Hardware-Algorithm Co-Design of Analog Recurrent Neural Networks for Sub-Microwatt Inference

J. Brandoit[(1)], A. Fyon[(1)], L. Mendolia[(1)], D. Ernst[(1)], J.-M. Redouté[(1)] and G. Drion[(1)]

[(1)] Department of Electrical Engineering and Computer Science, University of Liège, Belgium

Always-on AI applications require ultra-low power consumption, yet current approaches face a fundamental tradeoff: digital implementations consume excessive energy (mW), while analog circuits struggle with reliability and noise accumulation. We address this challenge through hardware-software co-design, selecting a recurrent neural network (RNN) architecture whose algorithmic properties align directly with ultra-low power analog primitives. First Quadrant Bistable Memory Recurrent Units (FQ BMRUs) are a class of RNNs with discrete-valued outputs and hysteretic dynamics [1]. We identify that these properties map naturally onto current-mode memory cells operating in the subthreshold regime [2]. Each learned parameter then corresponds directly to a tunable circuit element: output amplitudes and switching thresholds map to independently programmable bias currents (Fig. 1).

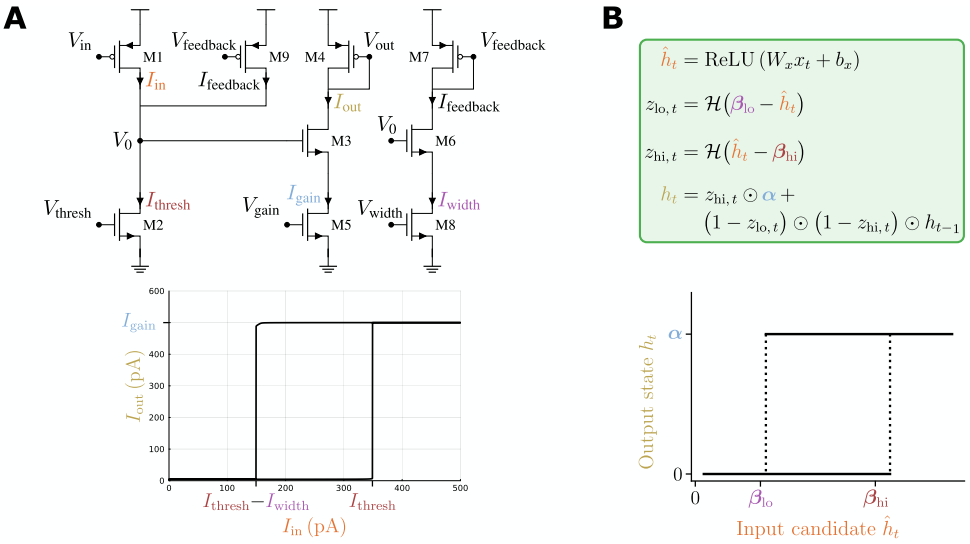

Figure 1: Hardware-software co-design. **A** Current-mode memory cell circuit and its input-output characteristic. **B** FQ BMRU equations and corresponding hysteresis curve, with color-coded parameter correspondence between learned parameters and circuit elements.

We validate the approach on single-word keyword spotting (KWS) using the Google Speech Commands dataset [3]. The complete analog circuit (768 transistors) is implemented in 180 nm CMOS (X-FAB) and simulated in Cadence Spectre. Across 50 test samples, hardware predictions match software in 49 cases, with the single discrepancy occurring at a near-tie decision boundary. The discrete FQ BMRU outputs suppress accumulated analog noise by at least 20-fold at each cell boundary, enabling robust deep architectures. The resulting implementation achieves >90% accuracy at approximately 100 nW average power consumption, at least three orders of magnitude below classical implementations. This work demonstrates that co-designing algorithms and circuits from shared primitive operations can yield efficiency gains unattainable by optimizing either domain in isolation.

[1] F. De Geeter et al., arXiv, 2601.09495, 2026.

[2] A. Fyon et al., patent application, in press, `https://hdl.handle.net/2268/338453`.

[3] P. Warden, arXiv, 1804.03209, 2018.