Auto-scheduler for TVM

Lianmin Zheng

Shanghai Jiao Tong University

Towards “full automation”

Towards “full automation”

Tensor Expression

$$C_{ij} = \sum_k A_{ik} B_{kj}$$

Towards “full automation”

Tensor Expression

$$C_{ij} = \sum_k A_{ik} B_{kj}$$

AutoTVM Schedule Template

Knob 1 …
Knob 2 …
Knob 3 …

Towards “full automation”

Tensor Expression

$$C_{ij} = \sum_k A_{ik} B_{kj}$$

AutoTVM Schedule Template

Knob 1 …
Knob 2 …
Knob 3 …

Optimizer

Towards “full automation”

Tensor Expression

$$C_{ij} = \sum_k A_{ik} B_{kj}$$

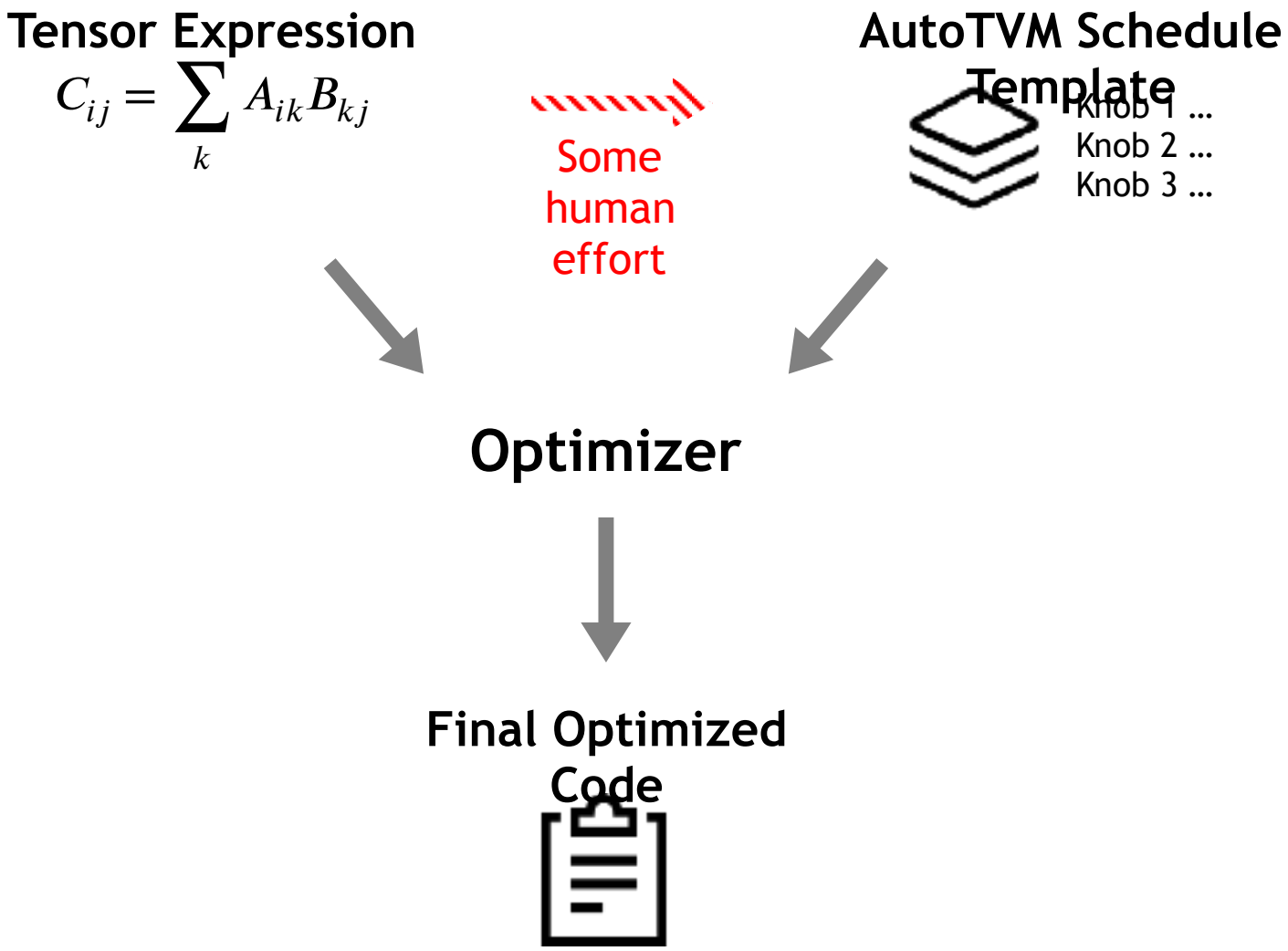
AutoTVM Schedule Template

Knob 1 …
Knob 2 …
Knob 3 …

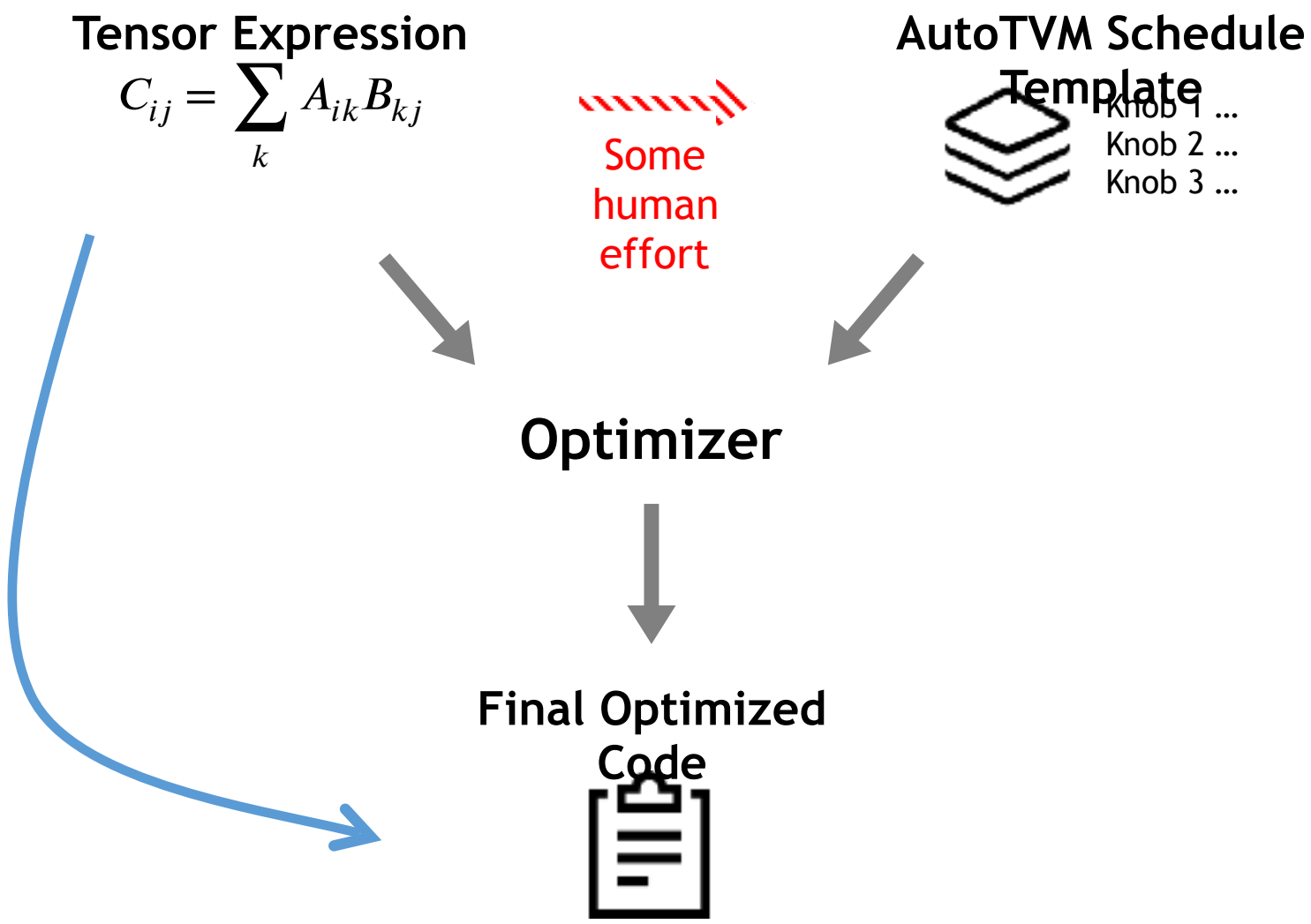
Optimizer

Final Optimized Code

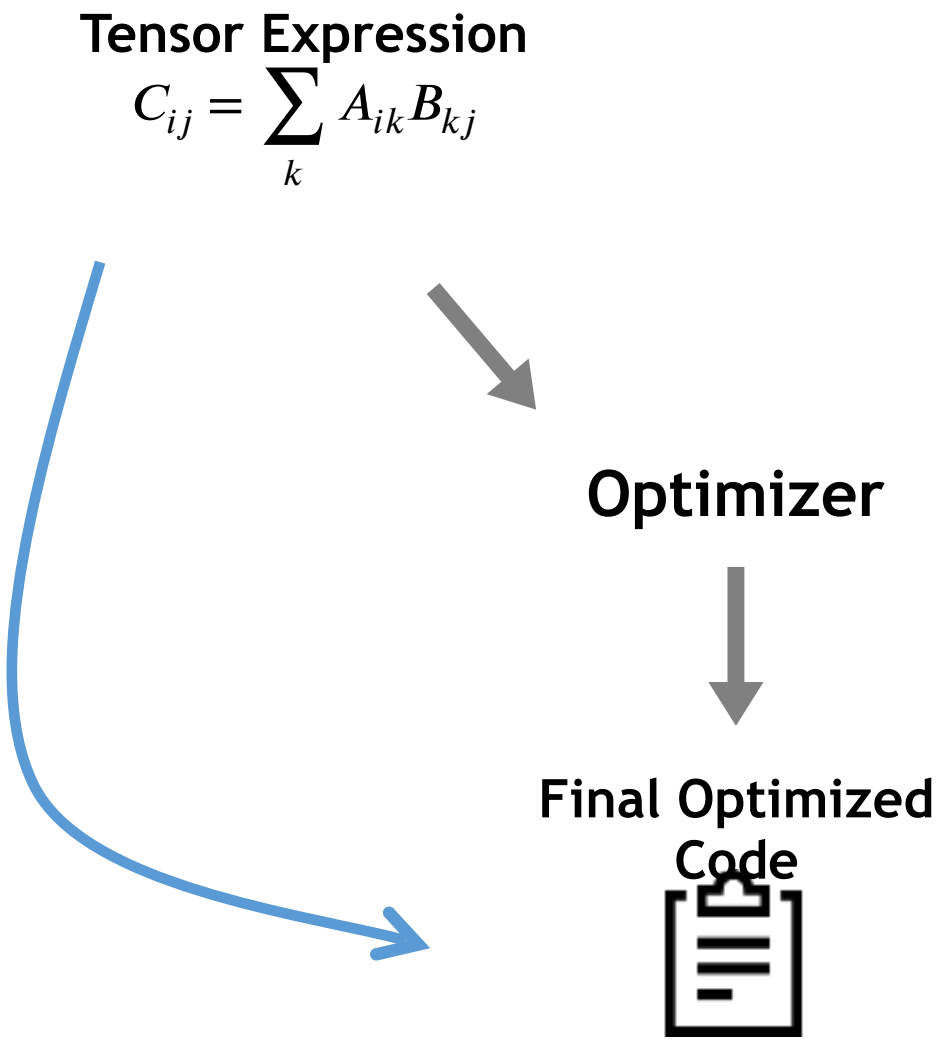
Towards “full automation”



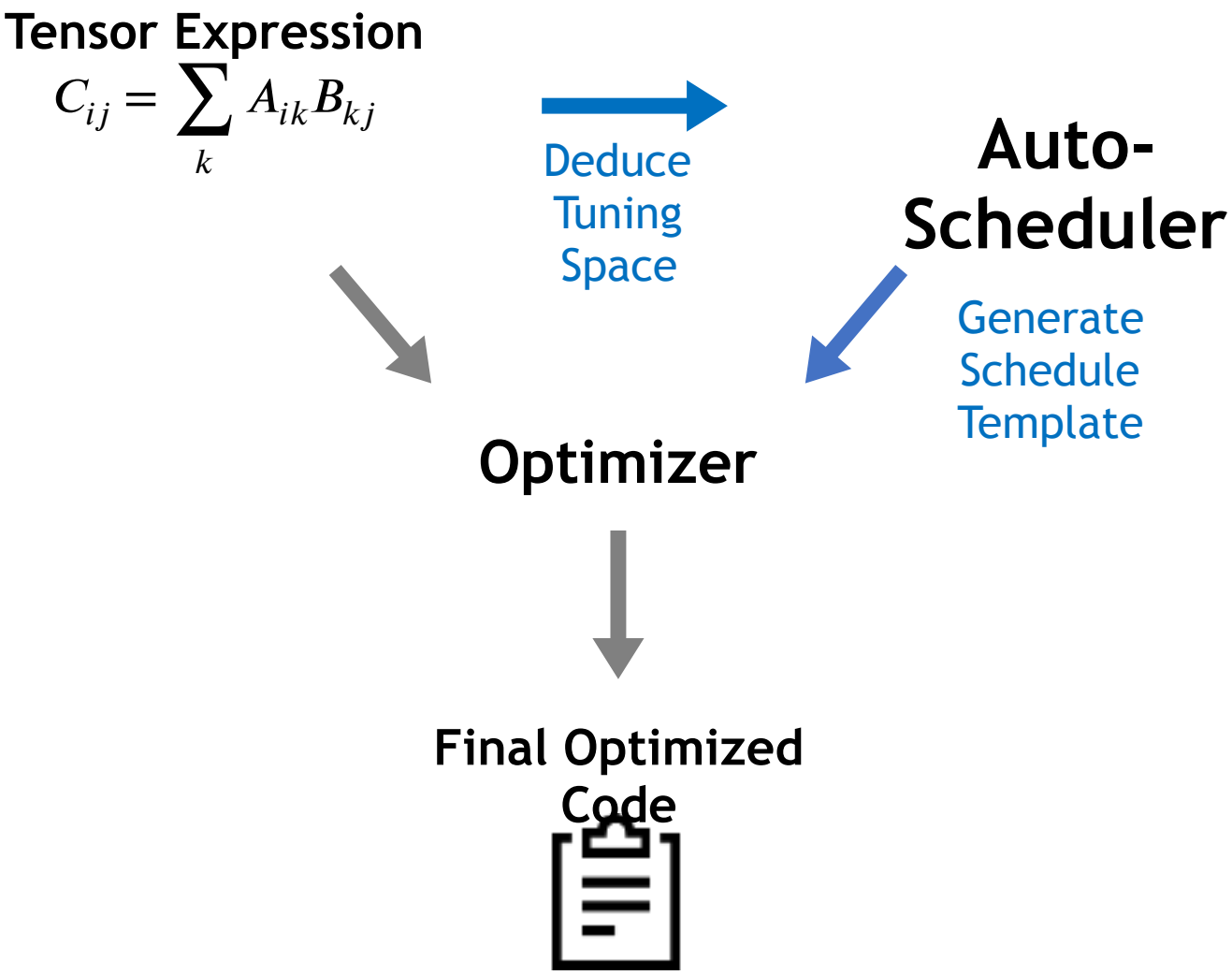
Towards “full automation”



Towards “full automation”



Auto-Scheduler



Auto-scheduler

Tensor Expression

$$C_{N,CO,H,W} = \sum_{CI,\ KH,\ KW} A_{N,CO,H+KH,\ W+KW}\, B_{CO,\ CI,KH,KW}$$

Auto-scheduler

Tensor Expression

$$C_{N,CO,H,W} = \sum_{CI,\ KH,\ KW} A_{N,CO,H+KH,\ W+KW}\, B_{CO,\ CI,KH,KW}$$

Auto-packing

- for vectorization
- for tensorization
- other packing strategies

Auto-scheduler

Tensor Expression

$$C_{N,CO,H,W} = \sum_{CI,\ KH,\ KW} A_{N,CO,H+KH,\ W+KW}\, B_{CO,\ CI,KH,KW}$$

Auto-packing

- for vectorization
- for tensorization
- other packing strategies

Tile-knob on axes

Tune multi-level tiling on N, H, W, CO, CI

Auto-scheduler

Tensor Expression

$$C_{N,CO,H,W} = \sum_{CI,\ KH,\ KW} A_{N,CO,H+KH,\ W+KW}\, B_{CO,\ CI,KH,KW}$$

Auto-packing

- for vectorization
- for tensorization
- other packing strategies

Tile-knob on axes

Tune multi-level tiling on N, H, W, CO, CI

Location-knob on buffers

Tune *compute_at* location for A, B, C and packing buffers

Auto-scheduler

Tensor Expression

$$C_{N,CO,H,W} = \sum_{CI,\ KH,\ KW} A_{N,CO,H+KH,\ W+KW}\, B_{CO,\ CI,KH,KW}$$

Auto-packing

- for vectorization
- for tensorization
- other packing strategies

Tile-knob on axes

Tune multi-level tiling on N, H, W, CO, CI

Location-knob on buffers

Tune *compute_at* location for A, B, C and packing buffers

Annotation-knob on axes

Tune *parallel*, *unroll*

Auto-scheduler

Tensor Expression

$$C_{N,CO,H,W} = \sum_{CI,\ KH,\ KW} A_{N,CO,H+KH,\ W+KW}\, B_{CO,\ CI,KH,KW}$$

Auto-packing

- for vectorization
- for tensorization
- other packing strategies

Tile-knob on axes

Tune multi-level tiling on N, H, W, CO, CI

Location-knob on buffers

Tune *compute_at* location for A, B, C and packing buffers

Annotation-knob on axes

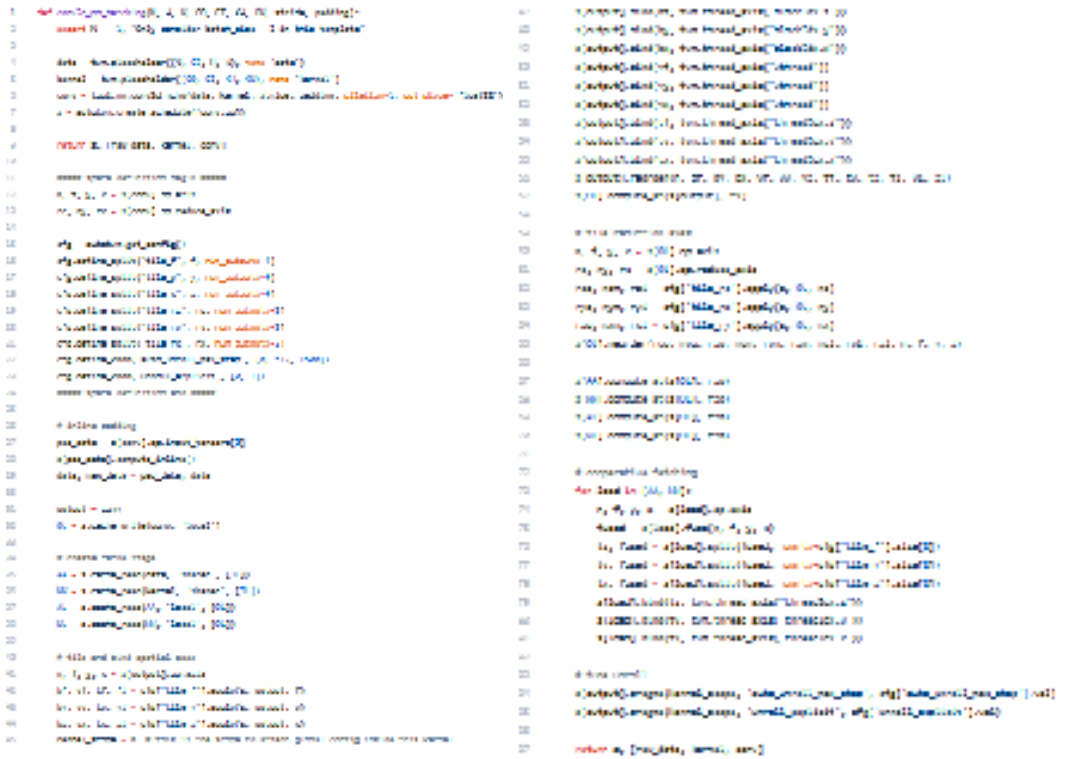
Tune *parallel*, *unroll*

Map to different backends

Map threading, memory scoping, tensorization

Usage

`tvm.create_schedule(C.op)` → `autotvm.create_schedule(C.op)`



87 Loc

```
def conv2d_nchw_padding(N, H, W, CO, CI, KH, KW, stride, padding):
    data = tvm.placeholder((N, CI, H, W), name='data')
    kernel = tvm.placeholder((CO, CI, KH, KW), name='kernel')
    conv = topi.nn.conv2d_nchw(data, kernel, stride, padding, dilation=1, out_dtype='float32')
    s = autotvm.create_schedule([conv.op])

    return s, [raw_data, kernel, conv]
```

7 Loc

Goal and progress