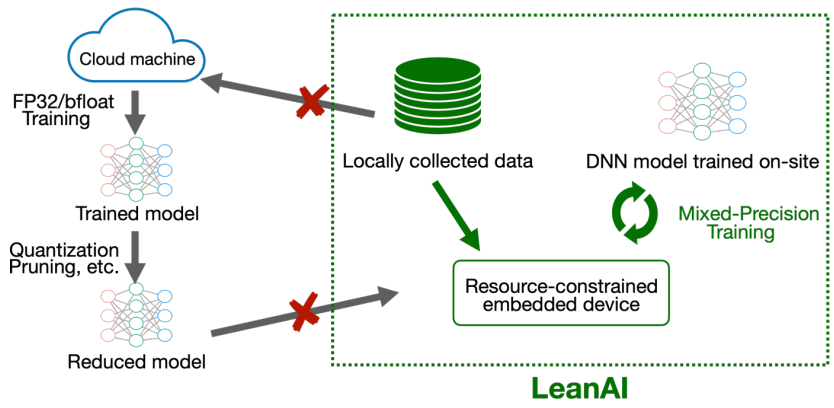
LeanAI: Dynamic Precision Training on the Edge

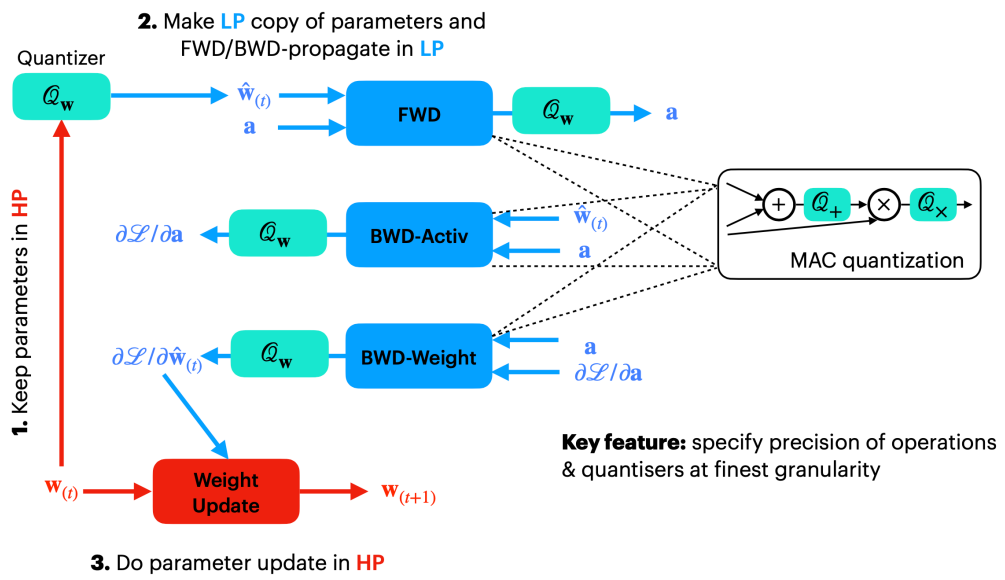
IRISA, TARAN
LS2N, OGRE
LIP, OKHAM

Context and Objectives



- Need for **learning acceleration** mechanism in both ***cloud*** (for large-scale models) and ***on-site*** settings (e.g. autonomous driving, privacy)
- Working on both arithmetic and algorithmic levels
- Design of dedicated HW operators

MPTorch: Mixed-Precision DNN Compute Simulator



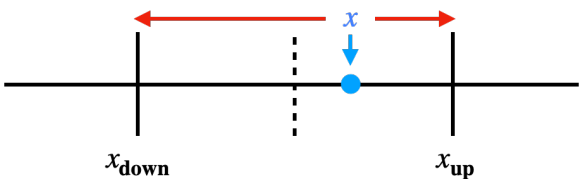
Key feature: specify precision of operations & quantisers at finest granularity

Support for CPU & GPU simulation + FPGA-based accelerator prototyping

Stochastic Rounding for DNN Training Acceleration

Stochastic Rounding (SR) can recapture information that is discarded when bits are rounded off in long computation chains (e.g. long summations or dot products)

$$p_{\text{down}} = \frac{x_{\text{up}} - x}{x_{\text{up}} - x_{\text{down}}} \qquad p_{\text{up}} = \frac{x - x_{\text{down}}}{x_{\text{up}} - x_{\text{down}}}$$



Results:

- Achieve up to 18.5% and 14.5% savings in area and energy
- Can report up to 32.2% delay reduction

Configuration	E	M	r	Delay (ns)	Area (μm^2)	Energy ($\mu W/MHz$)
SR eager W/O Sub	6	5	4	1.85	508.36	0.46
			7	1.87	540.19	0.49
			9	1.87	558.63	0.51
			11	1.93	579.19	0.53
			13	**1.93**	**601.71**	**0,56**
RN W/ Sub (FP16)	5	10	-	**2.73**	**692.62**	**0.65**
RN W/ Sub (FP32)	8	23	-	**4.71**	**1404.01**	**1.17**

Configuration: ResNet18 + CIFAR10

Configuration	E	M	r	Accuracy (%)
FP32 Baseline	8	23	-	**91.47**
RN W/ Sub	5	10	-	91.1
RN W/ Sub	8	7	-	88.79
RN W/ Sub	6	5	-	83.03
SR W/ Sub	6	5	4	43.11
SR W/ Sub	6	5	9	89.34
SR W/ Sub	6	5	11	90.7
SR W/ Sub	6	5	13	**91.39**
SR W/O Sub	6	5	11	90.67
SR W/O Sub	6	5	13	**91.39**

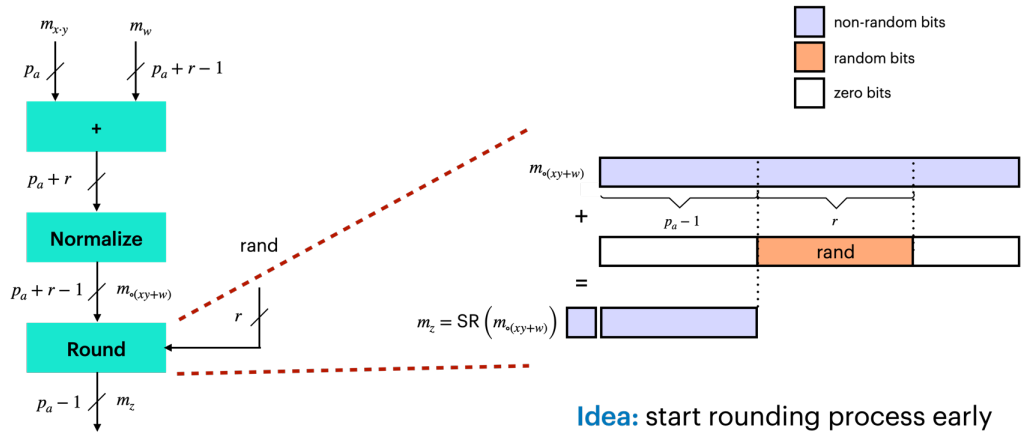
Model/Dataset	Configuration	E	M	r	Accuracy (%)
VGG16/CIFAR10	FP32 Baseline	8	23	-	93.46
	RN W/ Sub	5	10	-	93.06
	SR W/O Sub	6	5	13	93.11
ResNet50/Imagewoof	FP32 Baseline	8	23	-	80.94
	RN W/ Sub	5	10	-	80.3
	SR W/O Sub	6	5	13	80.33

Shown to be beneficial for DNN training acceleration

Challenge: not obvious how to optimize in hardware

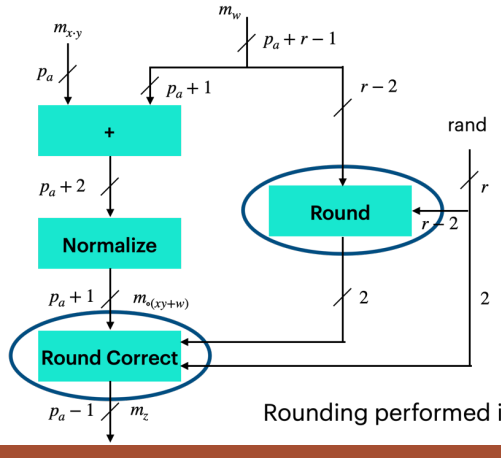
Basic building block is the multiply accumulate (**MAC**) unit: z = xy + z

Classic SR Design (lazy)



Idea: start rounding process early

Proposed SR Design (eager)



Rounding performed in two stages

Theoretical result: probabilistic error analysis on the number of required random bits r need to implement SR wrt the length n of the compute chain

$$r \approx \lceil \log_2(n)/2 \rceil$$

AdaQAT: Adaptive Quantization-Aware Training

Optimization-based method for **mixed-precision** (weights and activations) **DNN quantization**

Idea:

$$\mathcal{L}_{\text{total}} = \mathcal{L}\left(\lceil N_w \rceil, \lceil N_a \rceil\right) + \lambda \underbrace{\mathcal{L}_{\text{HW}}\left(\lceil N_w \rceil, \lceil N_a \rceil\right)}_{:=\lceil N_w \rceil \lceil N_a \rceil}$$

BitOps HW cost estimate

$$\frac{\partial \mathcal{L}}{\partial N_w} \approx \mathcal{L}\left(\lceil N_w \rceil, \lceil N_a \rceil\right) - \mathcal{L}\left(\lfloor N_w \rfloor, \lceil N_a \rceil\right) \qquad \frac{\partial \mathcal{L}}{\partial N_a} \approx \mathcal{L}\left(\lceil N_w \rceil, \lceil N_a \rceil\right) - \mathcal{L}\left(\lceil N_w \rceil, \lfloor N_a \rfloor\right)$$

$$\frac{\partial \mathcal{L}_{\text{total}}}{\partial N_w} \approx \frac{\partial \mathcal{L}}{\partial N_w} + \lambda \frac{\partial \mathcal{L}_{\text{HW}}}{\partial \lceil N_w \rceil} \qquad \frac{\partial \mathcal{L}_{\text{total}}}{\partial N_a} \approx \frac{\partial \mathcal{L}}{\partial N_a} + \lambda \frac{\partial \mathcal{L}_{\text{HW}}}{\partial \lceil N_a \rceil}$$

Model	Method	# exploration epochs	# total epochs epochs	Bit-width (W/A)	Accuracy (%) top-1	BitOPs (Gb)
ResNet-18	DQ [Uhl+19]	50	**50**	5.11/10.4	70.1	93.6
	FracBits [YJ21b]	120	50	4.00/4.00	70.6	34.7
	SDQ [Hua+22]	60	150	3.85/4	**71.7**	33.4
	Our	**<1**	100	**3.84/4.00**	71.4	**31.4**
MobileNet-V2	DQ [Uhl+19]	50	**50**	5.77/-	69.7	93.6
	FracBits [YJ21b]	120	150	4.00/4.00	71.3	5.35
	SDQ [Hua+22]	60	180	3.79/4	**72.0**	5.07
	Our	**<1**	150	**3.86/3.88**	71.3	**4.95**

Publications: