



Inference engine for custom neural networks with oneAPI

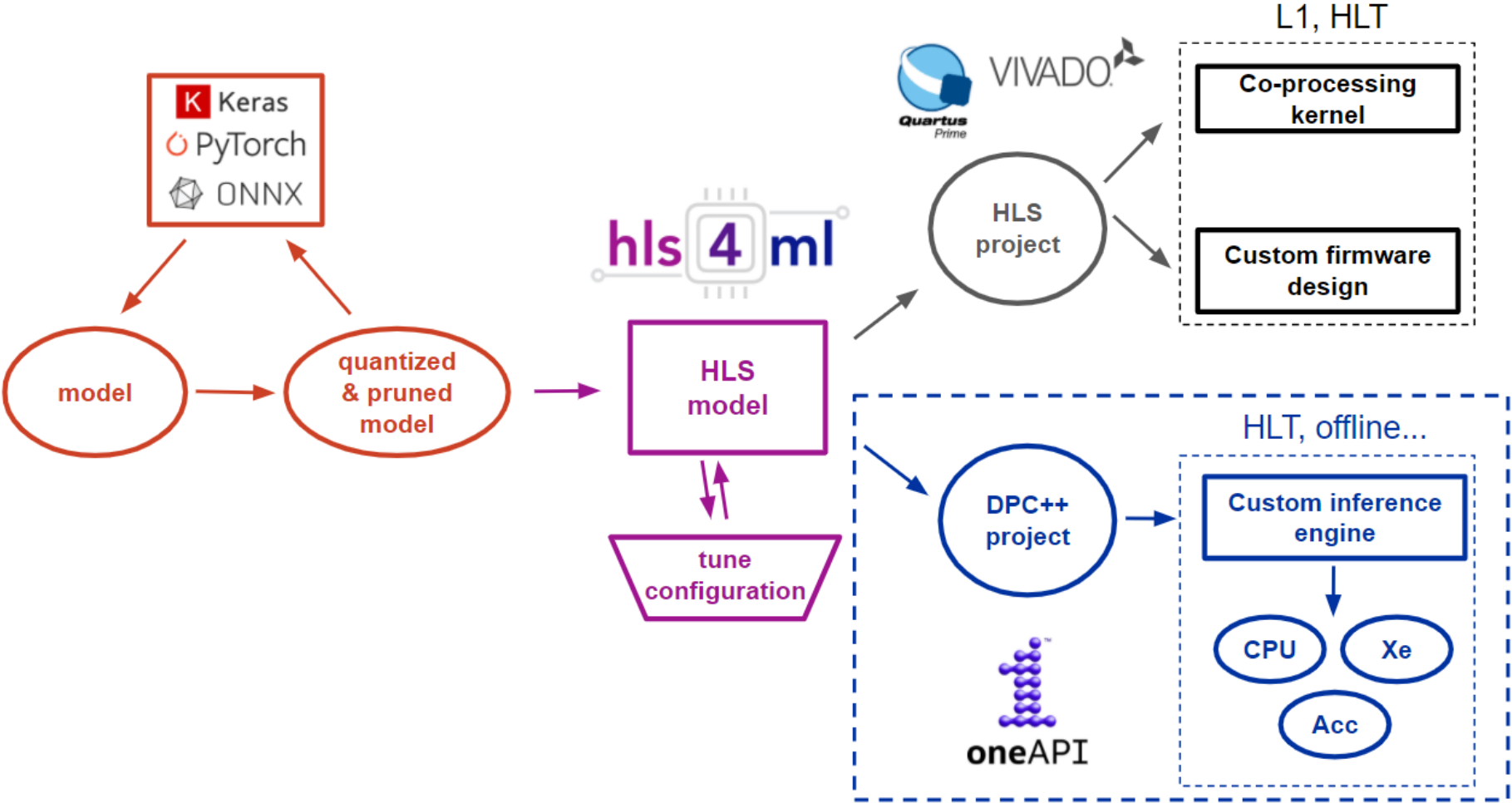
Marcin Świniarski

24 / 09 / 2020

Agenda

- Quick introduction to the project;
- Tools used in the project;
- Results;
- Summary.

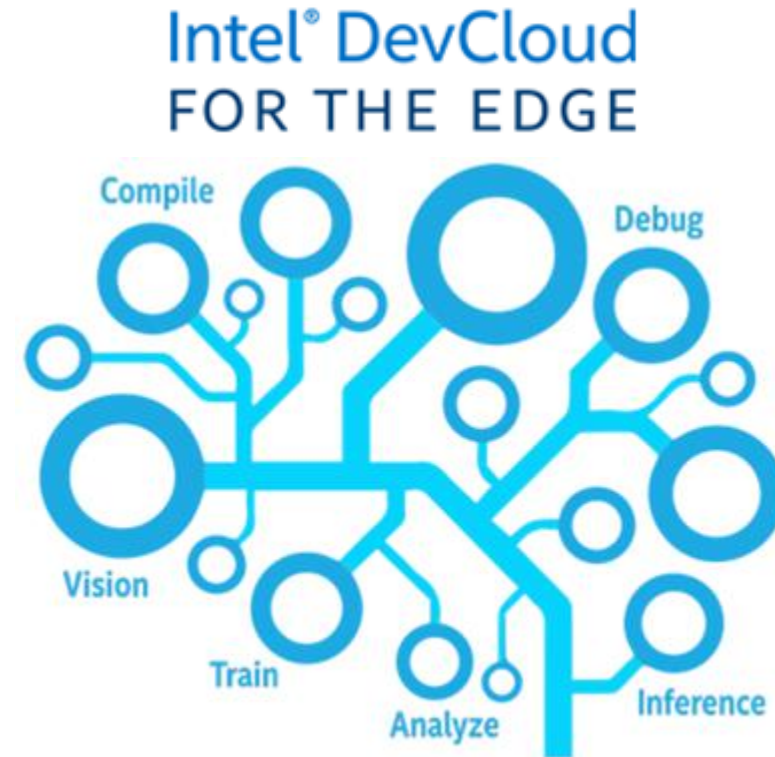
Quick introduction to the project



Tools used in the project

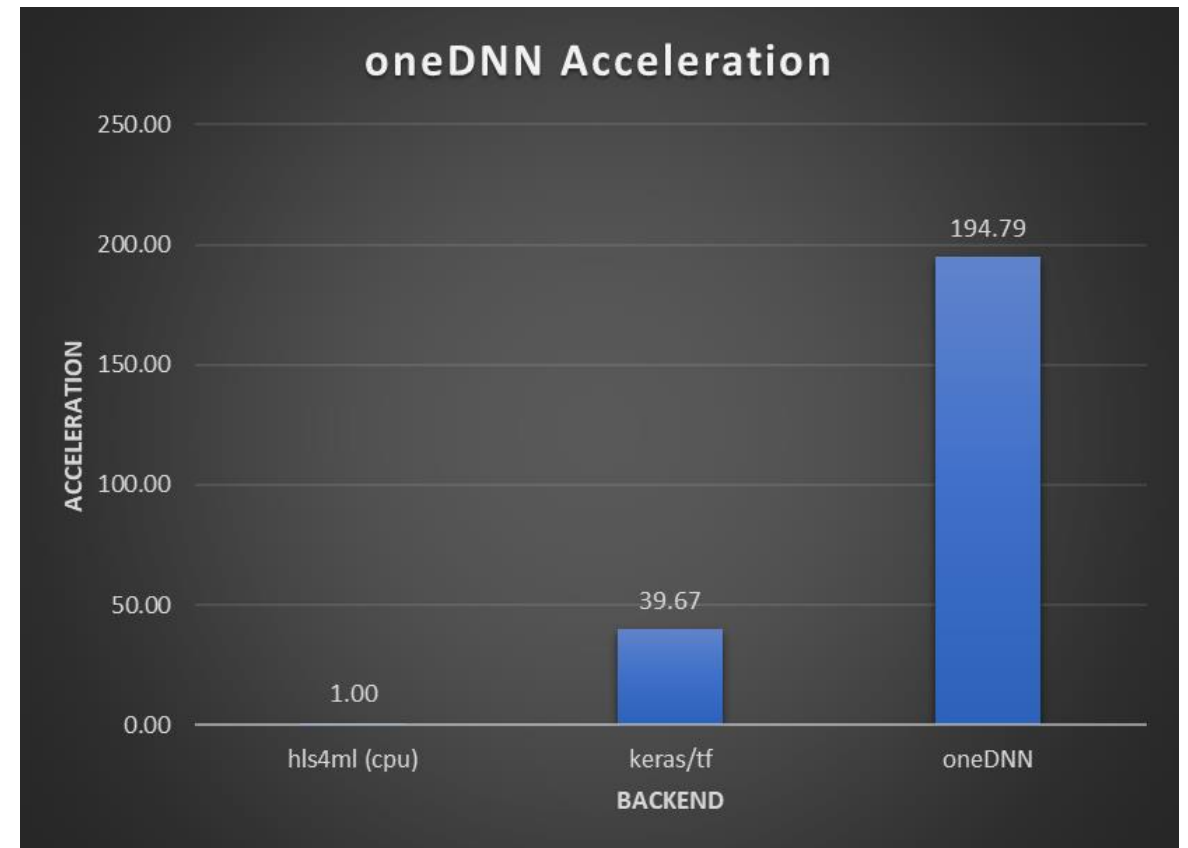
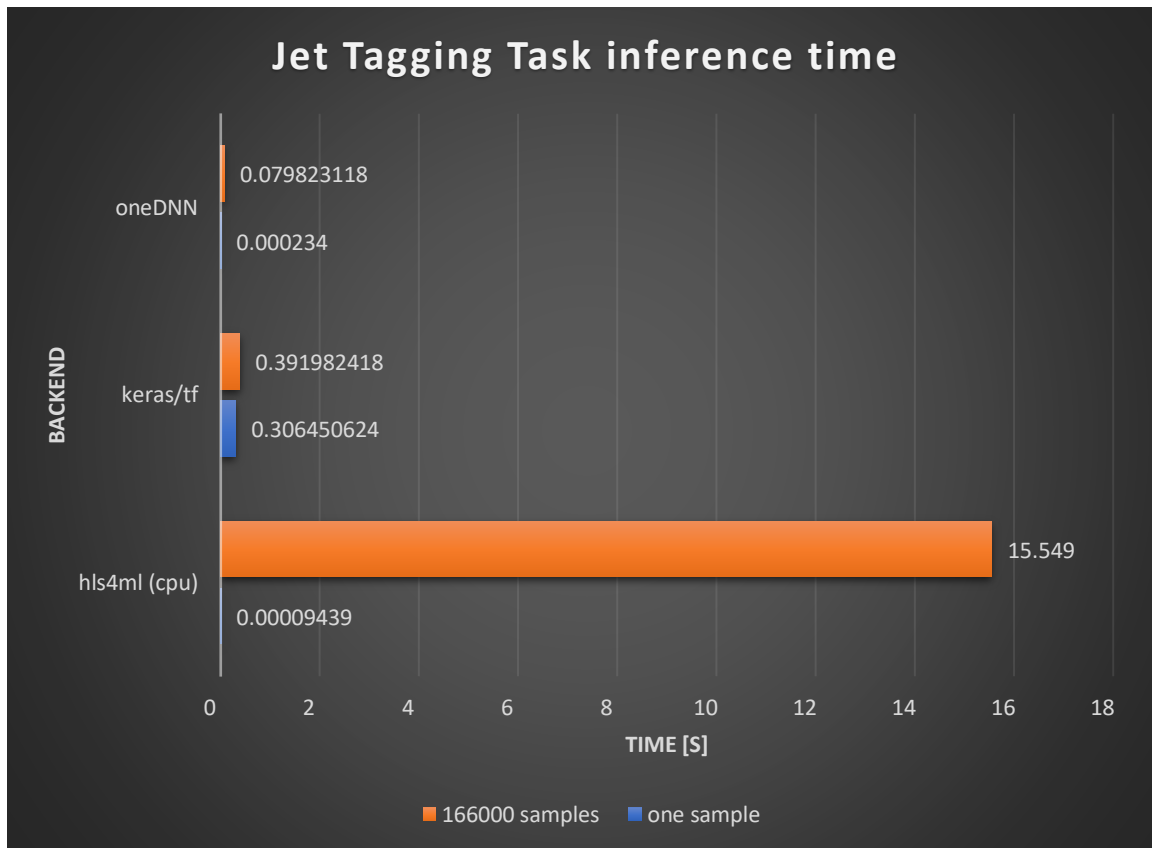


Intel oneAPI Deep Neural Network Library - oneDNN

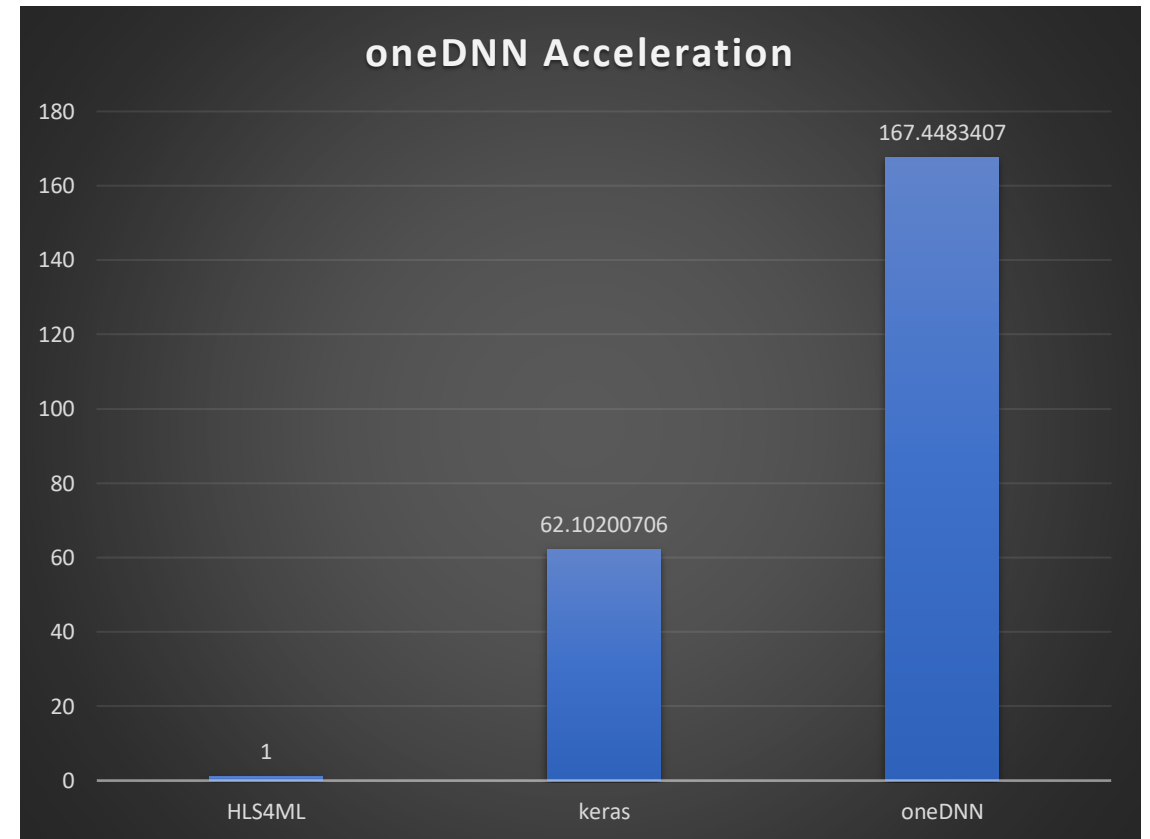
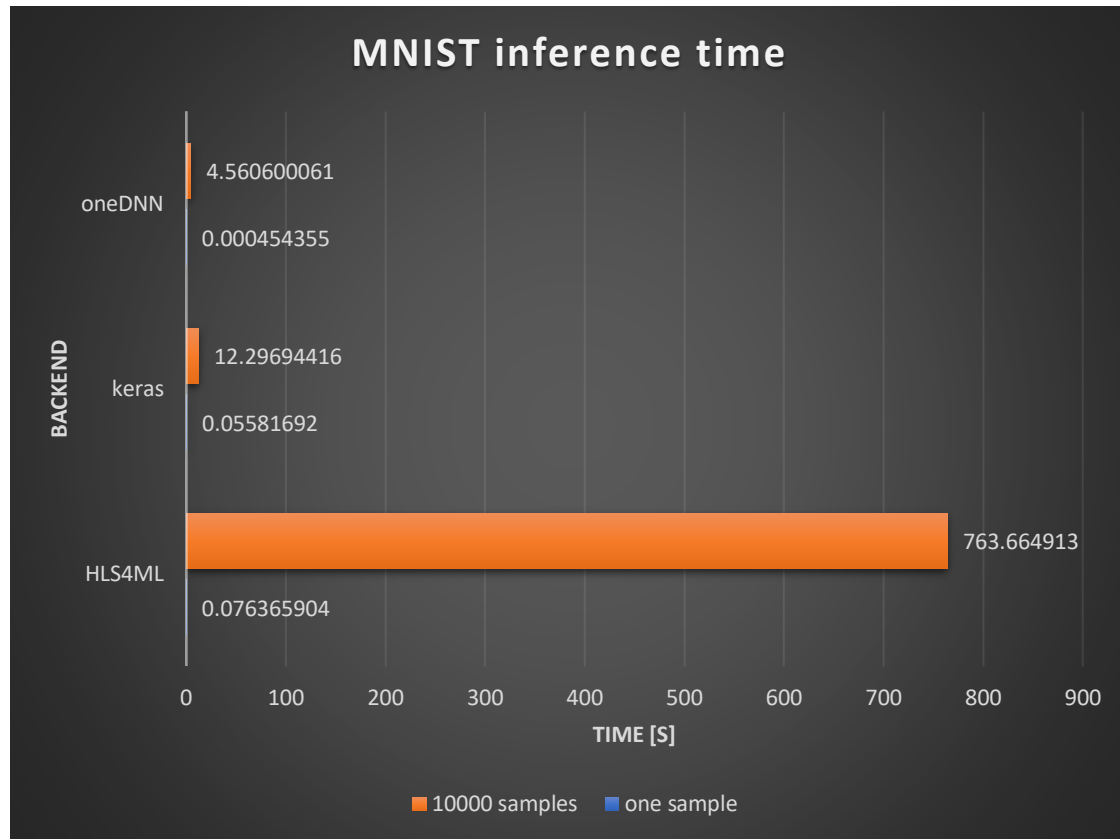


<https://software.intel.com/content/www/us/en/develop/tools/oneapi.html>

Jet Tagging Task with oneAPI backend

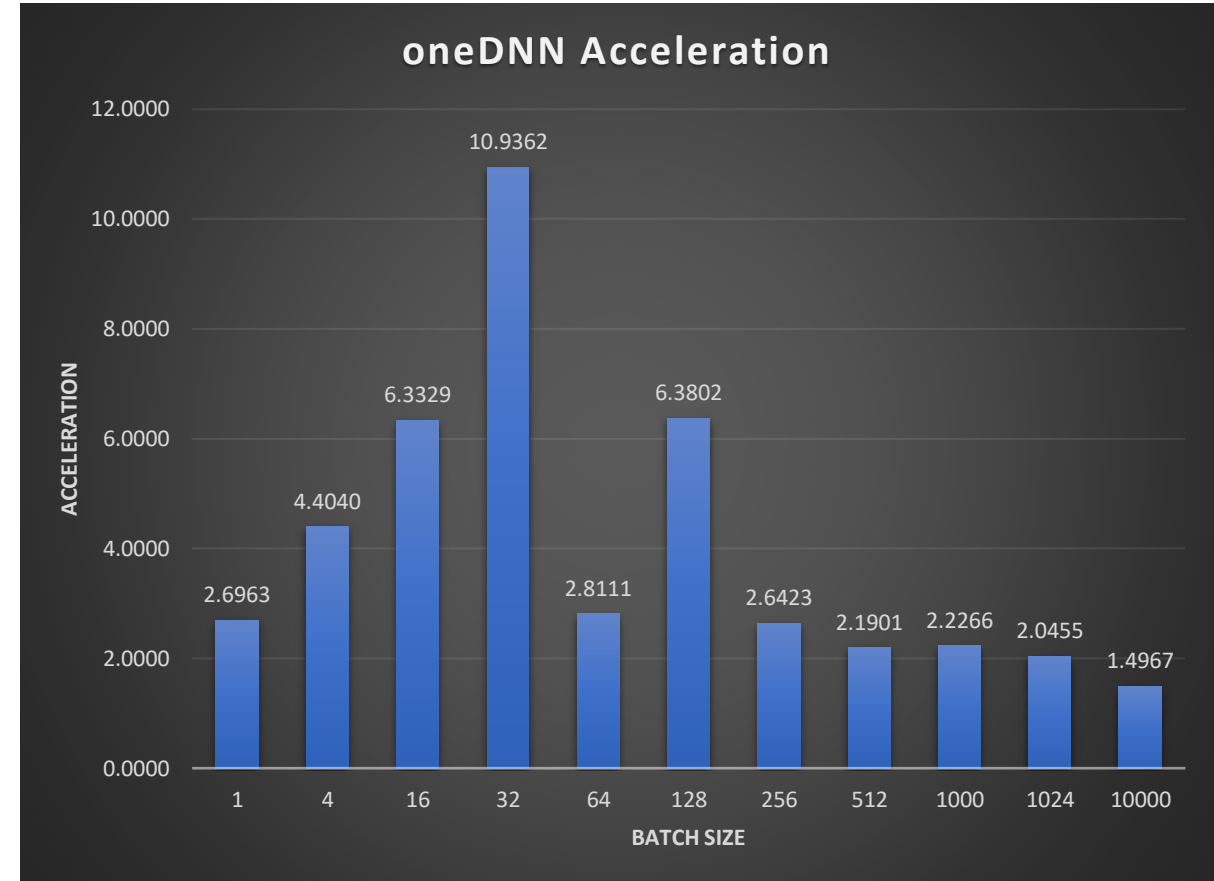
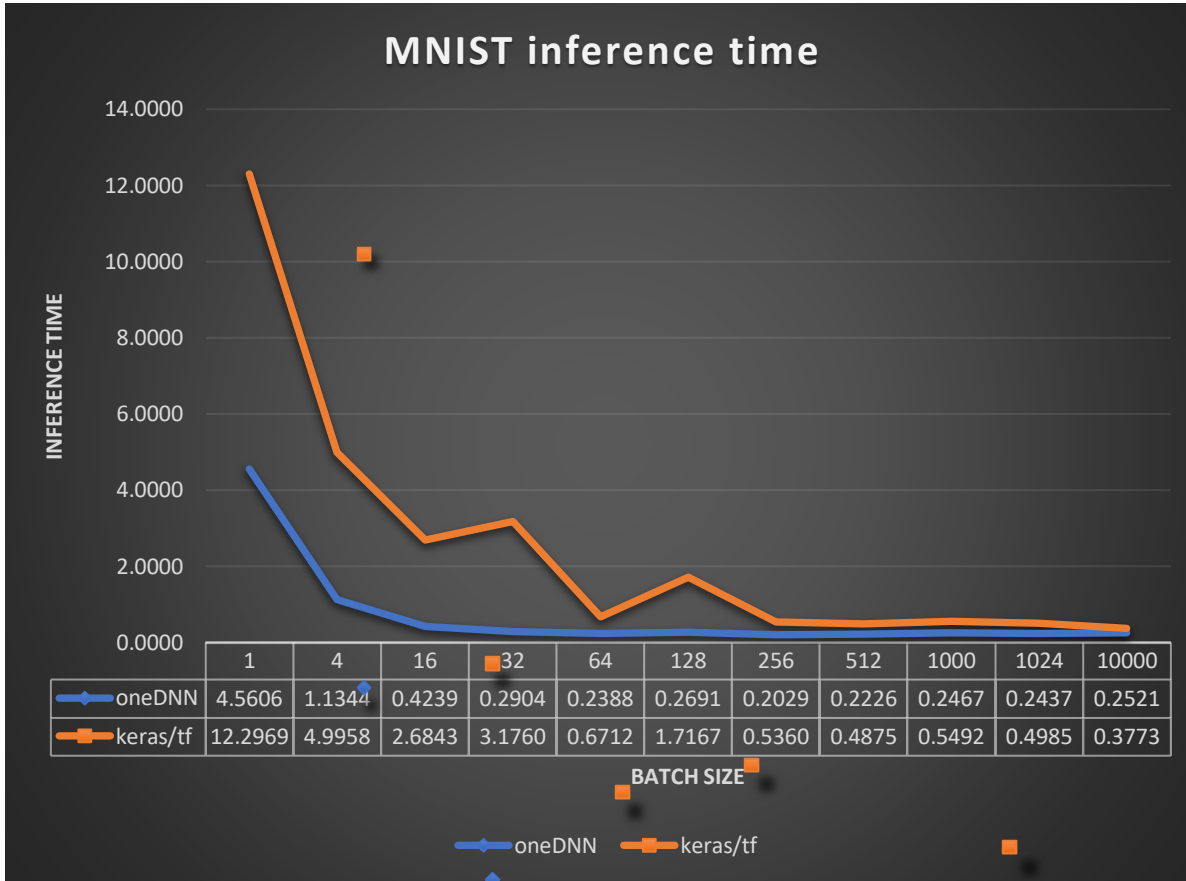


MNIST Digit Classification problem using oneDNN



Best case: 3763 x times of acceleration for batch size 256!

MNIST Digit Classification problem using oneDNN



Best case: 3763 x times of acceleration for batch size 256!

Summary

- OneAPI Programming model can speed up inference time even hundreds of times on CPUs.
- This software is focused on CPUs, but is not limited to them. The novel Xe architectures as well as Deep Learning Accelerators will also be supported without big changes.



QUESTIONS?

[Link to the report](#)

marcinswiniarski20@gmail.com

<https://www.linkedin.com/in/marcin-swiniarski-011bb2159>