Model-Based High-Precision Tuning of NOR Flash Memory Cells for Analog Computing Applications

F. Merrikh Bayat†, X. Guo†, M. Klachko†, N. Do‡, K. Likharev††, and D. Strukov†††

1 UC Santa Barbara, Santa Barbara, CA 93106-9560, U.S.A.
2 Silicon Storage Technology Inc., A Subsidiary of Microchip Technology Inc., San Jose, CA 95134, U.S.A.
3 Stony Brook University, Stony Brook, NY 11794-3800, U.S.A.

Email: †farnodmb@ece.ucsb.edu, ‡strukov@ece.ucsb.edu, ††konstantin.likarev@stonybrook.edu

High-precision individual cell tuning was experimentally demonstrated, for the first time, in analog integrated circuits redesigned from a commercial NOR flash memory. The tuning is fully automatic, and relies on a write-verify algorithm, with the optimal amplitude of each write pulse determined from runtime measurements, using a compact model of cell’s dynamics, fitted to experimental results. The algorithm has allowed tuning of each cell of a 100-cell array to any desired state within a 4-orders-of-magnitude dynamic range. With 10 write pulses, the average tuning accuracy is about 3%, while with 35 pulses the precision reaches ~0.3%. Taking into account the dynamic range, the last number is equivalent to ~1,500 levels, i.e. 10+ bits.

Analog weights are required for a variety of important applications including threshold logic circuits, AD/DA converters, analog trimming circuits, and most importantly artificial neural networks [1, 2]. In many such applications, the memory devices holding the analog weights are the most critical components, occupying the largest area of a circuit. (Since the weights are typically changed infrequently, their tuning time and energy are usually of a secondary importance.)

Flash memories are inherently analog, and offer the advantage of a mature fabrication technology, ready for large-scale integration. While being not as compact as some emerging nonvolatile analog memory devices [1], highly optimized commercial versions of flash memory have very high density. Yet, commercial flash memories are optimized for digital applications and require modifications to enable individual high-precision analog tuning of their cells.

Earlier, we have demonstrated a modification of the commercial ESF1 NOR memory from SST Inc. [3] without interfering with the highly optimized cell fabrication process to enable high precision tuning of individual memory cells [4]. While the cell area fabricated in a 180-nm process is about 1.5 μm², i.e. ~2x of the original area, it is still at least 10x smaller than that of prior analog flash memories [2]. Here we demonstrate high-precision tuning of cells of the modified ESF1 arrays. The experiments were performed within a 10×10 cell array, with additional two rows of peripheral cells, to implement a gate-coupled vector-by-matrix multiplier [5] (Fig. 1). The experiments have shown that the desired high tuning accuracy cannot be reached with just one write (either partial erasure or partial programming) pulse, and several alternating pulses are necessary. In order to minimize the pulse number, we have developed an algorithm whose main idea (Fig. 2) is to use runtime state measurements, together with a compact model of the erasure and programming dynamics, fitted to the experimental data (Fig. 3), to calculate the optimal amplitude of each next write pulse. Due to device-to-device variations in switching behavior, rather noticeable at this precision level (Fig. 4), model’s parameters are adjusted at the initial tuning stage of each specific cell. The algorithm also takes into account a very steep dynamics of the erasure process, governed by Fowler-Nordheim tunneling (in contrast with the programming process based on the hot-electron injection), by using smaller erase pulses than would be formally recommended by a deterministic model, to avoid frequent overshoots. The optimal algorithm parameters, giving the smallest average number of tuning pulses (and hence the fastest tuning time) have been found via an exhaustive data-based search. Tuning algorithm’s functionality has been successfully verified in variety of conditions (Figs. 5 and 6). Naturally, the tuning is faster when the necessary tuning precision is lower, with a roughly exponential increase in the number of pulses required to get higher precision (Fig. 5b). This is similar to the recently reported PCM memory cell tuning results [6, 7].

In conclusion, we have experimentally demonstrated high-precision individual cell tuning in modified commercial NOR flash memory arrays. Similarly to previous proposals [8-11], we use a feedback algorithm, yet unlike other model-based algorithms [9-10], our approach does not rely on the global erasure - the feature essential for tuning of large-scale arrays. Furthermore, to our knowledge this is the first time when a high-accuracy tuning algorithm was successfully applied to very dense commercial-grade memories. The demonstrated tuning precision is very encouraging and is likely sufficient for first analog-hardware implementation of practical artificial neural networks [12] (Figs. 7 and 8). The key element of such networks is the vector-by-matrix multiplier (Fig. 1) whose weights are set up according to the specific classification task. In addition, we believe that our approach may also be used for writing into multilevel digital flash memories.

This work was supported by DARPA under Contract No. HR0011-13-C-0051UPSIDE via BAE Systems.

Fig. 1. Fabricated gate-coupled vector-by-matrix multiplier: (a) circuit’s layout in a 180-nm process, and (b) its schematics. The first and the last row of supercells are the parts of the peripheral current-mirror circuitry [5], which converts input currents into voltages applied to the gates of the cells in array’s body.

Fig. 2. Flowchart of the optimized tuning algorithm.

Fig. 3. Measured and fitted switching dynamics data: output current as a function of the write pulse amplitude for different initial states for: (a) programming and (b) erasure.

Fig. 4. (a, b) Device-to-device variations for erase and programming voltage thresholds for 100 devices. (c) Cycle-to-cycle variations showing the spread of the final current after erasing (programming) a cell 500 times.

Fig. 5. (a) Measured vs. target final states for 100 devices, and (b) tuning error as a function of the required number of tuning pulses. (a) confirms that the disturbance of half-selected devices during tuning is negligible, while panel (b) shows that the number of necessary tuning pulses grows exponentially with the required accuracy.

Fig. 6. (a) Average number of required write pulses, and (b) tuning error for 100 devices, for different, log-spaced target states, and (c, d) histograms of the required number of erase and programming pulses.

Fig. 7. Preliminary retention data showing repeated current measurements over >1 day for a cell tuned to different states.

Fig. 8. Simulated results for the classification fidelity degradation of neural networks, tested on common benchmarks and trained ex-situ, as a function of weight import precision.