

Research Project

Acceleration of non-rigid image registration with Tensor Cores

Jonathan LEVY

June 14, 2019

Outline

- 1 My cursus
- 2 Research Project proposal

About me

- Jonathan LEVY
- MSc student in Computer Science
- Wide background in Engineering

Cursus Summary

- *Classe Préparatoire PTSI/PT**
- Ecole Normale Supérieure de Rennes (BSc, Master in Teaching)
- *Agrégation* in Engineering, CS track
- MSc Embedded Systems, TU Delft

Since September 2019:
GASAL2 : GPU-accelerated library for DNA alignment

When First as Extra Project, then MSc Thesis

Languages C/C++ and CUDA

Algorithm Smith-Waterman - optimal alignment for short pair

Goal: integrate in the *Burrough-Wheeler Aligner*, "**BWA**"

<https://github.com/j-levy/GASAL2>

<https://github.com/j-levy/bwa-gasal2> ← private repository

Acceleration of non-rigid image registration with Tensor Cores

- Image registration: aligning a *floating* image with a *reference*.
- *Non-rigid*: various deformations allowed
- Use GPU for parallel calculation

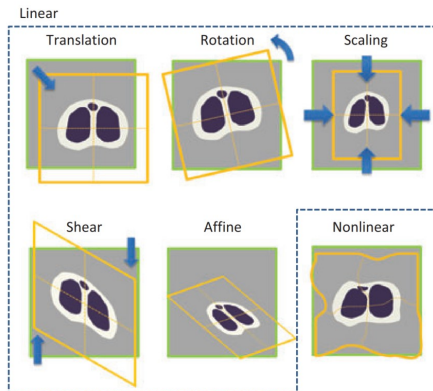


Figure 1: Different types of deformation.

The Volta Architecture

- NVIDIA GPUs' architecture (2017)
- Several changes:
 - HBM2 memory
 - Parallel FP/Integer calculation
 - Tensor Cores



Figure 2: The full GV100 architecture

The Volta Architecture

- NVIDIA GPUs' architecture (2017)
- Several changes:
 - HBM2 memory
 - Parallel FP/Integer calculation
 - Tensor Cores

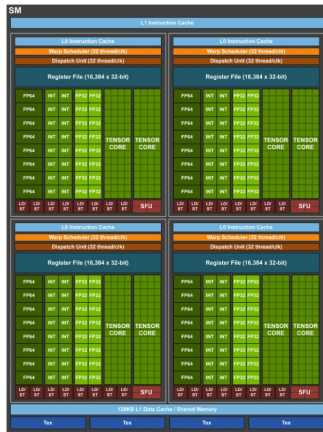


Figure 2: Volta Streaming Multiprocessor (80 units per GV100)

The Volta Architecture

- NVIDIA GPUs' architecture (2017)
- Several changes:
 - HBM2 memory
 - Parallel FP/Integer calculation
 - Tensor Cores



Figure 2: Volta Compute Module (4 units per SM)

Tensor Cores

WHAT Matrix-matrix multiplication

HOW Mixed precision

WHY Deep Learning

$$\mathbf{D} = \begin{pmatrix} A_{0,0} & A_{0,1} & A_{0,2} & A_{0,3} \\ A_{1,0} & A_{1,1} & A_{1,2} & A_{1,3} \\ A_{2,0} & A_{2,1} & A_{2,2} & A_{2,3} \\ A_{3,0} & A_{3,1} & A_{3,2} & A_{3,3} \end{pmatrix} \begin{pmatrix} B_{0,0} & B_{0,1} & B_{0,2} & B_{0,3} \\ B_{1,0} & B_{1,1} & B_{1,2} & B_{1,3} \\ B_{2,0} & B_{2,1} & B_{2,2} & B_{2,3} \\ B_{3,0} & B_{3,1} & B_{3,2} & B_{3,3} \end{pmatrix} + \begin{pmatrix} C_{0,0} & C_{0,1} & C_{0,2} & C_{0,3} \\ C_{1,0} & C_{1,1} & C_{1,2} & C_{1,3} \\ C_{2,0} & C_{2,1} & C_{2,2} & C_{2,3} \\ C_{3,0} & C_{3,1} & C_{3,2} & C_{3,3} \end{pmatrix}$$

FP16 or FP32 FP16 FP16 FP16 or FP32

Figure 3: Operation done by a Tensor Core