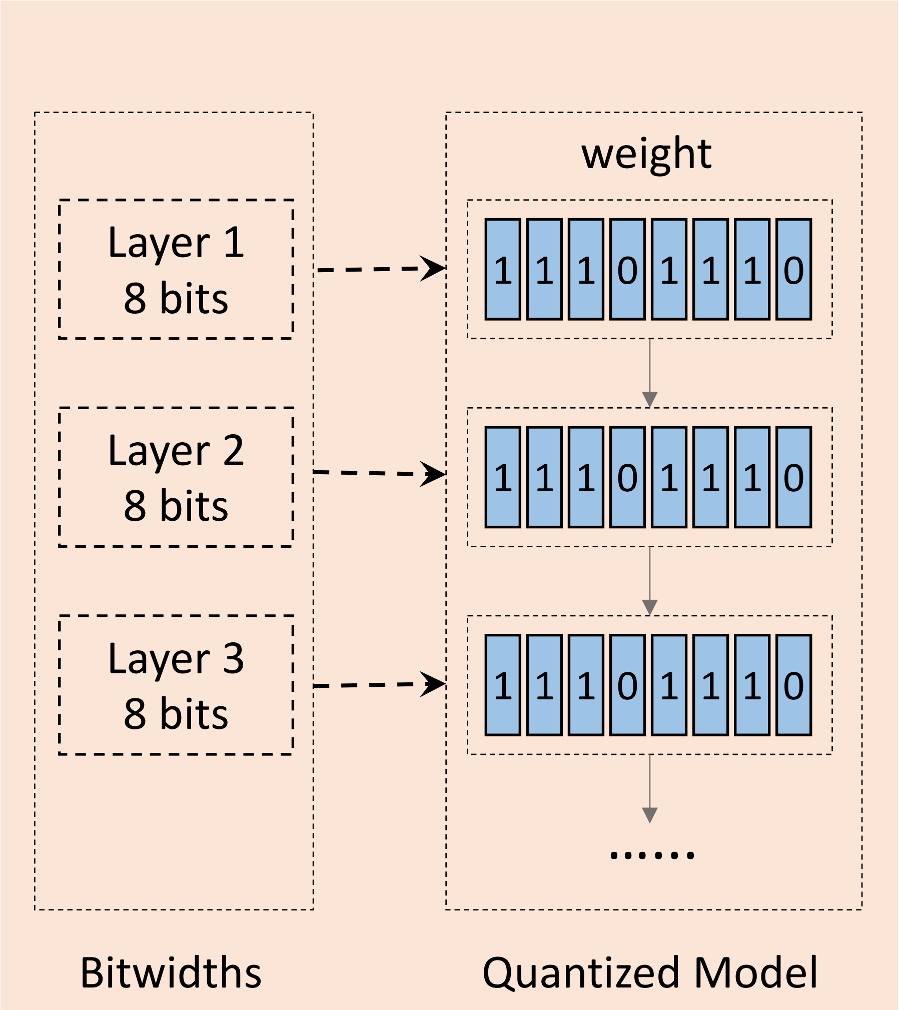
Simple Augmentation Goes A Long Way: ADRL for DNN Quantization

Lin Ning[1], Guoyang Chen[2], Weifeng Zhang[2], Xipeng Shen[1]

[1] North Carolina State University, Raleigh, NC, USA

[2] Alibaba Group, Sunnyvale, CA, USA

Mixed-Precision Quantization Problem



Layer 1
8 bits

Layer 2
8 bits

Layer 3
8 bits

Bitwidths

weight

1 1 1 0 1 1 1 0

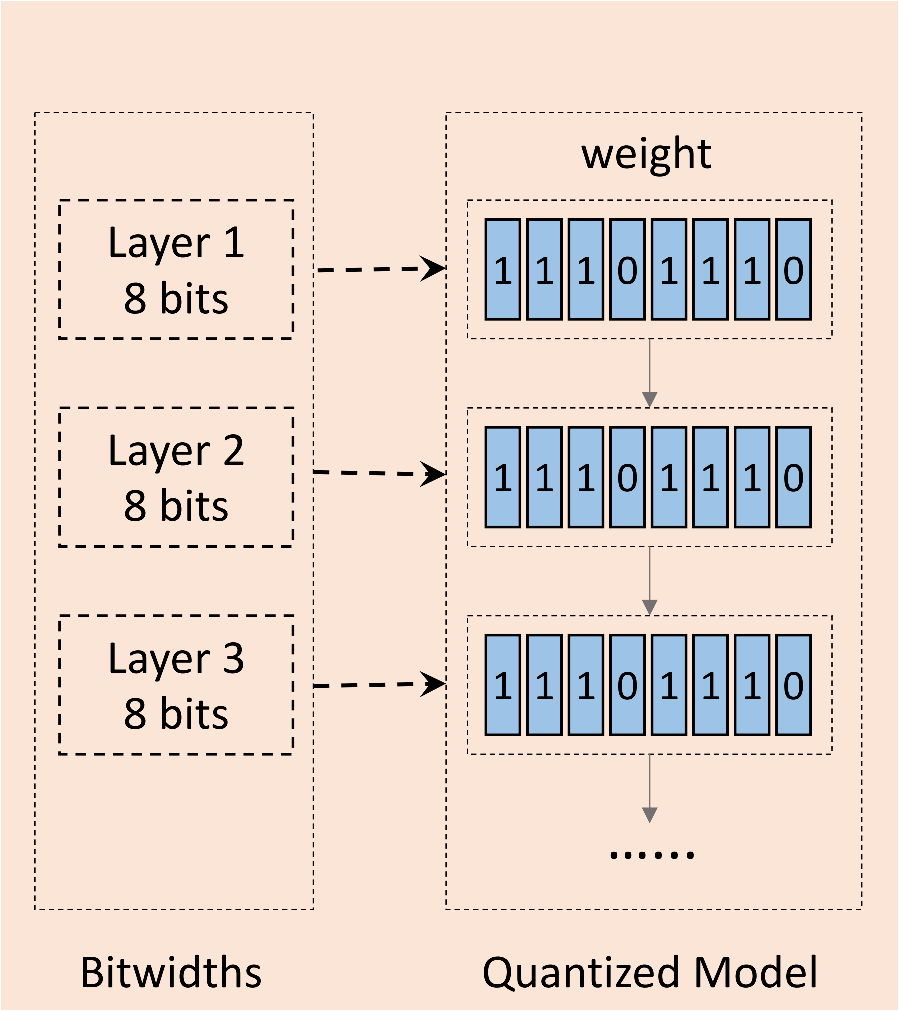
1 1 1 0 1 1 1 0

1 1 1 0 1 1 1 0

……

Quantized Model

Mixed-Precision Quantization Problem



Layer 1
8 bits

Layer 2
8 bits

Layer 3
8 bits

weight

1 1 1 0 1 1 1 0

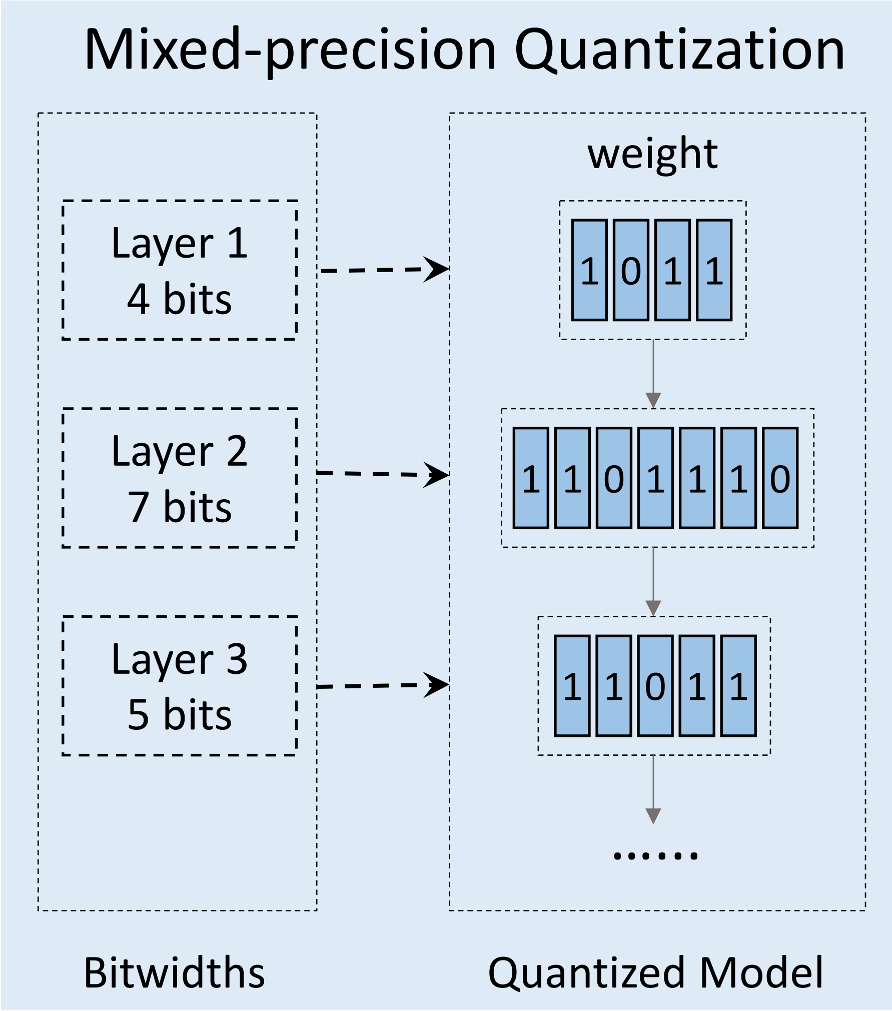
1 1 1 0 1 1 1 0

1 1 1 0 1 1 1 0

……

Bitwidths

Quantized Model



Mixed-precision Quantization

Layer 1
4 bits

Layer 2
7 bits

Layer 3
5 bits

weight

1 0 1 1

1 1 0 1 1 1 0

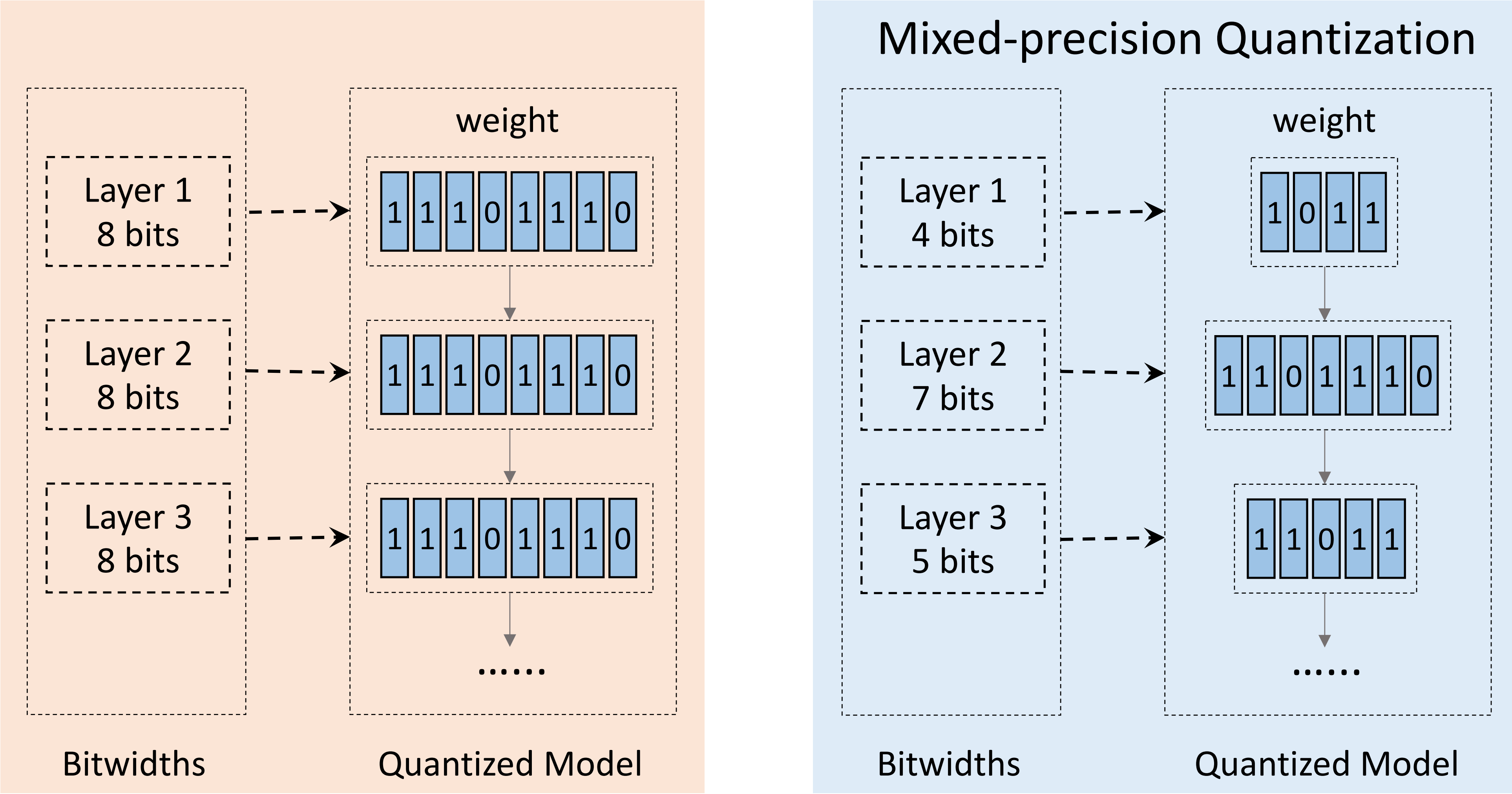
1 1 0 1 1

……

Bitwidths

Quantized Model

Mixed-Precision Quantization Problem



Layer 1
8 bits

Layer 2
8 bits

Layer 3
8 bits

weight

1 1 1 0 1 1 1 0

1 1 1 0 1 1 1 0

1 1 1 0 1 1 1 0

......

Bitwidths

Quantized Model

Mixed-precision Quantization

Layer 1
4 bits

Layer 2
7 bits

Layer 3
5 bits

weight

1 0 1 1

1 1 0 1 1 1 0

1 1 0 1 1

......

Bitwidths

Quantized Model

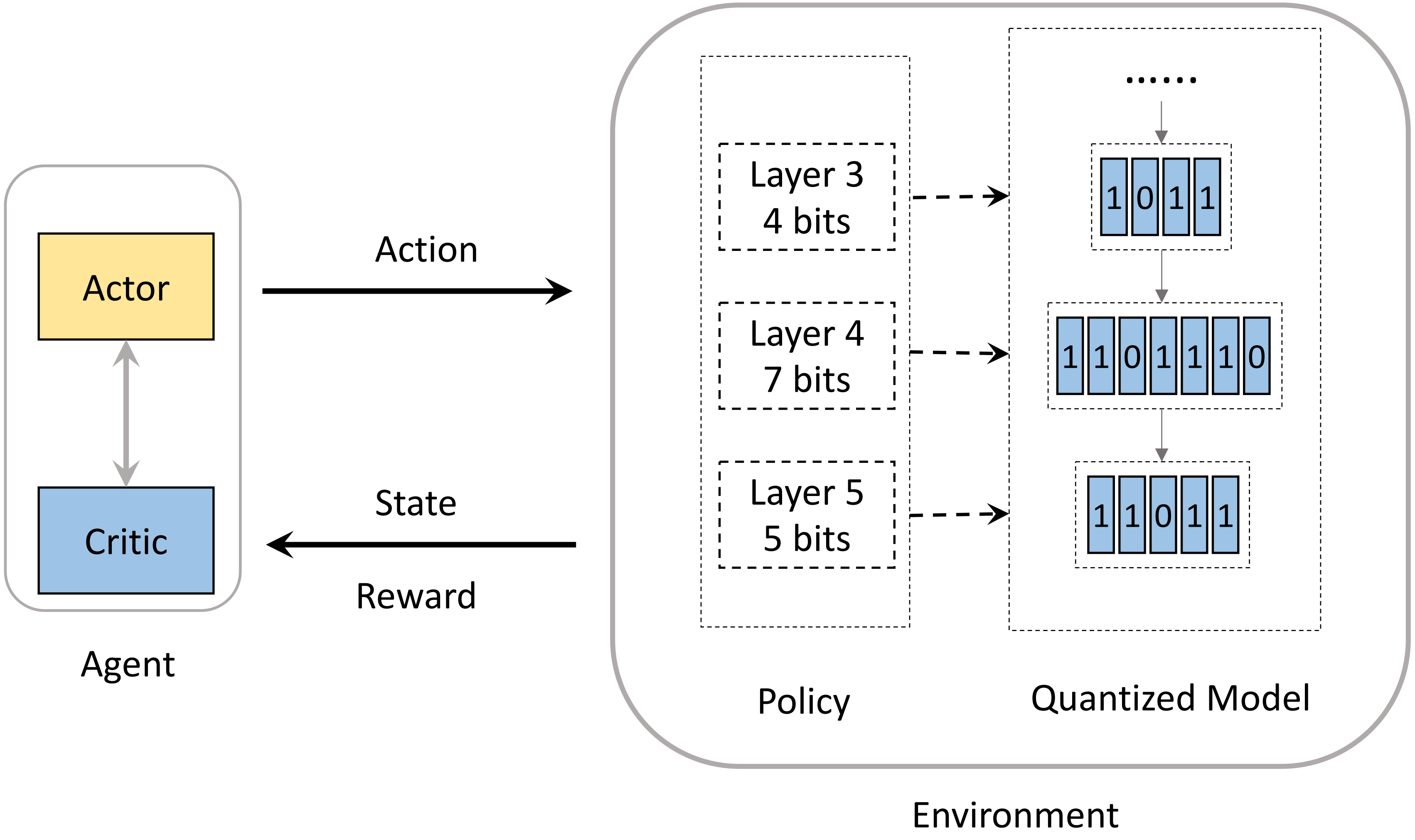
choices: 8

choices: 8

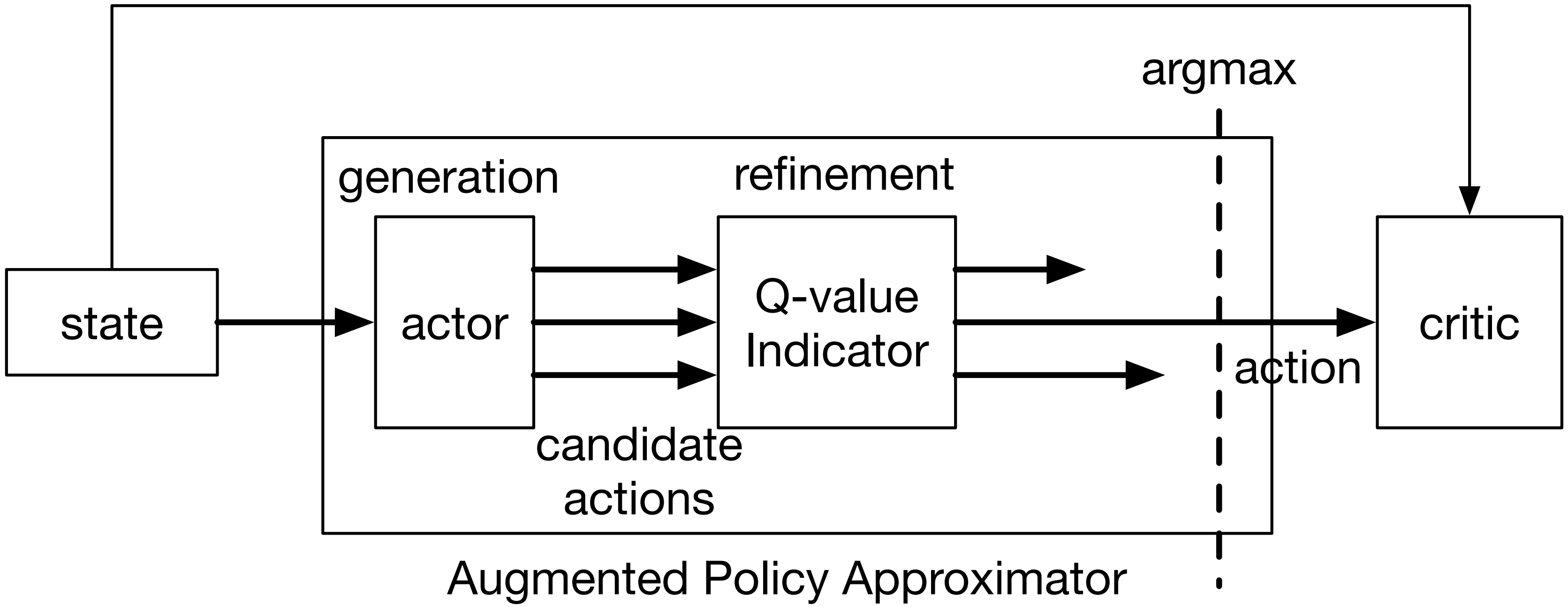
choices: 8

Design Space: 8^n

DRL for Mixed-Precision Quantization

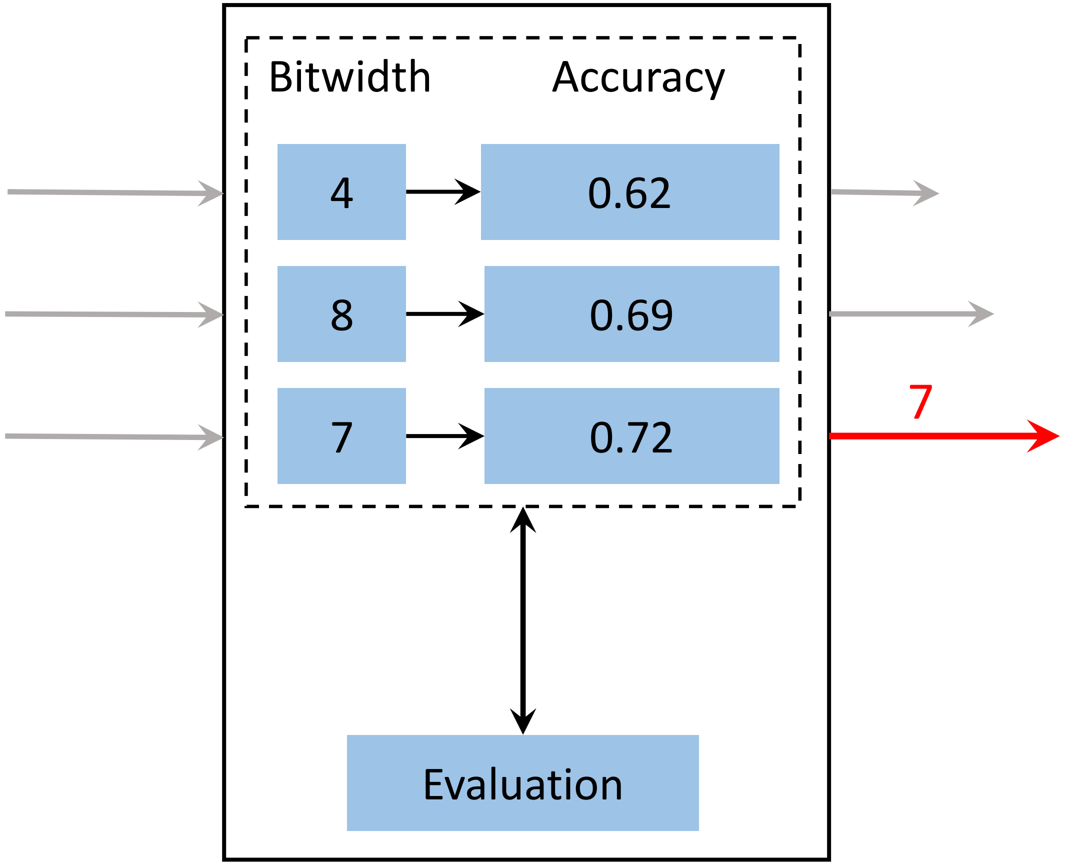


ADRL: Augmented Policy Approximator



ADRL: Q-value Indicator

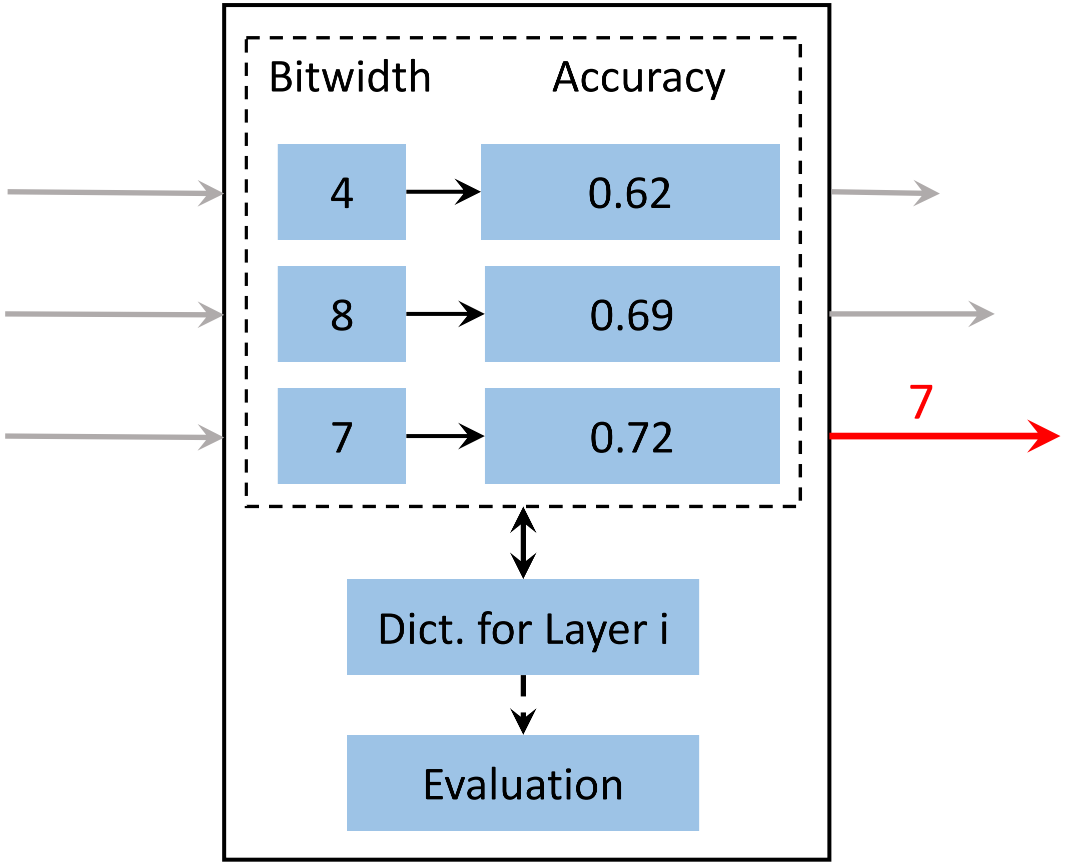
Profiling-Based Indicator



Bitwidth	Accuracy
4	0.62
8	0.69
7	0.72

ADRL: Q-value Indicator

Profiling-Based Indicator



- Memoization
- Early Termination

Dictionary for Layer i

1 : ?	5 : ?
2: ?	6 : 0.67
3 : 0.69	7 : ?
4 : 0.62	8 : 0.72

Bitwidth : Accuracy

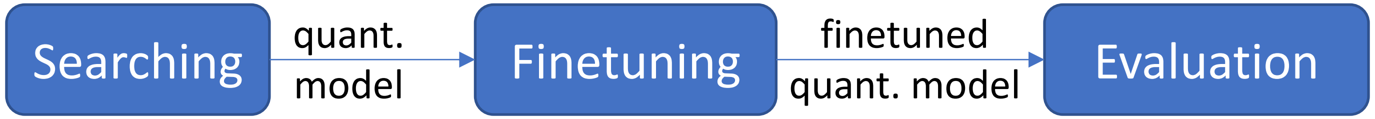
ADRL: Evaluation

Network	Dataset
CifarNet	Cifar10
ResNet20	Cifar10
AlexNet	ImageNet
ResNet50	ImageNet

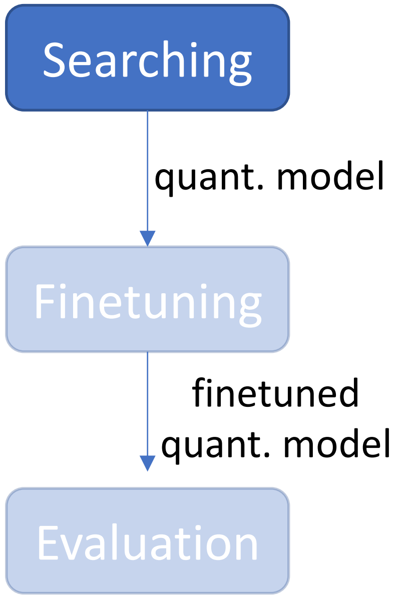
Server

- Intel(R) Xeon(R) Platinum 8168 Processor
- 32GB memory
- 4 NVIDIA Tesla V100 32GB GPUs.

DRL based mix-precision quantization:



ADRL: Evaluation



Searching

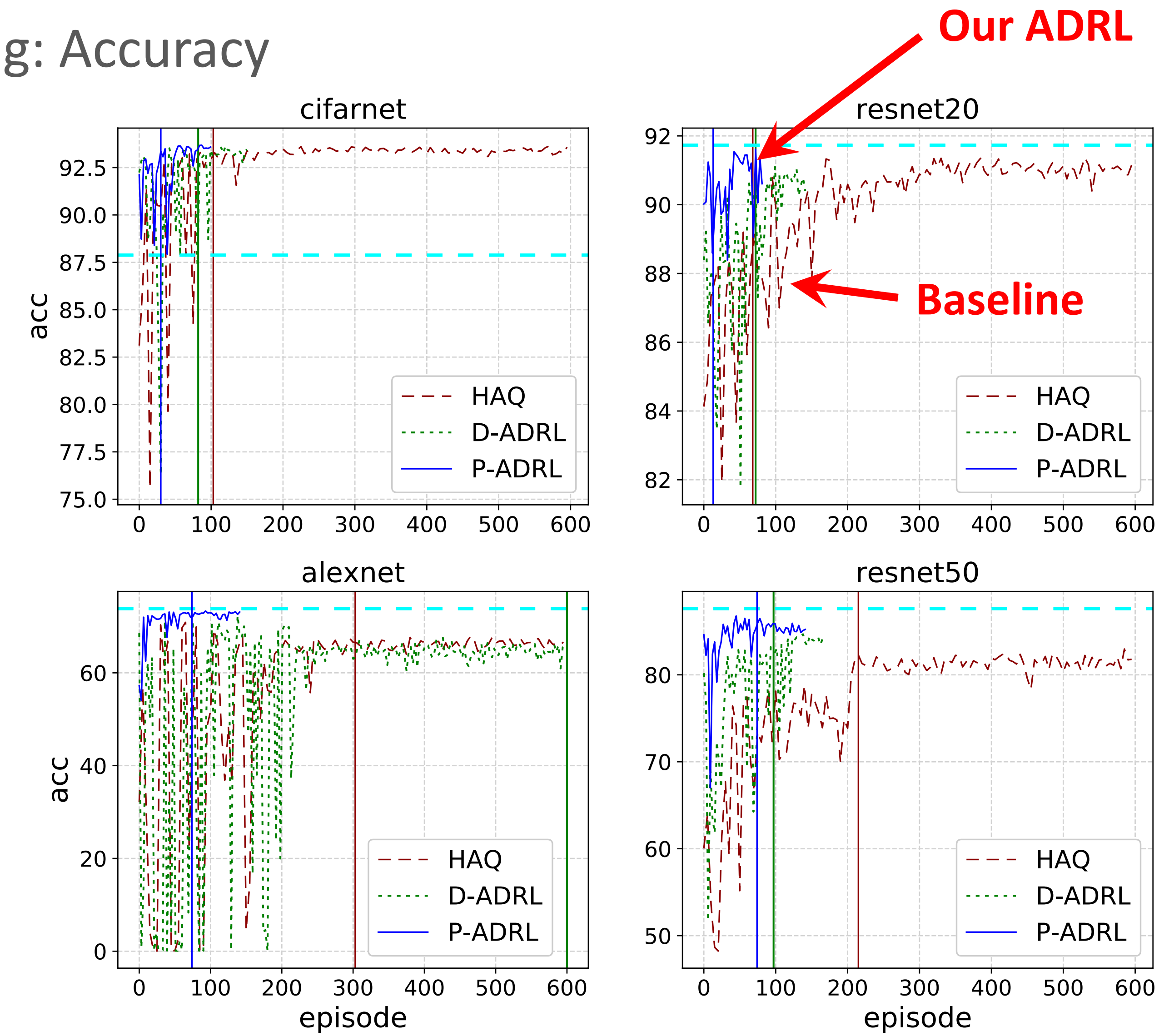
quant. model

Finetuning

finetuned quant. model

Evaluation

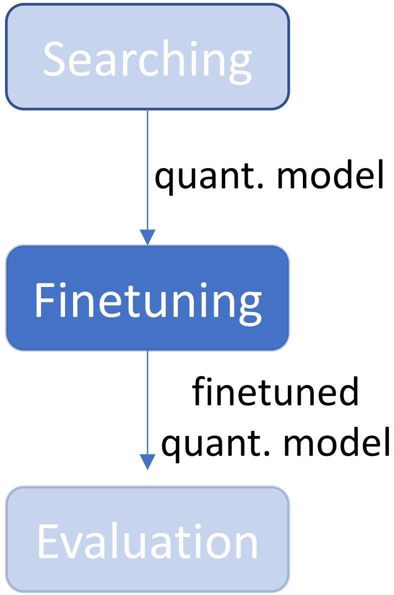
Searching: Accuracy



Our ADRL

Baseline

ADRL: Evaluation



Searching

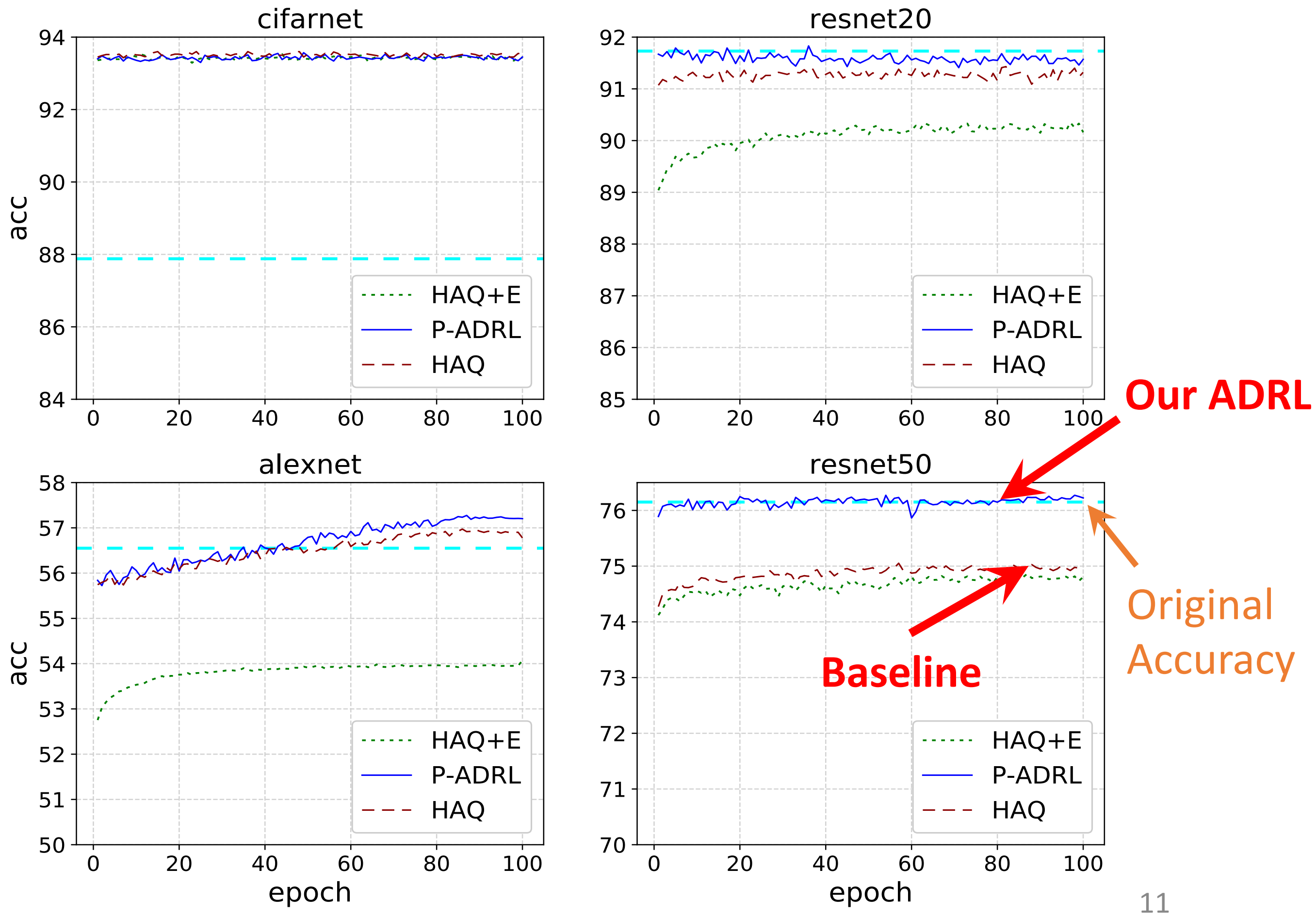
quant. model

Finetuning

finetuned quant. model

Evaluation

Finetuning: Accuracy

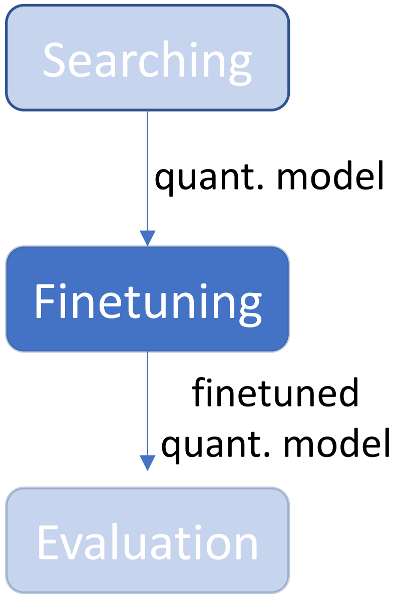


Our ADRL

Baseline

Original Accuracy

ADRL: Evaluation



Searching

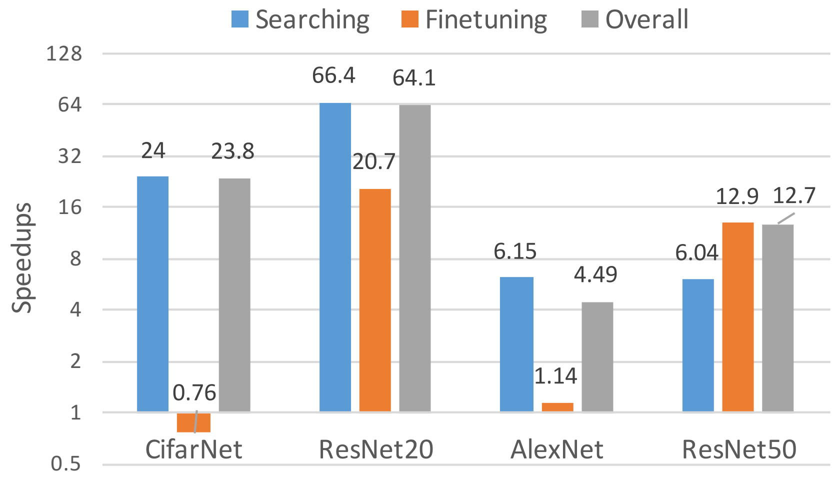
quant. model

Finetuning

finetuned quant. model

Evaluation

Speedup



Summarize

Augmented DRL

- ➢ Produces more accurate quantized models than the state of the art DRL-based quantization
- ➢ Improves the learning speed
- ➢ Significantly magnifies the potential of DRL for DNN quantization.