Amazon **Elastic Inference**

Amazon **Elastic Inference**

Add GPU acceleration to any Amazon EC2 instance for faster inference at much lower cost (up to 75% savings)

Allows you to attach just the right amount of GPU-powered acceleration to any Amazon EC2 and Amazon SageMaker instance to reduce the cost of running deep learning inference by up to 75%. Amazon Elastic Inference supports TensorFlow, Apache MXNet, and ONNX models, with more frameworks coming soon.