AWS **Inferentia**

AWS **Inferentia**

High performance machine learning inference chip, custom designed by AWS

AWS Inferentia is a machine learning inference chip, custom designed by AWS to deliver high throughput, low latency inference performance at an extremely low cost. AWS Inferentia will support the TensorFlow, Apache MXNet, and PyTorch deep learning frameworks, as well as models that use the ONNX format.