Extending Chameleon to tensors contraction

Working group
internship presentation

Brieuc NICOLAS

Summary

Inria

01

Background Information

Inria

Education

- CPGE - Maths/Physics/Chemistry
  - > Marcellin Berthelot, Paris region
- Enseirb - Matmeca
  - > CISD option

Internships

- 1st year - Inria
  - > Participated in implementing and benchmarking Mixed Precision in PaStiX
- 2nd year - ICL (Tennessee)
  - > Evaluating the performance of PaRSEC's schedulers
  - > Worked on features inside DPLASMA

Inria

02

Tensors in Chameleon

Inria

Data

- Rework or create new data descriptors to support more than 2 dimensions

Operations (focus contraction)

- Efficient use of data descriptors to allow efficient matricization/reordering/folding allowing:
- Efficient use of optimized lvl 1-3 BLAS and 2D linear algebra routines

Validation

- Many frameworks so we need to choose one (Pros/Cons)
  - > Tensorflow, Pytorch, Numpy, Scipy, Cupy
- Most are in Python so we need a Python Interface for Chameleon

Objective for the internship

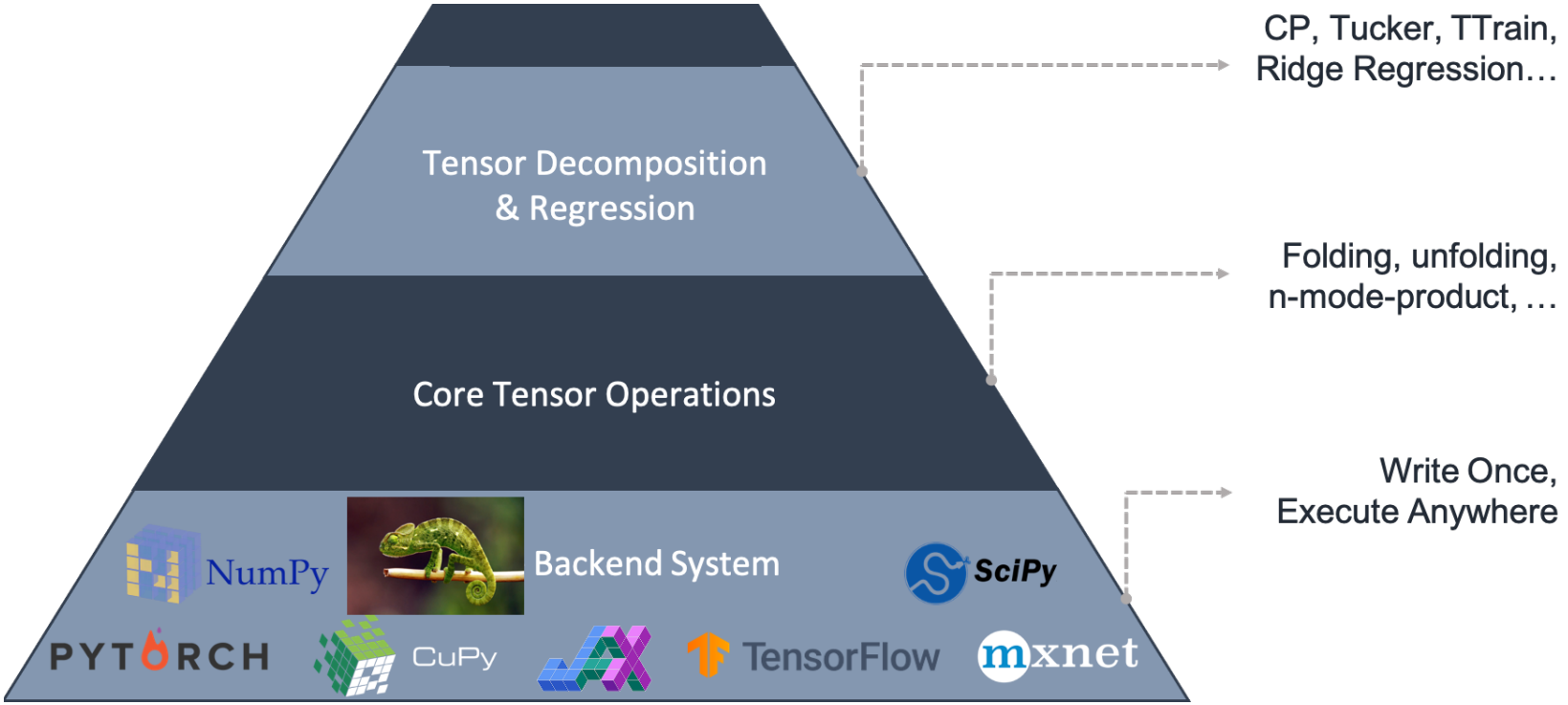

Figure: Pyramidal representation of the internship's current objective

Objective for the internship

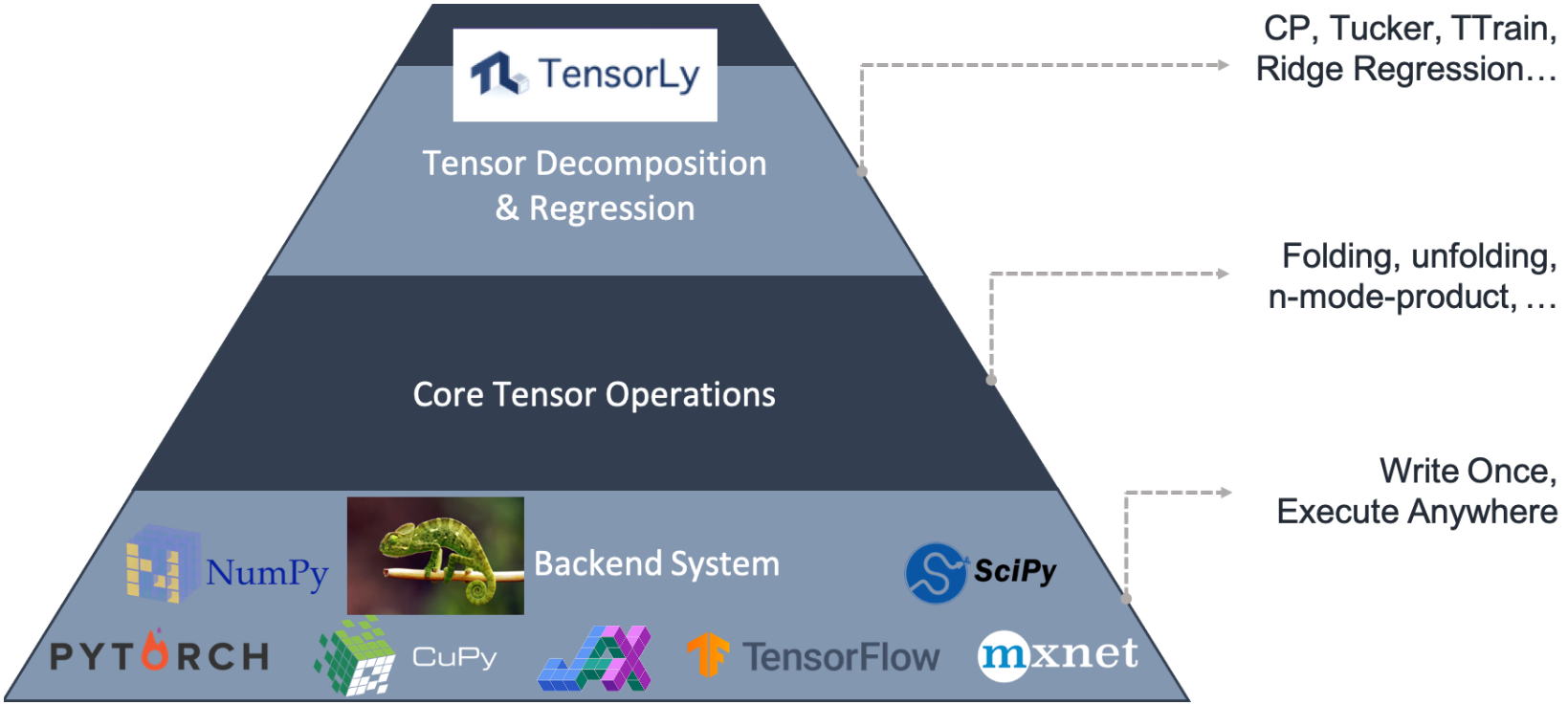

Figure: Pyramidal representation of the internship's current objective

Current progress

- Selection of the framework : Tensorly
- Python API for Chameleon:
  - > Well underway thanks to PaStiX wrapper generator

To be done soon

- Plug Chameleon into Tensorly for 2D operations
- Extend to multidimensional arrays : Chameleon algorithms and interface

Future Work

In conjunction with Ana, half/mixed-precision tensors

Inria