MLWeaving:A One-Size-Fits-All System for Any-precision Learning

Zeke Wang, Kaan Kara, Hantian Zhang, Gustavo Alonso, Onur Mutlu, Ce Zhang

Problem: Linear model training using low-precision SGD.

Existing Approach:

1, One hardware design for each precision.

2, One quantized dataset for each precision.

$\vec{x}$: model,
$\vec{g}$: gradient,
Γ: learning rate,
df: derivative of loss function

Stochastic Gradient Decent (SGD):

```
For e = 1 to E do                    /* E epochs*/
 For i = 1 to N do                   /*N samples (a_i, b_i)*/
   ax = Q_s(a_i)* x;                 /*dot product*/
   scale = γ*df(ax, b_i);            /*serial part*/
   g = scale*Q_s(a_i);               /*gradient comp*/
   x = x - g;                        /*model update*/
```

Our Approach (MLWwaving):

One hardware design and one memory layout for any precision.

MLWeaving: Software/Hardware Co-design

MLWeaving memory layout (software)

1, Original full-precision fixed-point table for training dataset *a*

1st row	ABCD	EFGH	IJKL	MNOP	RSTU	VXYZ	1010	0101
2nd row	abcd	efgh	ijkl	mnop	rstu	vxyz	1100	0011

A~Z (a~z) is binary, 0 or 1.

1st row

1st bit	AEIMRV10
2nd bit	BFJNSX01
3rd bit	CGKOTY10
4th bit	DHLPUZ01

If you want 3-bit precision

2nd row

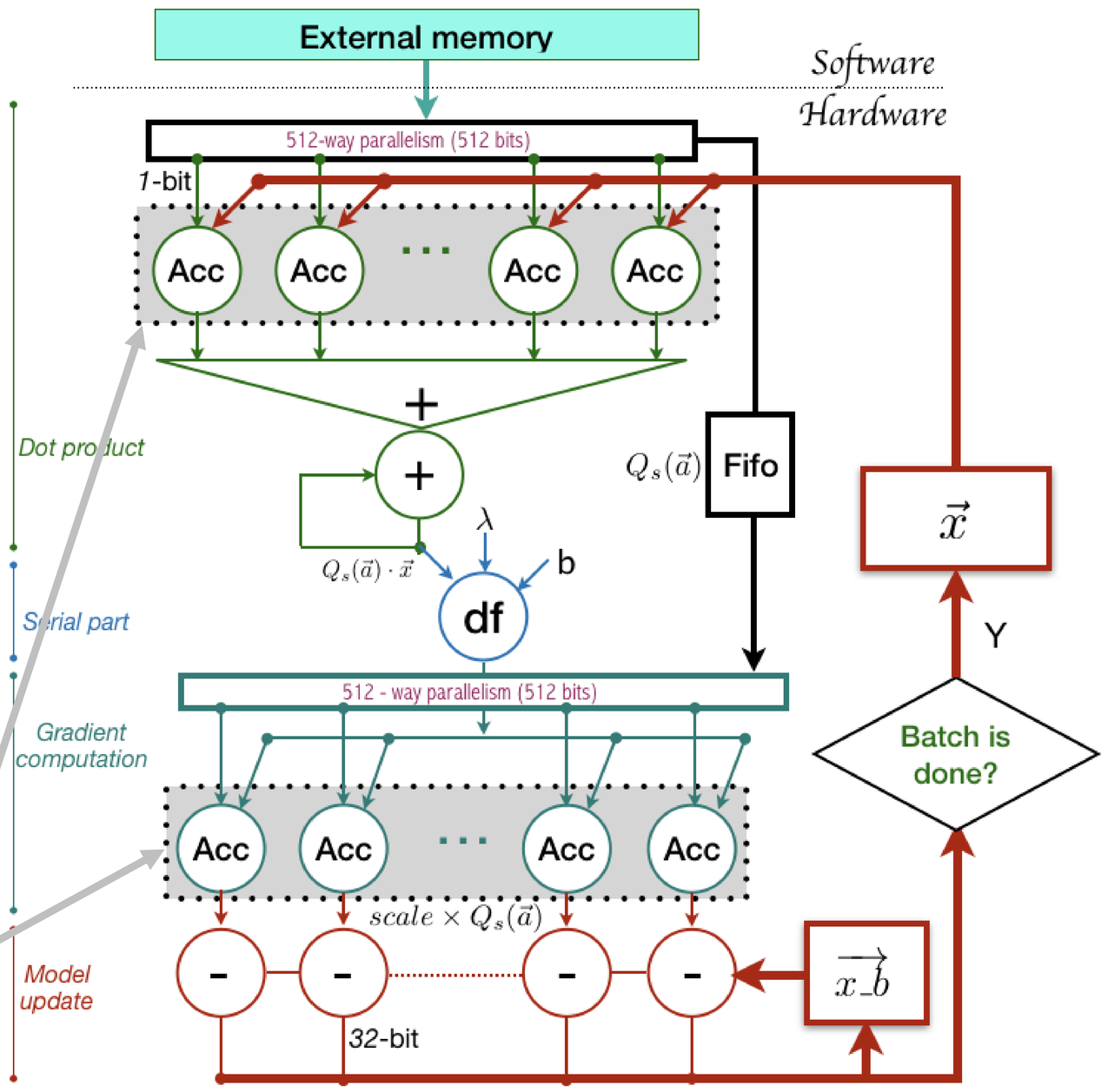
1st bit	aeimrv10
2nd bit	bfjnsx10
3rd bit	cgkoty01
4th bit	dhlpuz01

2, MLWeaving memory layout for training dataset a

Bit-level flexibility of precision from software side

Directly consume the data from MLWeaving memory layout with *bit-serial multipliers*.

MLWeaving arithmetic (hardware)

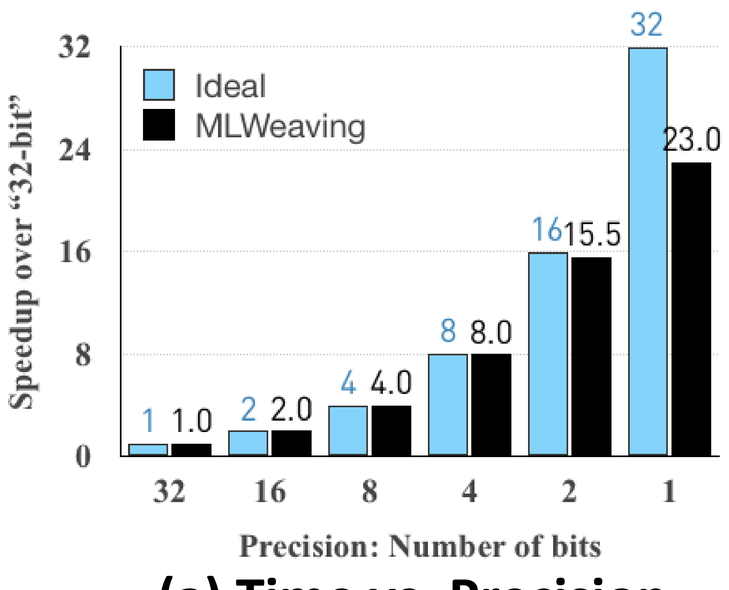


Experiment

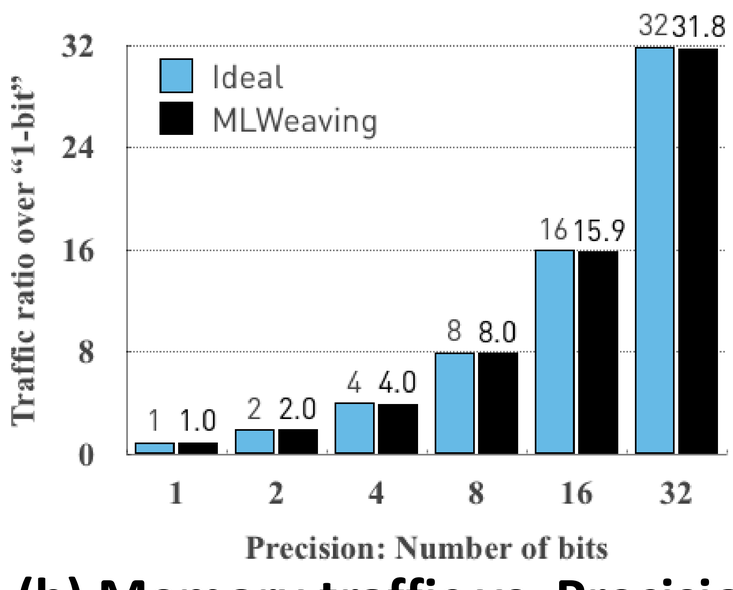
- **Hardware**: an Intel Broadwell CPU (14 cores, 35MB LLC, 60 GB/s memory bandwidth) and an Intel Arria 10 FPGA (directly access the CPU memory via one QPI and two PCIe, with memory bandwidth: 15GB/s.
- **Dataset**: Epsilon (40,000 samples, 2000 features).
- **Hogwild (ModelAverage)**: state-of-the-art parallel implementations of SGD on CPUs, using 14 cores, AVX2 and 8-bit dataset.

Findings:

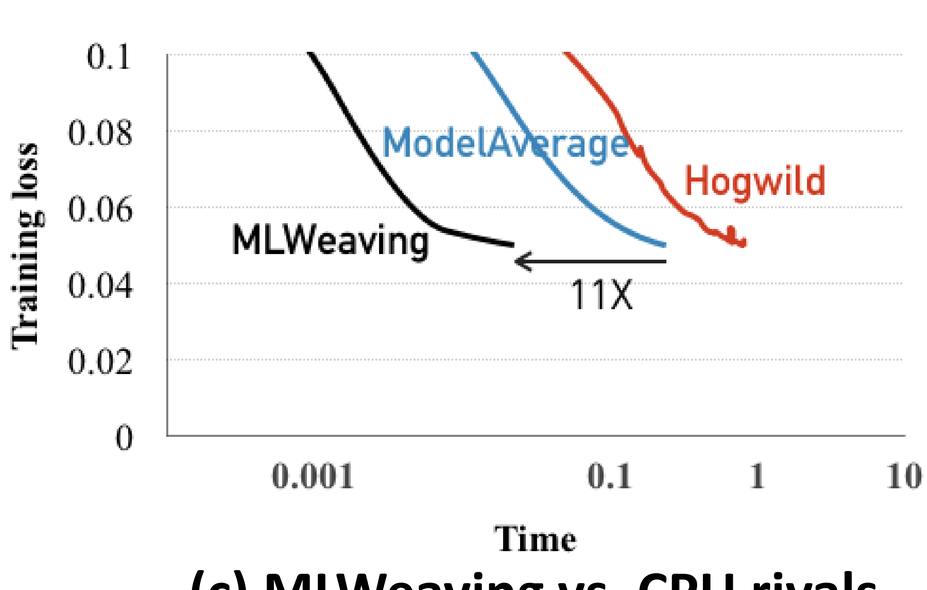
1, MLWeaving can roughly achieve linear speedup (time or memory traffic), when a lower number of bits is used, as shown in Figures a, b.
2, MLWeaving on an FPGA can achieve 11X speedup over its CPU rivals in Figure c.



(a) Time vs. Precision



(b) Memory traffic vs. Precision



(c) MLWeaving vs. CPU rivals