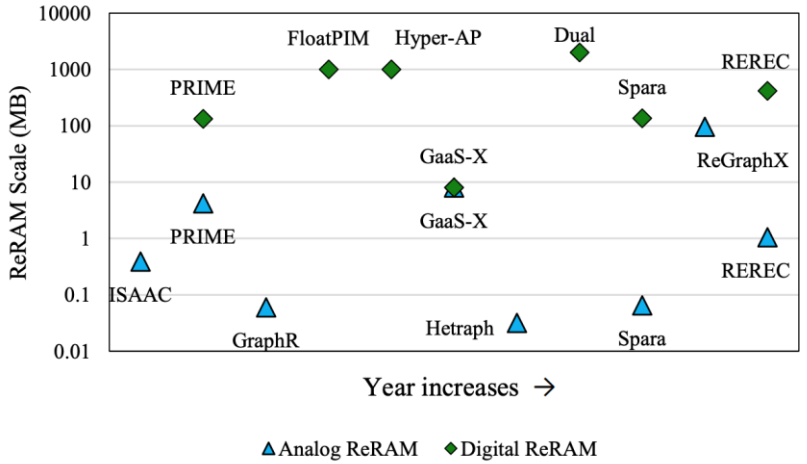
ARCHER: A ReRAM-based Accelerator for Compressed Recommendation Systems

Xinyan SHEN, Xiaofei LIAO, Long ZHENG, Yu HUANG, Dan CHEN, Hai JIN

Frontiers of Computer Science, DOI: 10.1007/s11704-023-3397-x

Problems & Ideas



- Memory-bound embedding lookup operation poses big challenges for processing recommendation systems on general computing architectures, which can be solved by ReRAM-based PIM.
- Existing monolithic ReRAM chips faces challenge while processing practical recommendation models due to chip size limits (several GBs ReRAM vs tens of GBs model, as shown in Figure 1).
- The simple lookup and pooling operations are not suitable for ReRAM crossbar.
- Therefore, we process the compressed (decomposed) the model on the chip by leveraging the tensor train decomposition, which can solve those two problems at the same time.

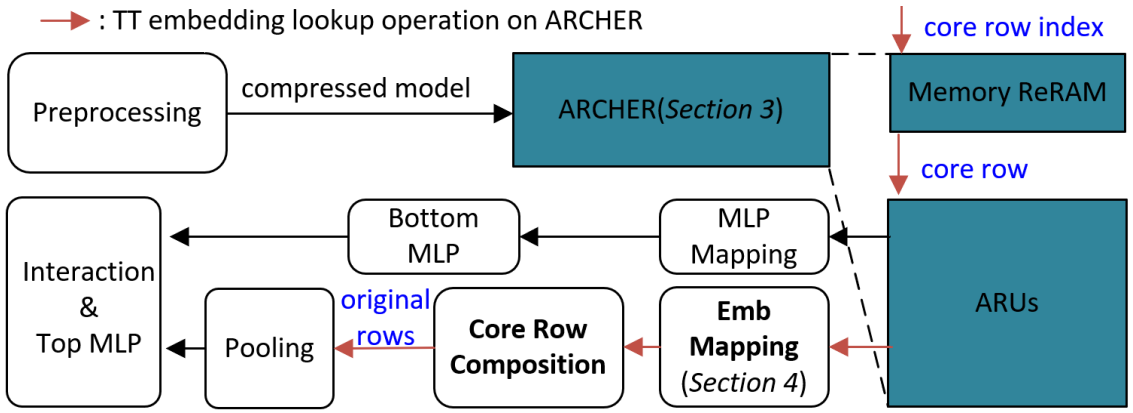
Ideas

➢Unified-MACs-for-all-layer computing schema.

- The entire decompression process comprises multiple vector-vector multiplications that operate on 32-width vectors, effectively implementing a series of MAC operations.
- Computing parallelism offered by ReRAM can offset the computational overhead associated with decompression.

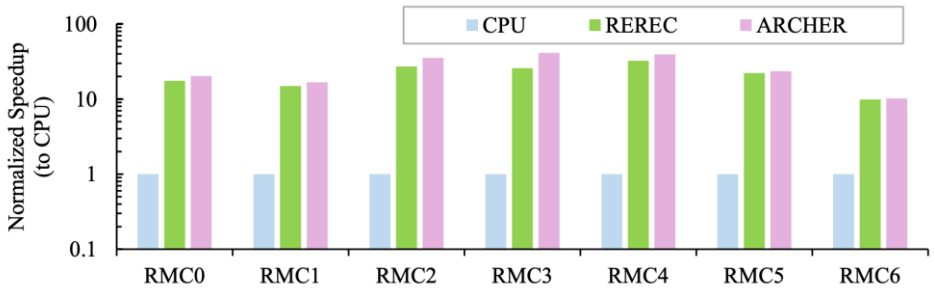
➢Hierarchical mapping schema.

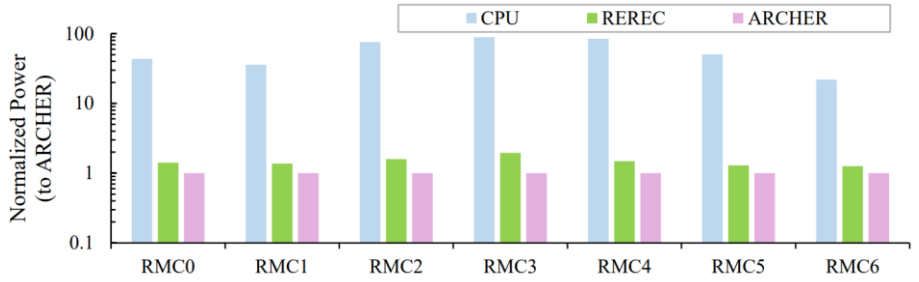
- Layer-wise Model Mapping: Hardware resource assignments among MLP, embedding and interaction layers.
- Index-based Core Mapping: Hardware resource assignment among tables of the same model and cores of the same table.
- Pattern-based Row Mapping: Intra-core assignments for each core row.



Main Contributions

- An in-depth investigation of existing ReRAM-based architectures and identify the mismatching gap between the recommendation model scale and limited crossbar architecture resources.
- Leveraging the tensor train decomposition technique to fill the gap by revising the properties of the decomposed recommendation model and breaking down the process of the composition into MAC operations.
- A novel model mapping scheme and efficient full-stage pipeline for the decomposed recommendation system to maximize hardware efficiency.





- ARCHER can support large practical recommendation model on monolithic ReRAM chip.
- Compared with an earlier ReRAM-based recommendation accelerator REREC, ARCHER achieves the average performance improvements of 1.21 ×, and the average energy savings of 1.71 ×. Despite extra composition process, ARCHER outperforms REREC due to the hierarchical mapping schema.