flash memory and the immutability of ROM. (Source: vendors and *MPR*)

sufficient for storing passwords and cryptographic keys, but they're usually inadequate for code storage, and the highest-security applications require on-die storage for secure software and authorization credentials.

Table 2 summarizes the various NVM technologies suitable for integration in SoCs. Gusto doesn't appear in this table as a distinct type because it's an improved version of existing antifuse OTP memory; it shares the characteristics of other antifuse OTP while expanding capacity to 4Mb using 2T bit cells.

Weighing the Trade-Offs

Overall, Gusto is a worthwhile enhancement of antifuse OTP memory. Quadrupling the maximum size of this type of NVM gives SoC designers more flexibility, especially when the embedded memory must be large enough to store the system's boot code or firmware. In some applications, Gusto is fast enough for XIP, eliminating the redundancy of storing executable code in DRAM. When security is essential—and almost all embedded developers are paying more attention to security these days—larger NVMs are necessary to store secure software on die, where the code is more difficult to subvert.

Gusto's biggest drawbacks are that it remains slower, less power efficient, more expensive, and less dense than old-fashioned mask ROM. The most highly optimized designs will continue to favor ROM over other types of NVM. The advantages of ROM come at a price, however: the high cost of new masks and respins if the programming must change. These costs are both direct (startling bills from the foundry) and indirect (lost market opportunity while waiting for turnaround). The unforgiving nature of ROM will keep embedded developers seeking reprogrammable alternatives. ♦