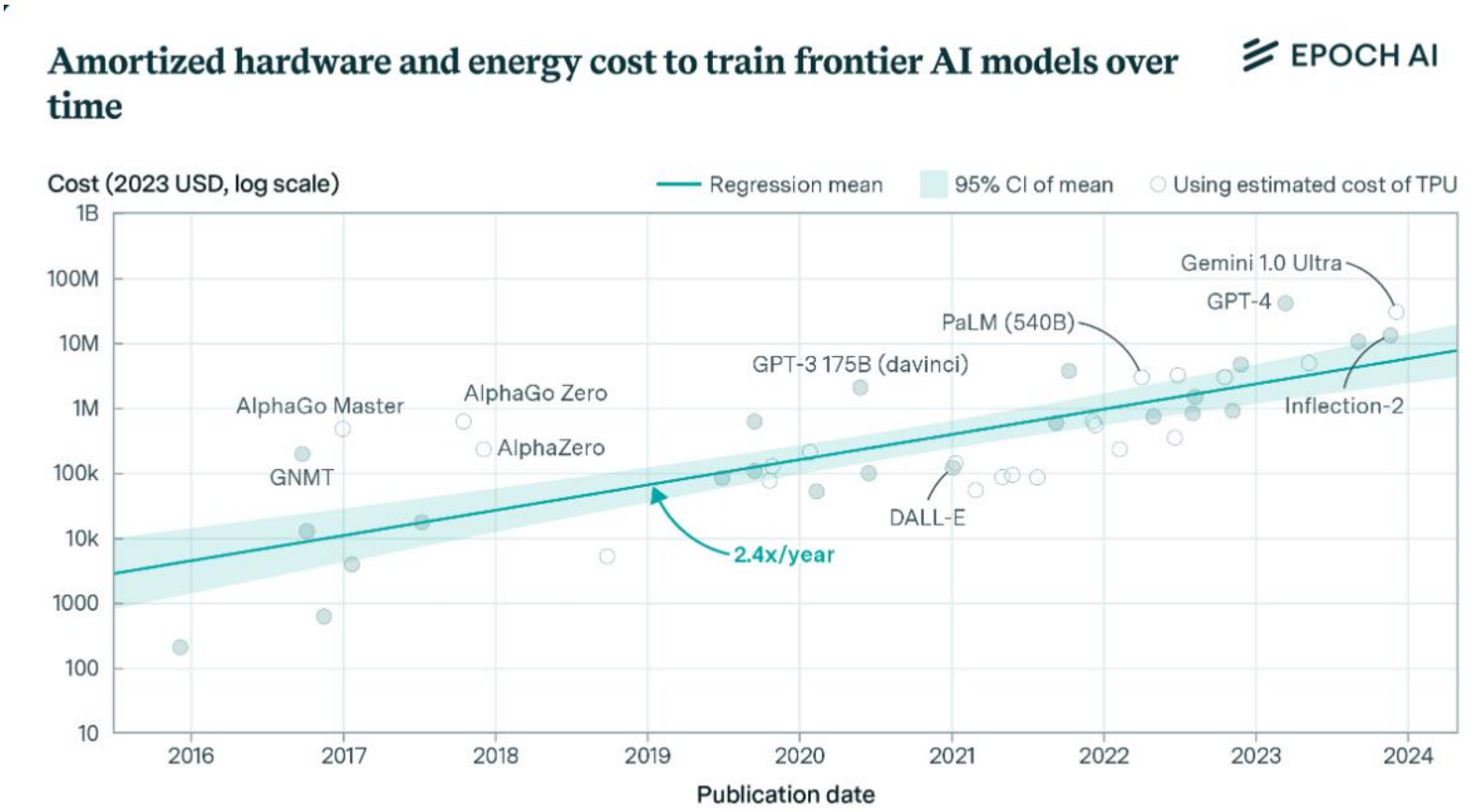
METODE INTELIGENTE DE REZOLVARE A PROBLEMELOR REALE

GPUs details

Laura Dioşan

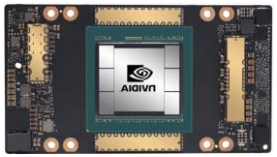
Costs for training DL models



- Check more details here [How Much Does It Cost to Train Frontier AI Models? | Epoch AI]

GPUs

FLOP = "Floating Point Operation"; one addition, multiplication, etc
TFLOP = 1 trillion FLOPs (10^{12})



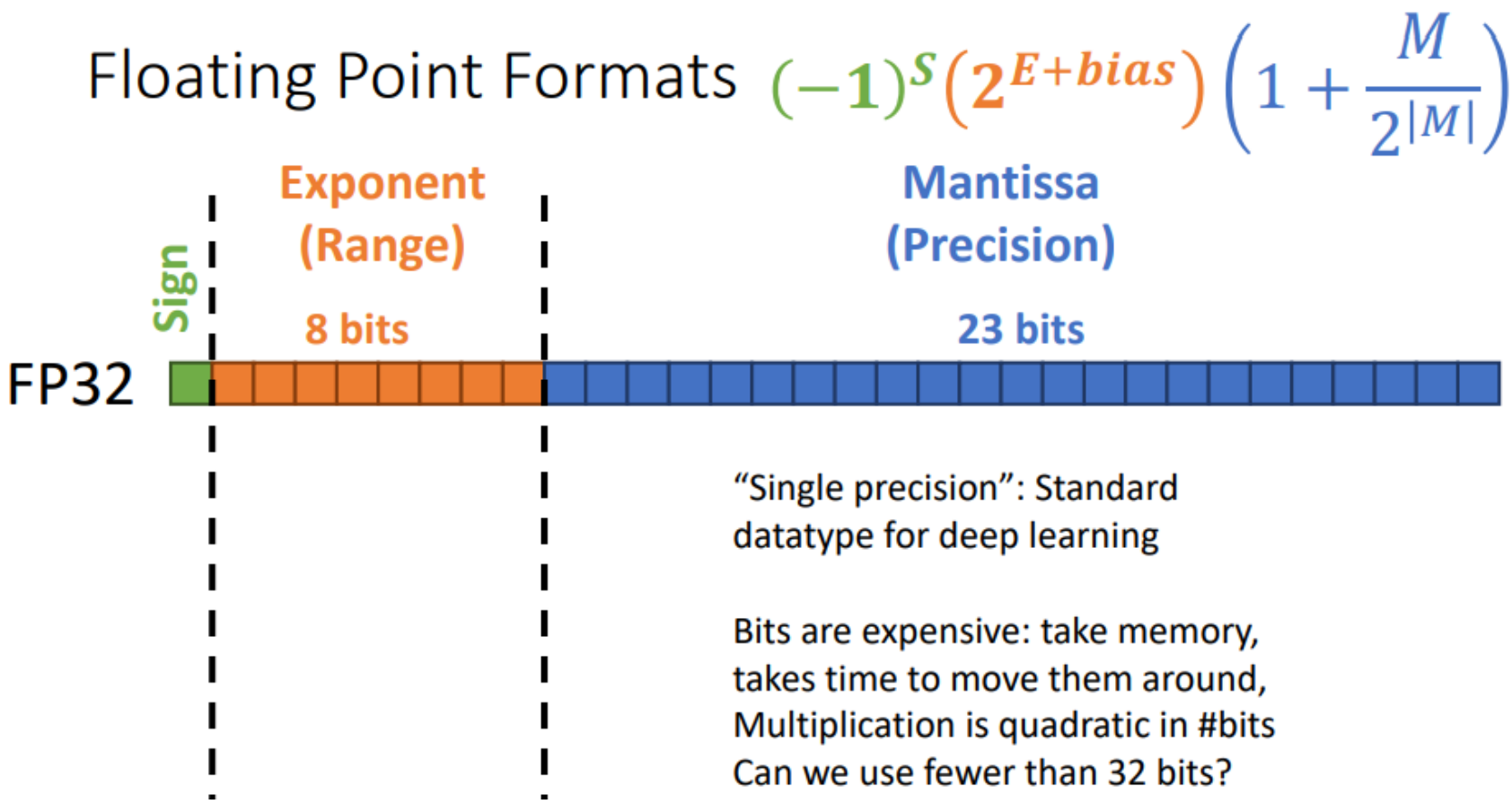
A100
- Memory
  - Capacity 40GB HBM2
  - Bandwidth 1.5 TB/sec
- Compute
  - FP64 9.7 TFLOPS/sec
  - FP32 19.5 TFLOPS/sec
  - BF16 39 FLOPS/sec
  - FP1678 TFLOPS/sec
- Tensor Cores
  - TF32: 156 TFLOPS/sec
  - FP16/BF16: 312 TFLOPS/sec

L40S
- Memory
  - Capacity 40GB HBM2
  - Bandwidth 1.6 TB/sec
- Compute
  - FP64 not supported
  - FP32 90 TFLOPS/sec
  - BF16 700 TFLOPS/sec
  - FP16 700 TFLOPS/sec
- Tensor Cores
  - TF32: 366 TFLOPS/sec
  - FP16/BF16: 733 TFLOPS/sec

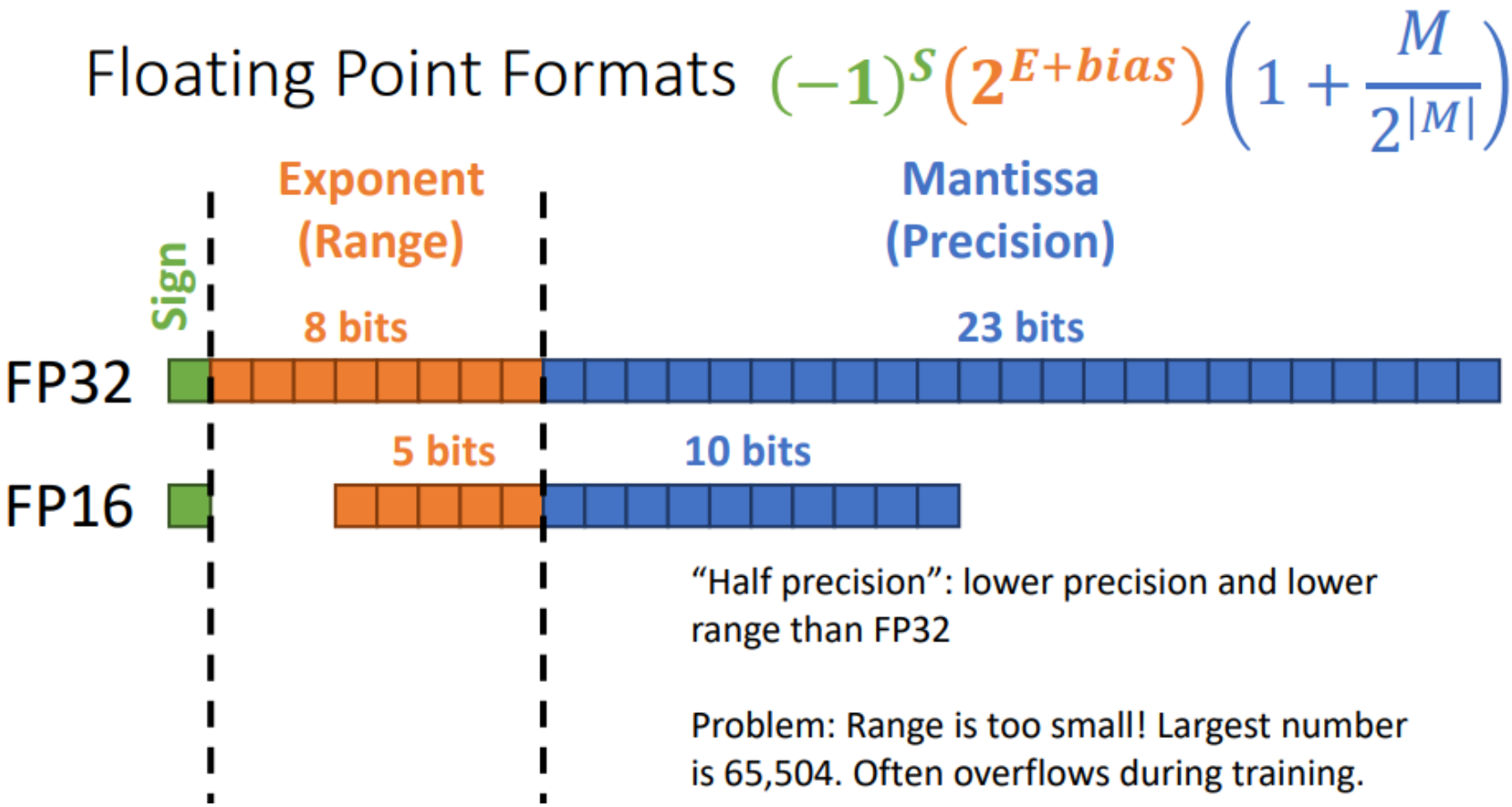
H100
- Memory
  - Capacity 40/80GB HBM2
  - Bandwidth 3.0 TB/sec
- Compute
  - FP64 30 TFLOPS/sec
  - FP32 60 TFLOPS/sec
  - BF16 120 TFLOPS/sec
  - FP16 120 TFLOPS/sec
- Tensor Cores
  - TF32: 500 TFLOPS/sec
  - FP16/BF16: 1000 TFLOPS/sec

H200
- Memory
  - Capacity 80GB HBM2
  - Bandwidth 2 TB/sec
- Compute
  - FP64 34 TFLOPS/sec
  - FP32 67 TFLOPS/sec
  - BF16 2000 TFLOPS/sec
  - FP16 2000 TFLOPS/sec
- Tensor Cores
  - TF32: 989 TFLOPS/sec
  - FP16/BF16: 1979 TFLOPS/sec

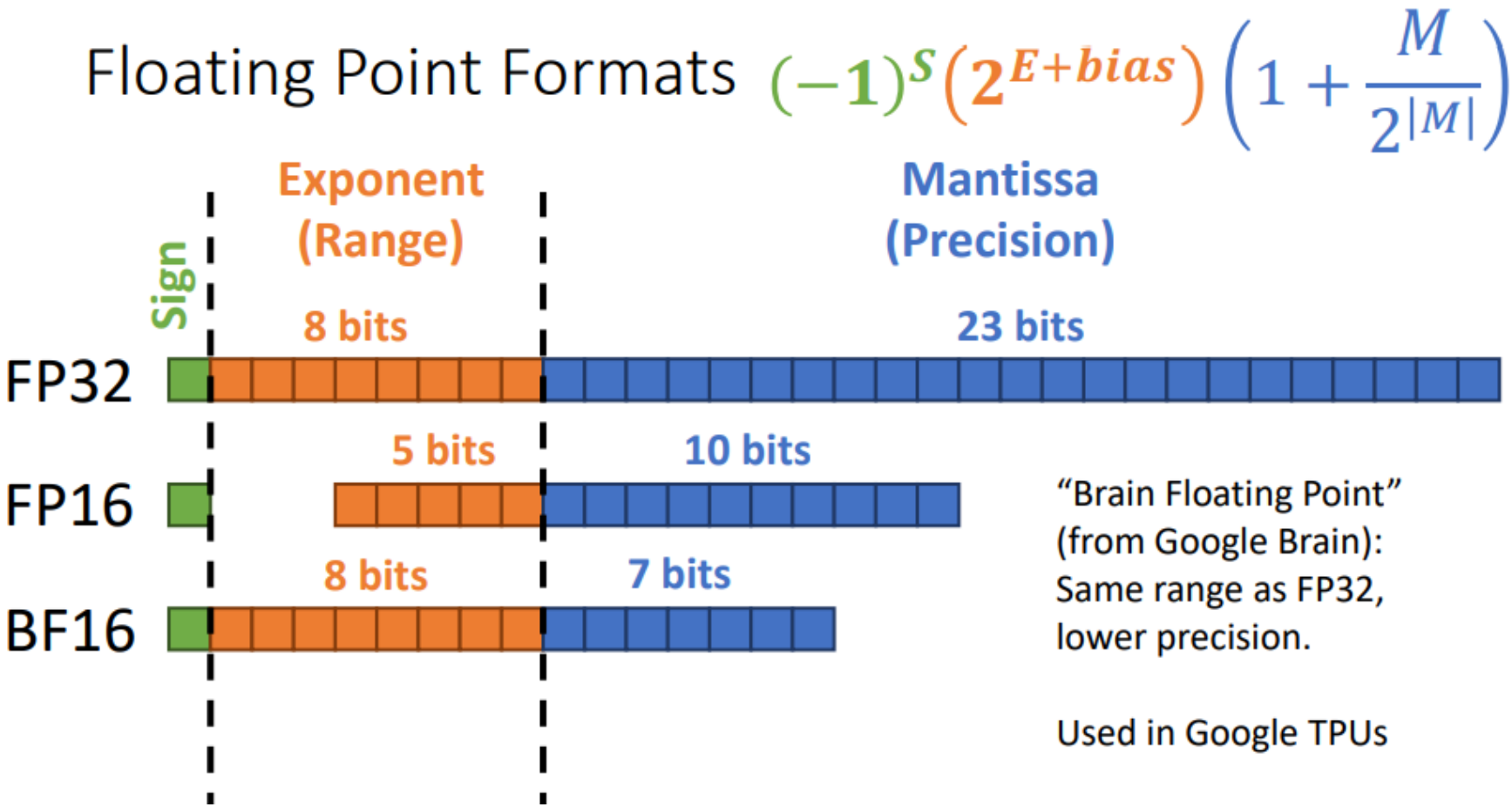
Floating Point details



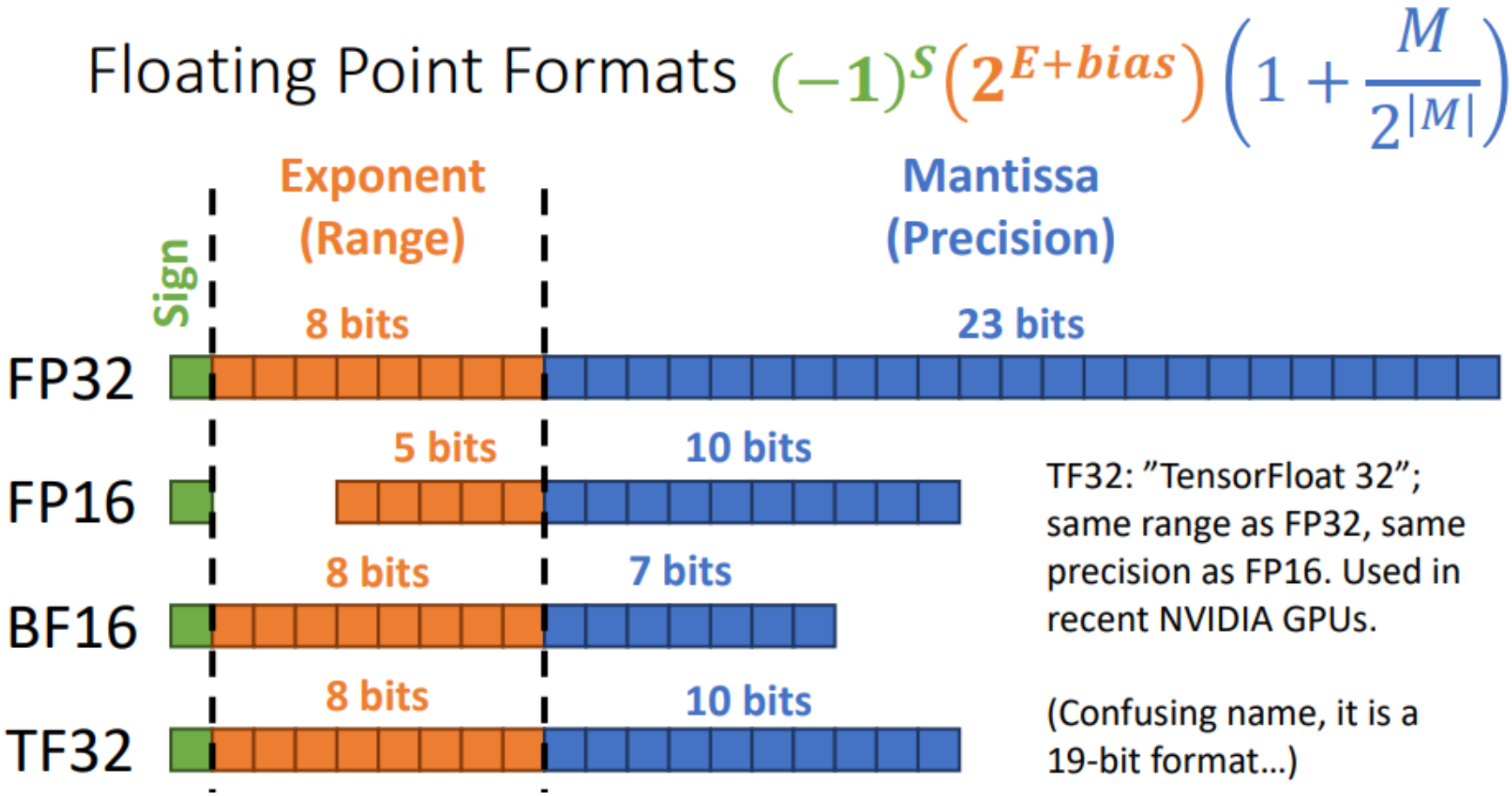
Floating Point details



Floating Point details



Floating Point details



Floating Point details

Mixed Precision

We often need to compute dot products (for matrix multiply, convolution, etc):

$$y = x_1 w_1 + x_2 w_2 + \cdots + x_n w_n$$

Multiplication is more expensive than addition
Idea: Multiply in low precision, add in high precision

Inputs: x_i, w_i in low precision (FP16, BF16, TF32)
Output: y in high precision (FP32)

$$y = FP32(x_1 w_1) + FP32(x_2 w_2) + \cdots + FP32(x_n w_n)$$

Tensor Cores in NVIDIA GPUs are special hardware for mixed-precision matrix multiplication with different low-precision formats (TF32, BF16 best for neural nets)