Mixed-Precision in Graphics Processing Units

Quentin Gallouédec

École Centrale de Lyon

Thursday, March 12, 2020

Introduction

- Using lower accuracy saves power, memory usage and speed

Introduction

- Using lower accuracy saves power, memory usage and speed
- Reducing precison means rouding error

Introduction

- Using lower accuracy saves power, memory usage and speed
- Reducing precison means rouding error

Main question

How to limit the loss of accuracy due to the use of reduced precision ?

Outline

Outline

IEEE 754 : Standard for Floating-Point Arithmetic

Example of single precision floating point format.

b_{31}	b_{30}	$\cdots$	b_{24}	b_{23}	$\cdots$	b_0
X	X	$\cdots$	X	X	$\cdots$	X
sign (1 bit)	exponent (8 bits)			fraction (23 bits)		

Table 1: Representation of `float32` format

IEEE 754 : Standard for Floating-Point Arithmetic

Example of single precision floating point format.

b_{31}	b_{30}	$\cdots$	b_{24}	b_{23}	$\cdots$	b_0
X	X	$\cdots$	X	X	$\cdots$	X
sign (1 bit)	exponent (8 bits)			fraction (23 bits)		

Table 1: Representation of `float32` format

$$\mathcal{F}^{32} : \mathbb{B}^{32} \longrightarrow \bar{\mathbb{R}}$$

$$(b_i)_{\mathbb{N}_{<32}} \mapsto \begin{cases} s \times 2^{-127} \times f & \text{if} \quad e = 0 \\ s \times 2^{e-127} \times (1+f) & \text{if} \quad e \in [\![1, 254]\!] \\ s \times +\infty & \text{if} \quad e = 255 \end{cases}$$

Outline

What is mixed-precision ?

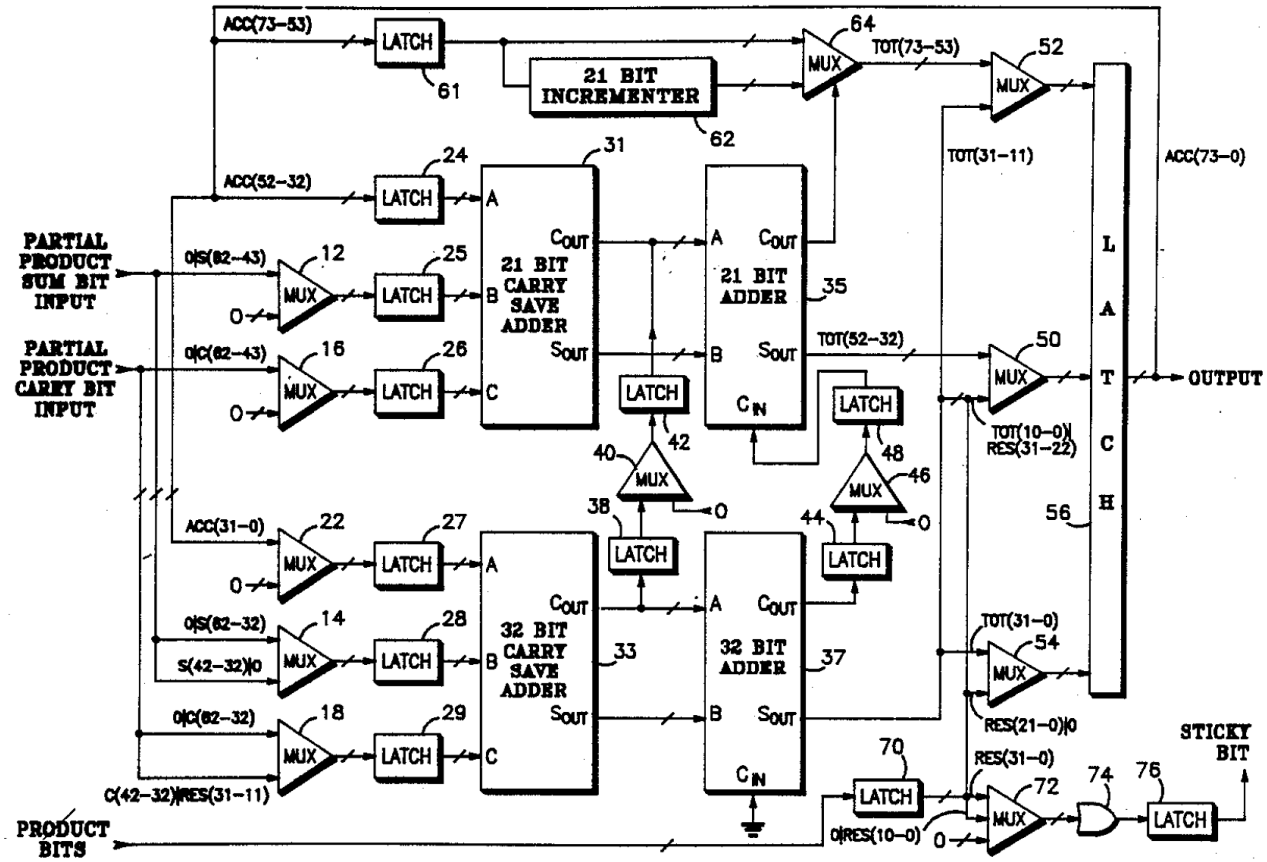


Figure 1: First implementation of the mixed-precision proposed by [Denman Jr et al., 1990]

What is mixed-precision ?

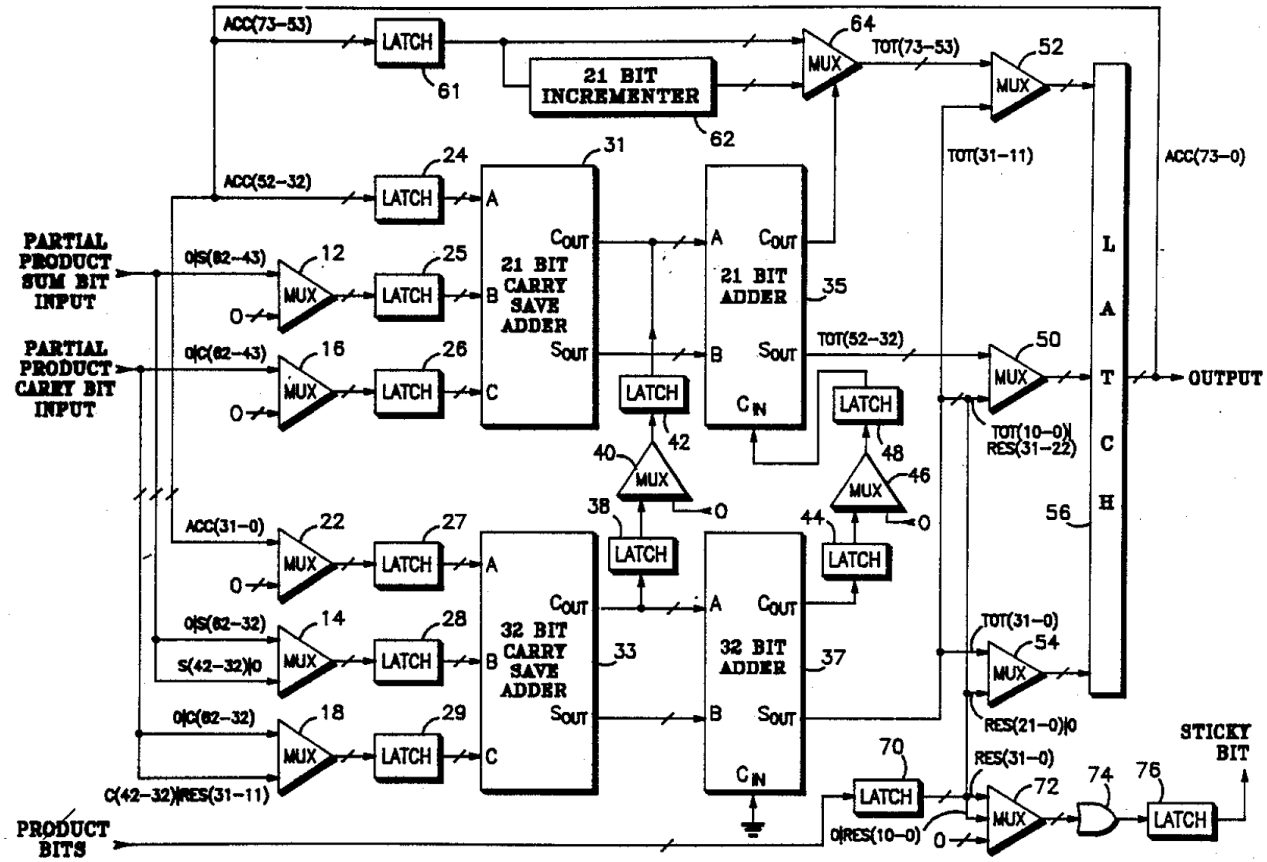


Figure 1: First implementation of the mixed-precision proposed by [Denman Jr et al., 1990]

Use lower precision to speed up calculations

Outline

Outline

Which applications can benefit from mixed-precision ?

Using fixed-point numbers with stochastic rounding

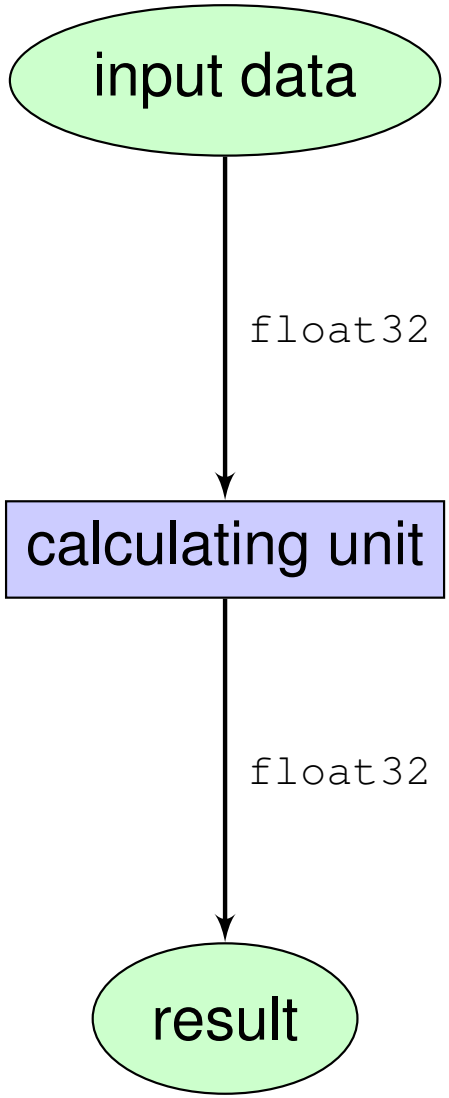


Figure 2: Fixed-point computation workflow

Which applications can benefit from mixed-precision ?

Using fixed-point numbers with stochastic rounding

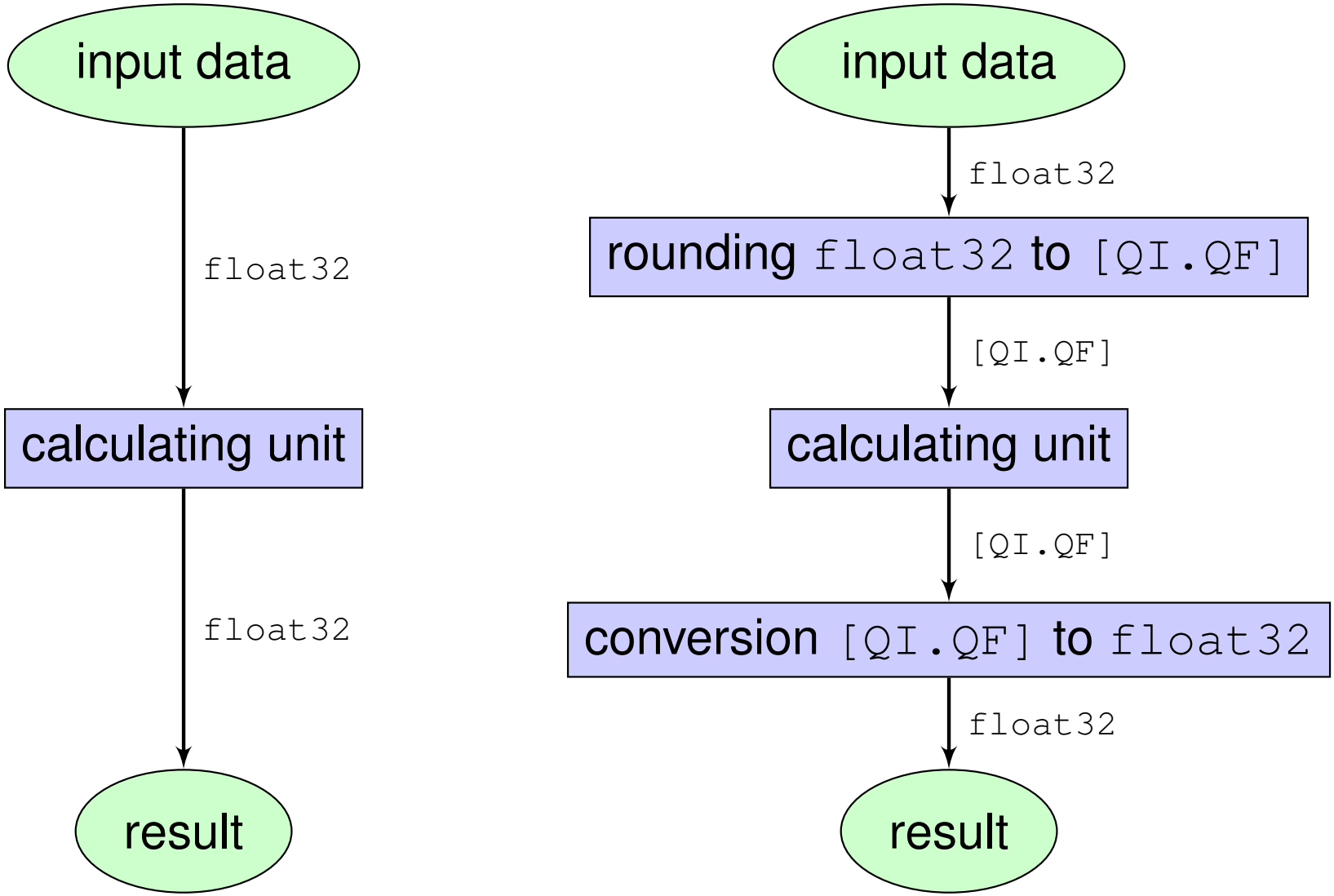


Figure 2: Fixed-point computation workflow

Which applications can benefit from mixed-precision ?

Using fixed-point numbers with stochastic rounding

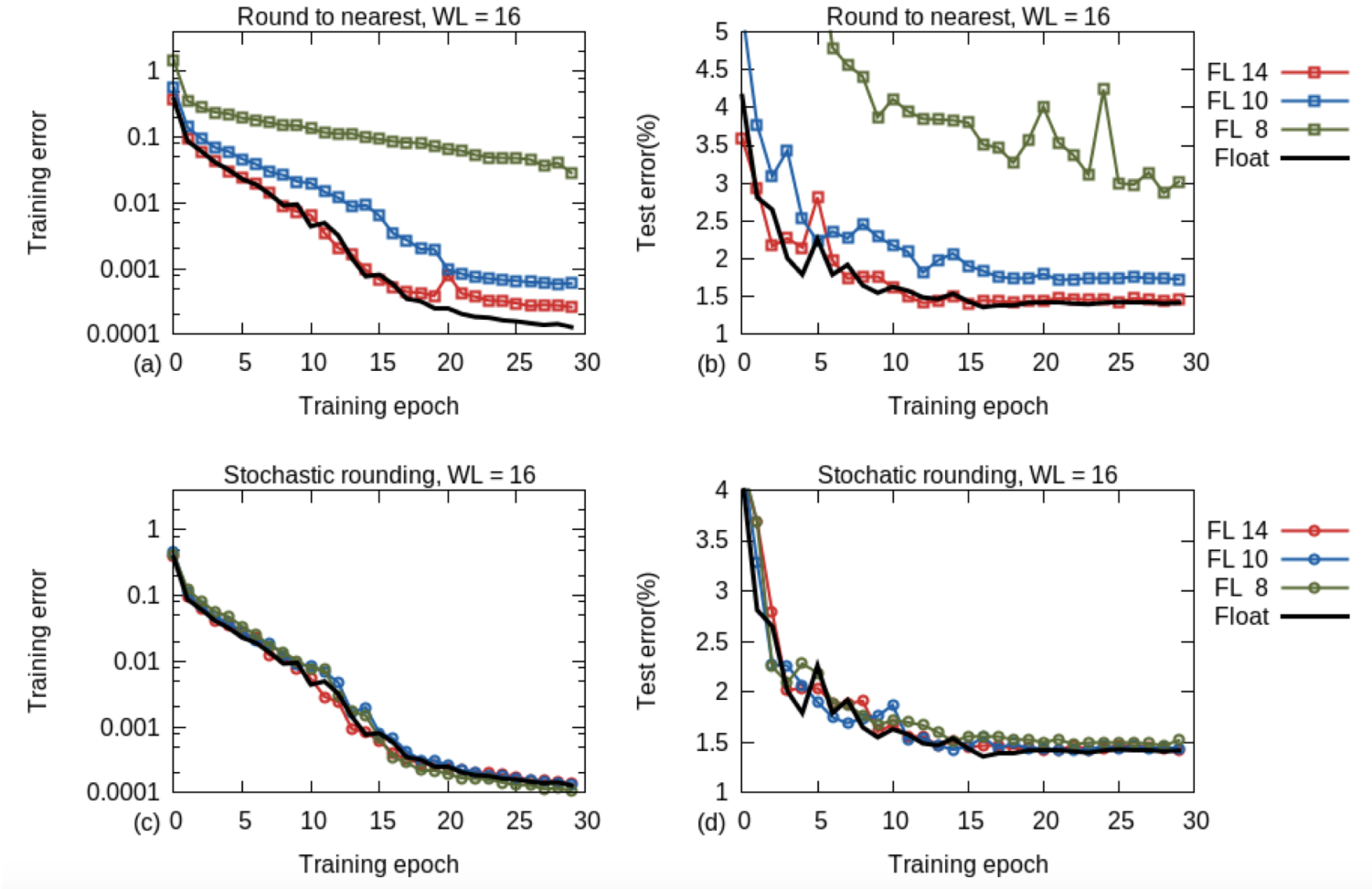


Figure 3: Compute fixed-point instead of floating-point : MNIST dataset using fully connected DNNs [Gupta et al., 2015]

Outline

Which applications can benefit from mixed-precision ?

NVIDIA GPU Tensor Core

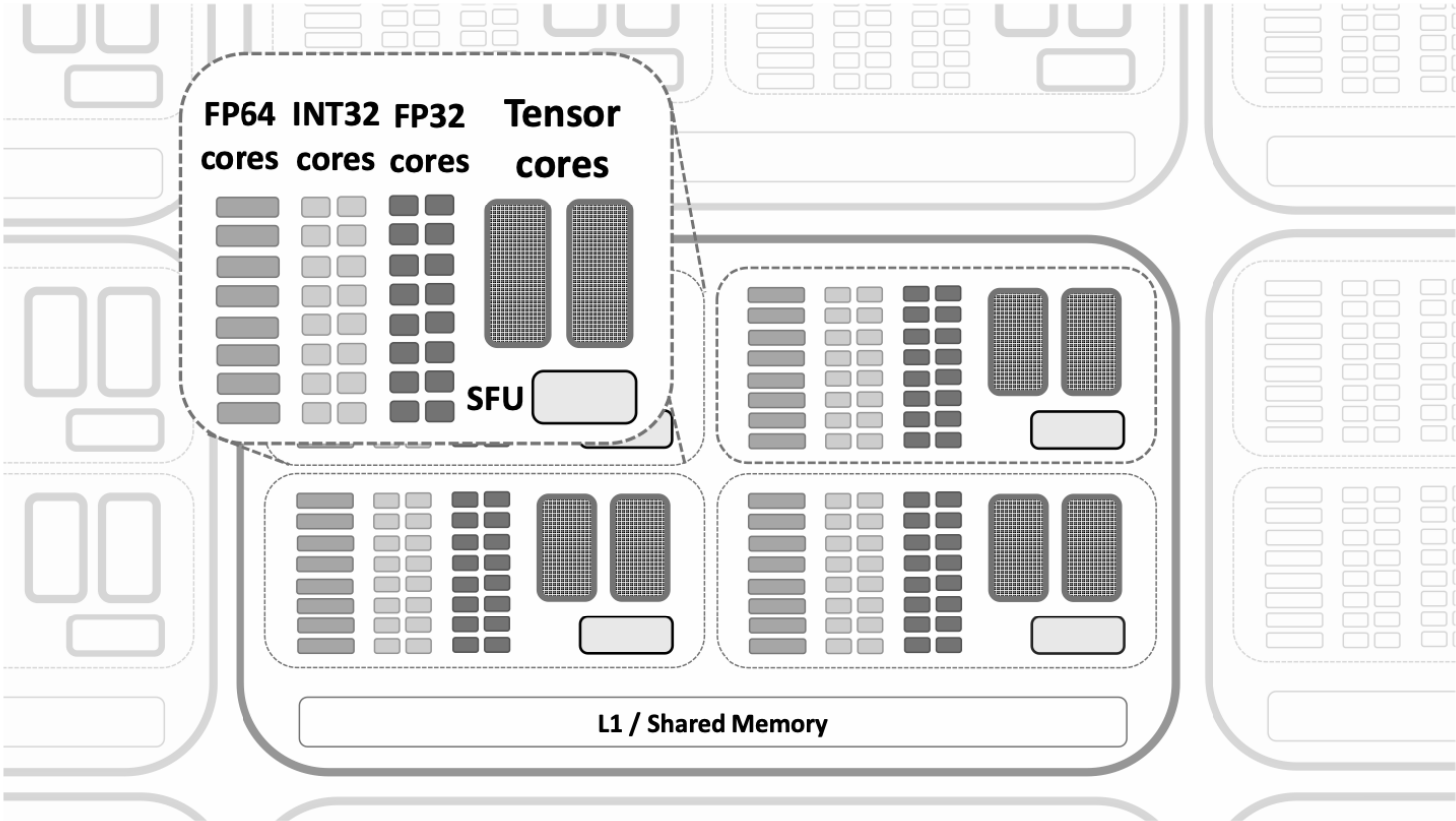


Figure 4: Simplified diagram of the Volta SM architecture. The NVIDIA Tesla V100 uses 80 SMs [Markidis et al., 2018]. In practice , NVIDIA Tesla V100 was able to deliver up to 83 Tflops/s in mixed-precision.

Which applications can benefit from mixed-precision ?

NVIDIA GPU Tensor Core

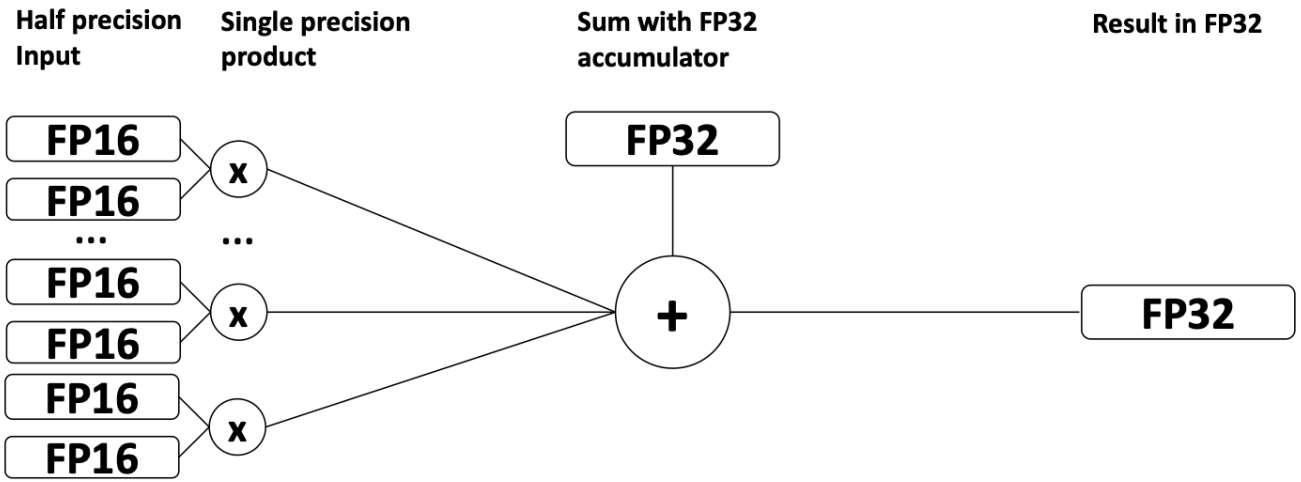


Figure 5: FMA (Fused-Multiply-Add) operation used in the NVIDIA Tesla V100 [Markidis et al., 2018].

Which applications can benefit from mixed-precision ?

NVIDIA GPU Tensor Core

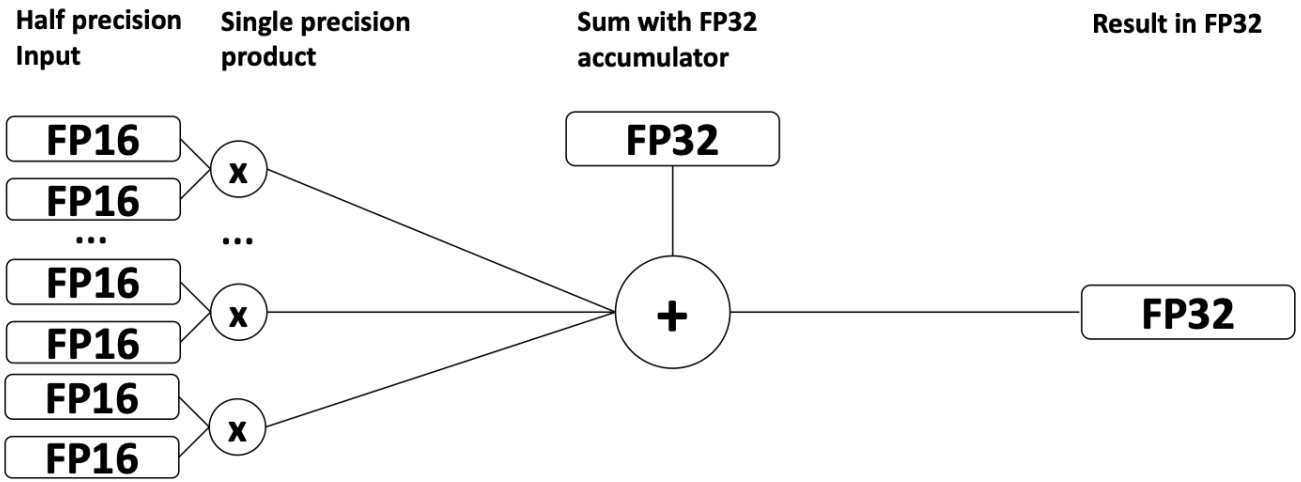


Figure 5: FMA (Fused-Multiply-Add) operation used in the NVIDIA Tesla V100 [Markidis et al., 2018].

`float16`	`float32`	`float64`	mixed-precision
31.4	15.7	7.8	125

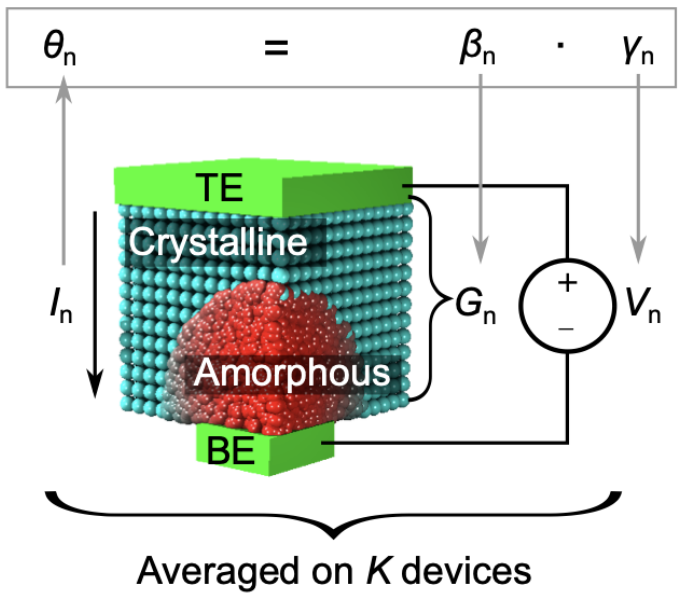
Table 2: Tesla V100 accelerator : theoretical maximum performance (Tflops/s)

Outline

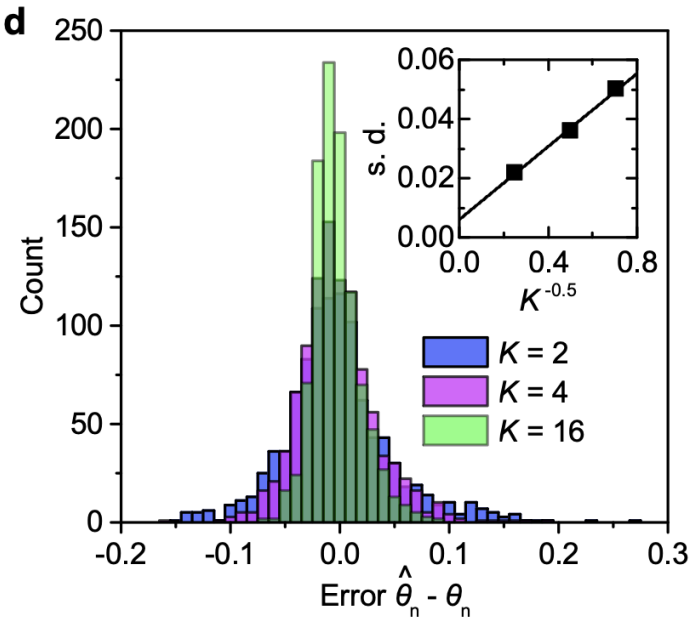
Which applications can benefit from mixed-precision ?

In-memory mixed-precision

Coarse part of a calculation realized "in-memory"



(a) Phase-Change Memory device for a scalar multiplication



(b) Distribution of scalar multiplication error using 1024 K PCM devices

Figure 6: Mixed-precision in-memory computing : scalar multiplication [Gallo et al., 2017]

Outline

Conclusion

- mixed-precision is faster, more energy efficient.
- mixed-precision is a real advantage, and should be required in many areas (eg. ML, Modeling).
- mixed-precision can be implemented at different levels of abstraction.
- NVIDIA already markets its tensor cores that enable mixed-precision.

Bibliography


Denman Jr, M. A., Young, J. M., and Alsup, M. K. (1990).
Circuit and method for accumulating partial products of a single, double or mixed precision multiplication.
US Patent 4,893,268.


Gallo, M. L., Sebastian, A., Mathis, R., Manica, M., Tuma, T., Bekas, C., Curioni, A., and Eleftheriou, E. (2017).
Mixed-precision memcomputing.
CoRR, abs/1701.04279.


Gallouédec, Q. (2020).
Mixed-precision in graphic processing units.


Gupta, S., Agrawal, A., Gopalakrishnan, K., and Narayanan, P. (2015).
Deep learning with limited numerical precision.
CoRR, abs/1502.02551.


Markidis, S., Chien, S. W. D., Laure, E., Peng, I. B., and Vetter, J. S. (2018).
NVIDIA tensor core programmability, performance & precision.
CoRR, abs/1803.04014.