SCALABLE EMBEDDED FLASH FOR MIXED-SIGNAL ASICS

Christian Binard

Product Development, AMI Semiconductor, J. Monnetlaan 1, 1804 Vilvoorde, Belgium

(www.amis.com; christian_binard@amis.com)

Non Volatile Memory, Embedded, Scalable, Topology, Testability

Fast time-to-market for mixed-signal ASICs with embedded non-volatile memory (NVM) requires robust IP blocks of different array sizes. This paper describes array topology, design architecture and digital wrapper trade-offs that must be considered in designing a scalable NVM. Specific time generator, sense amplifier and array topology are displayed as well as a BIST specific approach. Finally HIMOS® Flash realizations of different sizes (128 Bytes to 128 kBytes) based on the same IP block are depicted.

1. INTRODUCTION

The focus on System on Chip (SoC) applications continues to grow in the semiconductor industry. Mixed-signal SoCs embed microprocessors (uP) and non-volatile memories (NVM) on top of low and high voltage analog circuitry. Because the uP can execute code from an on-chip re-programmable memory, the circuits enable fast system customization for aggressive time-to-market applications. The NVM also allows users to store system status on chip.

The SoC demand is also growing for automotive applications. The NVM sizes requirements for this market (up to 64 kBytes) are smaller than for cutting edges arrays used in consumer or telecom applications, but the robustness and the cost effectiveness are much more important assets. Cost is kept under control by selecting a NVM technology requiring few additional masks over the base mixed-signal process and by providing sizes that fit the user's needs.

If a key to success is the array size, design for an adaptive size is a major concern. To reduce the design and characterization efforts, the analog core of the memory must be insensitive to array load. To reduce the layout effort, the array topology must be easy to rescale. To increase the memory portability, the core memory and its interfacing digital wrapper must form a flexible system.

2. ROBUST DESIGN AND ARRAY LOAD TOLERANCE

When the array capacitive load varies from a few nano-Farads to negligible values and the current consumption from tenths of milli-Amps to negligible values, robust wide-range design is a must. This is especially true for Sense Amplifiers (SA) and read access timings.

2.1 Time reference

Not only must the redesign cycle be short, the characterization effort to get a reliable model is also critical. To achieve both of these goals, parametric timing needs to be independent of array size. The Sense Amplifiers (SA) timings must then be constant rather than proportional to the AC array characteristics. One approach is to use a time reference. An available accurate time reference is the system clock. By using a time reference, the read access time is system clock frequency dependent and can be characterized on whatever desired array size. The clock period is not directly usable since the SA control signals are shorter that the period to enable one read per clock cycle. A phase-lock loop type of circuit is needed.

2.2 Sense Amplifiers

One critical step in data read out is the Bit Line (BL) precharge voltage control. The drain of the selected cell must be set high enough to get a reliable erase state current but without pinching the SA or causing read disturbs. Classically, a hard switch to the supply controlled by a threshold detection or BL to dummy BL capacitive equalization (2) constitute the precharge. Both methods are sensitive to array size: the first one through the switch reaction time and the second one by the capacitive contribution of the non-array proportional circuits (equalization switch, X decoders).

A robust SA is shown in Fig. 1. A source follower (SF) transistor T1 defines the BL voltage. The BL settling time will increase with the array size. If working with constant precharge time rather than settling level target, the sensing cycle will start with an unknown remaining BL capacitive current. To reduce this effect, a similar SF T2 is used in the reference branch loaded by a dummy BL. The drawbacks are rather large SF transistors and slow precharge (w/l = 685, 30ns precharge, 5pF BL). The reference voltage generator Vref must be able to drive the SF gate-source capacitive current during precharge.

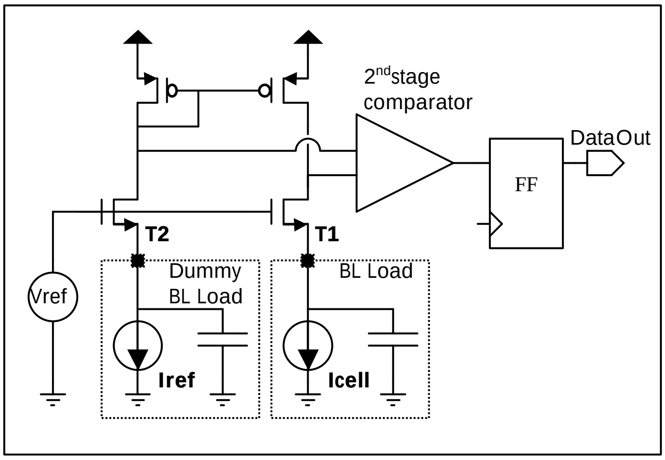

Fig. 1: Source follower sense amplifier robust architecture

3. REUSABLE LAYOUT AND CONVENIENT ARRAY TOPOLOGY

The proposed robust SA architecture is rather large. When considering medium sized arrays, scalability cannot come from a variable number of banks with local SA without large layout overhead. Only one set of SAs, one SA per bit in a word, is used.

The classical topology for an array with “n” bits per word and “m” words per line is depicted in Fig. 2_1. Bits of the same weight are grouped in column Bits Blocks (BB) to keep their common SA as close as possible. The SA and X Decoders (XD) pitch is “m” times the cell pitch.

Both X and Y array dimensions must be resized to keep the aspect ratio as square as possible for the decoders area optimization. Y direction scaling is straightforward: the array may be reduced by sectors (groups of rows driven by a Y driver). X direction scaling is more problematic. Since the number of bits “n” in a word is a system constant, only the number of words per line “m” modulates the X direction size. The SA and XD fixed pitch block the possible layout shrink according to an array down scale.

3.1 Words blocks topology

Grouping bits per words instead of bits per weights solves the layout issue. As shown in Fig. 2_2, when bits are grouped in Words Blocks (WB), dropping some WB rescales the array with no impact on the block pitch. From a layout point of view this gives the advantage that the XD pitch is constant. The “n” SAs are then better place in the memory corner since that area, defined by X and Y decoders height, is not impacted by resizing. The bits related to the same SA are then spread among the array. From an electrical point of view, this is a drawback that introduces parasitic capacitive and resistive BL loads mismatch. This electrical issue is overcome by the robust SA architecture.

In practical realizations, a WB is made of interleaved words of two different weights (words of different pages) due to inhibit conditions to allow BL sharing. The WB pitch is then “2n” times the cell pitch.

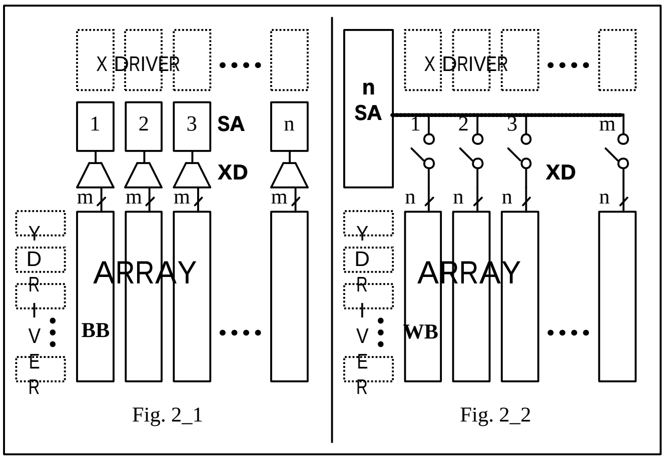

Fig. 2: Bits Blocks [1] and Words Blocks [2] topologies

WB topology makes Page Program Mode (PPM) data management easy: an “n” bits wide shift register is build using the X decoder block where each WB decoder has an element of the shift register. WB “n” holds the “nth” data word after shift-in.

3.2 Continuous address table

Let's divide the address bus into word, page, line and sector addresses. If sector addresses are the MSB, Y rescaling by removing sectors reduces the address table depth with no issues. But X rescaling makes holes in the address table by the lack of LSB locations. If the WB count is reduced by a power of two, a reduced set of word address bits code the locations with no holes. The address table is made continuous again when unused word address bits are ignored at the user interface. The ignored bits may remain in the memory core to limit changes.

4. FLEXIBLE SYSTEM AND DIGITAL WRAPPER

A digital wrapper will boost the memory flexibility by the way it interfaces with the user and by fitting the word sizes and clock periods.

4.1 Command interface

A simple but powerful set of commands is defined in the digital wrapper to interface with multiple types of uPs. It has to mask all the timings and low-level signal requirements. A Ready signal informs the uP about the memory availability.

4.2 Word width

The memory core is designed to support the longest desired words. The digital wrapper, interfacing with smaller words system, selects a portion of a designated word according to extra address bits value. In read cycle, a simple mux makes the function; in word program cycle, the full internal word is programmed with the unselected bits kept unchanged by cell inhibition ("program" the erased logical value). Such word program method results in more neighbor cells disturb because of the multiplied number of program cycles within a sector. In page program and erase cycle, the wrapper operation is obvious since more than one word is being addressed.

4.3Built-In Self-Test

Built-In Self-Test (BIST) is required to speed up the production test time and lower the test equipment requirements for NVM. However, not all ASICs require equal quality level and thus equal test-coverage. Therefore, performing the same BIST for all applications is not economical.

Instead of a fully automated BIST making all the checkerboards tests, a set of more elementary selectable self-tests (e.g., Table 1) provides the required flexibility.

Table 1: BIST elementary instruction set

Instruction	Description
BTE & {00, 01}	Erases the entire array or erases only sector 0.
BTP & {00, 01, 10, 11}	Programs checkerboards {00, 01, 10, 11}
BTPUP	Programs a unique pattern in each word
BTD 1, BTD2	Executes a program disturb (type 1 or 2)
BTR & {00, 01, 10, 11}	Verifies checkerboards {00, 01, 10, 11}
BTRUP	Verifies unique pattern
BTRS & {00, 01, 10, 11}	Verifies erased value in sector 0 and checkerboards {00, 01, 10, 11} in other sectors

The full test sequence is then defined by software on the test equipment. The same low-end type of tester may then be used for multiple applications. Indeed, the requirements are to set test modes and read test results; only the interaction periodicity is increased, which could reduce tests parallelism with other ASIC functions. Such flexible BIST set is also useful in the development phase to optimize the test strategy.

Variable address table sizes are handled by words per pages and sector count parameters for BIST address counters.

5. A 128 BYTES TO 128 KBYTES SIZABLE FLASH IP BLOCK

This paper presents a 128 Bytes to 128 kBytes sizable Flash realization based on the explained principles. It is targeted for automotive and industrial markets and is realized in AMIS eighty-volt I3T80 technology (Table 2). For future porting to other AMIS technologies, the Flash block uses only low-voltage and Flash-specific extended drain devices. There are no eighty-volt specific devices utilized and only three metal layers are used to serve all possible metallization schemes. The memory is based on the two poly, three gate HIMOS cell (1) in a NOR type array. The digital wrapper allows communication with ARM7 (16/32 bit words) and R8051 (8 bit words) uP.

5.1 Array topology

The core word size is twenty-two bits wide to get a sixteen bit data capability with one bit error correction and two bits true error detection. The Error Correction Code (ECC) uses five Hamming code bits and one extra parity bit for true two bits detection. Each line is made of two pages; each sector is made of thirty-two pages (sixteen lines). The largest array has thirty-two words in a page and sixty-four sectors. Smaller arrays have either one, two, four, eight, sixteen or thirty-two words per page and less than sixty-four sectors. The sector count must be even for layout reasons. Indeed, sectors are two by two interleaved and Y decoders are placed on both sides of the array as illustrated in Fig. 3. So virtually 192 different sizes exist out of which 99 give an acceptable aspect ratio with 71 different sizes.

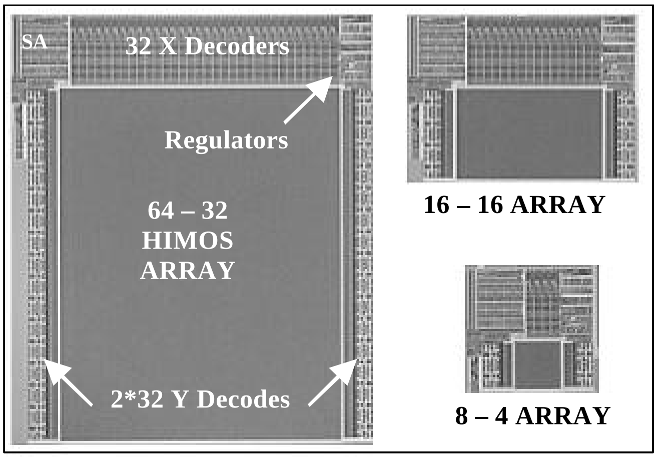

Fig. 3: Three different array sizes in AMIS I3T80 technology

5.2 Realizations

The layout is automatically made with a Programmable Cell (PCell). The user has just to set the desired array dimensions parameters. Fig. 3 shows a 128 kBytes array (64 sectors, 32 WB), a 16 kBytes array (16 sectors, 16 WB) and a 2 kBytes array (8 sectors, 4 WB). Other sizes have been successfully tested as well: 1 WB, 2 sectors; 32 WB, 16 sectors. Fig. 4 represents the Flash area relation to the memory depth and aspect ratio.

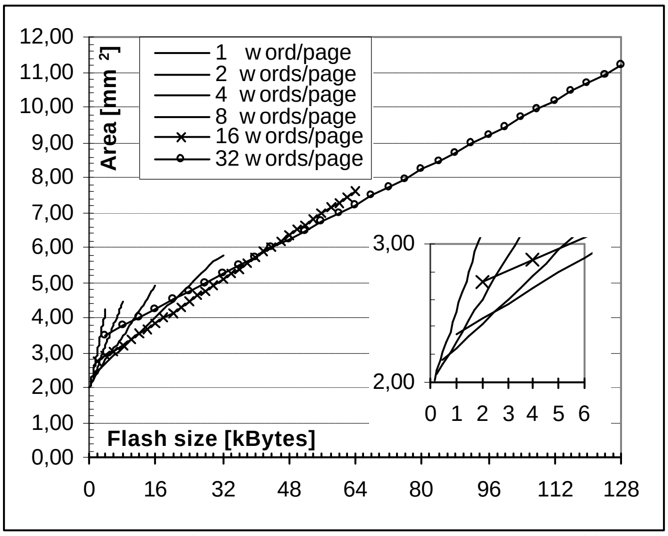

Fig. 4: AMIS flash area vs array depth for different X ratios

Table 2: Presented Flash characteristics

Feature	Value
AMIS Technology	I3T80-Flash: 0.35um, 2 poly, 3 Metals
Extra masks	3
Temperature ranges	-40 to 150 [°C]; 160 [°C] in read mode
Core Voltage	3.0 to 3.6 [V]
Wrapper Voltages	1.7 [V] to core voltage
Core Random access time	60 to 77[ns] (6 sigma measures)
App. guaranteed read time	100 [ns] (value used for qualification)
Program time	20 [us] (value used for qualification)
Page Pgm equivalent rate	0.31 [us/data byte] (32 words per page)
Erase time	500 [ms] (value used for qualification)
Full erase equivalent rate	4 [us/data byte] when 128 kBytes array
Wrapper: Interface count	2.0 [kGates]
Wrapper: BIST count	3.0 [kGates]
Wrapper: ECC count	1.0 [kGates]

6. REFERENCES

[1] J. Van Houdt and D. Wellekens, "HIMOS® Flash technology: the best choice for embedded nonvolatile memory applications," http://www.imec.be/wwwinter/business/himos_01.pdf, Feb. 2001.

[2] Pasotti, M. Rolandi, P.L. Canegallo, R. Gerna, D. Guaitini, G. Lhermet, F. Kramer, A, ”Analog sense amplifiers for high density NOR flash memories”, CICC, 1999. Proceedings of the IEEE 1999