Design of Embedded MRAM Macros for Memory-in-Logic Applications

Sumanta Chaudhuri1,2, Weisheng Zhao1,2, Jacques-Oliver Klein2,1, Claude Chappert1,2, Pascale Mazoyer3
1. IEF, CNRS, Orsay, 91405, France
2. Univ Paris-Sud 11, UMR 8622, Orsay, 91405, France
3. STMicroelectronics, 850 Rue Jean Monnet Crolles, Grenoble 38026, France
Tel: (+33)169156292, Fax: (+33)169154000
weisheng.zhao@u-psud.fr

ABSTRACT
In this article we present a design method for integrating non-volatile MRAM memory cells into standard CMOS design. The emphasis is on standard cell based flow for general purpose logic and automatic generation of MRAM macros suitable for the applications. We present a design space exploration for this purpose and transient simulation results of the hybrid MTJ/CMOS designs. We continue the article with examples of automatic macro generation, integration layout and a prototype in 130nm CMOS which is designed to test a large subset of this design space. In conclusion we show that a high 3D integration density with reasonable speed can be achieved with automatic flow by sharing the reading/writing circuitry among a number of MTJs.

Categories and Subject Descriptors
B.7.1

General Terms

Keywords
MRAM, Embedded, Thermally Assisted Switching (TAS), low power, System on Chip (SOC), Non-volatile, Architecture

1. INTRODUCTION
Among the emerging memory technologies, MRAM is one of the most promising candidates [1]. This is mainly due to its non-volatility, ease of integration with CMOS (requires only two or three additional masks), radiation tolerance etc. Indeed it has been called by many as the universal memory for both the standalone and embedded applications. Earlier MRAMs used Field Induced Magnetic Switching (FIMS) mechanism [2-3], which was rather demanding in terms of power consumption due to energy lost in radiation while generating a magnetic field. Recent discoveries in new switching methods such as Spin Transfer Torque (STT) [4-5] and newer techniques such as Temperature Assisted Switching (TAS) [6-7] now permit solid state implementation of MRAM and created a new wave of research in the highly competitive memory market. NEC has recently come out with a 32MB MRAM which uses FIMS switching with low (~1mA) switching current [8]. Several prototypes of STT-MRAM have been fabricated with switching currents in the uA range [9-10]. Meanwhile, advances in the semiconductor industry have entered into the 3D era. Several prototypes with stacked CMOS have been reported such as a processor and SRAM back to back [11]. First of all, 3D integration results in higher density, even better is the fact that due to close placement to the processor, memory transfers are much faster. MRAM is particularly suited for this application because of its ease of integration on top of CMOS at the back-end process. Several applications have thus studied based on embedded MRAM, which ranges from processors integrating with MRAM as cache memory [12], FPGAs with MRAM as configuration memory [13], or simple memory blocks enhanced with logic functions such as Content Addressable Memory (CAM) [14]. In this article we will present a method of integrating automatically generated MRAM macros with general purpose logic based on standard cell methodology.

2. DESIGN SPACE EXPLORATION
Fig.1(a) describes the conventional Magnetic Tunneling Junction (MTJ) stack, basic memory cell of MRAM [3]. The stack comprises of two ferromagnetic layers separated by a thin insulating layer, and is图案成 nanopillar shape. The resistance of the nanopillar depends on the relative orientations of the magnetizations of two ferromagnetic layers (represented as low resistance Rp and high resistance Rap). In standard applications the magnetization of one layer is fixed, while that of the other layer is able to take two directions parallel or anti-parallel to the magnetization of the fixed layer. Recent discoveries in new switching methods such as Spin Transfer Torque (STT) [4-5]...
Thermally Assisted Switching (TAS) is used for our MRAM development. In this method [6-7], the MTJ stack (see Fig. 1b) is slightly modified to be adapted to the TAS operation. An additional anti-ferromagnetic layer (AF2) with low blocking temperature $T_{B2}$ (~160°C) [6] is added in the stack above the ferromagnetic storage layer. As a result, the magnetization of the storage layer is pinned at the standby temperature and only becomes unpinned (free) when the MTJ is heated to a temperature above the blocking temperature $T_{B2}$ of this additional AF layer. In order to prevent any switching of the magnetization of the reference layer, the AF1 material which is used to pin this reference magnetization is chosen with a much higher blocking temperature $T_{B1}$ (~300°C) than $T_{B2}$. During the write operation, two low currents $I_{\text{heat}}$ (~100uA) and $I_{\text{write}}$ (~4mA) are required.

The major sense amplifiers considered in our MRAM macro design is Pre-charge Sense Amplifier (PCSA) [15]. It has been shown to present the high performances in terms of reliability, speed and power. Due to the overhead associated with the sense amplifiers it is necessary to share them among several MTJs in order to have a high density. In Fig. 2(a) each bit is stored in two MTJs with opposite configuration. The select transistors of both MTJs are turned on simultaneously, to select a cell among the cells sharing the same sense amplifier. We name this structure twin cell amplification. In Fig. 2(b) two MTJs are used as reference MTJs which can be designed with an appropriate size then have the resistance in between the $R_{P}$ and $R_{AP}$ states. We name this structure reference cell amplification. For each odd bit the even reference is chosen, and vice versa. This adds some overhead in the control circuitry while doubling the density. Another major problem with this scheme is the reduction of noise margin then improves the data sensing error rate.

![Figure 2. (a) Twin cell amplification structure, the two cells are with opposite configuration (b) Reference cell amplification](image)

### 3. SIMULATION RESULT

In this section we present the transient simulation results of a single bank of MRAM macro by using CMOS 65nm [16] and a TAS-MTJ spice model integrating the experimental parameters [17]. In Fig. 3(a) we can see the TAS writing sequence. The low heat current (~87 uA) passes through the MTJ and heats the junction (65nm × 65nm) up to the blocking temperature of 150°C. Then a bidirectional comparatively high current (~4mA) generating magnetic field is sent to write the junctions. In Fig. 3(b) we show the sensing operation, the control signal “SENSE” should be set to ‘0’ during the pre-charge phase and as soon as the “SENSE” is turned on, the “DATA” and “DATA_BAR” lines come to a stable value in as low as 200ps [15]. The signal “RD_SELECT/WR_SELECT_N” controls the switch of sensing and switching operations.

![Figure 3. Transient Simulation Results](image)

### 4. AUTOMATIC MACRO GENERATION

Fig. 4 shows the bank organisation of the memory macro, this allows the sharing of switching/heating drivers by different banks. Fig. 5 details the organization of MTJs inside a memory bank. Only one select transistor is used for each word with $n_b$ bits and there are $n_w$ words in a memory bank. The select transistor is driven by the $n_b$ addressing lines which are decoded from the $\log(n_w)$ LSBs of the memory address. The macro can be used both in twin cell mode and reference cell mode as explained in section 2. The even and odd reference cells (with slightly larger size) [18]
are placed in the word address 0 and 1 respectively. While reading in reference cell mode the word according to the given address and the alternate reference cells are selected. That is if the address is odd, the even reference is selected and vice versa. While reading in twin cell mode two consecutive addresses are considered to be one cell.

While writing in reference cell mode only, the references are initialized to $R_{AP}$ or value ‘1’, prior to any utilization of the memory. Thereafter only one word cell is selected and heat amplifiers are turned on. This will heat up the particular word to the blocking temperature, thus when the field current amplifiers are turned only that will be written according to the field direction. The sequencing and selection of mode is done via a memory controller which is a synchronous state machine, and it also generates the “Request” and “Wait” signals since write and read speed for this memory is not equal. In fig.4, the address bus to the macro consists of \( \log(N_{BANKS} \times N_{ROWS}) \) MSBs and \( \log(n_w) \) LSBs. The address bus to the macro consists of \( \log(N_{BANKS} \times N_{ROWS}) \) MSBs and \( \log(n_w) \) LSBs. The addresses are decoded to select one word in the bank. The \( \log(N_{BANKS} \times N_{ROWS}) \) MSBs are decoded to generate select signals for banks resulting in a unique selection of word in the macro. The banks are used in a row of length \( N_{BANKS}=2 \) to share the write field amplifiers and row select cells. Only two set of field amplifiers are used which are shared for all banks. The controller selects the row with the help of row select cells depending on the address, through which a short circuit current will circulate and generate the magnetic fields. Due to the high current required for switching, the row select cell transistors are very big for adequate drive strength. For this reason using two or more banks in a row will reduce the macro area. The field lines pass under the MTJs for TAS switching.

This memory organization allows the sensing, switching and heating circuits are shared globally inside the memory bank and the memory macro. A MTJ crossbar is generated at the center of the memory and the density can be easily increased if the pitch of MTJs can be reduced at the bank-end fabrication.

**5. INTEGRATION OF MTJ WITH CMOS**

**5.1 Standard cell design**

In this section we discuss the integration of MTJ technology into well developed CMOS design flow. In Fig. 6, we can see the layout of the MRAM macro which occupies a number of standard cell rows. The CMOS logic parts are generated by the CMOS Placement & Routing tool [19]. It doesn’t know about the MTJ junctions but sees routing and placement blockages instead. The MTJ (rose circular) part contact the CMOS circuit with a thin metal layer (blue lines), metal 4 (green lines) are connected with the select transistors, metal 5 (yellow lines) are the top layer to generate the magnetic switching field. In our layout design based on CMOS 130nm and MTJ 120nm [16], the standard cells representing arbitrary logic being placed in the above 4 rows. It can use 66% of metal levels 2, 3 and 4 for routing purposes. An automatically generated MRAM bank with \( n_w=512 \) words of \( n_b=8 \) bits occupies an area of 818.77x29.52um\(^2\). It amounts to a density of about 6um\(^2\)/bit, but this also includes the 4 standard row cells and 3 routing layers available for CMOS logic.

Fig. 7 describes the hybrid MTJ/CMOS prototype. It consists of a 128Kb memory macro with write drivers and the controller. The
drivers are programmable for strength, and pulse durations can also be modified. This strength and duration values are written to the registers inside the controller. This is done keeping in mind the need of experimentation, both heating duration and field duration can be modified.

6. CONCLUSION AND PERSPECTIVES
In this article, we addressed the issue of hybrid MTJ/CMOS design Methods for embedded MRAM development. An automatic design flow for this purpose will be very useful for applications of type Memory-in-Logic where non-volatile memory is integrated into CMOS logic circuits, thereby increasing the density and reducing the memory transfer bottleneck. Later in the article we have shown design of MTJs with very high reading speeds (~200ps) and high writing speeds (e.g. 40MHz), mainly limited by the speed of buffers providing this high current. However this is a process meant for academic research, many industrials have presented better processes. From the automatically generated layouts in this article we see that 66% of the area (CMOS and 5 routing layers) below the MRAM is available for CMOS logic design. Although the density of these macros may be less than SRAM, the additional logic capability must be taken into account for comparison. Finally the experiments on the prototype presented will provide us enough data for modelling and right design decisions. Future research directions include enhancing the reliability of MRAMS, embedding of error correcting codes, and design of non-volatile Flip-Flops [20] and use of MTJs with perpendicular anisotropy for long term scaling [21].

7. ACKNOWLEDGMENTS
The authors wish to acknowledge support from the French national project CILOMAG and the research contract NANO2012 with STMicroelectronics. We thank also Guillaume PRENAT and Bernard DIENY from SPINTEC laboratory for decisive inputs scientific discussions and crucial help with the simulation model, Olivier REDON from CEA LETI and Lucien PREJBEANU from Crocus technology for wafer design and fabrication, Kholdoun TORKI and Gregory DIPENDINA from CMP for the magnetic design kit and demonstrator assistance.

Figure. 6. The layout of MRAM bank including the CMOS standard library and MTJ standard cells (rose circular). Metal levels 2, 4 and 5 are used to route address lines, logic lines and field generation lines respectively.

Figure. 7. Hybrid embedded 128k MRAM prototype based on CMOS 130nm and MTJ 120nm.

8. REFERENCES