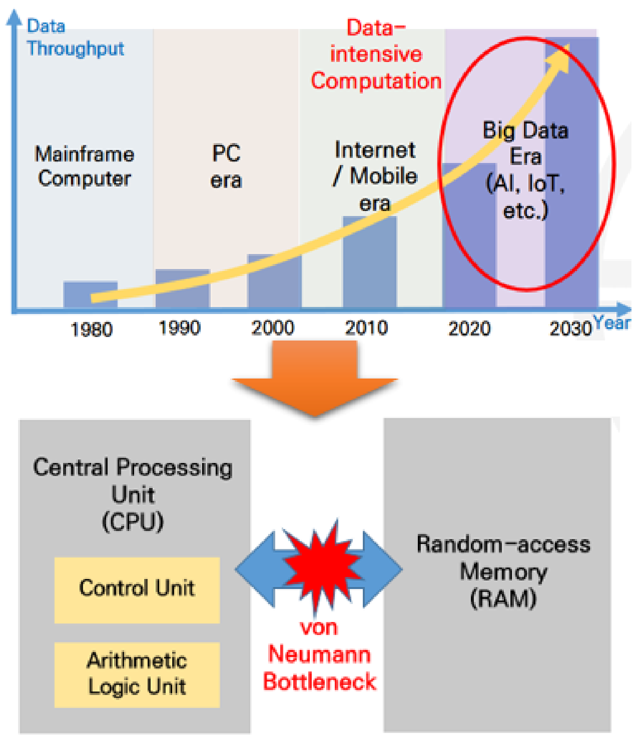
In-Memory Nearest Neighbor Search with Nanoelectromechanical Ternary Content-Addressable Memory

이재성 (삼차원 집적 및 소자 연구실)

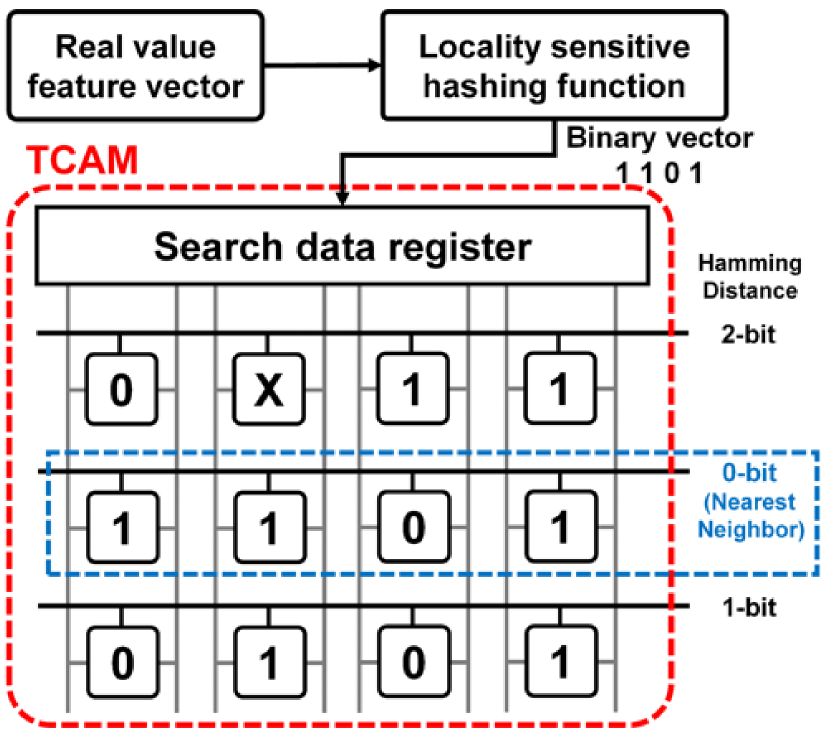
SOGANG UNIVERSITY

Motivation

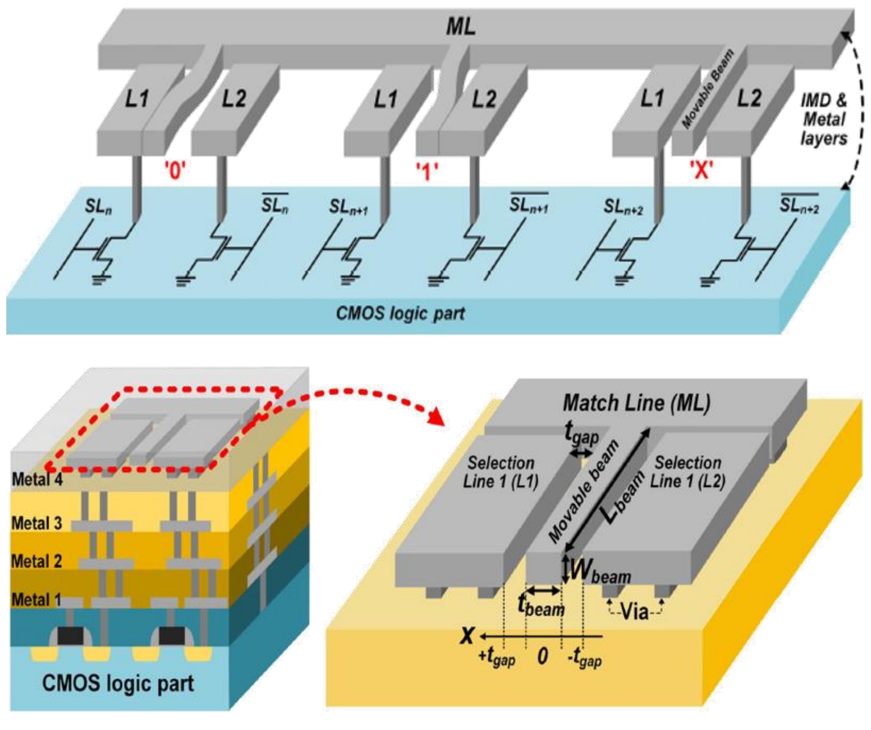
Limitation of conventional CPU-based Nearest Neighbor search



Ternary content-addressable memory-based Nearest Neighbor search method



Ideas

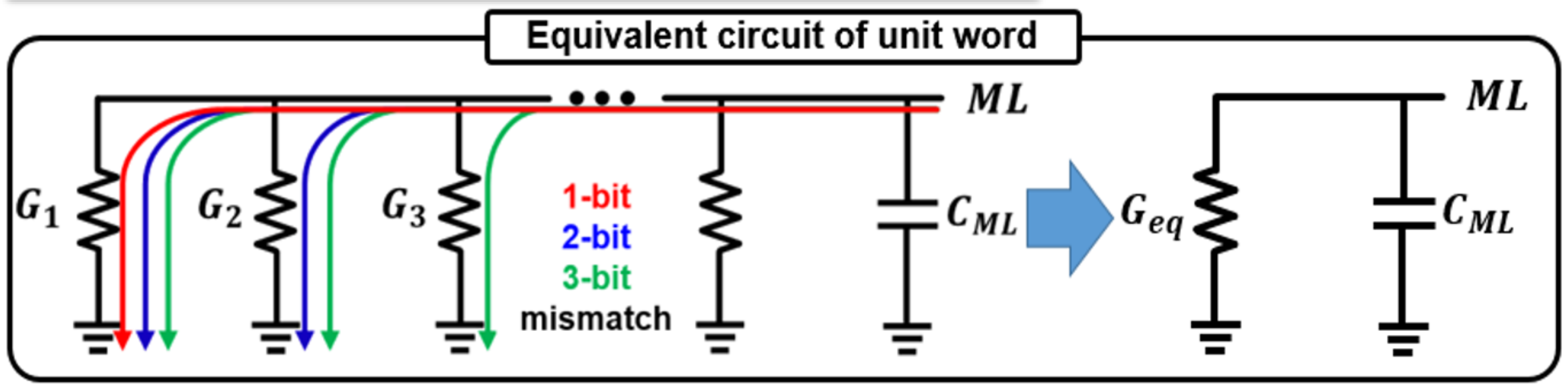


<J.S.Lee and W.Y.Choi, *IEEE TED*, 2021>

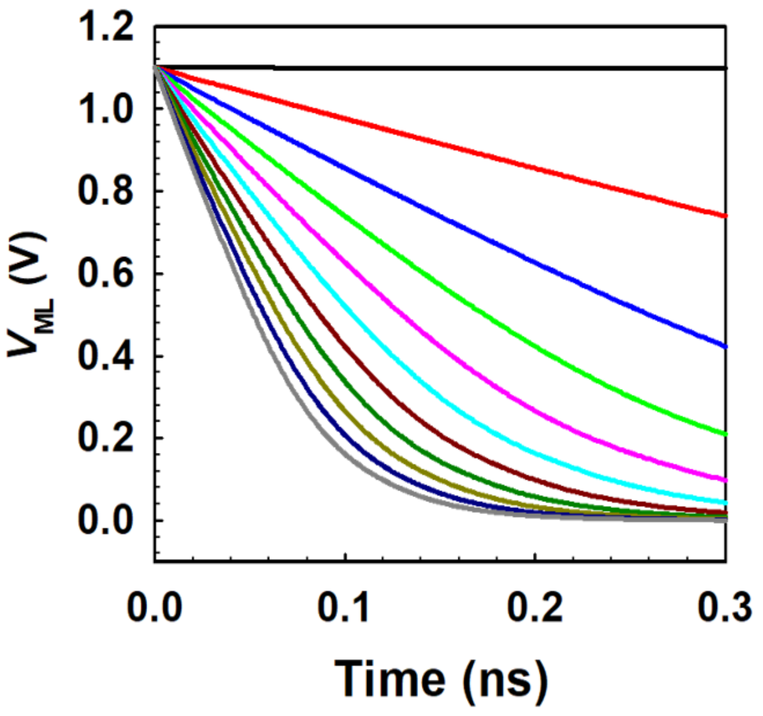
- NEMTCAM has been considered as a promising option of nonvolatile TCAM designs.
- NEMTCAM achieves the smallest cell area, highest speed, energy efficiency among other TCAM design.
- In this study, NEMTCAM was introduced as a novel NN classifier in memory-augmented neural network.

Simulation Result

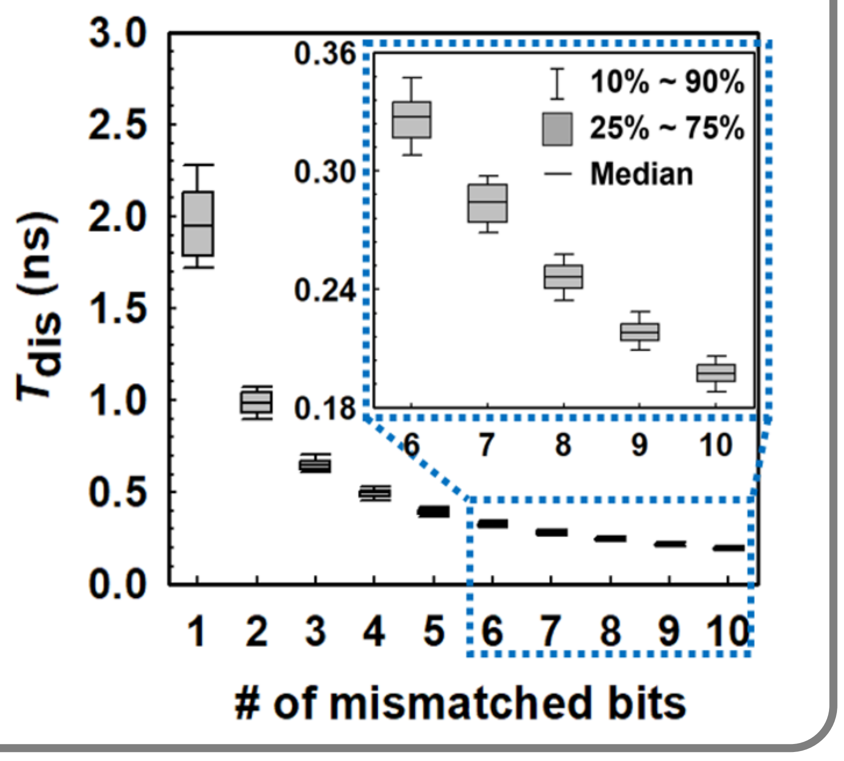
Equivalent circuit of unit word



Simulated ML voltage increasing with the number of mismatched bits

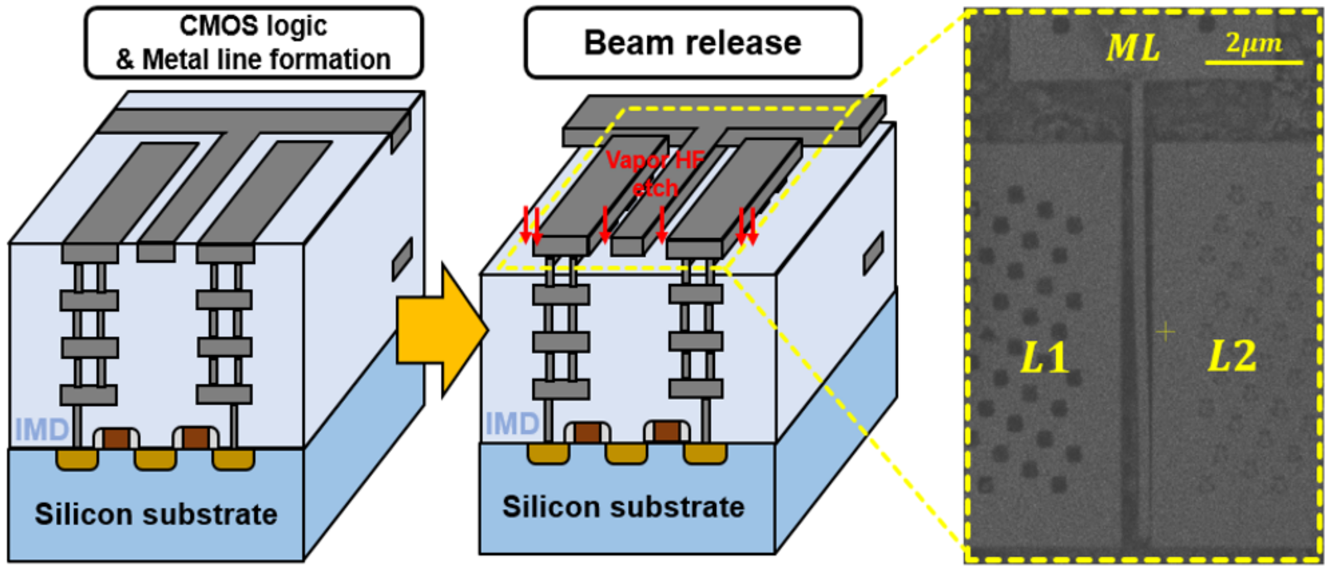


Calculated discharge delay (T_{dis}) with the effect of device-to-device variations

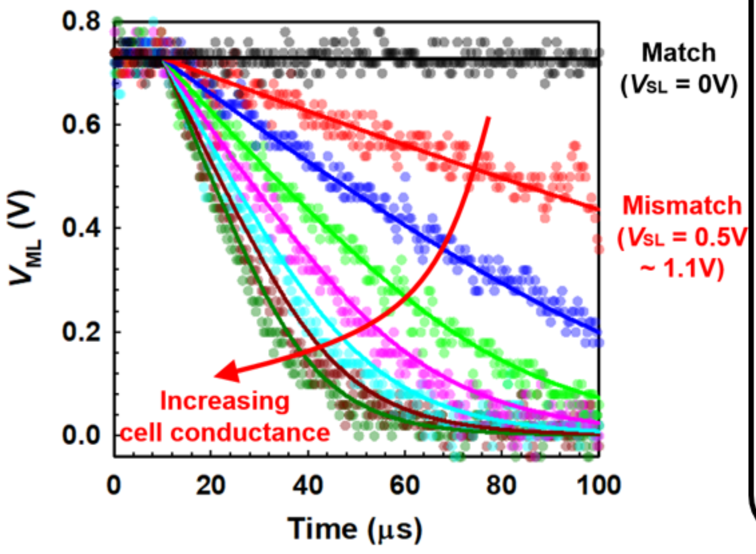


Experimental Result

NEMTCAM fabrication process



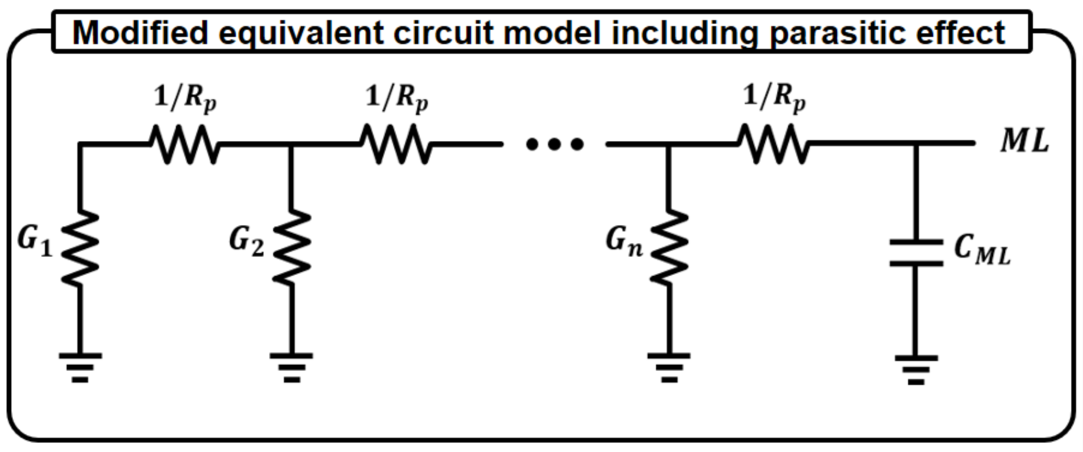
Measurement result

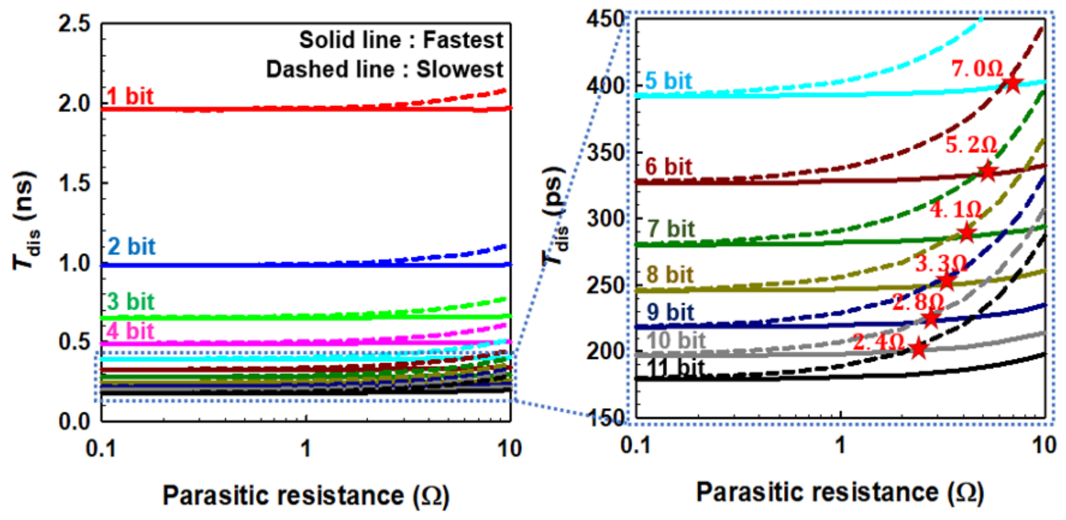


- NEM memory switches are patterned using a dual damascene process.
- IMD layer was selectively removed using vapor HF and fluorine-based plasma etch process.
- It was observed that a higher conductance led to a faster ML voltage drop.

Parasitic Resistance Effect

Modified equivalent circuit model including parasitic effect





Analytic model analysis

- Total conductance (G_{eq}) can be calculated as follows:

$$G_{\mathrm{eq}} \approx \frac{\sum_{i=1}^{n} G_i}{1+\sum_{i=1}^{n} iG_iR_p} \quad \left(R_pG_i \ll 1\right).$$

- G_{eq}'s when T_{dis} can is the smallest and the largest are derived as follows:

$$G_{eq,fastest}(k) = \frac{2kG_{miss}}{2+k(k+1)R_pG_{miss}},$$

$$G_{eq,slowest}(k) = \frac{2kG_{miss}}{2+k(2n-k+1)R_pG_{miss}}.$$

- To avoid the overlap between *k*-bit and (*k+1*)-bit mismatched case, the following condition should be satisfied:

$$G_{eq,fastest}(k) < G_{eq,slowest}(k+1).$$

- As a result, the condition of R_p is derived as follows:

$$R_p < \frac{2R_{miss}}{k(k+1)(2n-2k-1)}.$$

Analysis

- NEMTCAM using the 65-nm node can discriminate up to 10 Hamming distance in a 32-bit word, which is applicable to generic NN search.
- For higher accuracy, when a larger bit-width is needed, G_{miss} can be decreased.

Conclusion

- An in-memory NN search operation using the NEMTCAM was successfully confirmed by both simulations and experiments.
- This will enable next-generation CAM architectures to transcend the existing neural network system.

서강대학교
SOGANG UNIVERSITY

2021퀄컴논문대전 최우수논문