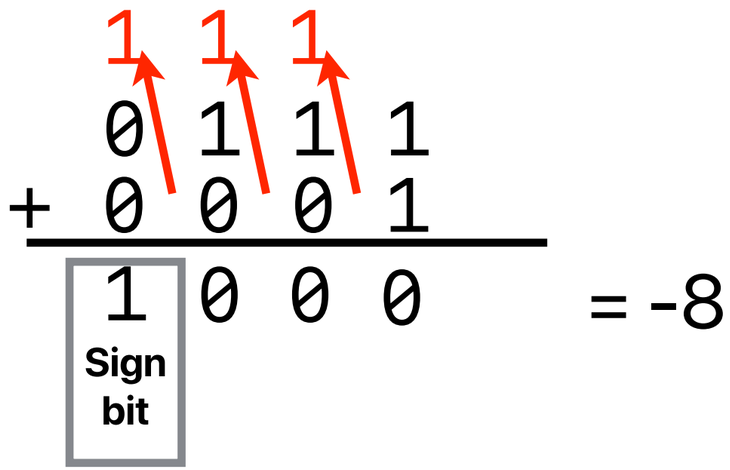
Floating Point Numbers

Prof. Usagi

Floating point numbers

Let's revisit the 4-bit binary adding

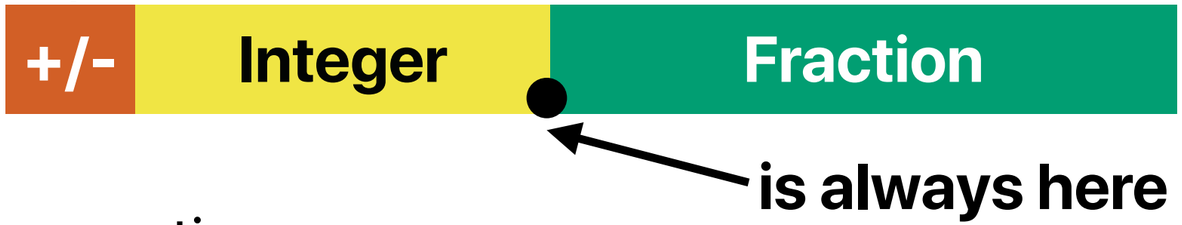
- 7 + 1 = ?



- If you add the largest integer with 1, the result will become the smallest integer.

"Floating" v.s. "Fixed" point

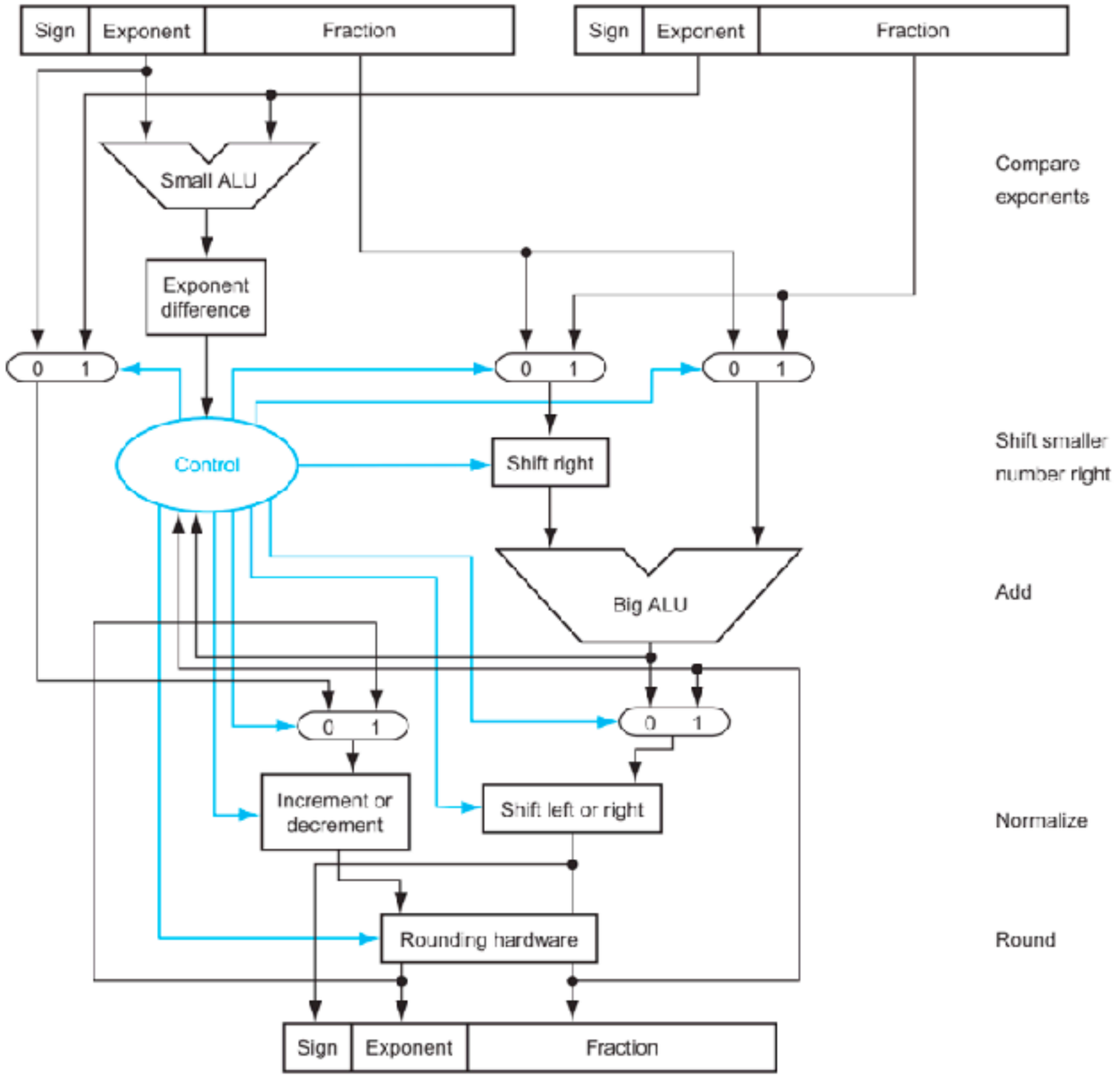
- We want to express both a relational number's "integer" and "fraction" parts
- Fixed point
  - One bit is used for representing positive or negative
  - Fixed number of bits is used for the integer part
  - Fixed number of bits is used for the fraction part
  - Therefore, the decimal point is **fixed**
- Floating point
  - One bit is used for representing positive or negative
  - A fixed number of bits is used for exponent
  - A fixed number of bits is used for fraction
  - Therefore, the decimal point is **floating — depending on the value of exponent**



IEEE 754 format

- Realign the number into 1.**F** * 2^{e}
- Exponent stores **e** + 127
- Fraction only stores **F**

Floating point adder

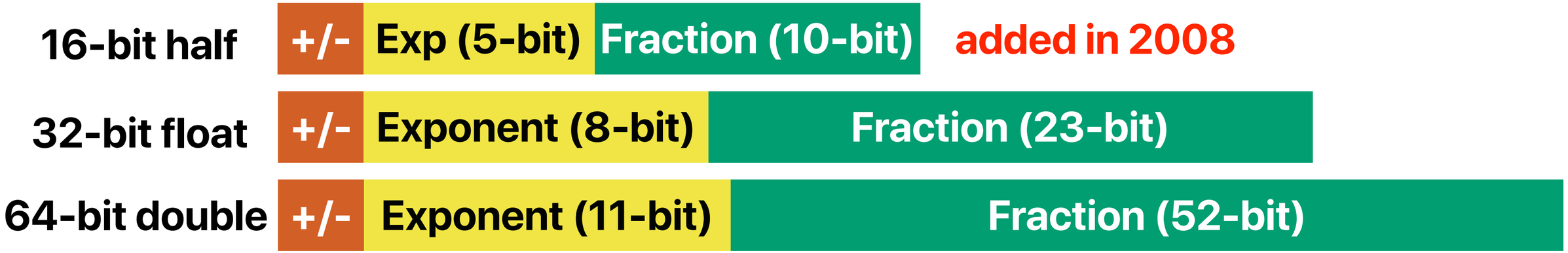


Demo — what's in c?

```c
#include <stdio.h>

int main(int argc, char **argv)
{
    float a, b, c;
    a = 1280.245;
    b = 0.0004;
    c = a + b;
    printf("1280.245 + 0.0004 = %f\n",c);
    return 0;
}
```

Other floating point formats



- Not all applications require “high precision”
- Deep neural networks are surprisingly error tolerable

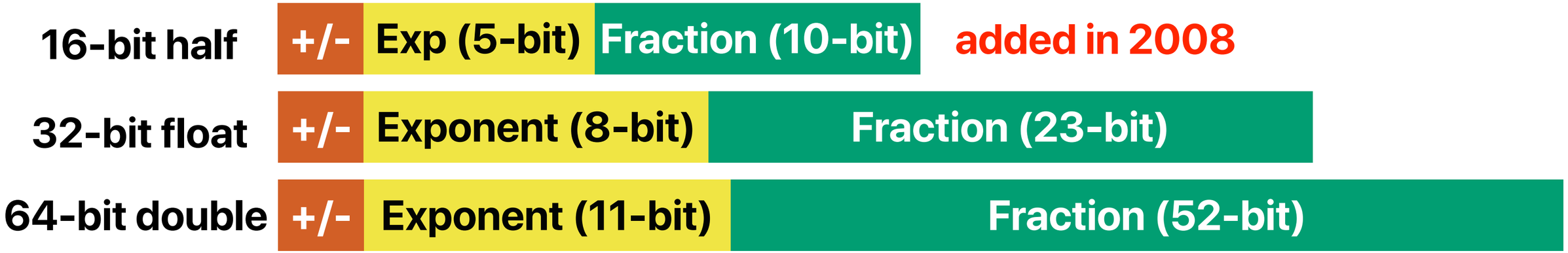
Can you tell the difference?

Higher resolution

But we all can tell they are our mascots!

How about this?

Other floating point formats



- Not all applications require “high precision”
- Deep neural networks are surprisingly error tolerable

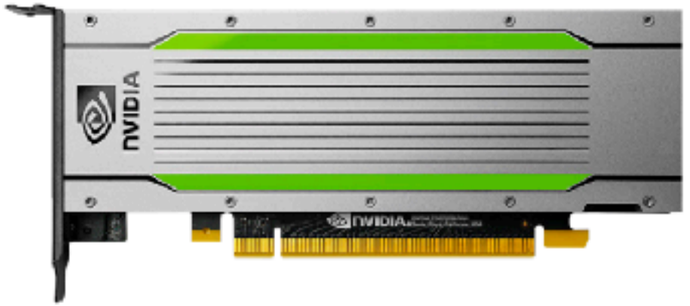
Mixed-precision

Double Precision Results

GPU	Tesla T4	Tesla V100	Tesla P100
Max Flops (GFLOPS)	253.38	7072.86	4736.76
Fast Fourier Transform (GFLOPS)	132.60	1148.75	756.29
Matrix Multiplication (GFLOPS)	249.57	5920.01	4256.08
Molecular Dynamics (GFLOPS)	105.26	908.62	402.96
S3D (GFLOPS)	59.97	227.85	161.54

Single Precision Results

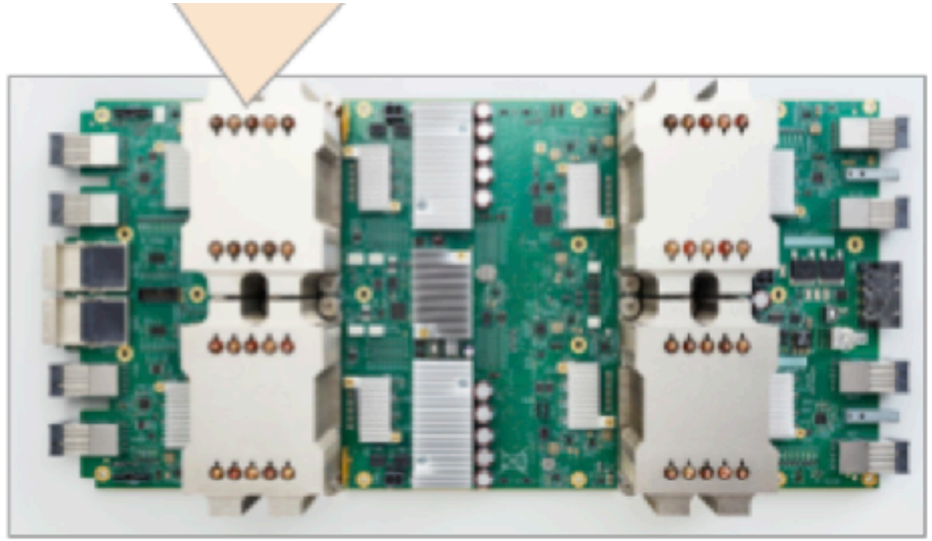
GPU	Tesla T4	Tesla V100	Tesla P100
Max Flops (GFLOPS)	8073.26	14016.50	9322.46
Fast Fourier Transform (GFLOPS)	660.05	2301.32	1510.49
Matrix Multiplication (GFLOPS)	3290.94	13480.40	8793.33
Molecular Dynamics (GFLOPS)	572.91	997.61	480.02
S3D (GFLOPS)	99.42	434.78	295.20



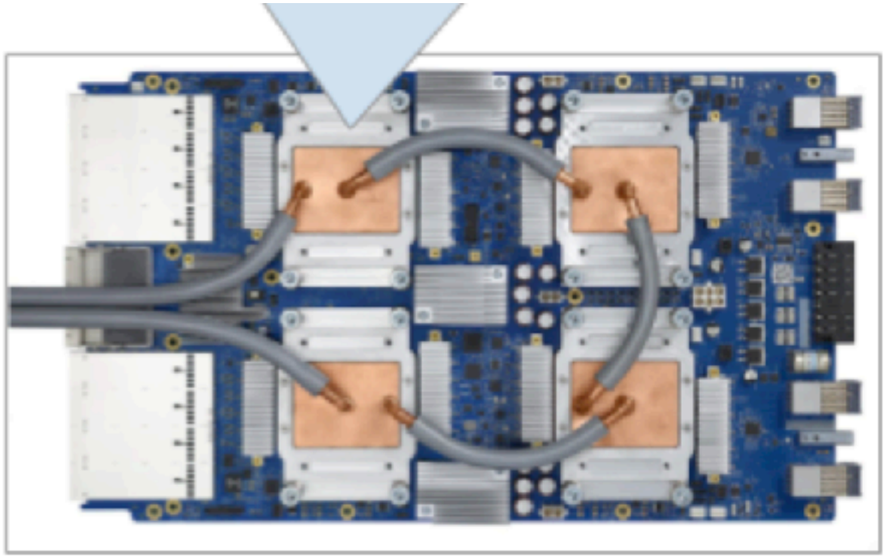
SPECIFICATIONS

GPU Architecture	**NVIDIA Turing**
NVIDIA Turing Tensor Cores	**320**
NVIDIA CUDA® Cores	**2,560**
Single-Precision	**8.1 TFLOPS**
Mixed-Precision (FP16/FP32)	**65 TFLOPS**
INT8	**130 TOPS**
INT4	**260 TOPS**
GPU Memory	**16 GB GDDR6 300 GB/sec**
ECC	**Yes**
Interconnect Bandwidth	**32 GB/sec**
System Interface	**x16 PCIe Gen3**
Form Factor	**Low-Profile PCIe**
Thermal Solution	**Passive**
Compute APIs	**CUDA, NVIDIA TensorRT™, ONNX**

Google's Tensor Processing Units



TPU v2 - 4 chips, 2 cores per chip



TPU v3 - 4 chips, 2 cores per chip

Each TPU core has scalar, vector, and matrix units (MXU). The MXU provides the bulk of the compute power in a TPU chip. Each MXU is capable of performing 16K multiply-accumulate operations in each cycle. While the MXU inputs and outputs are 32-bit floating point values, the MXU performs multiplies at reduced bfloat16 precision. Bfloat16 is a 16-bit floating point representation that provides better training and model accuracy than the IEEE half-precision representation.

https://cloud.google.com/tpu/docs/system-architecture

EdgeTPU

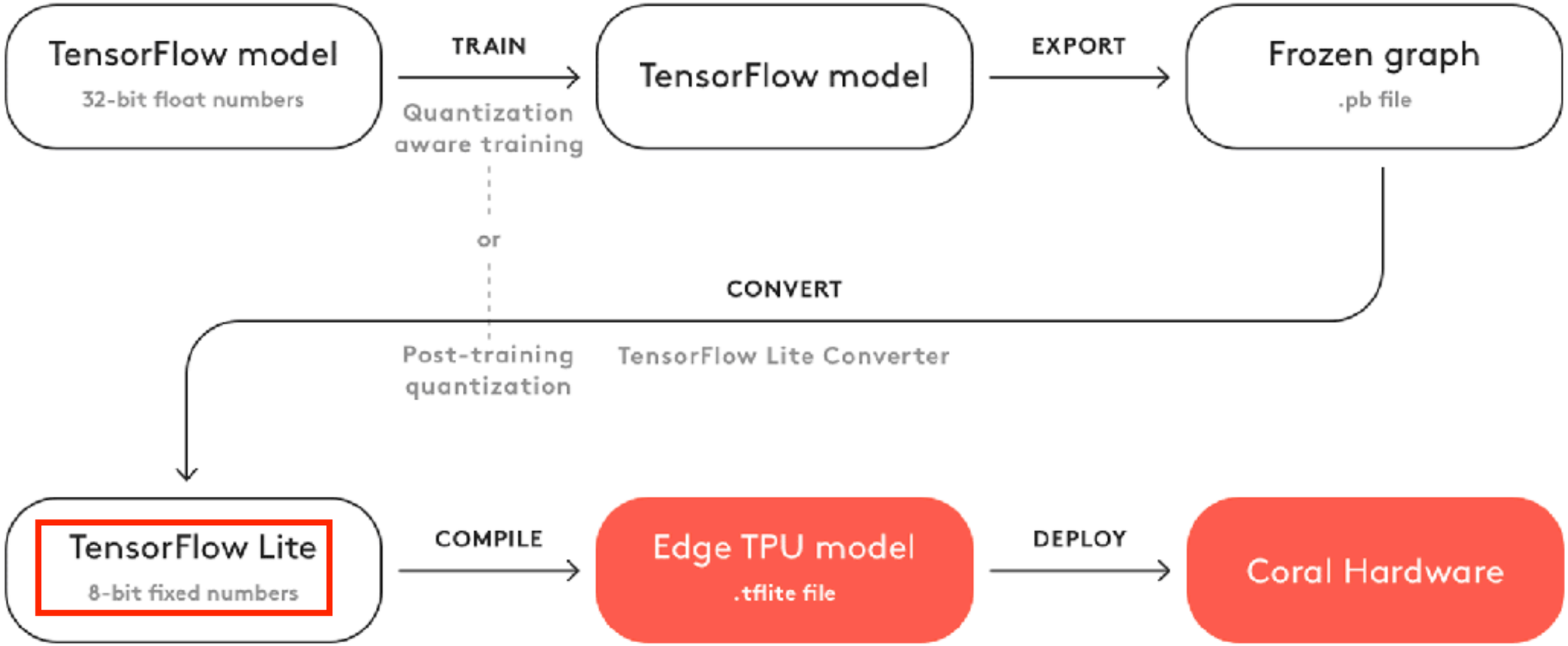


Figure 1. The basic workflow to create a model for the Edge TPU

Announcement

- Reading quiz 5 due 4/28 **BEFORE** the lecture
  - Under iLearn > reading quizzes
- Lab 3 due 4/30
  - Watch the video and read the instruction BEFORE your session
  - There are links on both course webpage and iLearn lab section
  - Submit through iLearn > Labs
- Midterm on 5/7 during the lecture time, access through iLearn — no late submission is allowed — make sure you will be able to take that at the time
- Check your grades in iLearn

Electrical
Computer Science
Engineering

120A

つづく