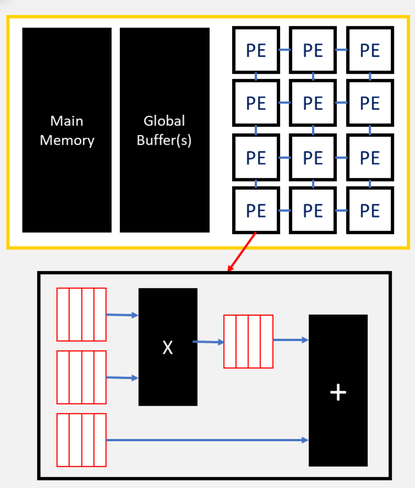
Mixed-Signal In-Memory Multi-bit Matrix-Vector Multiplication

Kévin Hérissé, Benoit Larras, Bruno Stefanelli, Antoine Frappé, Andreas Kaiser

Mixed-Signal In-Memory Multi-bit Matrix-Vector Multiplication

Kévin Hérissé, Benoit Larras, Bruno Stefanelli, Antoine Frappé, Andreas Kaiser
Univ. Lille, CNRS, Centrale Lille, Junia, Univ. Polytechnique Hauts-de-France, UMR 8520 - IEMN, Lille, France
kevin.herisse@junia.com

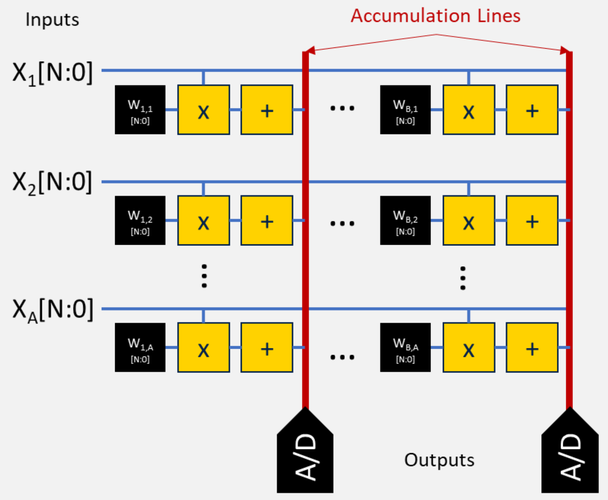
Memory Wall



Matrix-Vector Multiplication (MVM) is used in all State-of-the-Art (SoA) neural network implementations. MVM performs Multiplication and Accumulation (MAC) of input and weight vectors.
Single MAC operation in a Von Neumann architecture [Murmann ESSCIRC 2020] :

- Data is fetched 4 times
- Register access cost = 50fJ/byte
- Total access cost = 200fJ/byte
- **System maximum efficiency = 10TOPS/W**

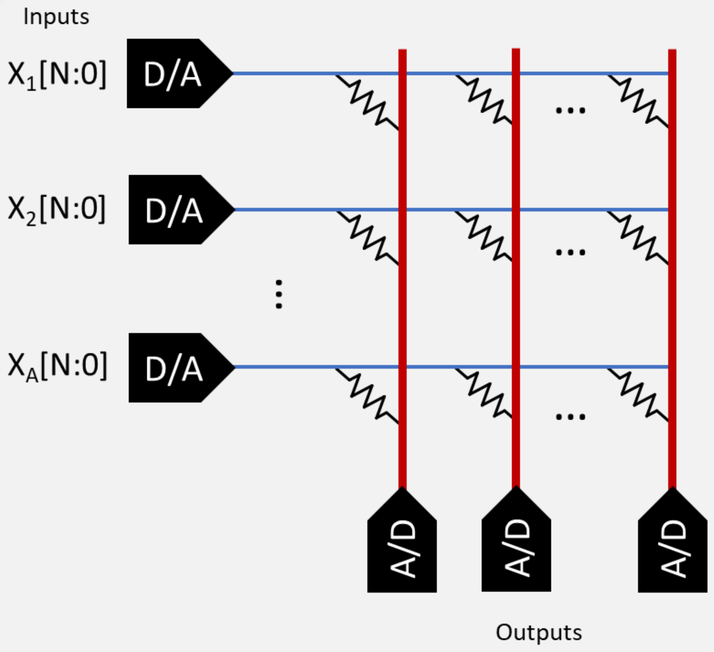
In-Memory Computing



- The processing elements are integrated into the memory.
- The energy efficiency upper limit is set by the ADC.
- The energy cost of the ADC is shared with multiple MACs across an Accumulation Line (AL).
- **ADC-limited efficiency of 200TOPS/W (arrays > 100 MACs).**

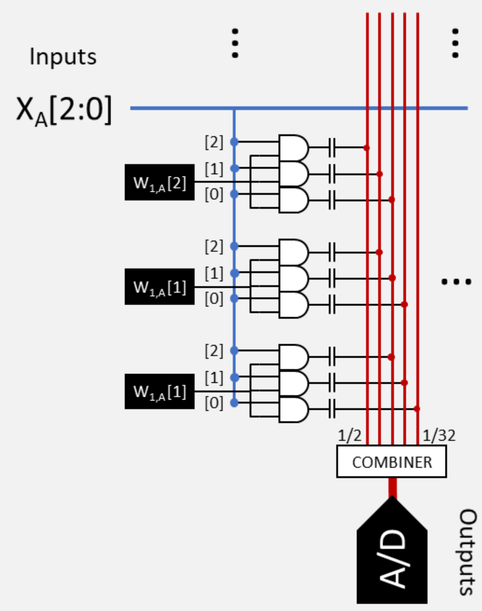
Comparison of computation methods for multi-bit MVM

ReRAM



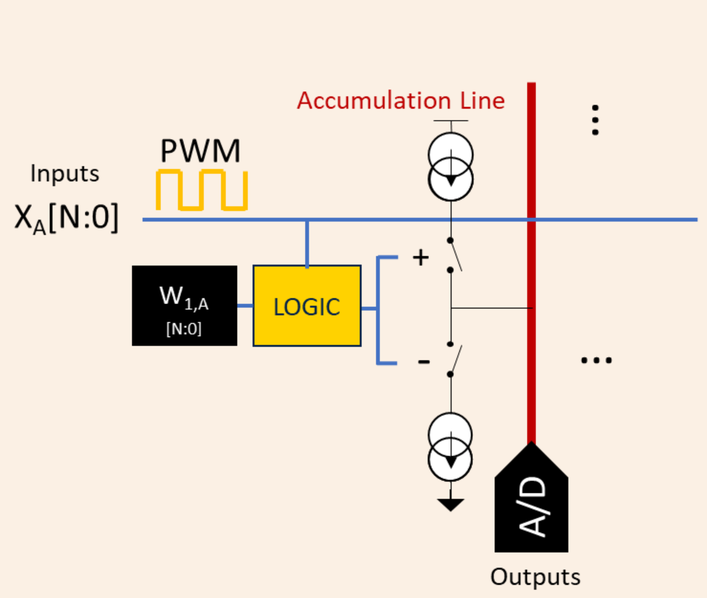
- Multi-bit operations up to 16 states of conductance (4 bits) [Hsieh EDL 2021]
- Process variations prevent reaching a higher number of bits.
- Writing necessitates high current spikes (around 6µA) [Dao & Koch ICEIC 2020]
- Inference cost of approximately 250fJ/MAC [Murmann TVLSI 2021]
- A trade-off exists between energy consumption and variability.

Switched Capacitor



- Needs additional digital circuitry to combine all the line's results. Only adequate for a reduced number of lines (<5) but dominates the power consumption for higher numbers of bits [Murmann TVLSI 2021]

Switched Current Source



- Take advantage of the time to charge the capacitor with ultra-small currents (thanks to recent CMOS technology)
- **Envisioned core energy consumption of 0.5fJ/MAC – 4000TOPS/W** (dominated by the ADC 10fJ/MAC – 200 TOPS/W for an 8-bit SoA ADC)
- PWM period and the current level can be tuned to address many applications.

Conclusion

Switched current sources offer great opportunities for configurable, low-energy multi-bit MVM. Enabled by ultra-low currents in advanced CMOS technologies, this contribution proposes to trade off time to reduce energy, being specifically recommended for applications with throughput lower than 50GOPS, such as Key Word Spotting (KWS), Voice Activity Detection (VAD), or sound recognition. It allows reaching the theoretical ADC-limited consumption of 200TOPS/W for applications with low throughput requirements.

This work was supported in part by the French National Research Agency under Grant ANR-18-CE24-0006-01 LEOPAR

	Memristor	Switched Capacitor	Switched Current Source
Configurability	✗	✗	✓
Technology	Non-standard (CMOS + RRAM)	Standard CMOS	**Standard CMOS**
Typical energy efficiency	21.9 TOPS/W [Xue ISSCC 2019]	10 TOPS/W [Murmann TVLSI 2021]	**200 TOPS/W**
Max number of bits	3 – 4	8	**8**