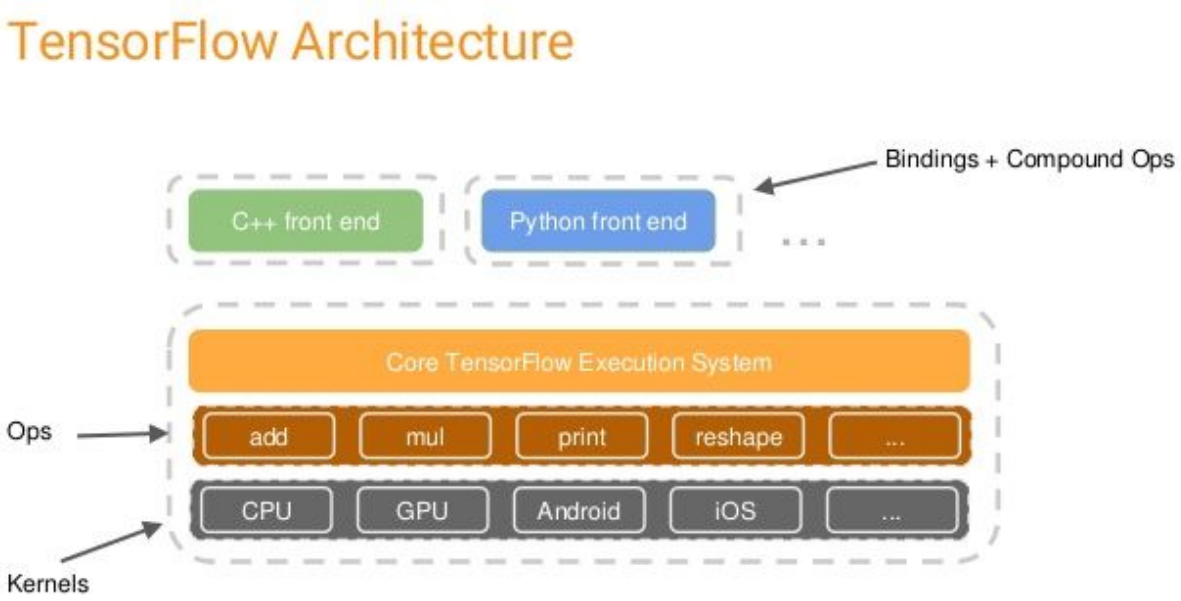

Tensorflow provides Client APIs in Python and C++. Python client supports more higher level API methods which is especially helpful in ML and Deep Learning tasks.

Tensorflow use Python or C++ to build a tf.GraphDef protocol buffer that represents the computation as a dataflow graph. The standard TensorFlow runtime will interpret the graph and issue **pre-compiled CUDA kernels** for each operation.

XLA (Accelerated Linear Algebra) is an experimental domain-specific compiler for linear algebra that optimizes TensorFlow computations. XLA takes graphs ("computations") defined in HLO and compiles them into machine instructions for various architectures. XLA is modular in the sense that it is easy to slot in an alternative backend to target some novel HW architecture. With XLA Ahead of Time Compilation Tensorflow graphs can be executed independently without using the Tensorflow runtime.