



AKADEMIA GÓRNICZO-HUTNICZA  
IM. STANISŁAWA STASZICA W KRAKOWIE  
AGH UNIVERSITY OF KRAKOW

# Klasyfikacja obrazów na FPGA z użyciem Vitis AI

# Problem

- Klasifikacja obrazów z datasetu CIFAR-10
  - 60000 obrazków 32x32 piksele
  - 10 klas – Samochody, samoloty itd.
- Celem jest sprawdzenie jak kwantyzacja i implementacja na FPGA wpływa na czas inferencji i dokładność modelu.

airplane



dog



automobile



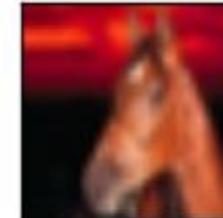
frog



bird



horse



cat



ship



deer

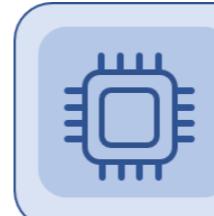


truck



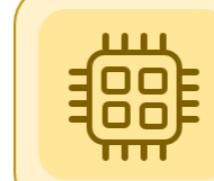
# Zalety Machine Learningu na FPGA

- Niskie opóźnienia
- Bardzo dobra wydajność
- Dobry stosunek cena/jakość
- Niskie zużycie energii



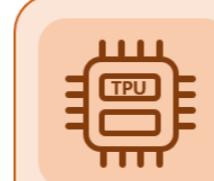
## CPU

- Small models
- Small datasets
- Useful for design space exploration



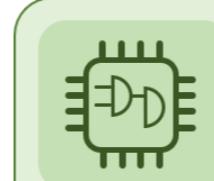
## GPU

- Medium-to-large models, datasets
- Image, video processing
- Application on CUDA or OpenCL



## TPU

- Matrix computations
- Dense vector processing
- No custom TensorFlow operations

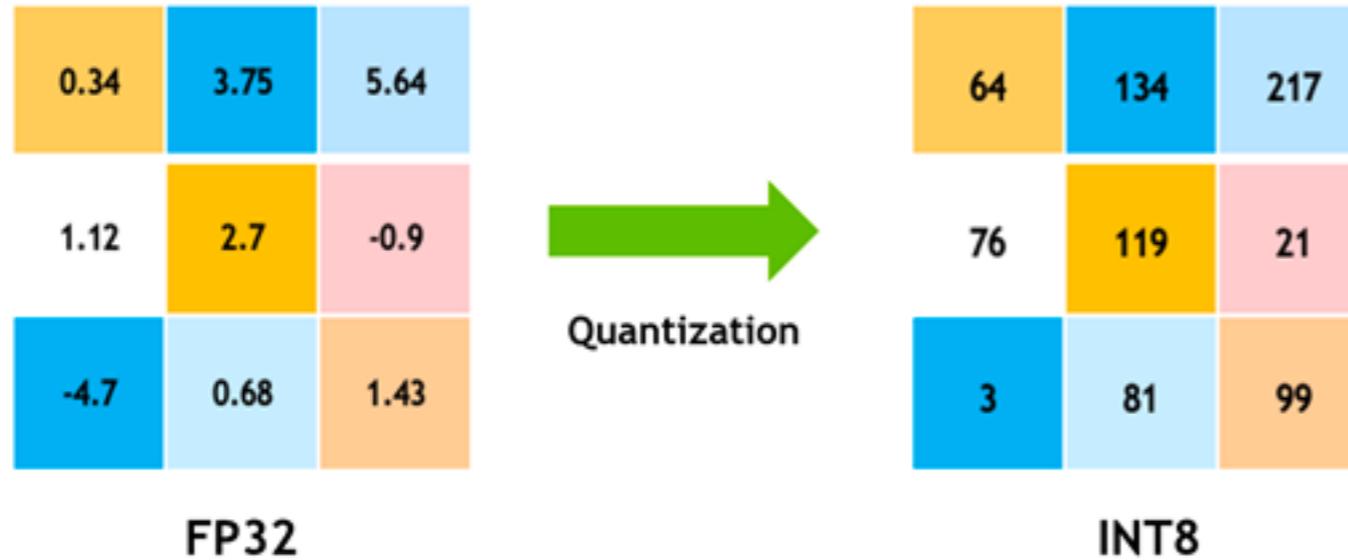


## FPGA

- Large datasets, models
- Compute intensive applications
- High performance, high perf./cost ratio

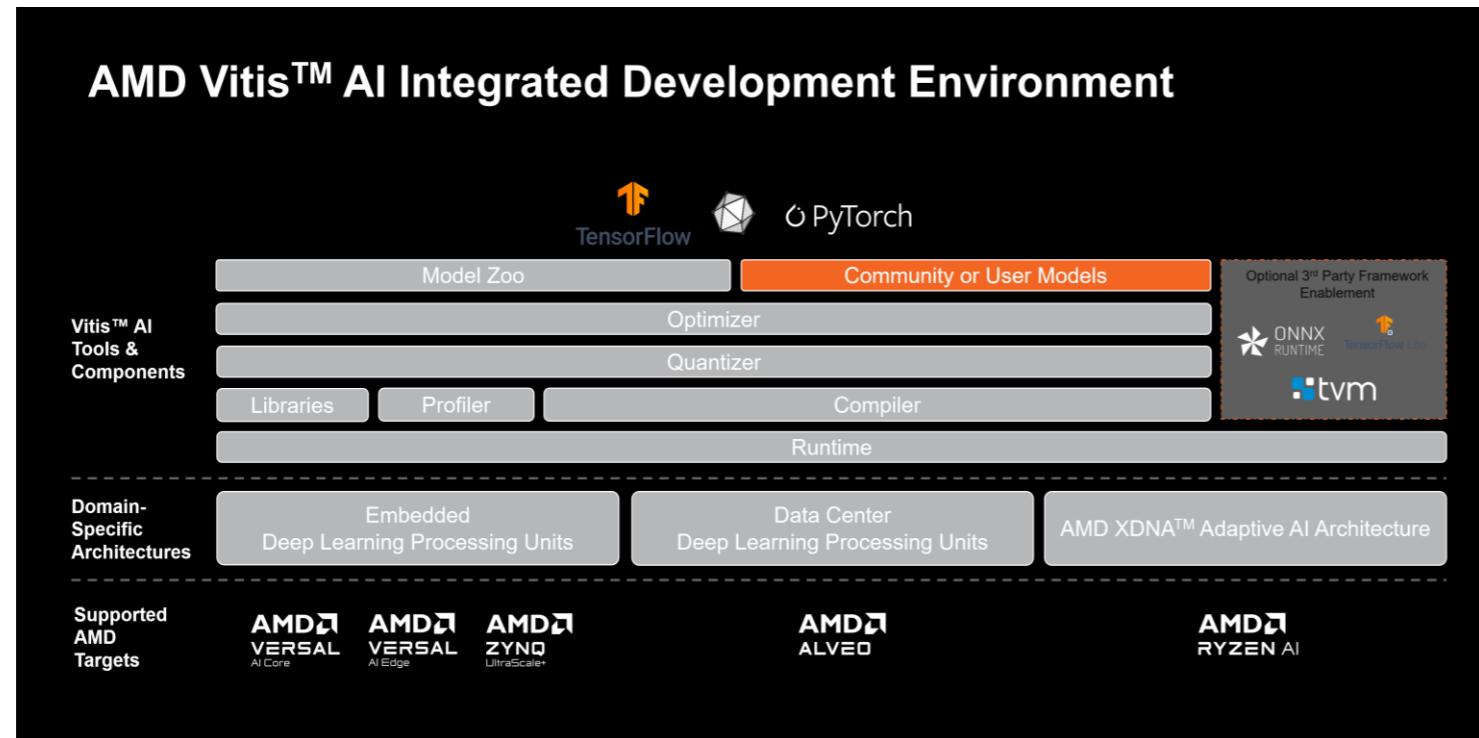
# Kwantyzacja

- Kwantyzacja pozwala na znaczne zmniejszenie modelu i jego przyśpieszenie kosztem minimalnego spadku na dokładności.



# Vitis AI

- Zintegrowane środowisko do rozwoju aplikacji AI na platformach AMD
- Zawiera narzędzia, biblioteki oraz gotowe modele i przykłady



# Etapy pracy

- Trenowanie modelu w PyTorch
  - MobileNetV2 na zbiorze CIFAR-10
- Eksport do ONNX
  - `torch.onnx.export(model, input, "model.onnx")`
- Kwantyzacja
  - Użycie Vitis AI Quantizer (`vai_q_pytorch`) do konwersji na INT8
- Kompilacja
  - Kompilacja modelem w Vitis AI Compiler pod konkretną platformę FPGA
- Uruchomienie inferencji
  - Wykorzystanie Vitis AI Runtime (VART) do wykonania na FPGA

# Źródła

- <https://my.avnet.com/silica/resources/article/fpga-vs-gpu-vs-cpu-hardware-options-for-ai-applications/>
- <https://xilinx.github.io/Vitis-AI/3.5/html/index.html>