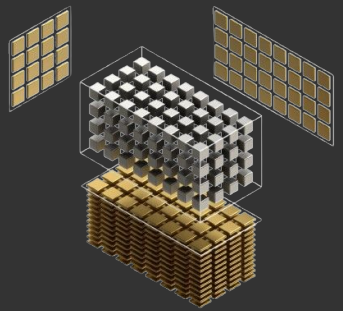
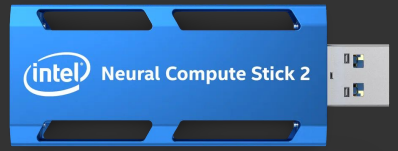
Deep Learning Tools for Edge Devices





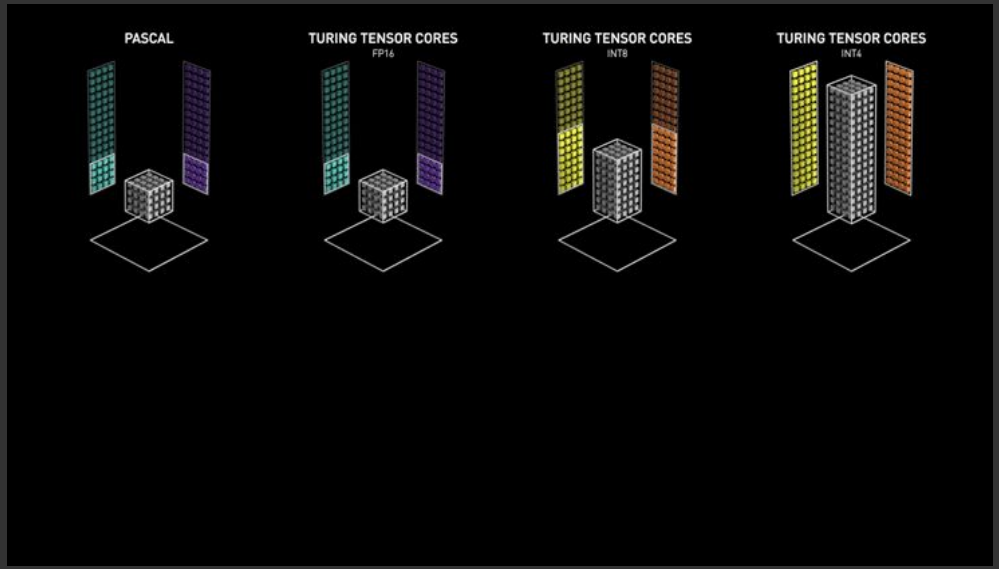
Lecture 2: Tensor Cores

What are Tensor Cores?

Dedicated Specialized Hardware Accelerator Cores

Used for Matrix Multiplication and Convolution Operations

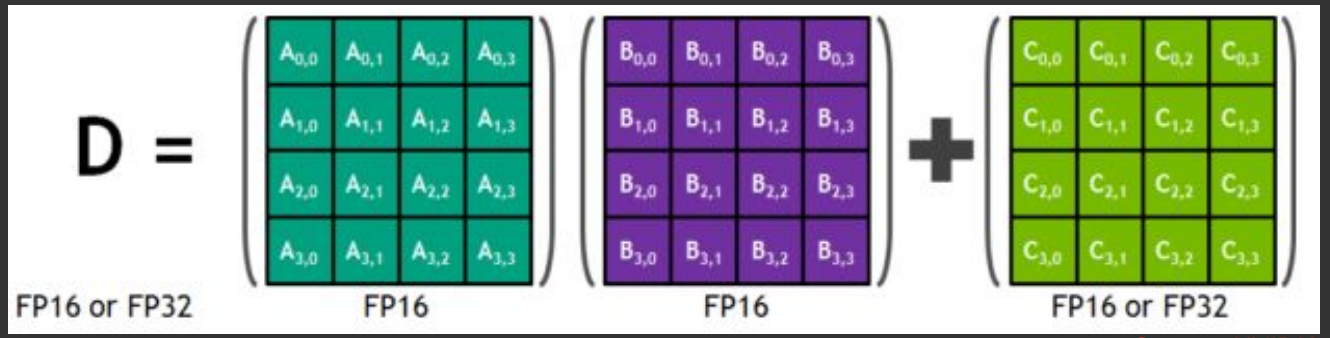
Capable of Performing 4x4 matrix multiplication in one go



Source: NVIDIA

SAXPY

Stands for Single Precision A.X + Y



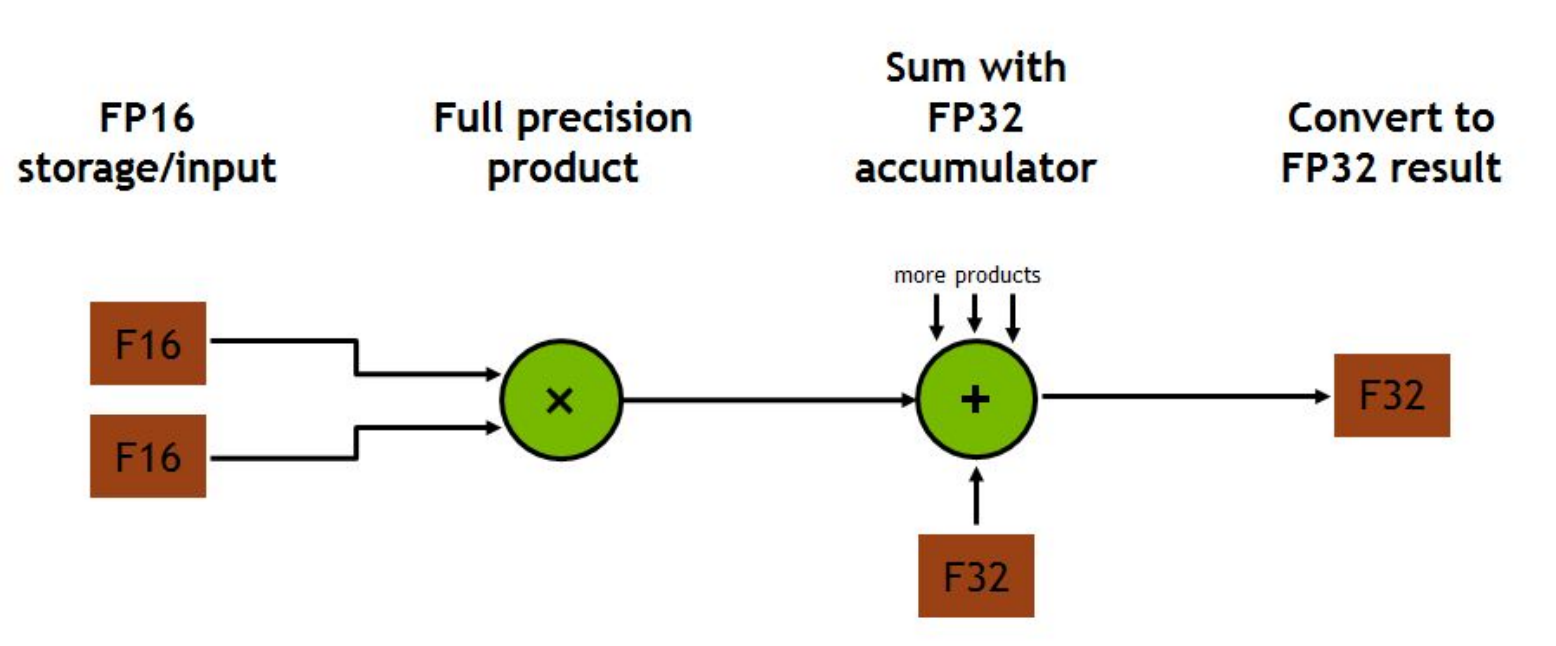
Source: NVIDIA

A combination of scalar multiplication and vector/matrix addition

4x4 Matrix Multiplication -> 64 Scalar multiplications

Requires 64 Cuda core vs 1 Tensor core

Automatic Mixed Precision



Source: NVIDIA

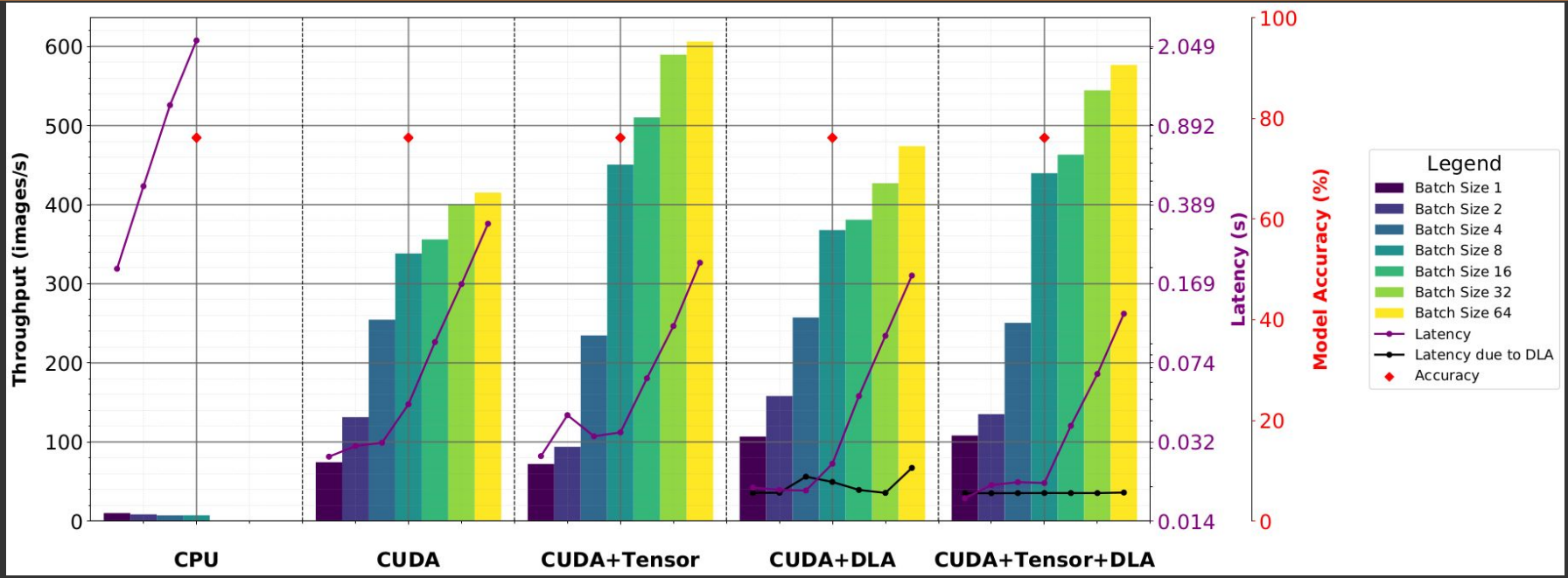
Example

Now, let us consider an example,

Model:	ResNet-50
Framework:	Pytorch
Precision:	FP16, FP32, AMP
Batch Sizes:	1, 2, 4, 8, 16, 32, 64
Concurrency:	2

Evaluating Multi-Instance DNN Inferencing on Multiple Accelerators of an Edge Device, by M. Tayal, Y. Simmhan
(IEEE International Conference on High Performance Computing, Data, and Analytics)

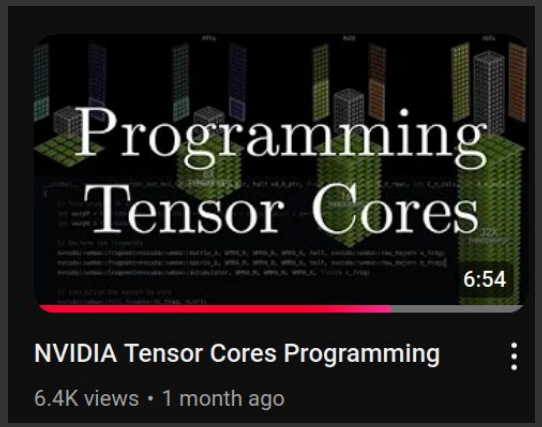
Example Results



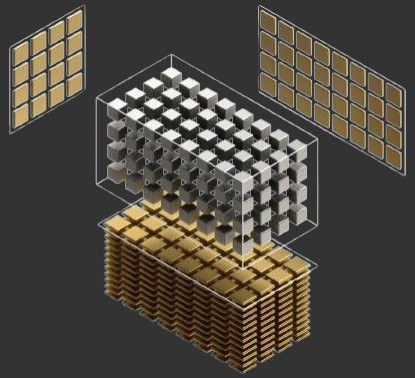
Evaluating Multi-Instance DNN Inferencing on Multiple Accelerators of an Edge Device, by M. Tayal, Y. Simmhan
(IEEE International Conference on High Performance Computing, Data, and Analytics)

Any References?

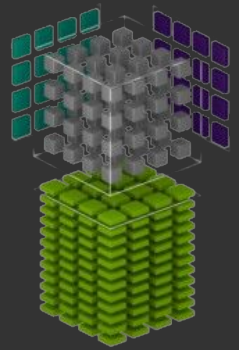
Useful Articles and Videos



YT Channel: 0Mean1Sigma [Video]



Article: Programming Tensor Cores in CUDA 9 [Article]



Article: Tips for Optimizing GPU Performance Using Tensor Cores [Article]

Thank You