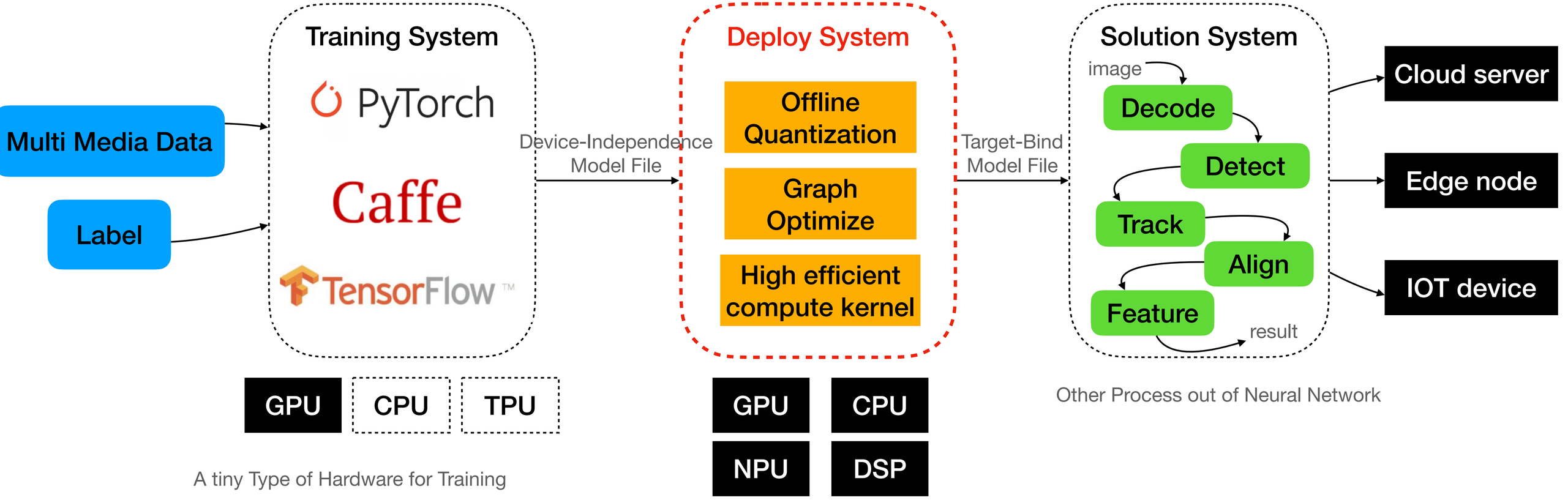
Spring Project：Multi Backend Neural Network Auto Quantization and Deployment over ONNX

Fengwei Yu (Sensetime-Chain)

Contents

- What is Neural Network Deployment?
- What feature we want of a Neural Network Deployment Solution/Archtect?
- Neural Network Quantization
- From Caffe to Onnx

What is Neural Network Deployment?



Neural Network Deployment

：**From** Device-Independence Model File to Target-Bind Model File