ELSA: Enabling SOT-MTJ Crossbars for ML using Sparsity-Aware Device-Circuit Co-design

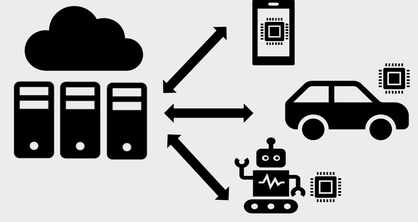
PURDUE UNIVERSITY

Elmore Family School of Electrical and Computer Engineering

Tanvi Sharma*, Cheng Wang*, Amogh Agrawal and Kaushik Roy

*equal contributors

Motivation: Enable AI at edge

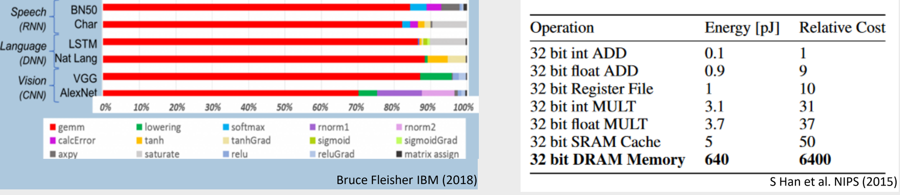


- High performance
- High latency/energy to transfer data

- Data available at edge
- Low performance
- Limited energy/area budget

Trade-offs between edge and cloud computing.
Need to enable AI in resource constrained platforms.

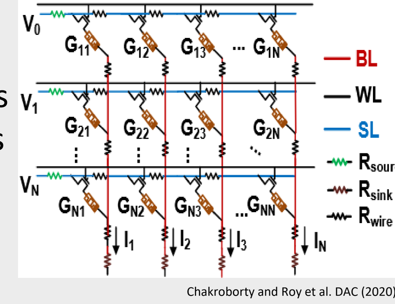
ML workloads dominated by matrix-vector multiplications. → In-memory computing provides a promising solution



In-Memory Computing: Challenges

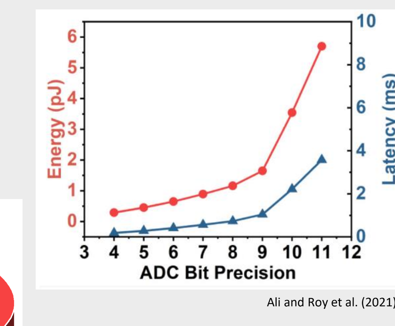
Functionality: non-idealities (NI)

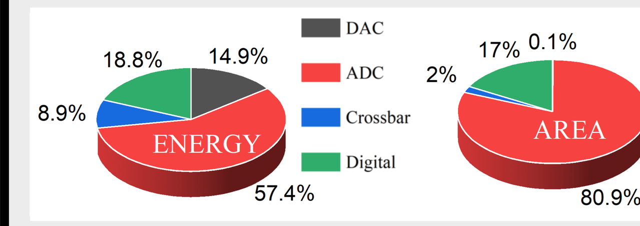
- Parasitic IR (voltage) Drops
- Non-linearity in NVM I-V characteristics
- Leakage current from access transistors
- Noises from device variation and ADC quantization



Performance: high cost of analog-digital conversion (ADC)

- Dominant in energy, latency, and area
- Increases exponentially with bit precision





Device Consideration

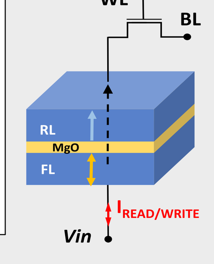
Characteristics	ReRAM	PCM	STT-MRAM
Storage	Multi-bit	Multi-bit	Single-bit
Endurance	Low	Low	High
Write energy	Medium	High	Low
Conductance Drift	High	High	Low

ReRAM: resistive RAM, PCM: phase change memory

MRAM has the potential to provide a more robust solution.

STT-MRAM
- High Endurance (>10^15)
- Binary storage
- Shared read-write path
- Low On/Off distinguishability
- High density (1T-1R, ~2-4x SRAM)
- Production Ready

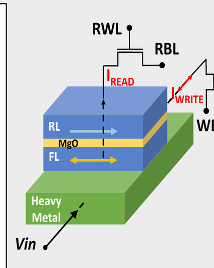
STT = spin transfer torque



Challenges with STT
- Small R_{ON}
- Limited room to vary R_{ON} due to writing constraint
- Difficult to sense with small ON/OFF

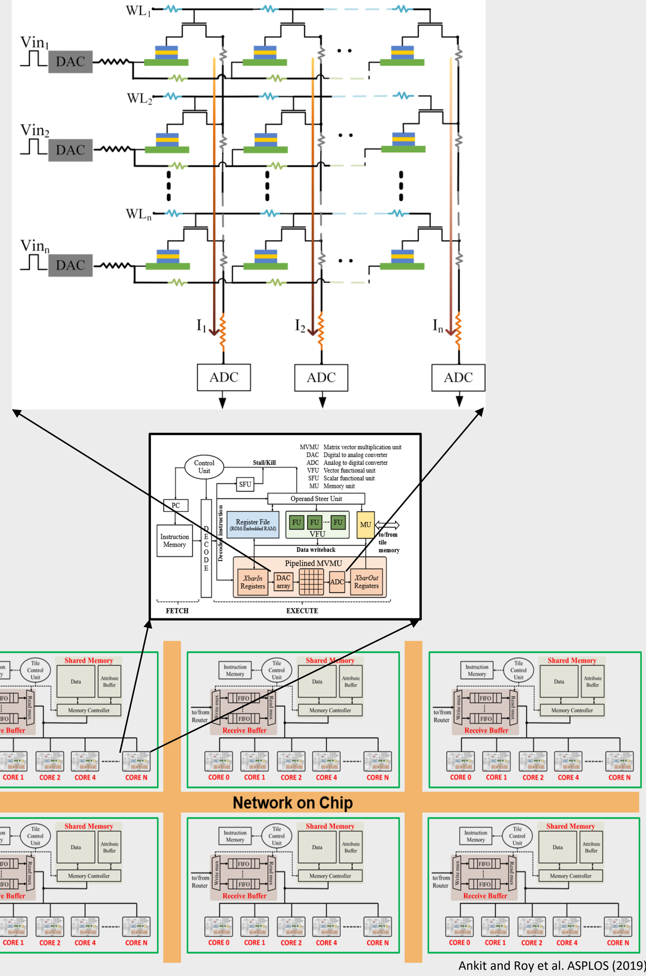
SOT-MRAM
- Higher endurance
- Binary storage
- Decoupled read-write path
- Improved On/Off possible (with device optimization)
- Large area (2T-1R)
- In development

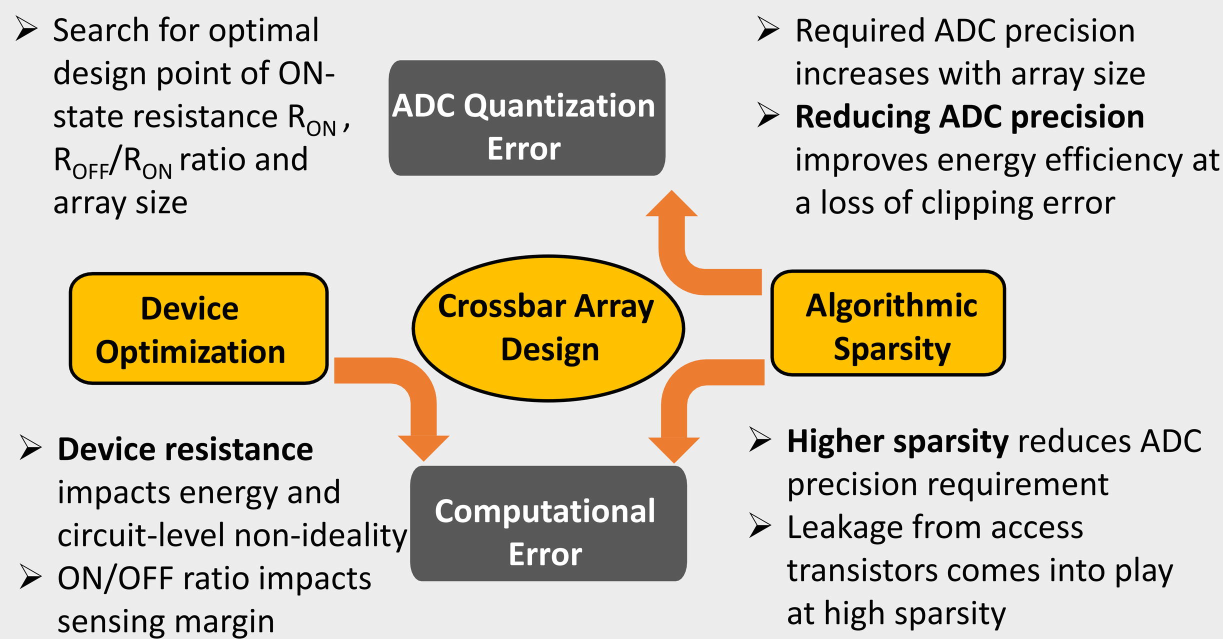
SOT = spin orbit torque



Opportunities with SOT
- Flexible design with R/W separated
- Device can be adjusted to reduce crossbar non-ideality

ELSA: Approach and Co-Design Strategy





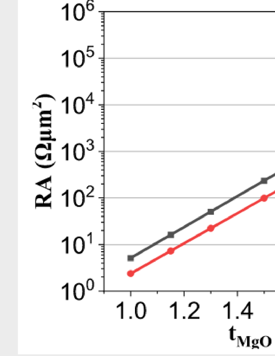
- Search for optimal design point of ON-state resistance R_{ON}, R_{OFF}/R_{ON} ratio and array size
- Device resistance impacts energy and circuit-level non-ideality
- ON/OFF ratio impacts sensing margin
- Required ADC precision increases with array size
- Reducing ADC precision improves energy efficiency at a loss of clipping error
- Higher sparsity reduces ADC precision requirement
- Leakage from access transistors comes into play at high sparsity

Approach involved: ***device/circuit simulation*** (array-level error), ***functional simulation*** (ML inference error) and ***performance modeling*** (energy/area)

SOT-MTJ Device Design

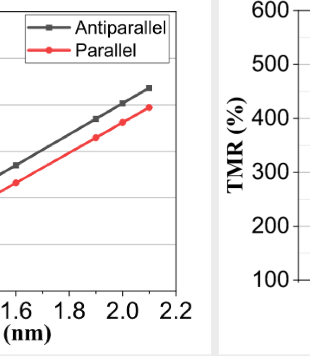
Tuning R_{ON} by varying tunnel barrier thickness (t_{MgO})

Varying t_{MgO} within the tunneling transport regime $lnR \propto t_{MgO}$



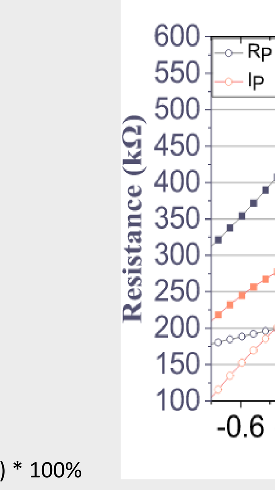
Enhancing R_{OFF}/R_{ON} by incorporating high temperature annealing (T_{anneal})

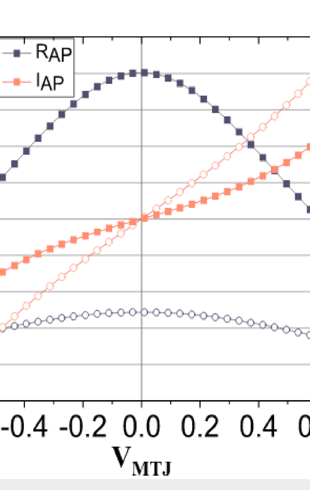
Calibrated with experimental observation in CoFeB/MgO based devices



Modified device magnetic configuration to ensure high R_{OFF}/R_{ON} and high R_{ON}

In-plane anisotropy with thicker magnetic layer CoFeB (to accommodate high-T annealing)





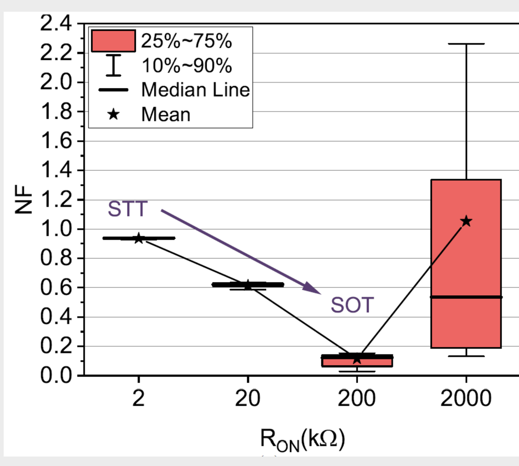
Non-linear device current-voltage (I-V) characteristics included in crossbar-level SPICE simulation

Tunneling Magnetoresistance (TMR): $(R_{OFF}/R_{ON} - 1) * 100\%$

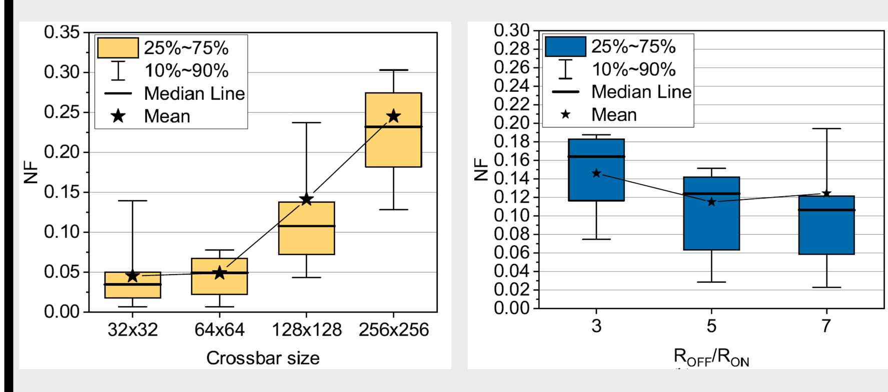
Circuit-Level Exploration

Functional error evaluated using Non-ideality factor:

$$NF = \left|\frac{I_{ideal} - I_{real}}{I_{ideal}}\right|$$

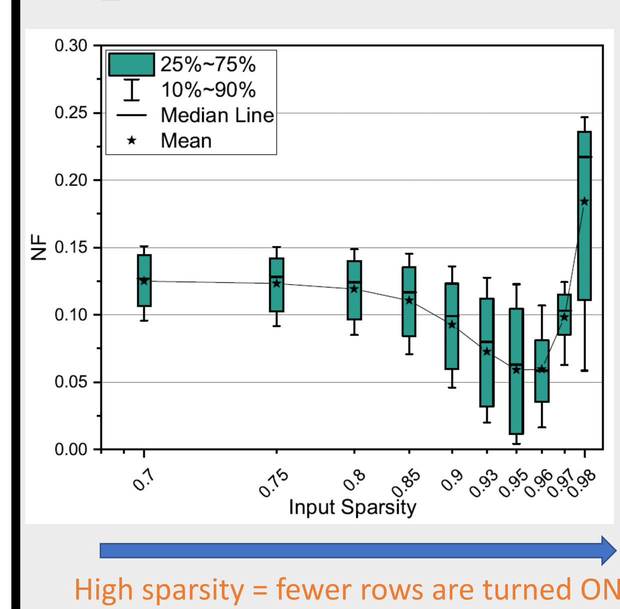


- IR drops reduce at high R_{ON}
- Large error observed under MegaOhm R_{ON} due to leakage current from access transistors.
- R_{OFF}/R_{ON} has comparatively lower impact on NF
- Larger crossbar sizes are error-prone



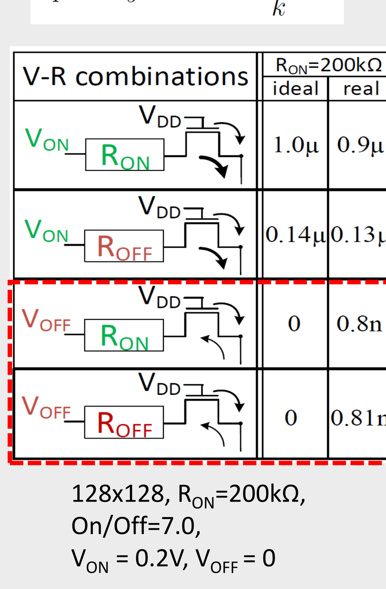
Impact of Workload Sparsity

$$Sparsity = 1 - \frac{\sum_{i=0}^{i=k} M_i}{k}$$

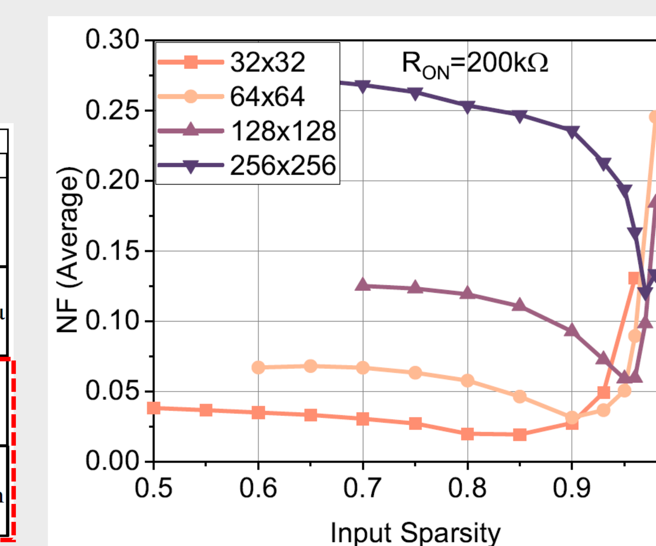


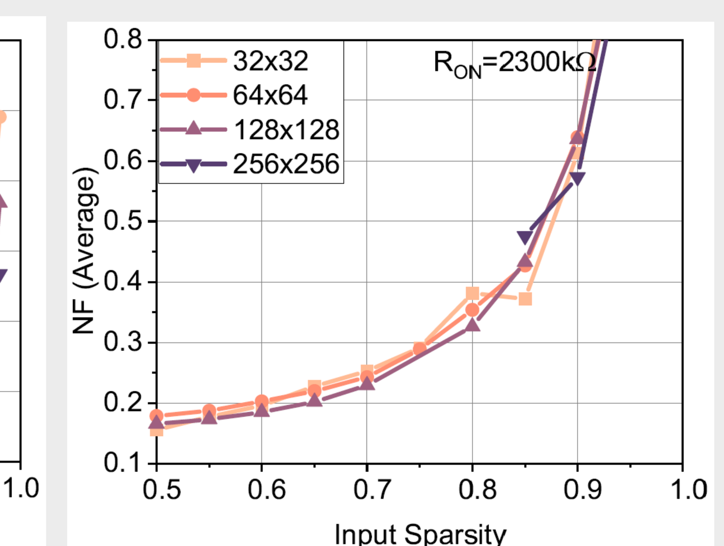
High sparsity = fewer rows are turned ON

- Sparsity increases -> total bit line current decreases -> IR drop effect reduces
- At high sparsity, leakage current accumulates and dominates the total current.



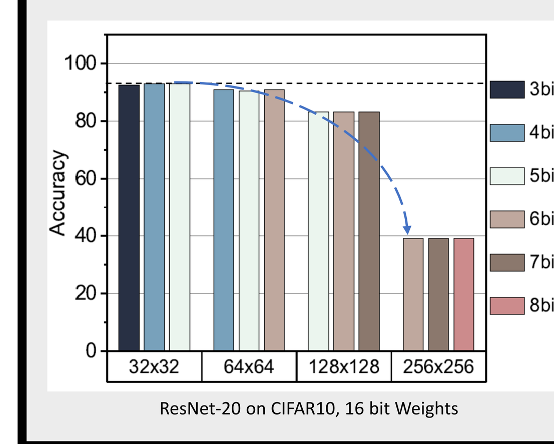
128x128, R_{ON}=200kΩ, On/Off=7.0, V_{ON} = 0.2V, V_{OFF} = 0



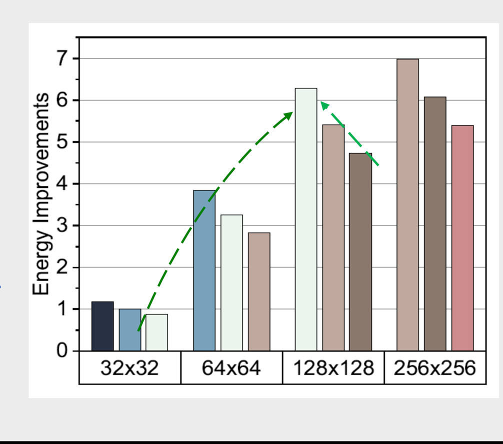


- Tradeoff in NF trends for sparsity ↓ vs crossbar size ↑
- Strong dependence on array size and device resistance
- Very high R_{ON} (~MΩ) leads to significant deterioration
- Sparsity helps to lower NF only if optimal R_{ON}

System Level Results and Key Learnings



Leveraging sparsity could help in achieving the optimal accuracy-energy tradeoff by lowering the ADC precision.



- SOT-MRAM can be better optimized for crossbars (compared to STT-MRAM)
- Higher R_{ON} is desirable to overcome the IR drops impact, but limited by the sensing and technology constraints, especially the transistor gate leakage current
- Impact of non-idealities is highly data dependent and exacerbated at extremely high input sparsity.

Scan for full paper at ISLPED'21:

C-BRIC