Extending Auto-scheduler to Support Performance Evaluation with Timing Model

Minchun Liao, NTHU

Yaohua Chen, ITRI

Chungta King, NTHU

ITRI
Industrial Technology
Research Institute

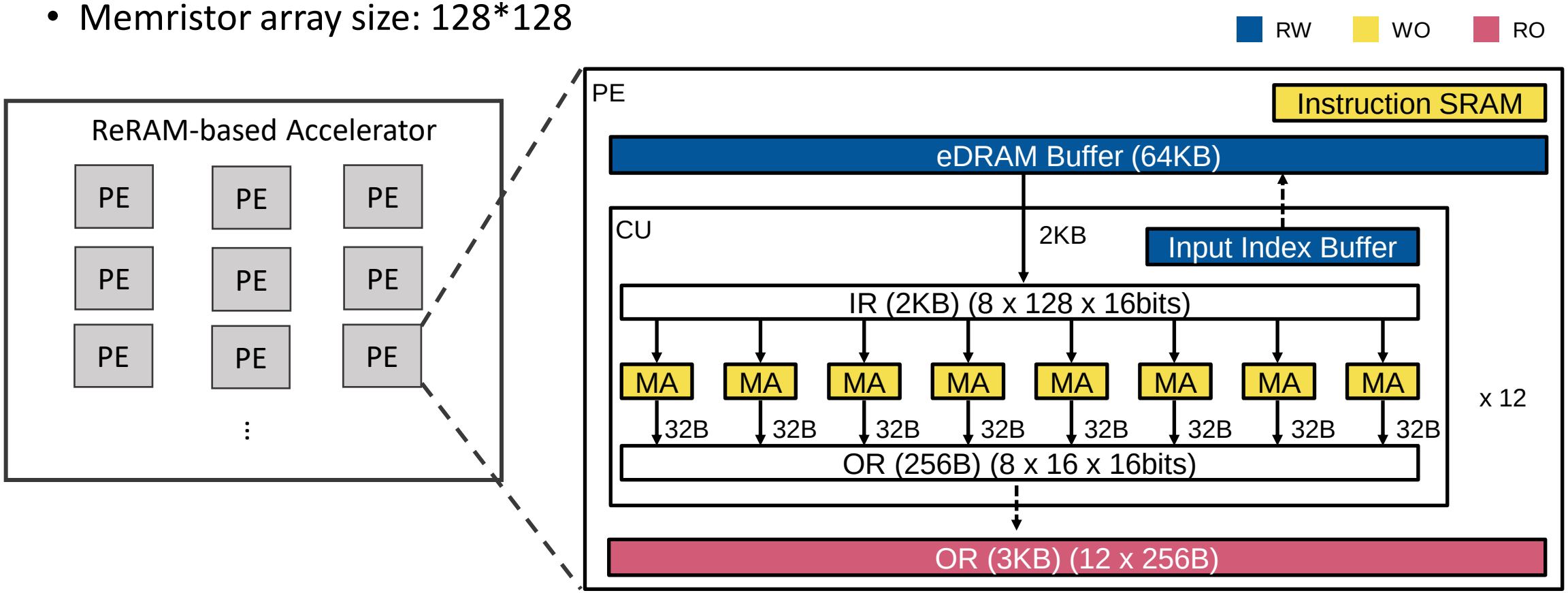
國立清華大學
NATIONAL TSING HUA UNIVERSITY

Motivation

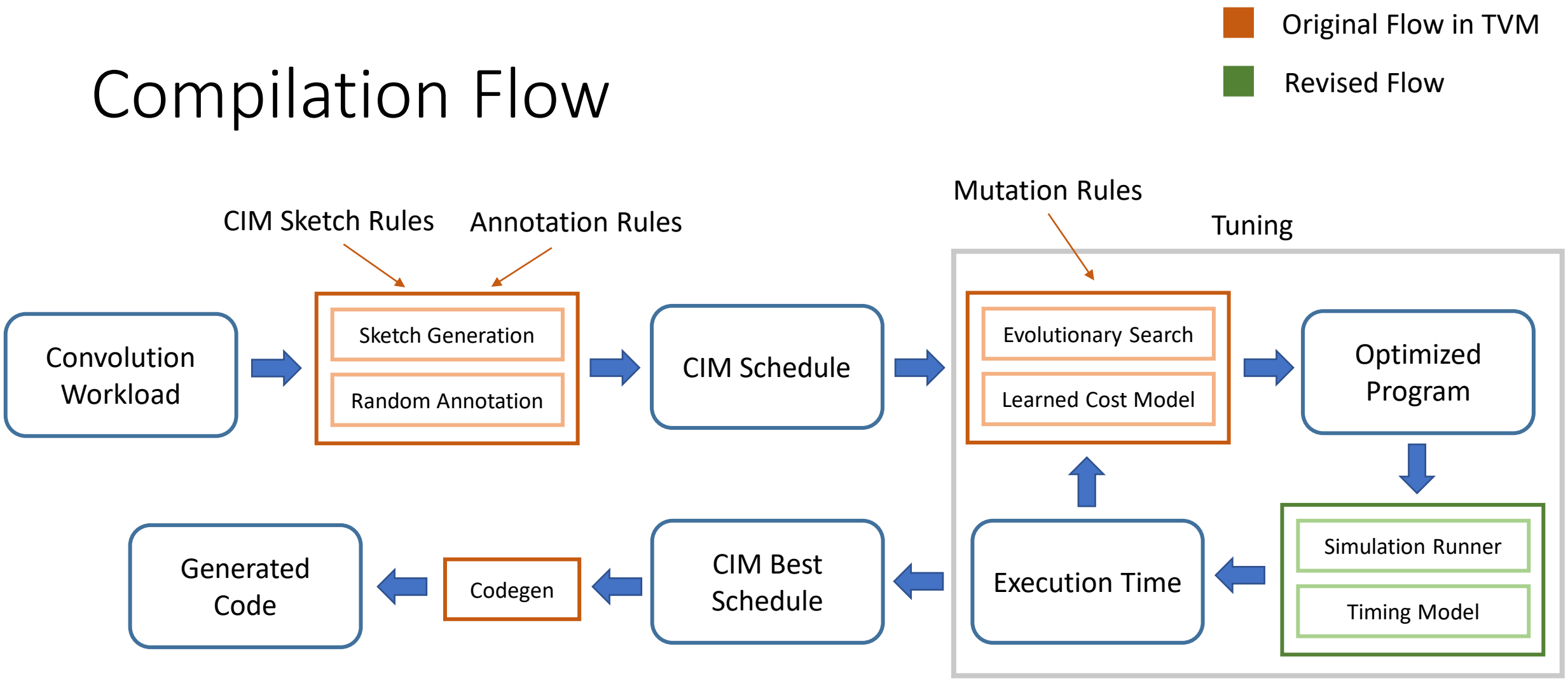
- Current TVM Auto-scheduler needs real devices or simulators to provide timing information for optimized code scheduling.
- If the device and/or its simulator are still in development, it is difficult to get the execution time.
- In the mean time, we still want the compiler to generate optimized code for the device in development for testing and design optimization.
- This work proposes the use of a timing model in TVM to approximate timing.
  - As a demonstration, we extend TVM Auto-scheduler to generate schedule for Computing-in-Memory (CIM) devices, like ReRAM-based accelerators.

CIM Behavior Model

- 168 PEs/chip, 12 CUs/PE, 8 Memristor arrays/CU
- Memristor array size: 128*128



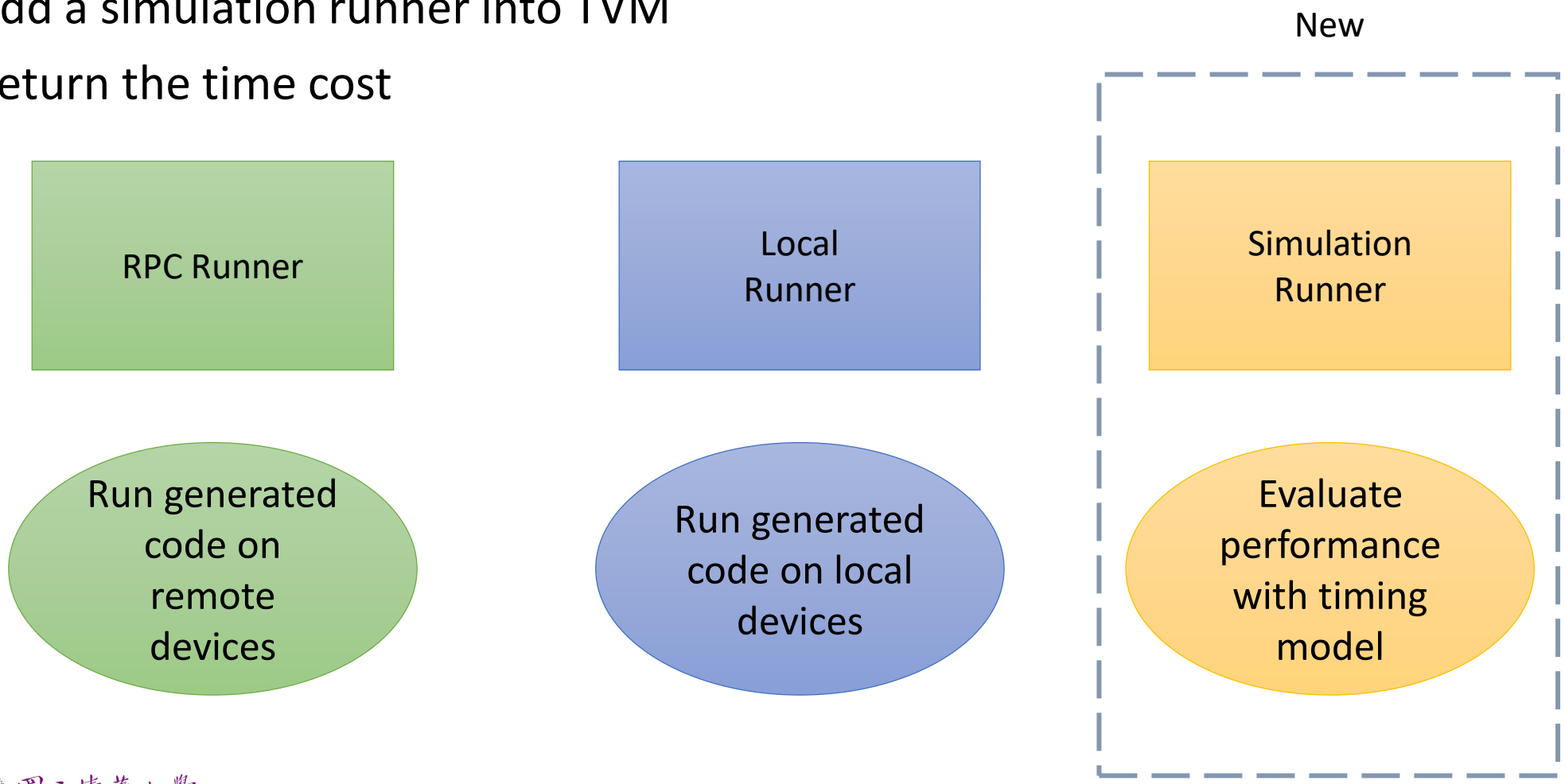
Compilation Flow



- Design sketch rules, annotation rules and mutation rules for CIM
- Leverage Auto-scheduler to generate schedule and perform schedule tuning

Runner

- Add a simulation runner into TVM
- Return the time cost



Timing model

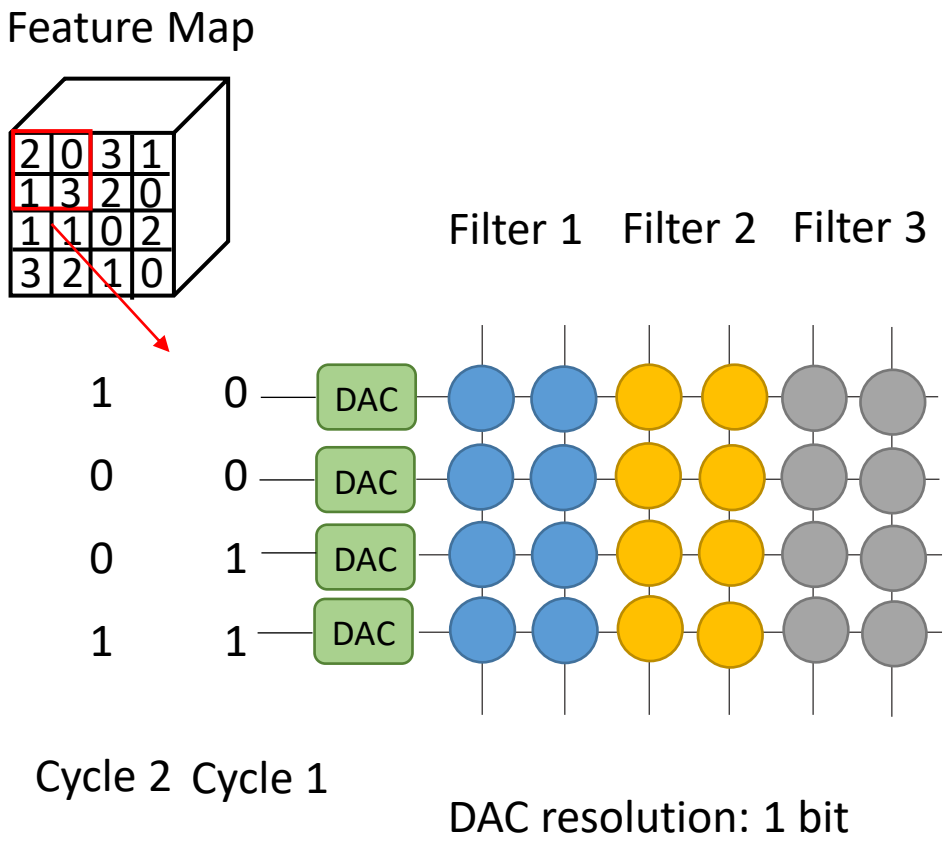
- Approximate the convolution execution time on CIM device

$$total_{cycle} = number_{batch} * cycle_{batch} + 6$$

$$cycle_{batch} = feature\ map\ resolution\ /\ DAC\ resolution$$

$$number_{batch} = \lceil \frac{\frac{kernel\_size^2 * IC}{\#uesd\_WL} * \frac{OC}{\#filters\_for\_MA}}{\#MA * \#CU} * W * H * N \rceil$$

- OC : Output channels
- IC : Input channels
- W : Width of input feature map over stride
- H : Height of input feature map over stride
- N : Number of input feature maps
- #used_WL : How many wordlines are used in a memristor array
- #filters_for_MA : How many filters are mapped in a memristor array
- #MA : Number of memristor arrays in one CU
- #CU : Number of computation units in one PE



Summary

- Use a timing model to work with TVM Auto-scheduler if the device or the simulator is not ready yet.
- Extend TVM Auto-scheduler to get the time cost through the timing model.
- During development of the device, the optimized code generated by TVM can be used to facilitate the development, e.g., for testing and design optimization.

Thank You!

Video Download Link

- https://drive.google.com/file/d/1BKlGeTSIATv1YJhxg2ND4_5Dv8wrJbTh/view?usp=sharing