Alibaba A.I. Labs
阿里巴巴人工智能实验室

AILabs & TVM

CONTENT

PART 1 ： ARM32 CPU

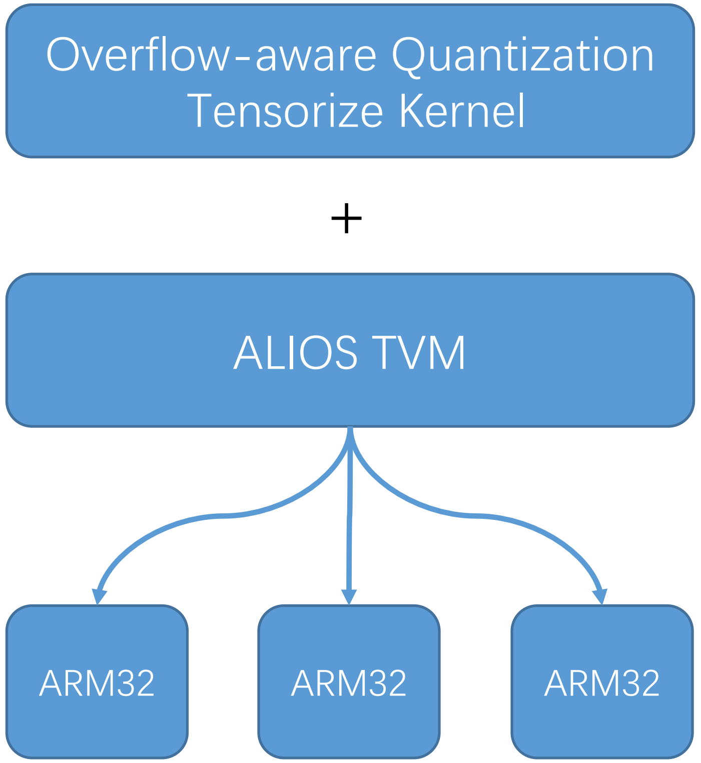
PART 2 ： HIFI4 DSP

PART 3 ： PowerVR GPU

Alibaba A.I. Labs
阿里巴巴人工智能实验室

ARM 32 CPU

Resolution



$$r = S(q - Z)$$

$$r_3^{(m,n)} = \sum_{i=1}^{k} r_1^{(m,i)} r_2^{(i,n)}$$

$$q_3^{(m,n)} = Z_3 + 2^{-n} * M_0 \left[\sum_{i=1}^{k} \left(q_1^{(m,i)} (q_2^{(i,n)} - Z_2)\right) + C\right]$$

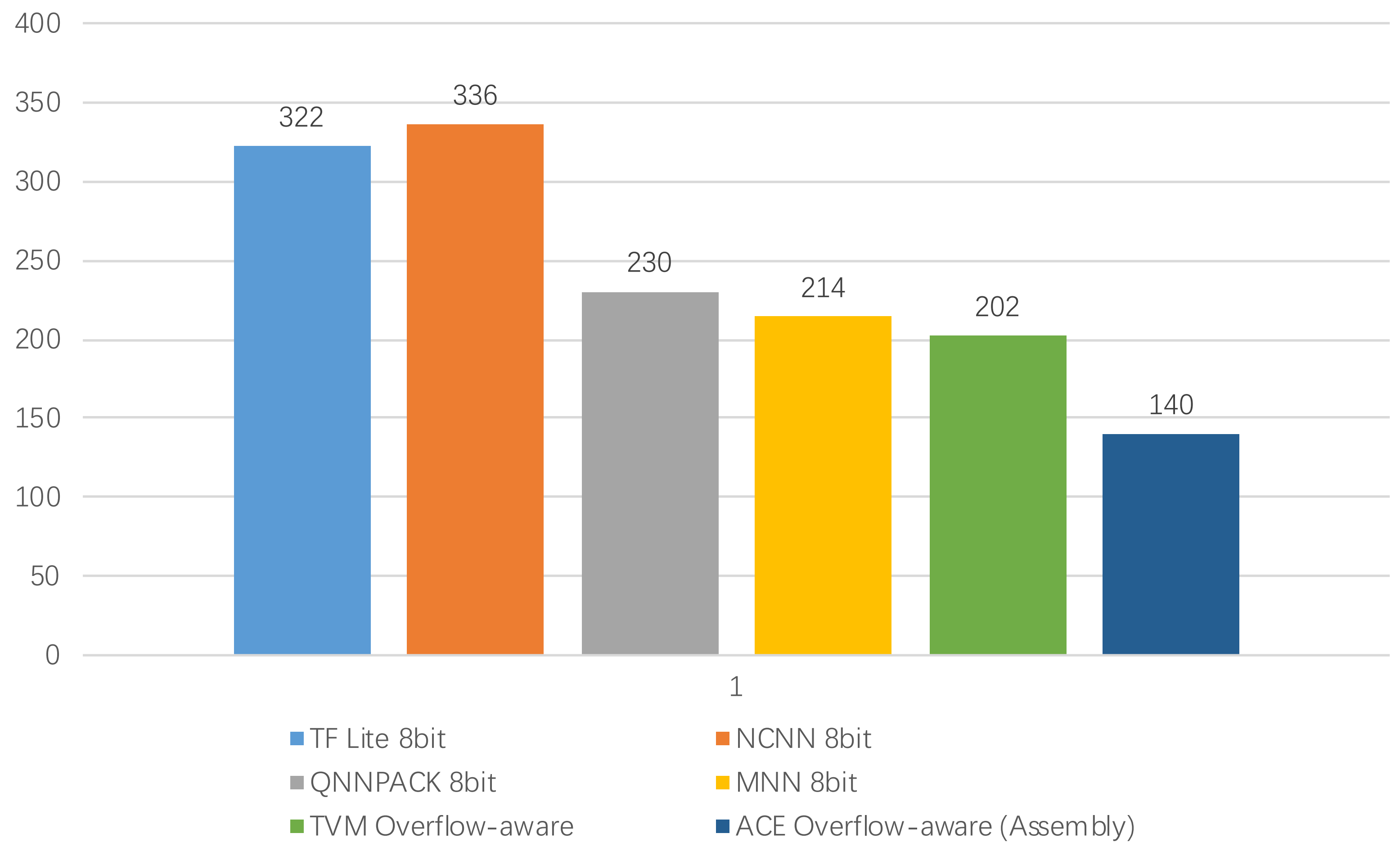
Current plan 1
int32 = int16 * int16

Current plan 2
int16_1 = int8 * int8
int16_2 = int8 * int8
int32 = int16_1 + int16_2

Overflow-aware
int16 = int8 * int8

Result

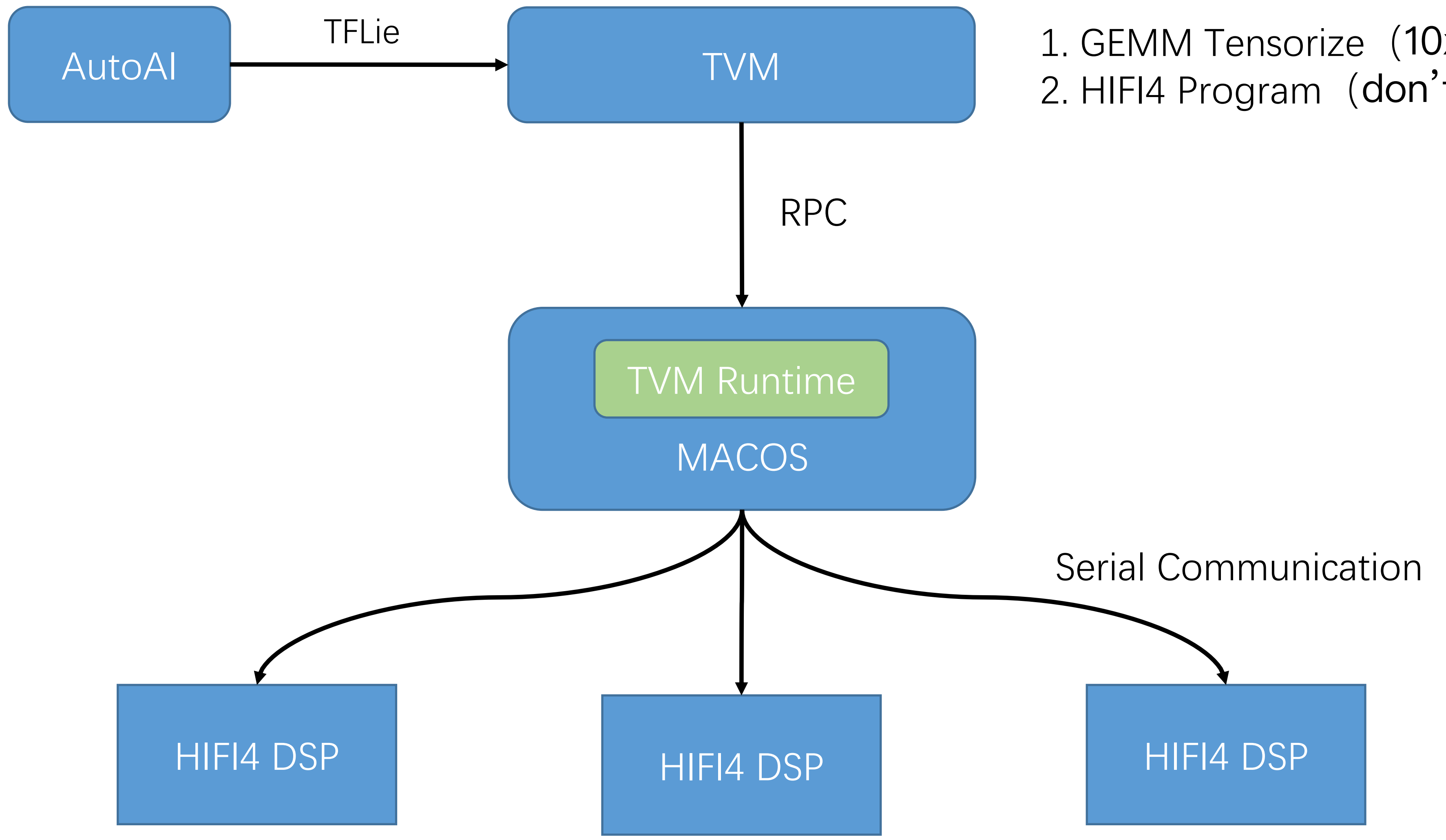
CPU：MTK8167S（ARM32 A35 1.5GHz）
Model：MobileNetV2_1.0_224



Alibaba A.I. Labs
阿里巴巴人工智能实验室

HIFI 4

Resolution



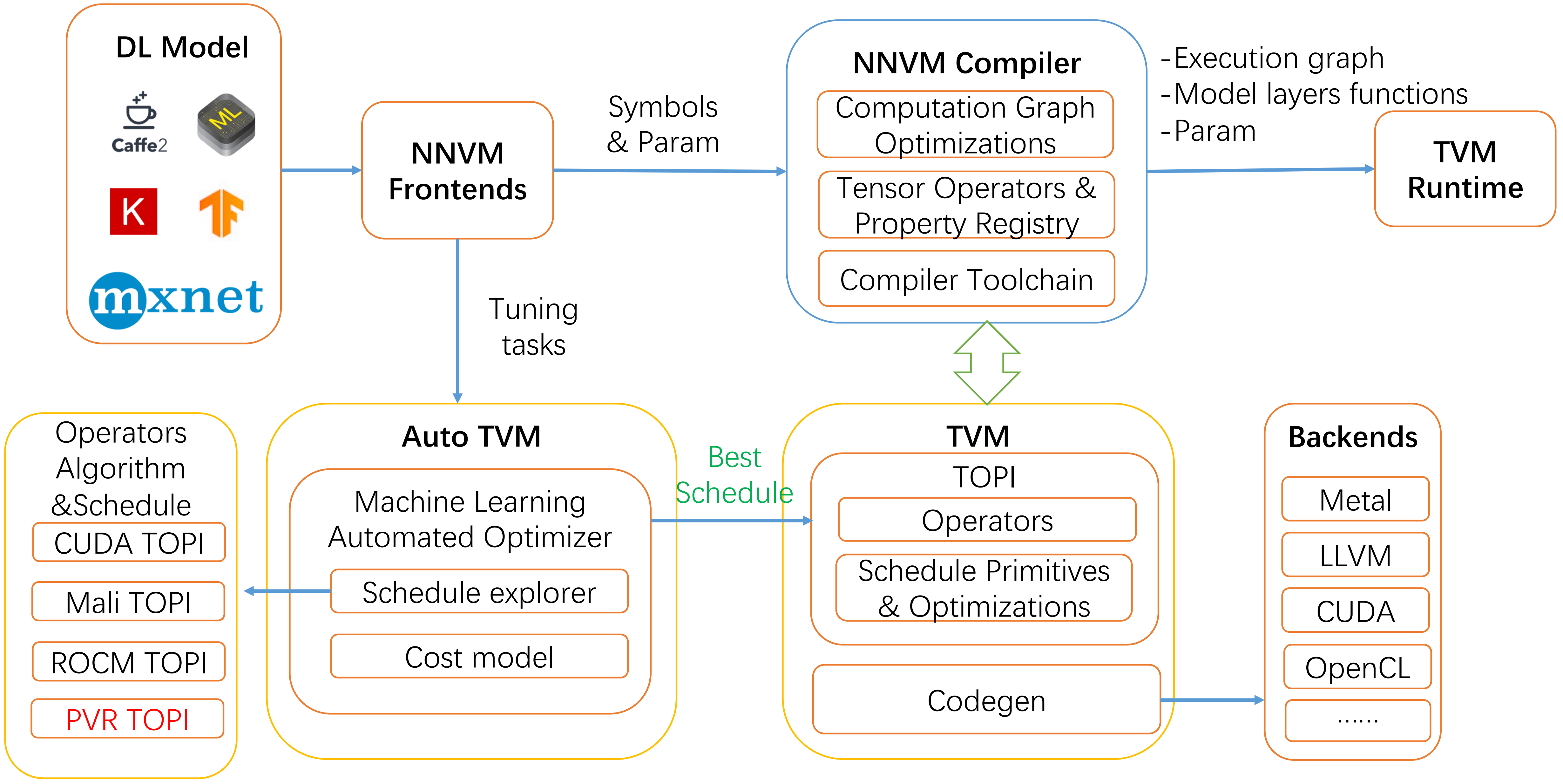
1. GEMM Tensorize（10x speed up）
2. HIFI4 Program（don’t need dlopen）

Alibaba A.I. Labs
阿里巴巴人工智能实验室

PowerVR GPU

PowerVR support by TVM

Alibaba A.I. Labs
阿里巴巴人工智能实验室



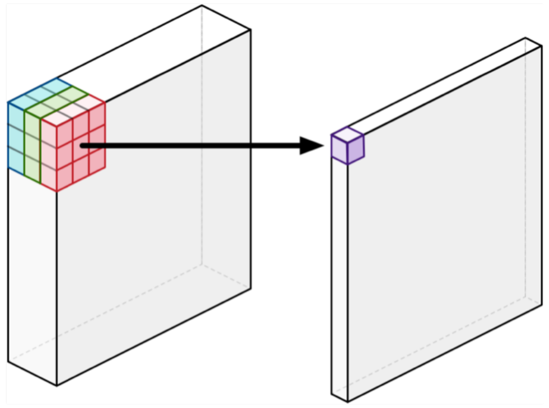
PVR TOPI

Alibaba A.I. Labs
阿里巴巴人工智能实验室

➢ TOPI for PVR, including what you want to compute and how to compute.

✓ What you want to compute

```
@autotvm.register_topi_compute(conv2d, 'pvr', ['direct'])
def conv2d_pvr(cfg, data, kernel, strides, padding, dilation, layout, out_dtype):
    #Describe algorithm with tensor expression language;
    #Return the out operation
```



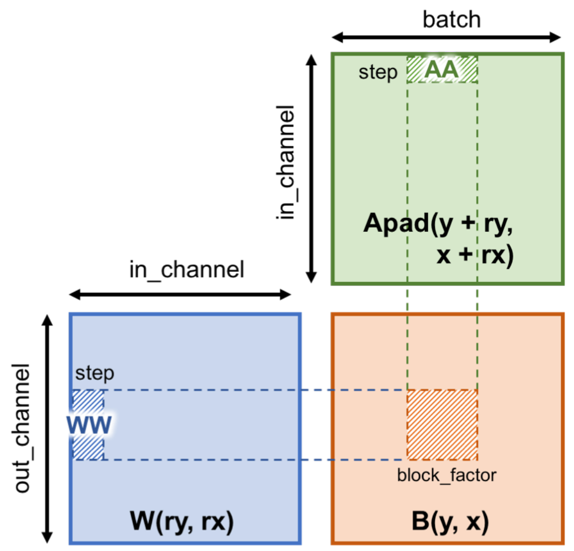
convolution

✓ How to compute.

```
@autotvm.register_topi_schedule(schedule_conv2d_nchw, 'pvr', ['direct'])
def schedule_conv2d_nchw_pvr(cfg, outs):
    #Describe how to compute output by primitive
```

Blocking
splits the workload into thread blocks and individual threads

Cooperative Fetching
lets threads in the same thread block cooperatively fetch dependent data



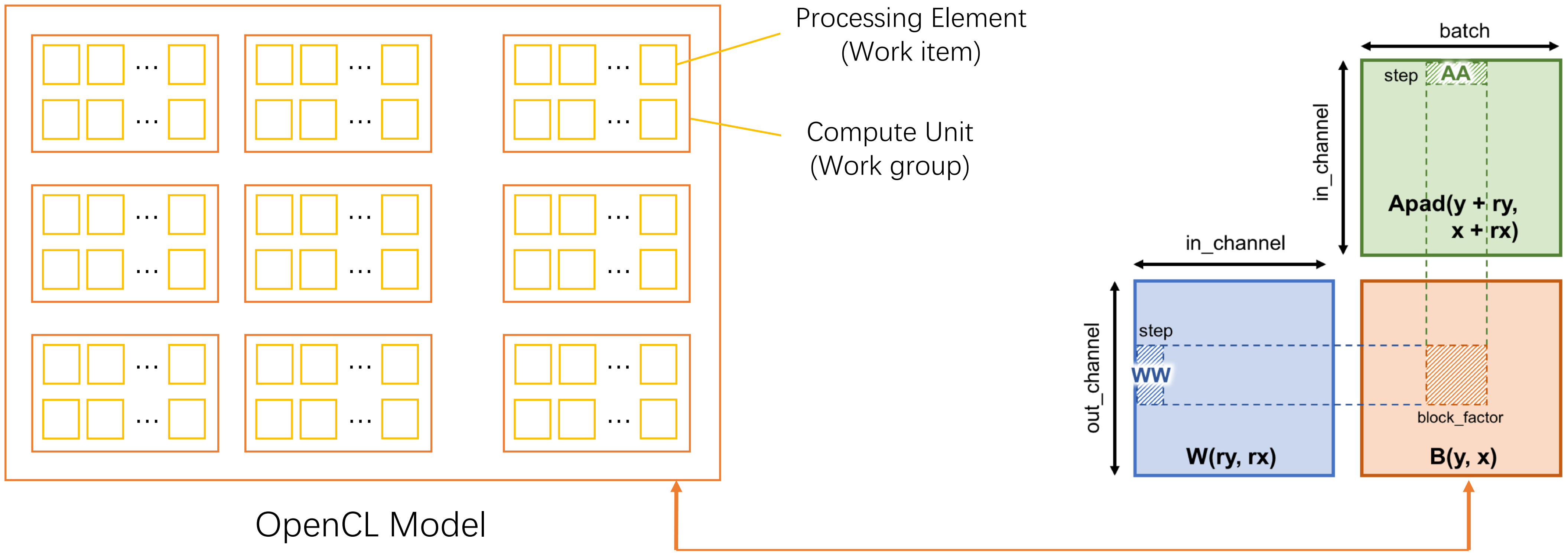
https://docs.tvm.ai/

PVR TOPI

Alibaba A.I. Labs
阿里巴巴人工智能实验室

Blocking

Splits the workload into thread blocks (work groups) and individual threads (work items)



Cooperative Fetching

Lets threads (work item) in the same thread block (work group) cooperatively fetch dependent data

https://www.khronos.org/registry/OpenCL/specs/opencl-1.2.pdf

Alibaba A.I. Labs
阿里巴巴人工智能实验室

Thanks