## RANA: Towards Efficient Neural Acceleration with Refresh-Optimized Embedded DRAM

### Corresponding author

Fengbin Tu, Tsinghua University

Shaojun Wei, Tsinghua University

### Keywords

Neural Network; Embedded DRAM (eDRAM); Refresh Optimization; Retention Time

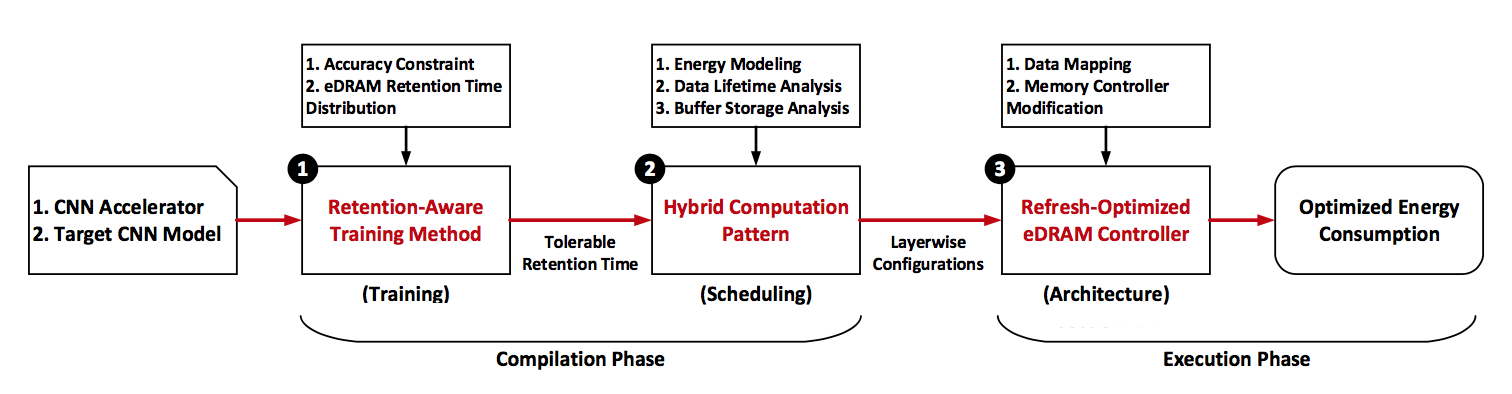
### Summary

#### Challenge

The growing size of convolutional neural networks (CNNs) requires large amounts of on-chip storage. In many CNN accelerators, their limited on-chip memory capacity causes massive off-chip memory access and leads to very high system energy consumption. Embedded DRAM (eDRAM), with higher density than SRAM, can be used to improve on-chip buffer capacity and reduce off-chip access. However, eDRAM requires periodic refresh to maintain data retention, which costs much energy consumption.

#### Contribution

A retention-aware neural acceleration (RANA) framework has been designed, which strengthens DNN accelerators with refresh-optimized eDRAM to save total system energy. RANA includes three techniques from the training, scheduling, architecture levels respectively as shown in the following figure.



* **Training Level**: A retention-aware training method is proposed to improve eDRAM's tolerable retention time with no accuracy loss. Bit-level retention errors are injected during training, so the network' s tolerance to retention failures is improved. A higher tolerable failure rate leads to longer tolerable retention time, so more refresh can be removed.
* **Scheduling Level**: A system energy consumption model is built in consideration of computing energy, on-chip buffer access energy, refresh energy and off-chip memory access energy. RANA schedules networks in a hybrid computation pattern based on this model. Each layer is assigned with the computation pattern that costs the lowest energy.
* **Architecture Level**: RANA independently disables refresh to eDRAM banks based on their storing data's lifetime, saving more refresh energy. A programmable eDRAM controller is proposed to enable the above fine-grained refresh controls.

#### Result

Owing to the RANA framework, 99.7% eDRAM refresh operations can be removed with negligible performance and accuracy loss. Compared with the conventional SRAM-based CNN accelerator, an eDRAM- based CNN accelerator strengthened by RANA can save 41.7% off-chip memory access and 66.2% system energy consumption, with the same area cost.