

17.3 A Current-Saving Match-Line Sensing Scheme for Content-Addressable Memories

Igor Arsovski, Ali Sheikholeslami

University of Toronto, Toronto, Canada

A Content-Addressable Memory (CAM) searches for data by its content and returns the address of the matching data. This feature is used extensively in applications such as internet routers to channel incoming packets towards their destination addresses contained in the packet header. Energy per search and search speed are two important metrics used to evaluate CAM performance [1]. A current-saving match-line (ML) sensing scheme is proposed that substantially reduces the energy per search without compromising search speed. The proposed sensing scheme consumes only 1.3fJ/bit/search, a 60% power reduction compared to previously reported sensing schemes [2,3], to achieve a search time of less than 2ns in a 256x144b array implemented in a 0.13 μ m CMOS process.

Figure 17.3.1 shows a general CAM architecture where search-lines (SL) run perpendicular to MLs. The search data is presented to the SLs which are connected bit-by-bit to all the words stored in the memory. A NAND-based ML architecture is well known [4] for its low power consumption (due to switching of only one ML) and relatively long search time (due to having several transistors in series). In contrast, a NOR-based CAM provides a much faster search by pulling down a precharged-high ML through parallel NMOS transistors that form the NORs. This speed however comes at the price of higher switching activities of the MLs, i.e., all MLs are charged to VDD and then to ground.

Figure 17.3.2 shows the circuit details of the current-saving ML sensing scheme. To explain the circuit operations of this scheme, ignore the effects of the current-saving block and assume the VAR node is grounded [2]. Prior to a search operation, search data is applied to the SLs while all MLs are precharged to ground and all 'sn' nodes are precharged high. A CAM search begins by lowering ML_EN and allowing the PMOS transistors to provide identical currents (I_{ML}) to all MLs. An ML0 (ML with no mismatch) is charged faster than any ML having a one-bit mismatch (ML1) or more (MLn for an ML with n-bit miss). When the Dummy ML (DML, equivalent to an ML0) reaches the threshold voltage of the NMOS transistor (V_{Th}) in the ML sense circuit, the 'sn' node is discharged, flagging ML_EN to turn off all current sources, hence preventing further power consumption in the array. The signal of the DML is slightly delayed, in fact, to let ML0 go above the threshold voltage (V_{Th}) while all other MLs stay below V_{Th} . Without the current-saving block, this architecture consumes power uniformly-distributed across the MLs. That is, all MLs will consume the same amount of power independent of the number of mismatches in the ML. The current-saving block is added to reduce the power consumed by the MLs having one-bit miss or more. This is achieved through dynamically allocating less current to slower-rising MLs as next described.

The VAR node is precharged to ground to guarantee all MLs initially receive the same level of current. This node is then gradually pulled up by a threshold current (I_{Th}) or pulled down by a current proportional to the ML voltage (I_{sink}). In case of an ML0 (or DML), I_{sink} is initially less than I_{Th} , but gradually rises above I_{Th} , pulling VAR to ground and providing a maximum I_{ML} to the corresponding ML. In case of an ML1 (or MLn with n higher than 1), I_{sink} remains less than I_{Th} , allowing the difference to charge VAR and cut off I_{ML} to the corresponding ML.

Figures 17.3.3 and 17.3.4 show the simulation results of the current-saving ML sensing scheme. Figure 17.3.3 compares the voltages developing on ML0 and ML1, both initially at ground. In less than 0.5ns, VAR0 and VAR1, corresponding to ML0 and ML1, respectively, begin to separate, helping ML0 to rise faster than ML1. In less than 2ns, the voltage difference between ML0 and ML1 reaches at least 200mV. This is far larger than any threshold-voltage mismatch of the NMOS transistors signaling a global shut-down of all current sources with ML_EN. Figure 17.3.4 illustrates the current supplied to the MLs during the search operation. The amount of current saving increases monotonically with the number of mismatches in a ML. This current saving reaches nearly its maximum for ML7, for which the total charge delivered by the current source is 48% less than the total charge delivered to ML0. All other MLs (MLn with n > 7) save similar amounts of current as ML7.

To verify the circuit operation under various process conditions, Fig. 17.3.5 shows the comparison of the voltage developed on ML0 with the voltage developed on a 'fast' ML1 (ML1f). To model an ML1f, its capacitance is reduced by 20% to $0.8C_{ML}$ and the size of its corresponding PMOS transistor is increased by 20% to $1.2W_{ML}$. As seen in the figure, ML1f does rise faster than ML0 initially, but falls below ML0 by at least 160mV in less than 2.5ns, producing correct search results. This is made possible partly by increasing I_{Th} in Fig. 17.3.2 through three programmable bits (provided off-chip) and partly by ensuring that under all process conditions, the product of I_{ML} (max) and the resistance to ground of ML1 is less than the minimum V_{Th} . This guarantees that the voltage of an MLn with n \geq 1 never reaches V_{Th} . Increasing I_{Th} , however, decreases I_{ML} , hence increasing the search time by 50% to 3ns.

To compare the energy saved in the proposed scheme versus those of the conventional precharge-high scheme and the current-race scheme, all three methods are simulated in an array of 256 rows by 144b. Figure 17.3.6 shows that the current-save sensing scheme cuts the energy/bit/search by 60% when compared to the precharge-high scheme and by 40% when compared to the current-race sensing scheme.

The proposed scheme is implemented along with current-race sensing scheme in an array of 256x144b for direct comparison of search speed and search energy in a 1.2V, 0.13 μ m CMOS process. The testchip layout, having a total area of 1.6mmx1.8mm, is shown in Fig. 17.3.7.

Acknowledgments

Authors thank Trevis Chandler, Kostas Pagiamtzis, and Marcus van Ierssel for insightful discussions on this work. Authors also thank Canadian Microelectronics Corporation (CMC) for testchip fabrication, and Natural Sciences and Engineering Research Council of Canada (NSERC) for funding.

References:

- [1] I. Y. L. Hsiao, D. H. Wang and C. W. Jen, "Power Modeling and Low-Power Design of Content Addressable Memories," *IEEE ISCAS*, Vol. 4, pp. 926-929, 2001.
- [2] I. Arsovski, T. Chandler, and A. Sheikholeslami, "A Ternary Content-Addressable Memory (TCAM) Based on 4T Static Storage and Including a Current-Race Sensing Scheme," to appear in the *IEEE J. Solid State Circuits*, Jan. 2003.
- [3] P. Lin and J. Kuo, "A 1-V 128-kb Four-Set-Associative CMOS Cache Memory Using Wordline-Oriented Tag Compare (WLOT) Structure with Content-Addressable Memory (CAM) 10-Transistor Tag Cell," *IEEE J. Solid State Circuits*, Vol. 36, No. 4, pp. 666-676, Apr. 2001.
- [4] F. Shafai, K. J. Schultz, R. Gibson, et al., "Fully Parallel 30-MHz, 2.5-Mb CAM," *IEEE J. Solid State Circuits*, Vol. 33, No. 11, pp. 1690-1696, Nov. 1998.

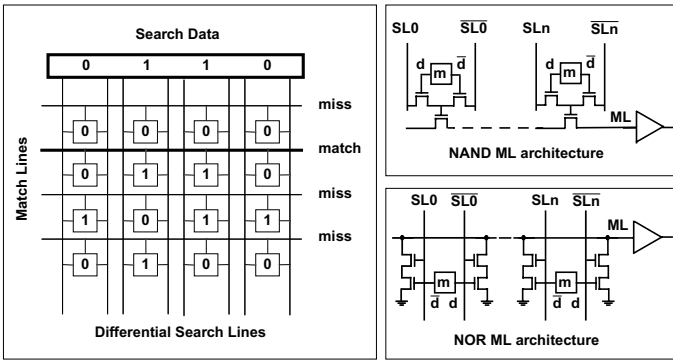


Figure 17.3.1: CAM ML architectures.

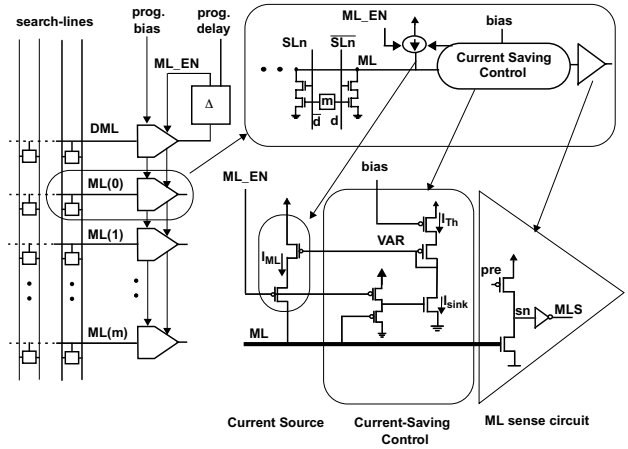


Figure 17.3.2: Current-saving match-line sensing scheme.

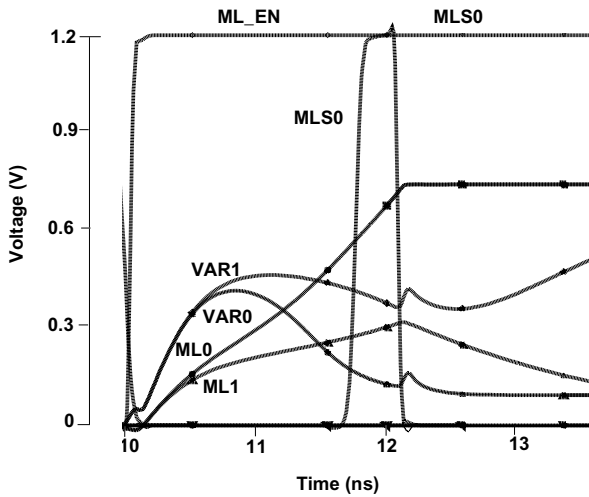


Figure 17.3.3: Voltage development on MLO (fully-matched) and ML1 (one-bit miss).

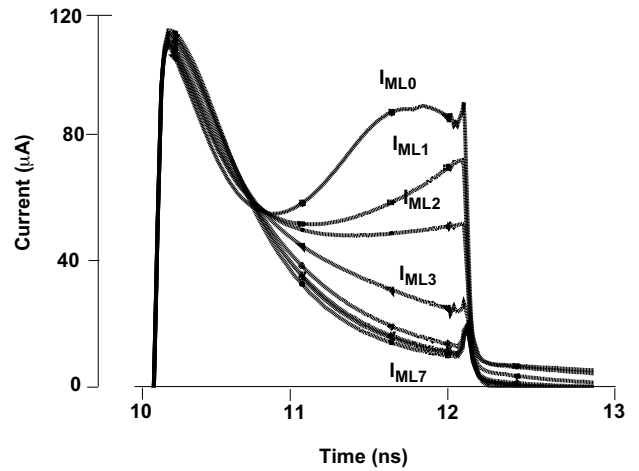


Figure 17.3.4: Current supplied to MLO, ML1, ..., and ML7. (MLn is an ML with n-bit mismatch).

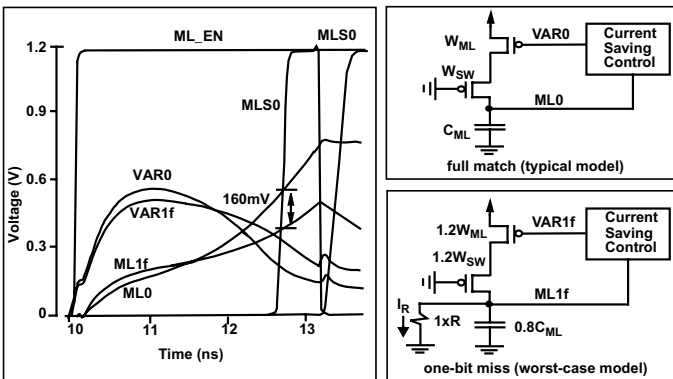


Figure 17.3.5: Simulation results comparing the voltage on a typical MLO against the voltage on a fast-rising ML1 (worst-case one-bit miss).

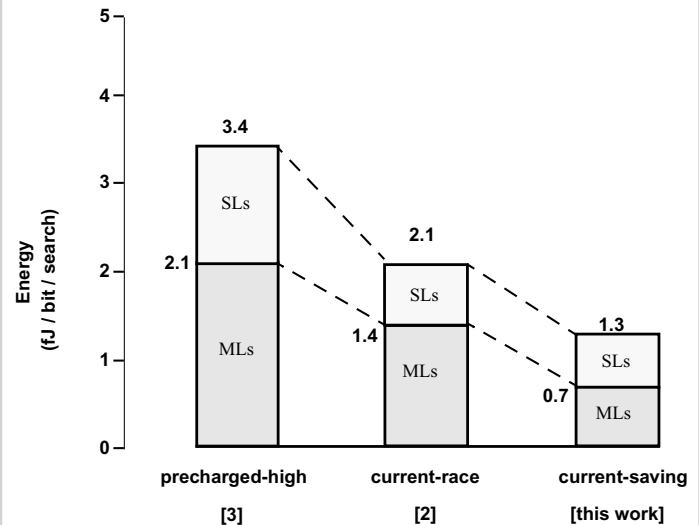


Figure 17.3.6: NOR architecture: energy-per-search comparison.

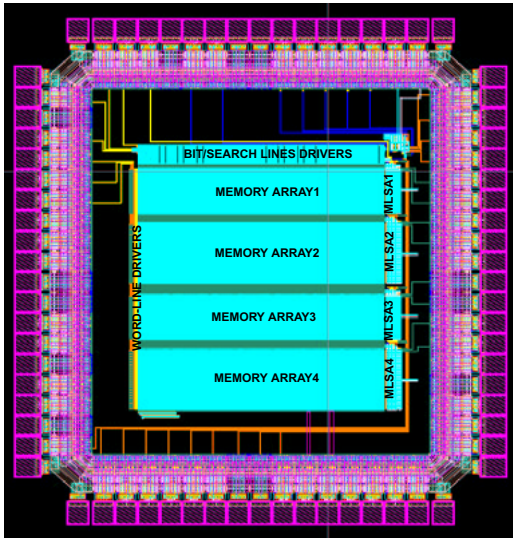


Figure 17.3.7: Chip Layout.

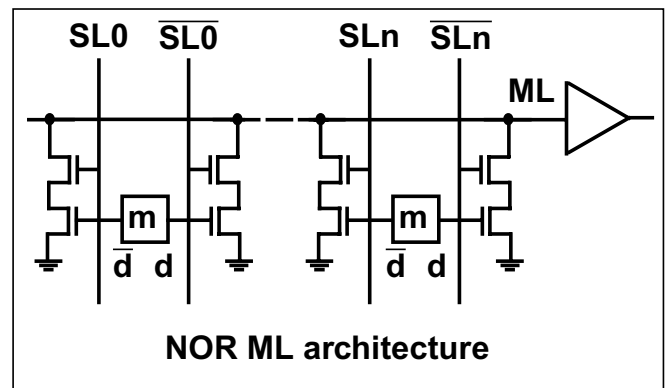
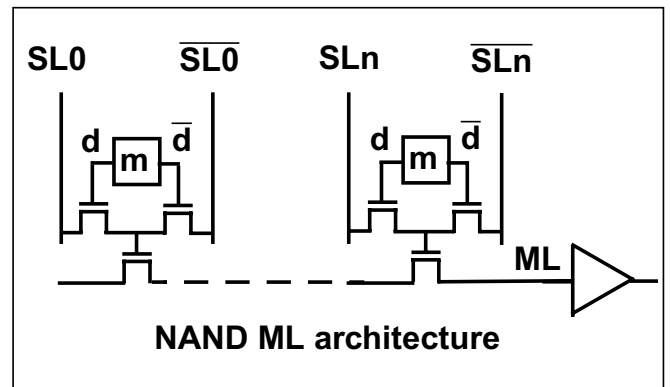
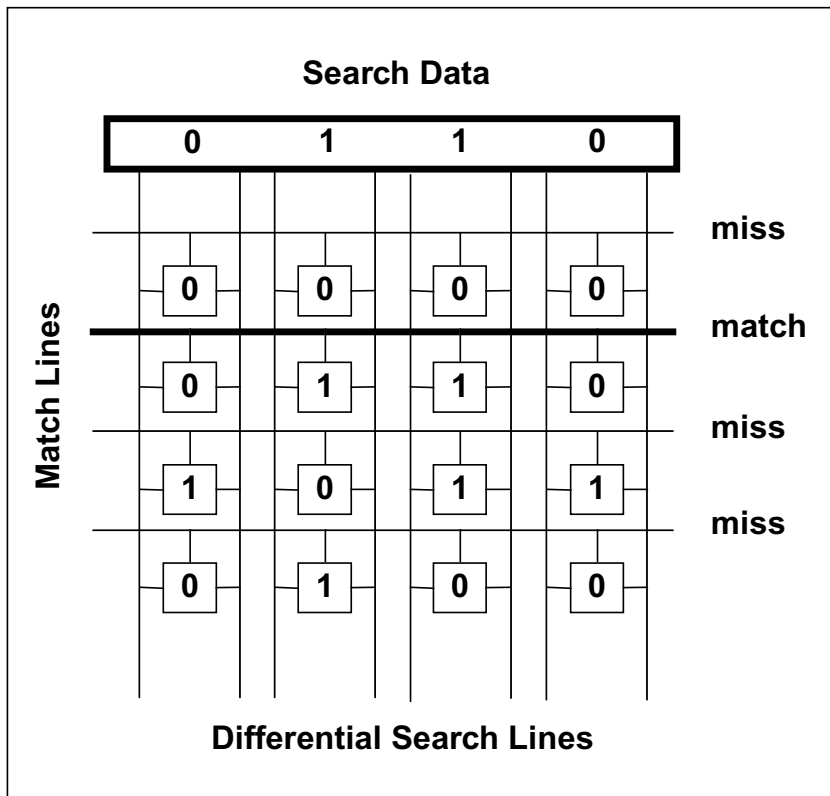


Figure 17.3.1: CAM ML architectures.

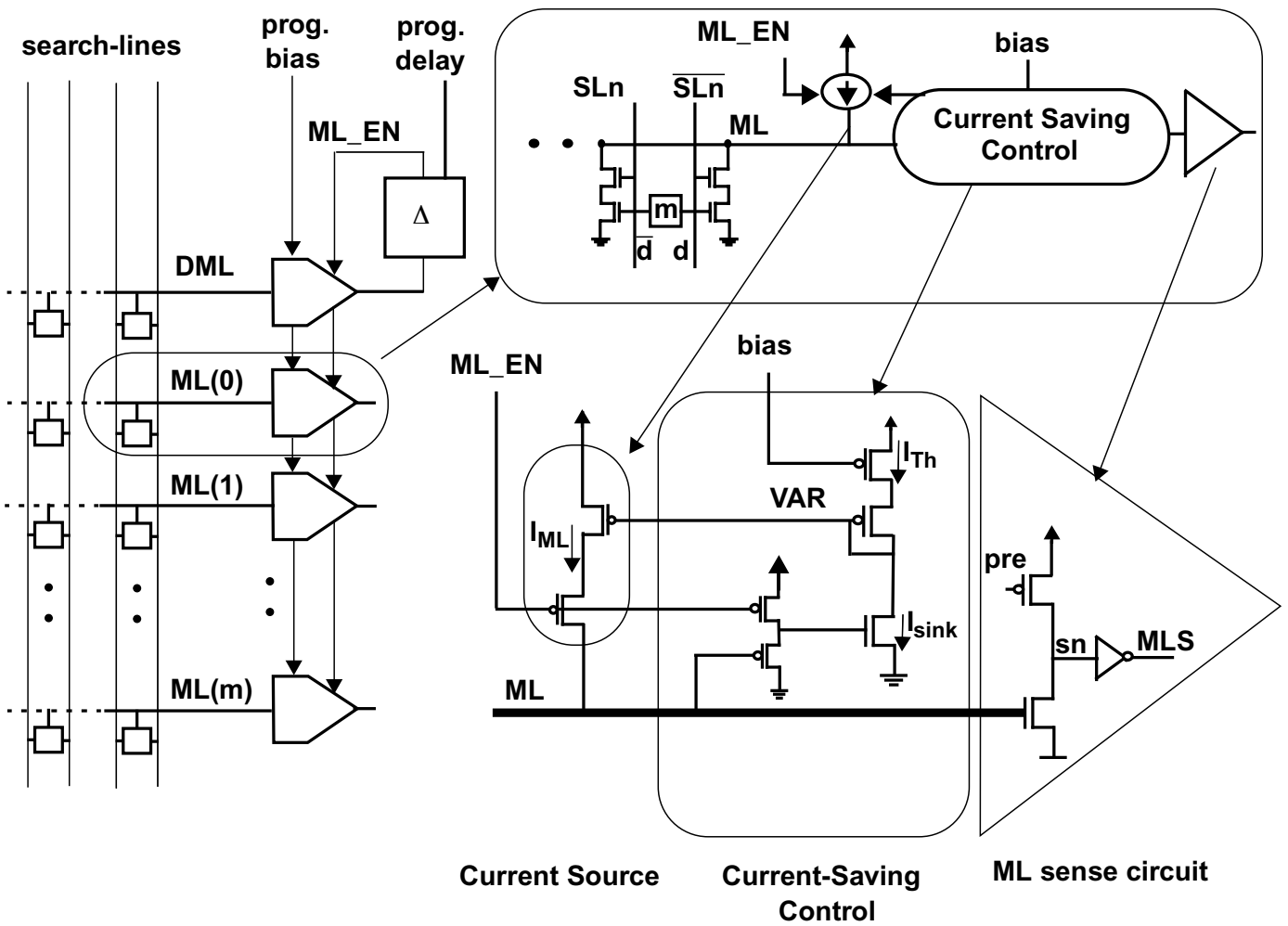


Figure 17.3.2: Current-saving match-line sensing scheme.

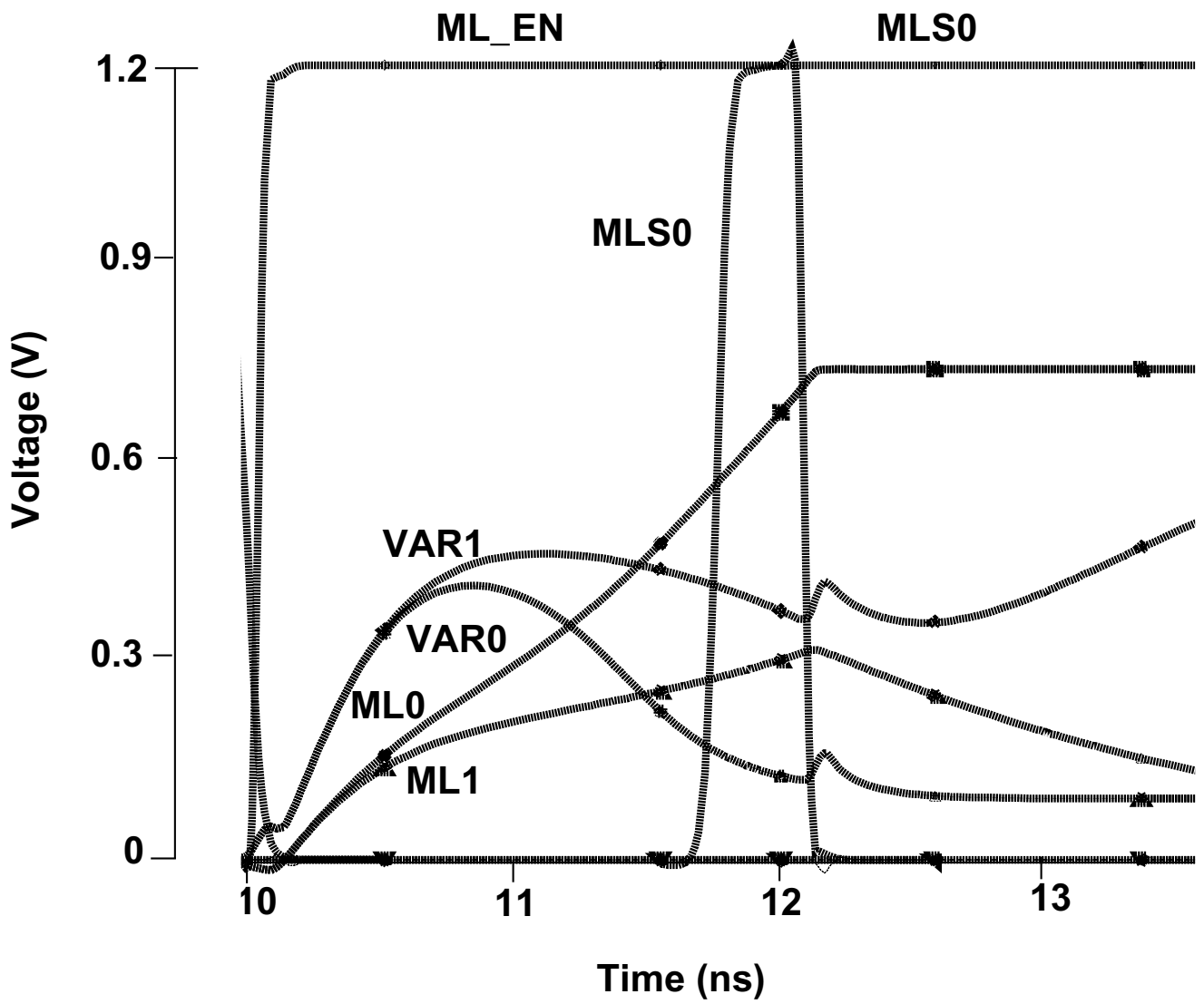


Figure 17.3.3: Voltage development on ML0 (fully-matched) and ML1 (one-bit miss).

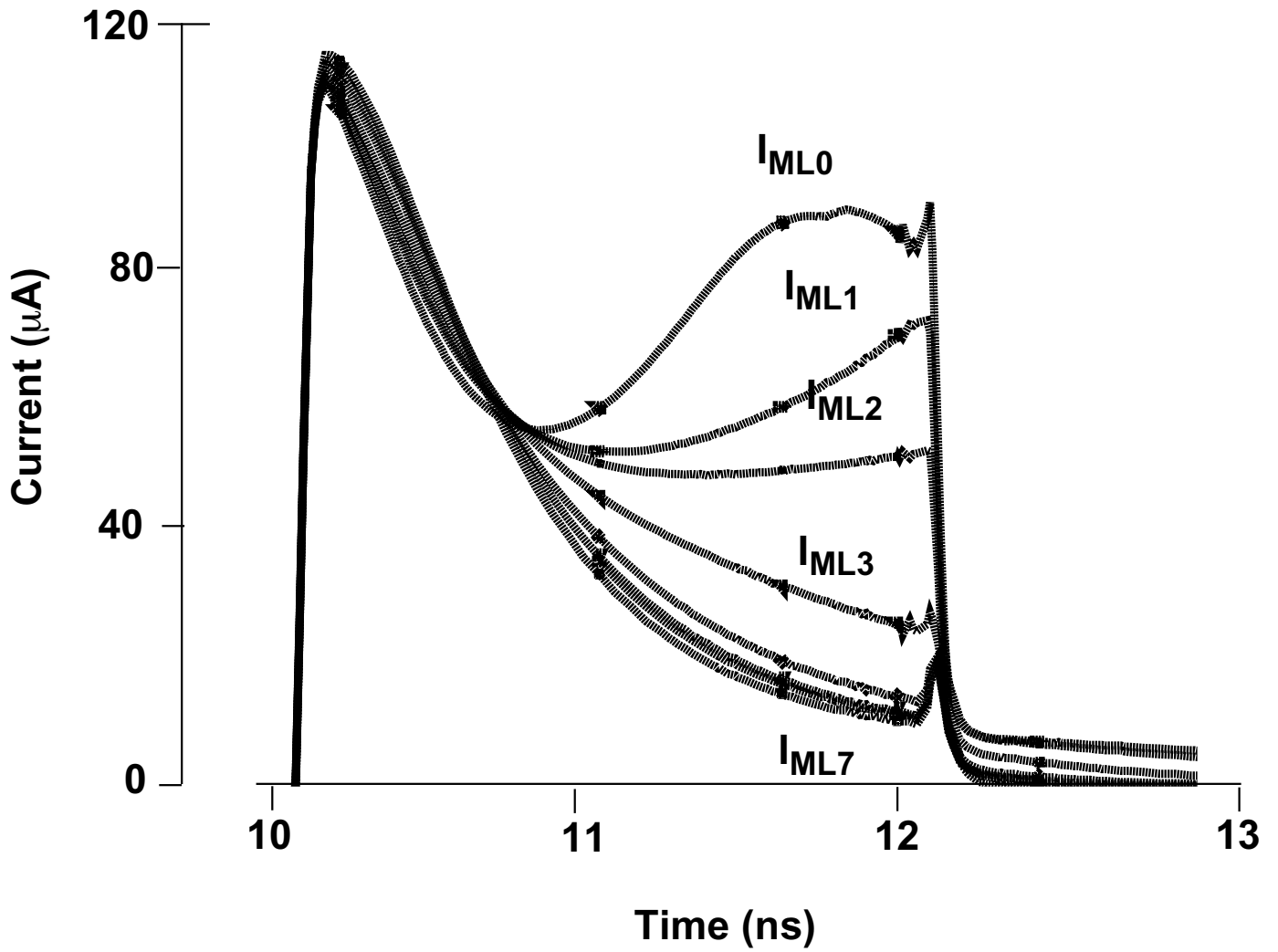


Figure 17.3.4: Current supplied to ML0, ML1,...,and ML7. (MLn is an ML with n-bit mismatch).

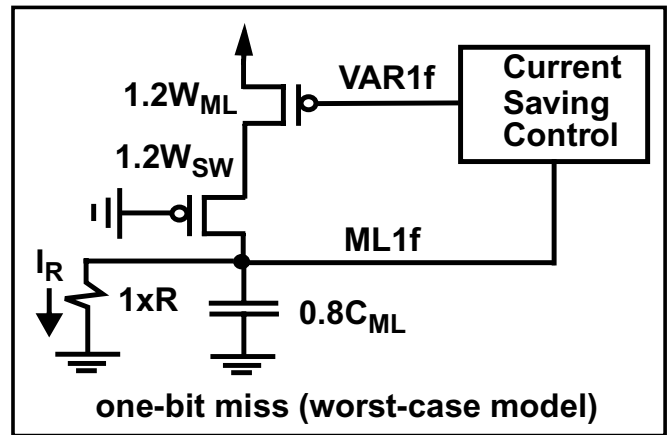
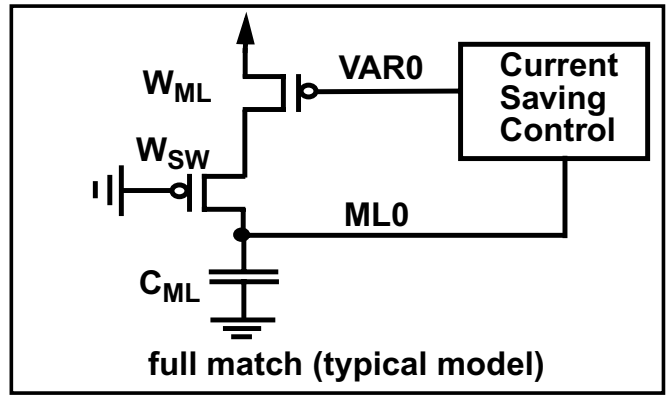
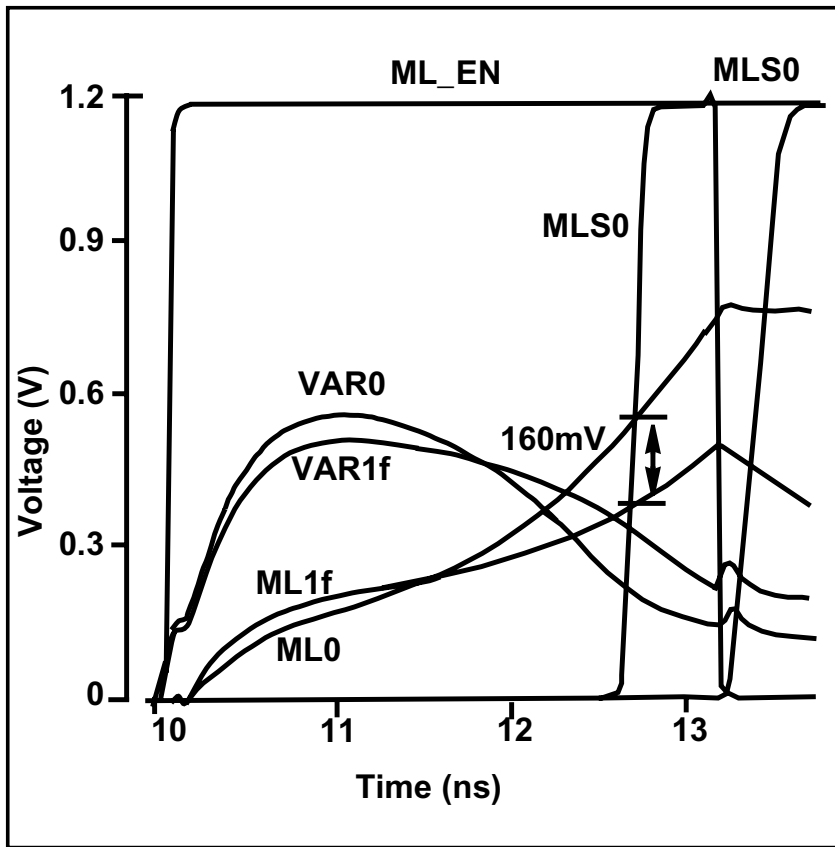


Figure 17.3.5: Simulation results comparing the voltage on a typical ML0 against the voltage on a fast-rising ML1 (worst-case one-bit miss).

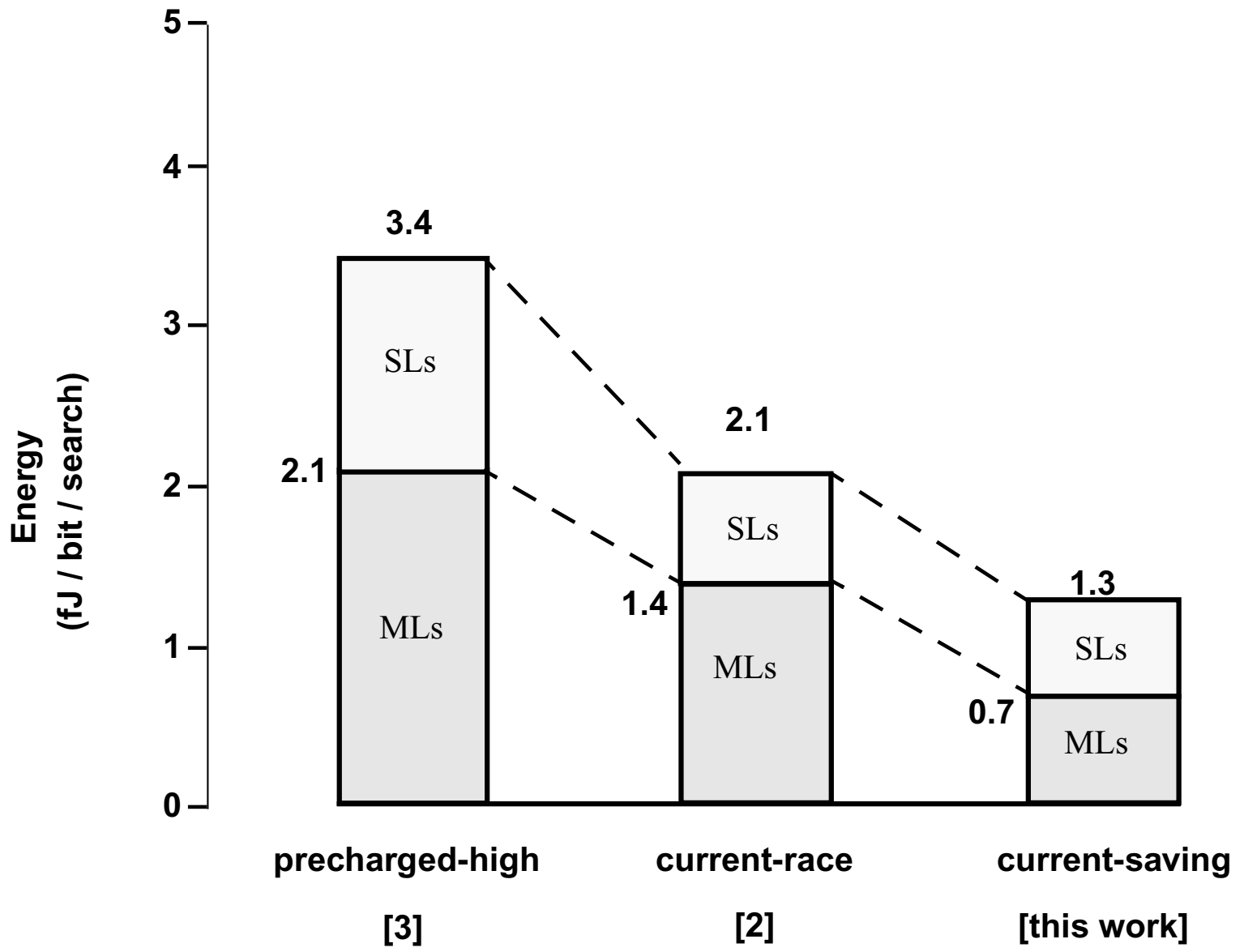


Figure 17.3.6: NOR architecture: energy-per-search comparison.

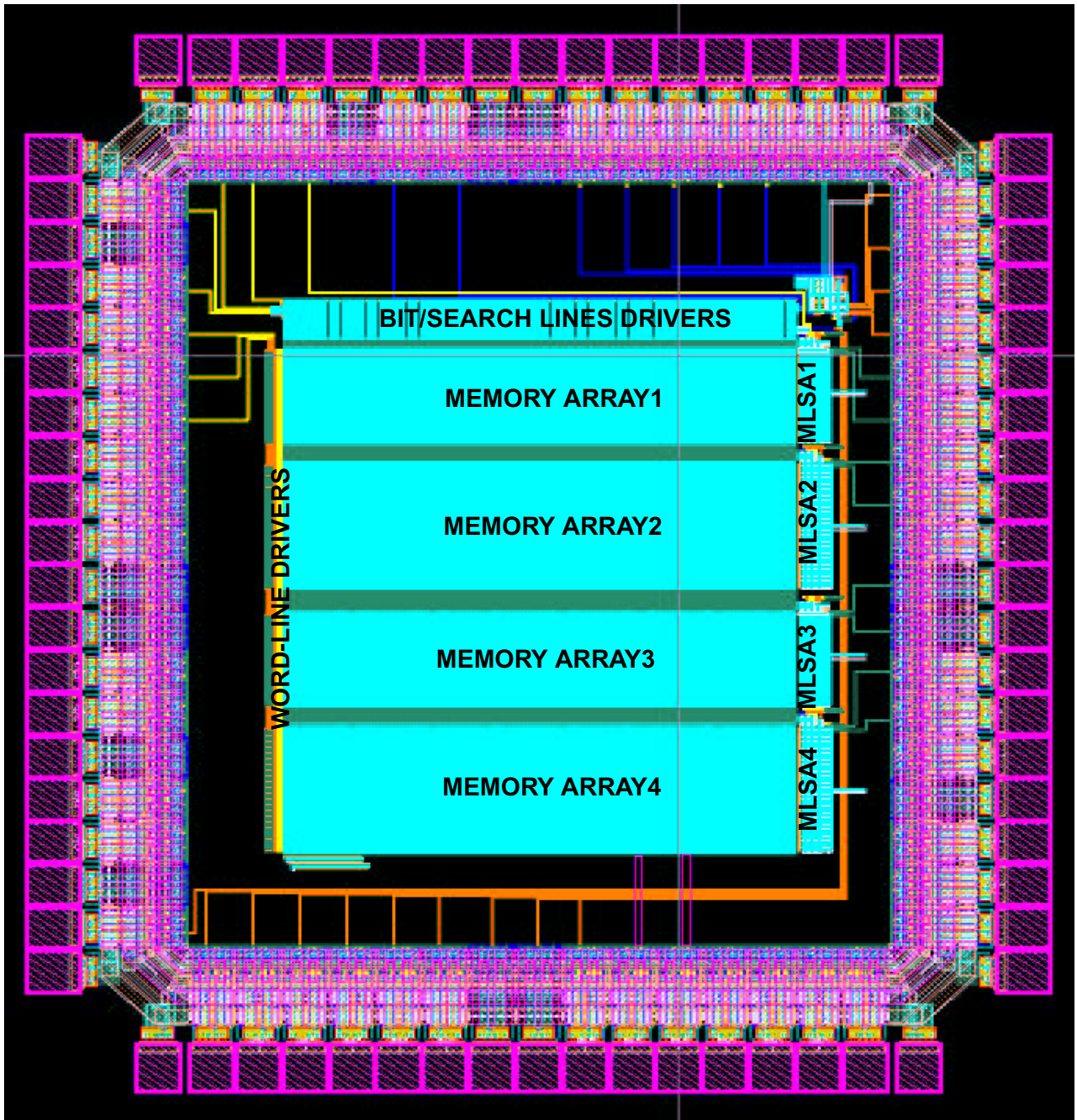


Figure 17.3.7: Chip Layout.