Microelectronic Engineering 147 (2015) 171–175

Contents lists available at ScienceDirect

Microelectronic Engineering

journal homepage: www.elsevier.com/locate/mee

Causes and consequences of the stochastic aspect of filamentary RRAM

R. Degraeve [a,*], A. Fantini [a], N. Raghavan [b], L. Goux [a], S. Clima [a], B. Govoreanu [a], A. Belmonte [a,c], D. Linten [a], M. Jurczak [a]

[a] IMEC, Kapeldreef 75, B-3001 Leuven, Belgium
[b] Singapore University of Technology & Design (SUTD), Singapore
[c] ESAT Dept. KULeuven, Leuven, Belgium

ARTICLE INFO

Article history:
Received 20 February 2015
Received in revised form 1 April 2015
Accepted 6 April 2015
Available online 16 April 2015

Keywords:
RRAM
RTN
Stochastic events

ABSTRACT

Filamentary resistive RAM devices are affected by various sources of instability and variability. Stochastic fluctuations, caused by variability of the RRAM-driving transistor and process-induced device-to-device variations, affect both forming and set/reset transients. Also the intrinsic stochastic mechanisms that control the filament growth and shrinkage complicate the set and reset operation. In particular, RRAM operated at low current shows very broad distributions of both the low and high resistive state. Furthermore, at low read voltage a large amplitude Random Telegraph Noise is measured, caused by both electrostatic modulation of the constricted current flow and by random perturbation of the vacancies in the filament. Additional variability arises from the stochastic probability of finding an intact or a ruptured filament in high resistive state at low current compliance.

1. Introduction – variability in scaled devices

Coping with variability is one the main challenges when scaling nano-technological devices [1–5]. When decreasing the dimensions of devices on chip, it has become increasingly difficult to make them identical to one another. As a consequence, both operational and reliability characteristics are nowadays expressed as distributions with a tolerated standard deviation rather than as sharply defined targets or thresholds [3,4].

Variability can be categorized in two groups corresponding to two root causes.

First, the *extrinsic* variability is caused by variations in device fabrication. This includes geometric device-to-device variations as for example length and width variations, oxide thickness variations, line edge roughness, etc. Also compositional variations of alloys, dopant concentration variations, etc. come under this category. All these variability sources have in common that they can be improved by a tighter process control. In other words, this variability is technology-controlled, and is – to a certain extent – manageable by optimizing the fabrication process.

Second, the *intrinsic* variability is related to the discrete nature of atoms and atomic defects. The issue of random dopant fluctuations [2,5] is probably one of the best known examples: for a fixed doping concentration, the mean number of dopant elements in an aggressively scaled transistor becomes a small integer value with stochastic device-to-device variations. As a result, the source-to-drain current flow cannot be considered uniform and different configurations of percolating paths arise in every transistor.

In scaled devices, also reliability-controlling defects like for example oxide/semiconductor interface states or volume defects in thin layers, have become denumerable integer quantities with strong stochastic device-to-device variations [6,8]. Controlling this type of variability would require advanced atom-by-atom handling which is far out of reach of conventional large-volume processing.

Variability has become prominently visible in all scaled devices, be it high-performance MOSFETs [1,2,7] or integrated memories [9,10]. Understanding the origin of variability is a major step into limiting or controlling it as far as this is physically possible. In this paper, we look in particular into the causes and consequences of stochastic fluctuations in resistive RAM devices.

2. Filamentary resistive RRAM

Resistive RAM has been proposed as a potential future memory candidate [11–15]. In general, an RRAM device is a metal–insulator–metal (MIM) device with a tunable insulator resistance value that can be set to a selected programmed state. Depending on the dimensionality, three types of RRAM are considered [16]. (i) 3D bulk devices have a cell resistance that changes in the total volume of the memory element (for example: memories based on phase change or on insulator–metal transition). (ii) 2D devices have a surface switching mechanism (for instance moving or

* Corresponding author.

http://dx.doi.org/10.1016/j.mee.2015.04.025
0167-9317/

modulating the metal–insulator interface). (iii) 1D devices use filamentary switching as operation mechanism. In these devices, a narrow (~1D) path for electronic conduction is formed connecting the electrodes in an MIM stack. The resistance of the filament is modulated by applying bias (different in polarity or magnitude or both) to the electrodes.

In the following we focus on the 1D filamentary switching type of RRAM. The filament could consist of small in-diffused interstitial metal atoms resulting from migrated cations (usually small radius ions: Cu^+, Ag^+ or Ni^{2+}), or of oxygen-deficient sites (i.e., oxygen vacancies) resulting from out-diffusion of O^{2-} anions [17].

1D-filamentary RRAM shows the promise of perfect area scaling [14]. Indeed, since the filament occupies only a small part of the area of the device, scaling the 2D device dimensions can be done without changing the 1D operation performance. From the viewpoint of variability, however, this scaling property is exactly the main problem: since only a very small part of the device, consisting of a few atom-sized particles, controls the memory function, 1D-filamentary RRAM is preeminently vulnerable to device-to-device variations.

Unfortunately, since the constitutive particles of the filament move around during every set and reset transient, reconstructing the same filament for every write/erase cycle is just as challenging as controlling the device-to-device variation. Especially below 10 μA operating current, when the number of particles in the filament becomes very low, the cycle-to-cycle variability rises to a level that is no longer acceptable to guarantee a proper memory function. At this current level, the high resistance state (HRS) distribution and low resistance state (LRS) distribution are largely overlapping [25].

In the following sections we aim at giving a detailed overview of the multitude of causes that result in variability of filamentary RRAM. Where possible, we indicate to what extent the variability might be controlled and to what extent it is due to intrinsic stochastic processes. First, we discuss the stochastic components in the forming process (Section 3), followed by the stochastic variations in the set and reset transients (Section 4). In Section 5, dynamic instabilities at read voltage (RTN) are explained and finally the stochastic probability for observing a ruptured filament at low current is shown (Section 6).

3. Stochastic components in the forming process

When the RRAM device is completely fabricated, an electrical voltage pulse is needed to form the filament. One could consider this as the last processing step, needed to prepare the RRAM for its normal operation. In oxygen vacancy-driven RRAM stacks (=OxRAM devices), the forming operation is akin to dielectric breakdown. Since this phenomenon has been studied for many decades as a prominent reliability topic, the learning from the field of dielectric breakdown has helped to understand RRAM [18].

Nevertheless, there is an important difference between an SiO_2 gate breakdown and a forming process in a transition metal oxide. The former has been related to stress-induced *generation* of oxygen vacancies [19], while the latter involves both *moving* pre-existing vacancies as well as creating new ones [17].

This is because well-designed 1D-RRAM devices are highly asymmetric and have a reservoir of atomic 'building blocks' at one electrode that serve to form the filament. If the building blocks are oxygen vacancies, a capping layer of for instance Hf or Ti can be used to create a sub-stoichiometric (vacancy-rich) transition layer at one side of the transition metal oxide [17,21] as illustrated in Fig. 1a. Similarly, if the building blocks are metal atoms, a metal layer with diffusion barrier can be used as a reservoir [20]. From

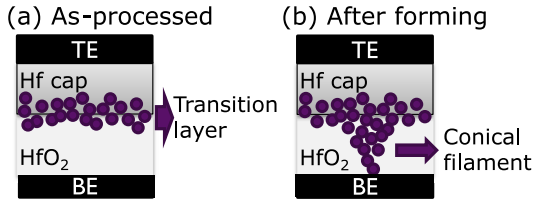

Fig. 1. Schematic representation of an OxRAM device showing (a) the sub-stoichiometric transition layer that serves as a reservoir of oxygen vacancies and (b) the conical shaped filament after the forming step. Forming is done by applying a positive voltage pulse to the top electrode (TE).

these reservoirs, the filament can directionally grow towards the opposite electrode during the forming pulse (Fig. 1b).

An example of an OxRAM device is shown in Fig. 2. HfO_2 is the insulating matrix in which the filament is formed. It is capped by Hf and the bottom and top electrodes are TiN. Device scaling down to lateral dimensions of ~10 nm without impact on the RRAM properties has been demonstrated for this stack using a crossbar device layout [14].

Although the forming step is crucial, tight forming *control* is very challenging. There are several reasons for this, both extrinsic and intrinsic in nature.

First, the RRAM device is driven through an on-chip external periphery as shown in Fig. 2b. In its simplest shape, a load resistor is used (2R configuration), but mostly a driving transistor is implemented (1T-1R configuration). The latter configuration is used throughout this paper. Since the forming step is an abrupt change from insulating to conductive state, parasitic capacitive discharge events will result in a loss of control over the filament growth. For this reason, all device dimensions (resistive element, transistor as well as the connecting lines) must be as small as possible [23].

Note that the presence of a load resistor or a driving transistor limits the experimental accessibility of the RRAM device itself. All RRAM characteristics must be measured through the periphery. Connecting as little as a bonding pad directly to the RRAM changes its performance drastically because of capacitive discharge during the forming. For the same reason, filaments formed in large area MIM capacitors – often used to evaluate new material options in research environments – will never yield identical filaments as in small capacitors with the same stack.

The forming process is very critical as it controls the number of particles (vacancies or metal atoms) available for the subsequent switching events [24,28]. The main determining parameter in the

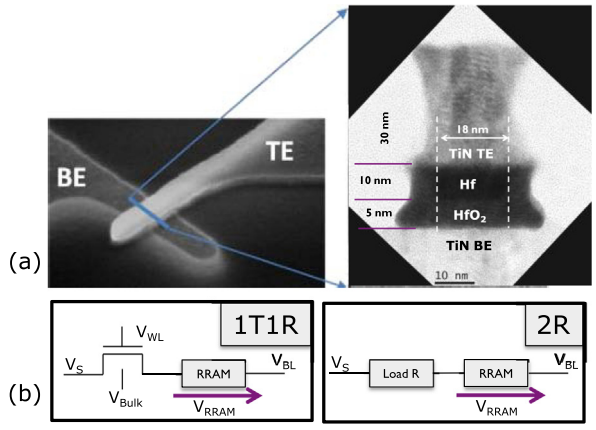

Fig. 2. (a) Cross-bar RRAM element with a TiN/HfO_2/Hf/TiN stack. Excellent performance down to $10 \times 10\ nm^2$ has been demonstrated [14]. Process details are in [22]. (b) Two configurations for driving the RRAM device. Either a transistor is used (1 transistor–1 resistor or 1T1R configuration) or a load resistor is connected to the RRAM (2R configuration).