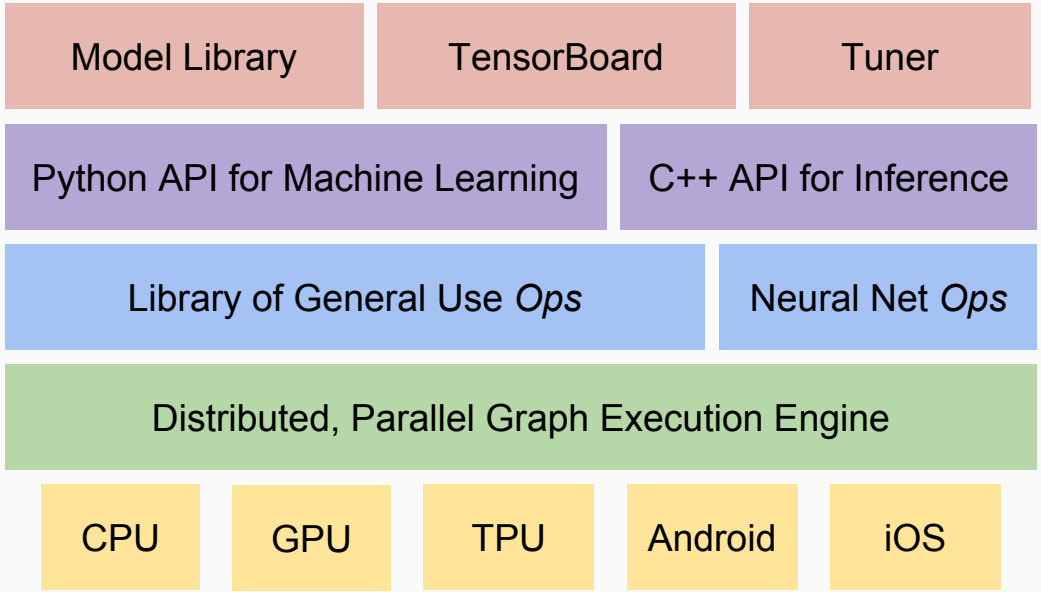
TensorFlow

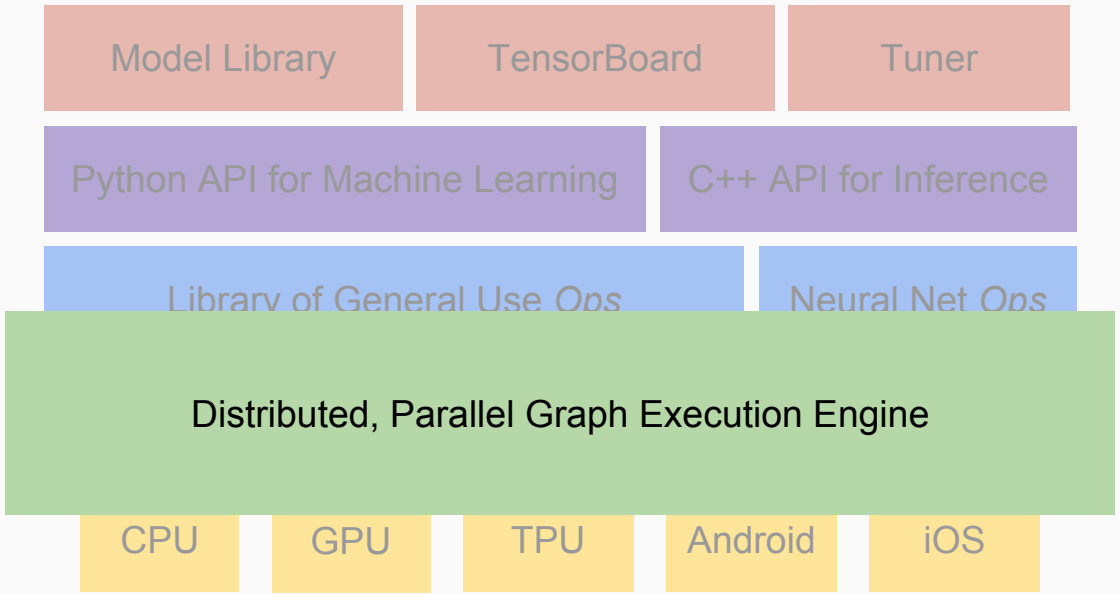
13.01.2017 Maciek Chociej @WFAIS

What is TF ?

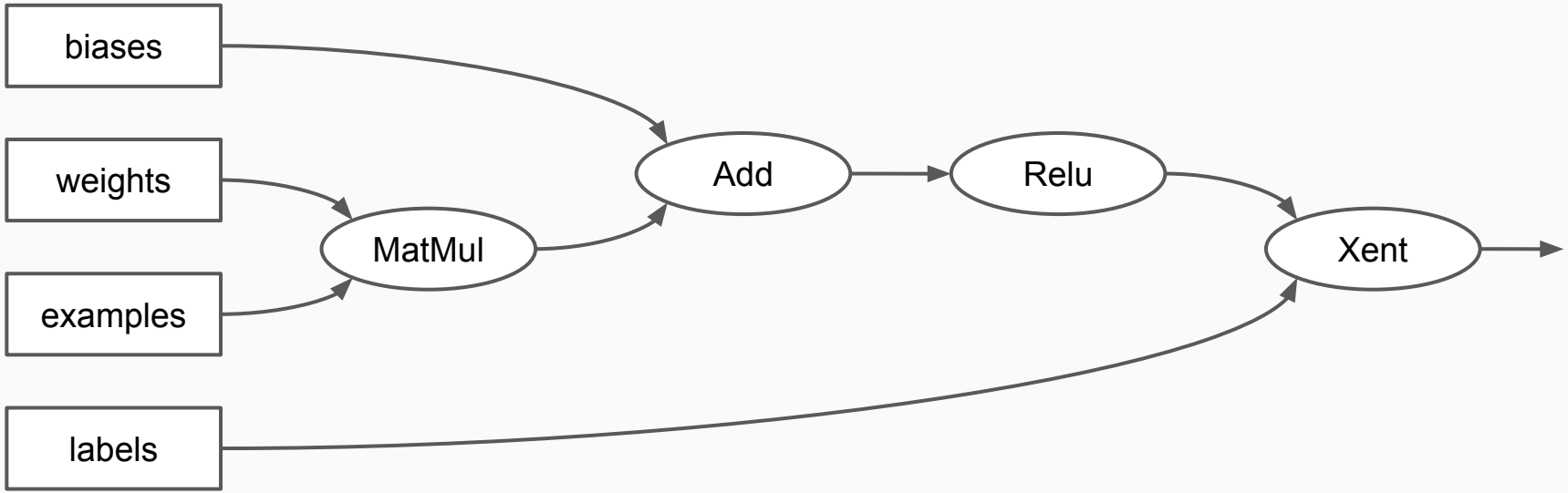
- Yet another dataflow system
- For distributed, parallel machine learning
- Rapid prototyping & high performance in one
- Targeting heterogeneous devices: embedded, mobile, CPU, GPU, TPU
- Multi-platform and multi-language: C++/python/go/Java
- Open source (Apache 2.0 license)

TF Stack

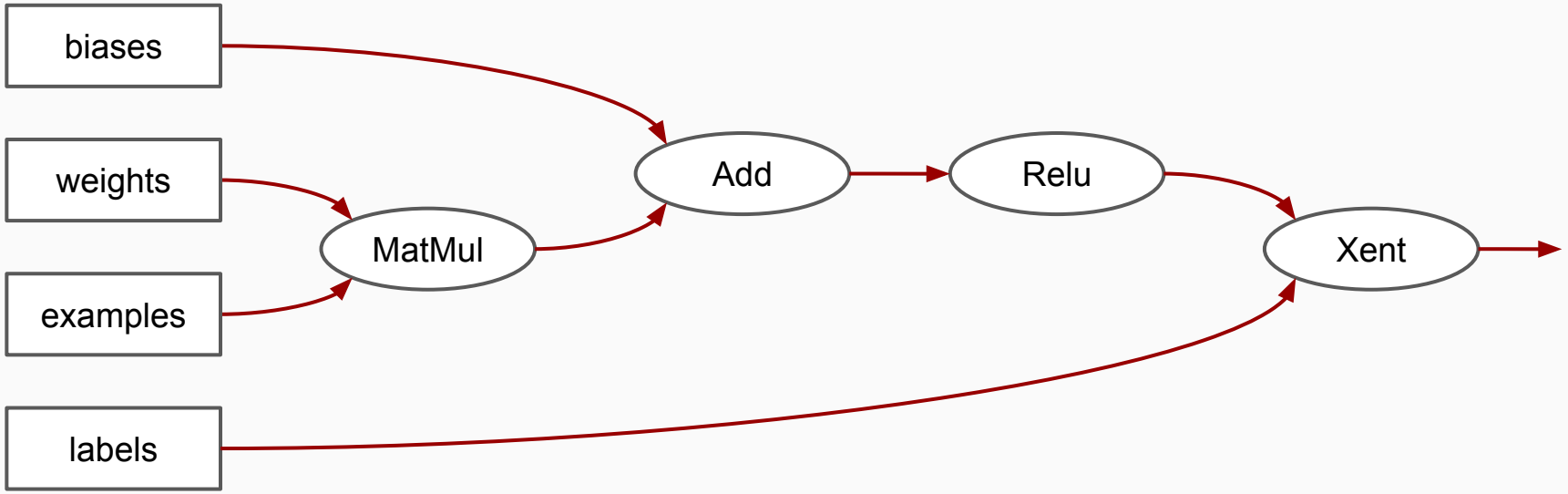

Model Library
TensorBoard
Tuner
Python API for Machine Learning
C++ API for Inference
Library of General Use *Ops*
Neural Net *Ops*
Distributed, Parallel Graph Execution Engine
CPU
GPU
TPU
Android
iOS


Model Library
TensorBoard
Tuner
Python API for Machine Learning
C++ API for Inference
Library of General Use Ops
Neural Net Ops
Distributed, Parallel Graph Execution Engine
CPU
GPU
TPU
Android
iOS

Graph of operations

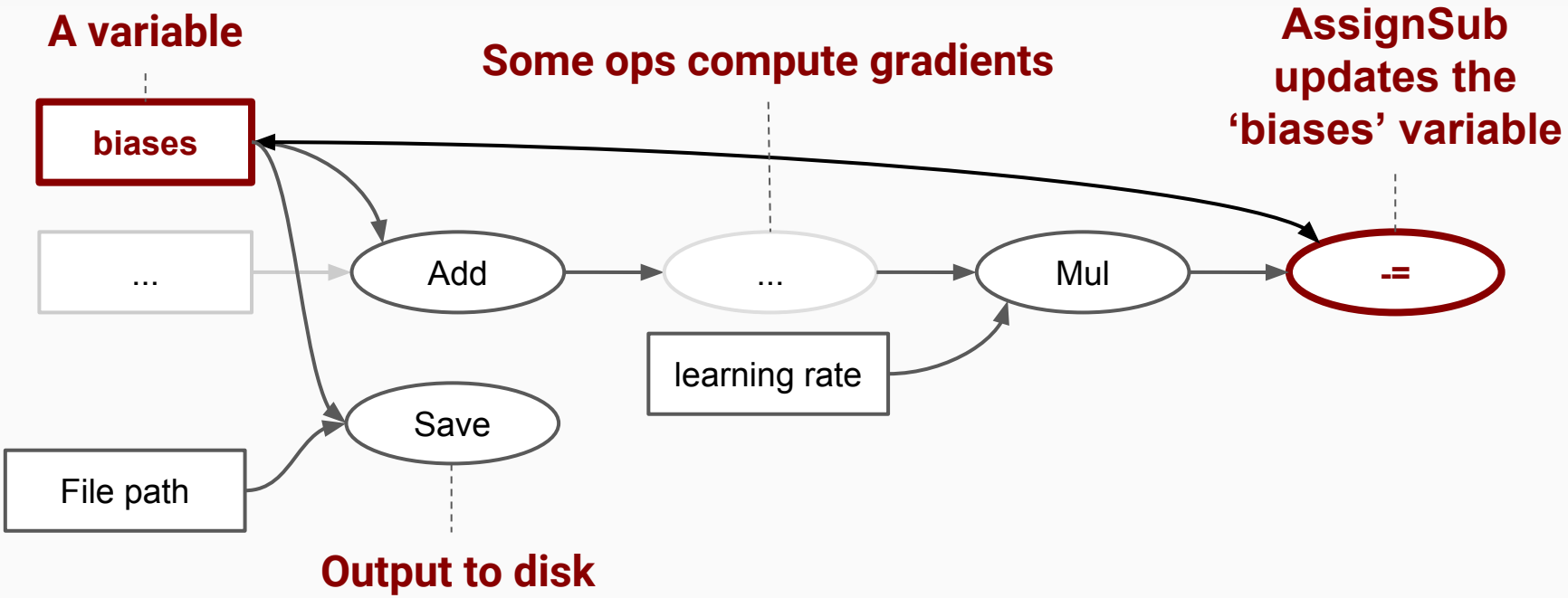


Edges are N-dimensional arrays



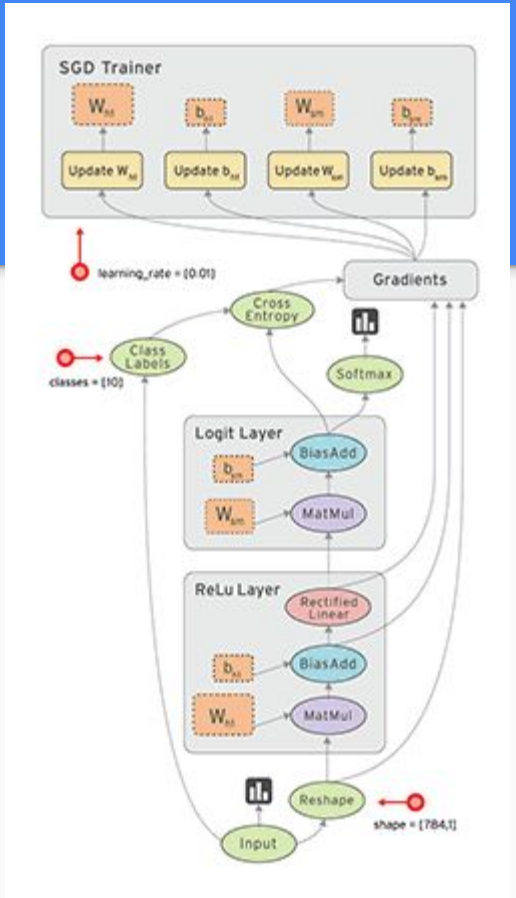
Types: floats, doubles, halves, integers, string, bools, quantized, complex

Variables for stateful computation

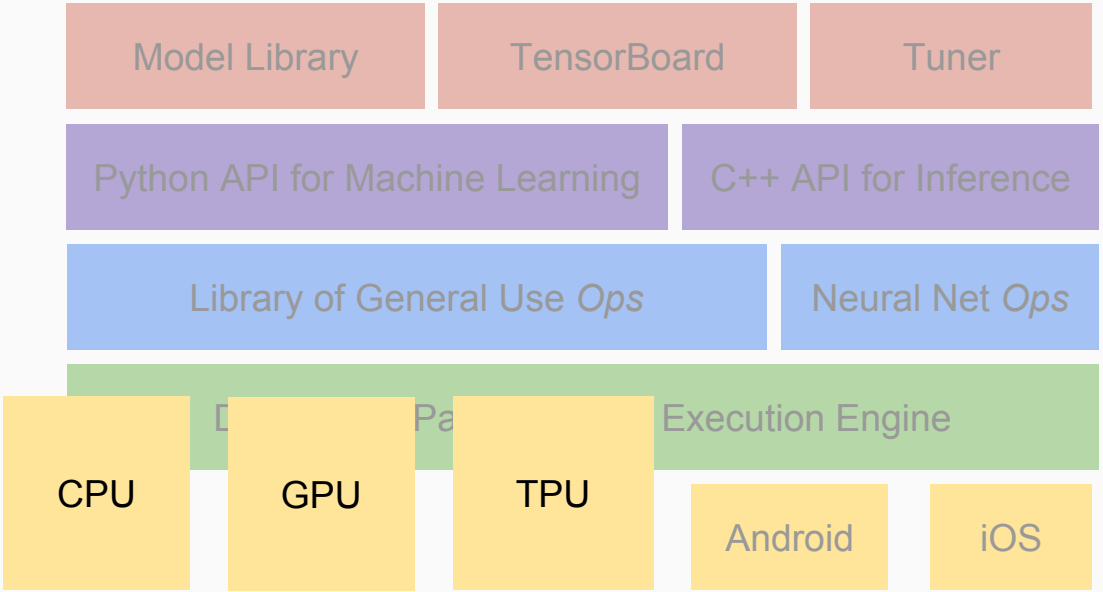


Parallel execution of a graph

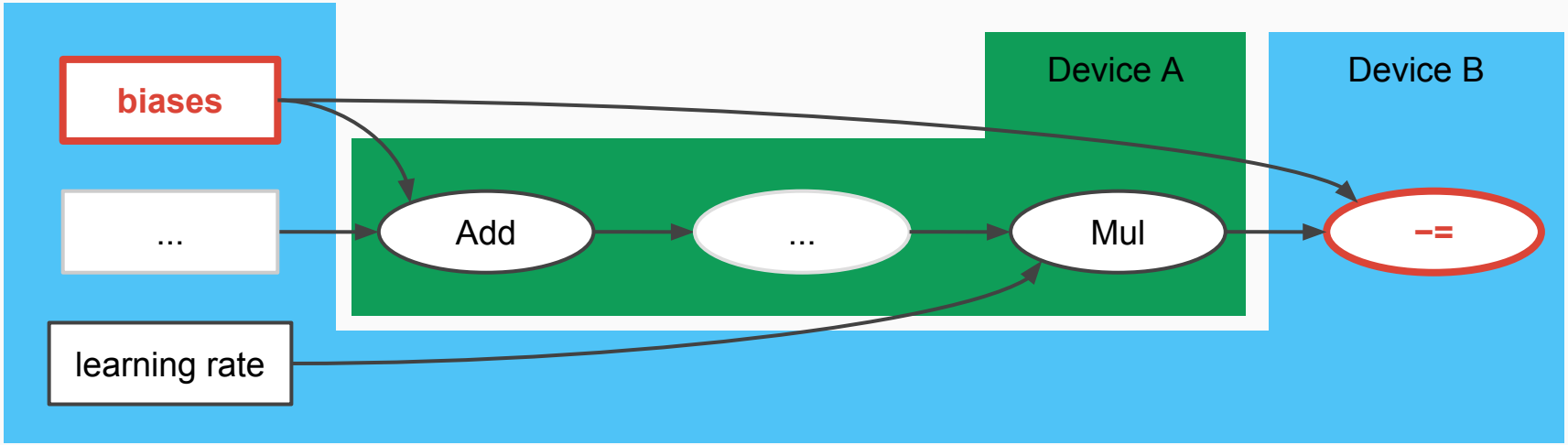
- launch a graph in a sync/async session
- request the output of some ops
- TensorFlow computes the set of ops that must complete in order to compute the requested outputs
- ops execute, in parallel, as soon as their inputs are available



Where do the ops run and tensors reside?



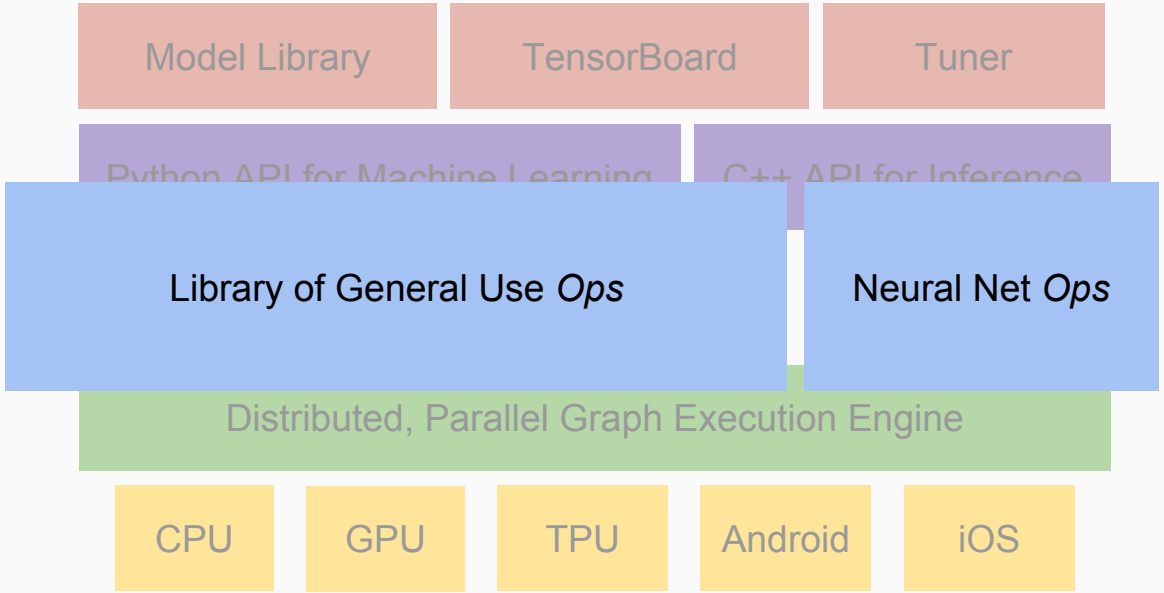
Explicitly pick devices, or let TF decide



Run graphs on multiple machines

- easily set up clusters of workers and shared data servers
- explicitly or automatically divide work between them
  - data-parallelism
  - graph-parallelism
- seamless marshalling of data between physical machines and devices (GPUs/CPUs)
- synchronization primitives for building complex distributed algorithms

The library of operations



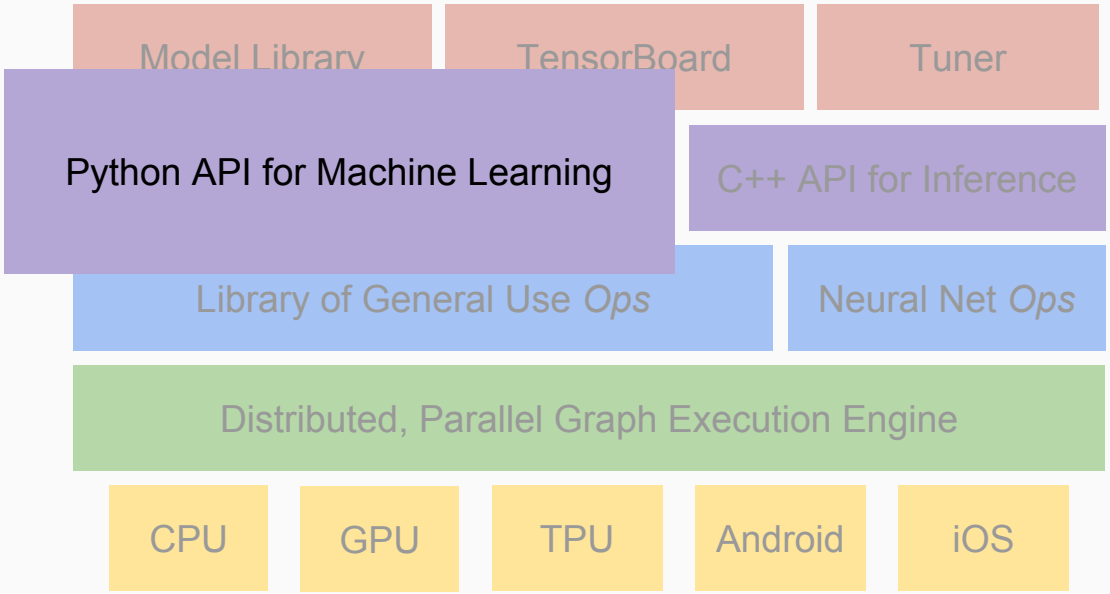
Basic operations

- **Basics**: constant, random, placeholder, cast, shape
- **Variables**: assign, assign_sub, assign_add
- **Queues**: enqueue, enqueue_batch, dequeue, blocking or not.
- **Logical**: equal, greater, less, where, min, max, argmin, argmax.
- **Tensor computations**: all math ops, matmul, determinant, inverse, cholesky.
- **Images**: encode, decode, crop, pad, resize, color spaces, random perturbations.
- **Sparse tensors**: represented as 3 tensors.
- **File io**: file readers & parsers
- **Control flow**: control dependencies, conditionals, loops, functions.

Neural Network and Deep Learning operations

- **Activations**: logistic, sigmoid, relu
- **Pooling**: max, avg, depthwise
- **Convolutions**: 2D/3D stacked and depthwise
- **Concatenations**
- **Normalization**: local, batch, moving averages
- **Classification**: softmax, topk
- **Embeddings**: distributed lookups/gather, scatter/updates
- **Losses:** Cross-entropy, logistic, L1, L2
- **Recurent NN building blocks**
- **LSTM NN building blocks**
- **Quantized versions of NN layers**

Training


Model Library
TensorBoard
Tuner
Python API for Machine Learning
C++ API for Inference
Library of General Use *Ops*
Neural Net *Ops*
Distributed, Parallel Graph Execution Engine
CPU
GPU
TPU
Android
iOS

Training utilities

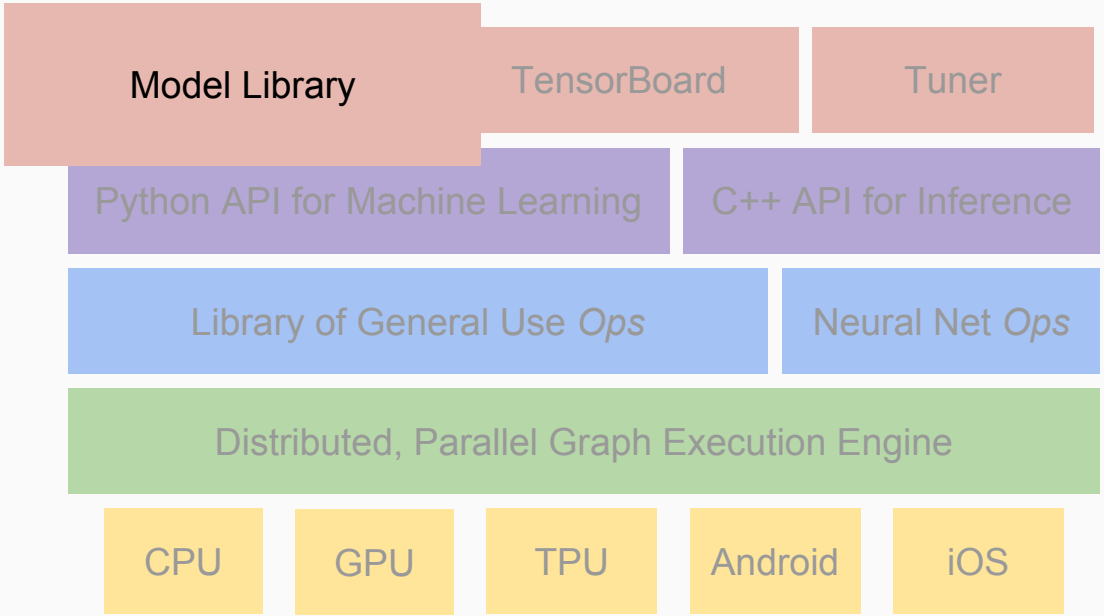
- Automatically build the back-propagation graph of operation gradients
- Use a number of optimizers to fit the parameters
  - SGD
  - AdaGrad
  - AdaDelta
  - ADAM
  - FTRL
  - ...

Minimal MNIST Neural Network

```
import tensorflow as tf
from tensorflow.examples.tutorials.mnist import input_data

mnist = input_data.read_data_sets('MNIST_data', one_hot=True)
x = tf.placeholder("float", shape=[None, 784])
W = tf.Variable(tf.zeros([784,10]))
b = tf.Variable(tf.zeros([10]))
output = tf.nn.softmax(tf.matmul(x, W) + b)
truth = tf.placeholder(tf.float32, [None, 10])
```

Minimal MNIST Neural Network

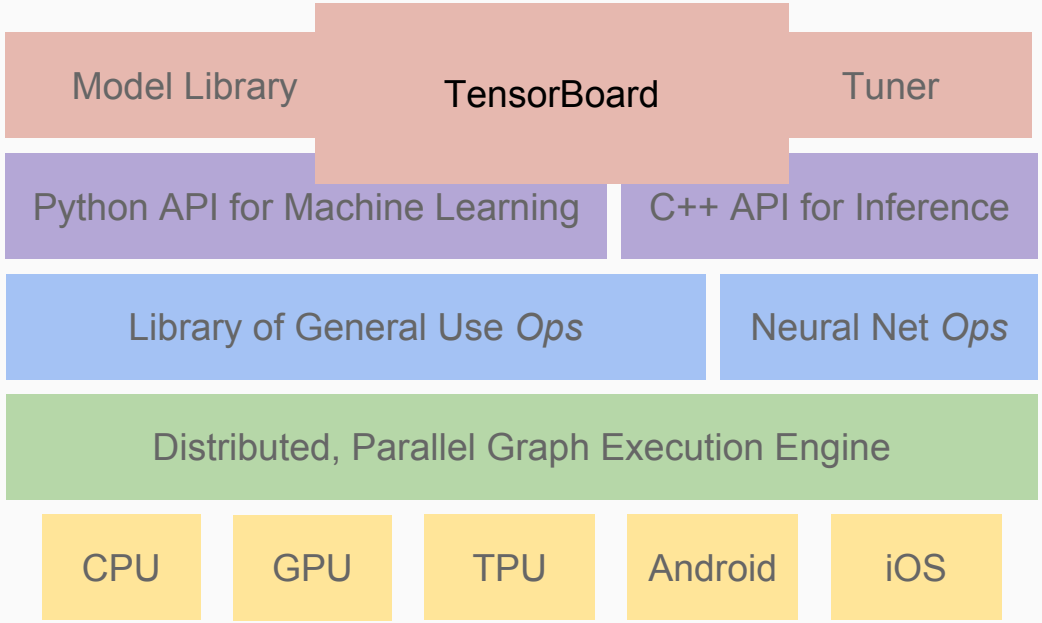
```
cross_entropy = -tf.reduce_sum(truth * tf.log(output))
optimizer = tf.train.GradientDescentOptimizer(0.01)
train_iteration = optimizer.minimize(cross_entropy)

session = tf.Session()
for i in range(1000):
  batch_xs, batch_ys = mnist.train.next_batch(100)
  session.run(train_iteration,
              feed_dict={x: batch_xs, y_: batch_ys})
```


Model Library
TensorBoard
Tuner
Python API for Machine Learning
C++ API for Inference
Library of General Use *Ops*
Neural Net *Ops*
Distributed, Parallel Graph Execution Engine
CPU
GPU
TPU
Android
iOS

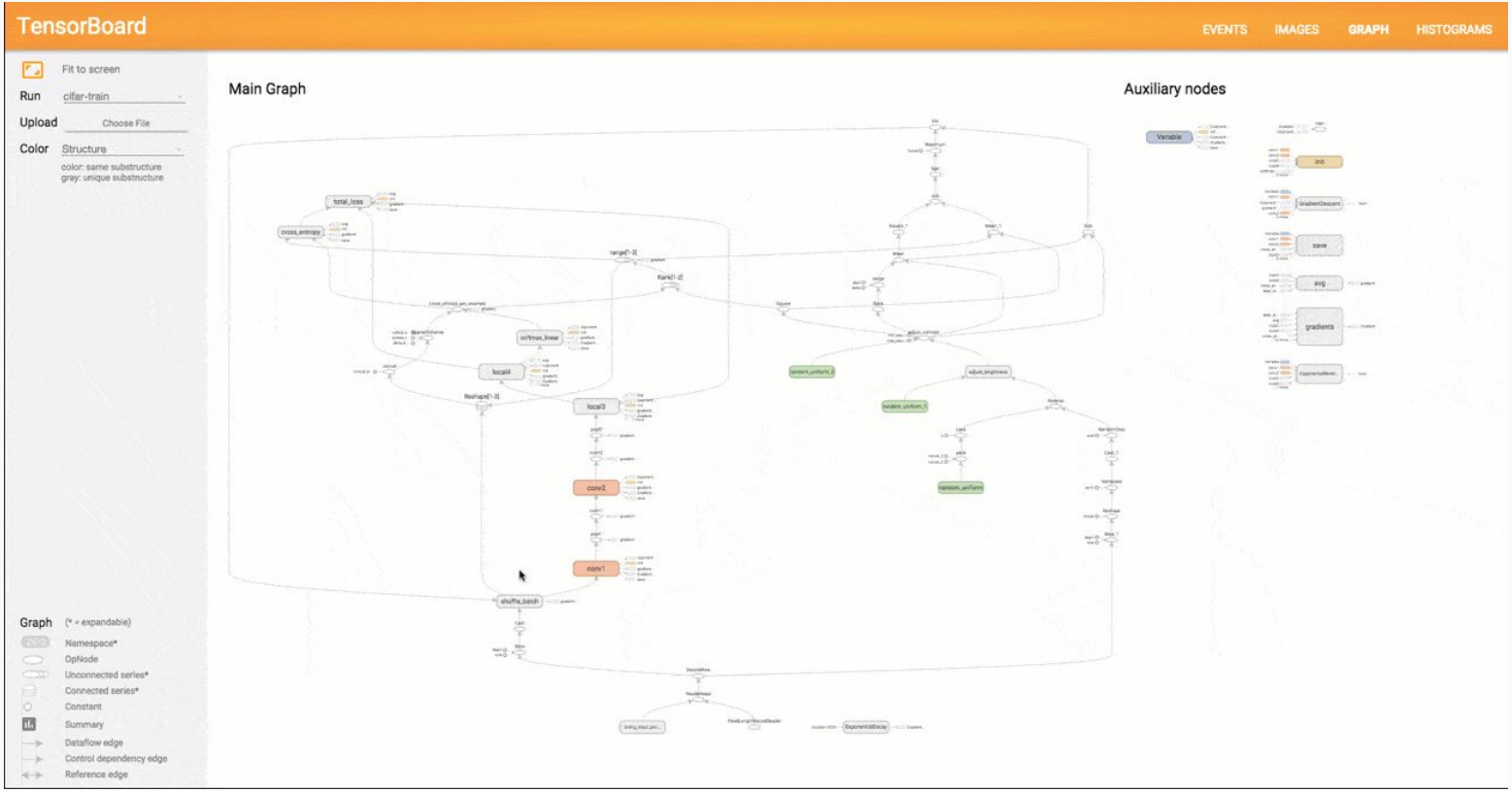
Freebie models

- In the TensorFlow tutorials at tensorflow.org
  - Conv net for Cifar10 dataset
  - Word embeddings
  - Sequence to sequence model
  - A pre-trained Inception model for ImageNet
  - SyntaxNex

Graph & computation visualization



Tensorboard in action



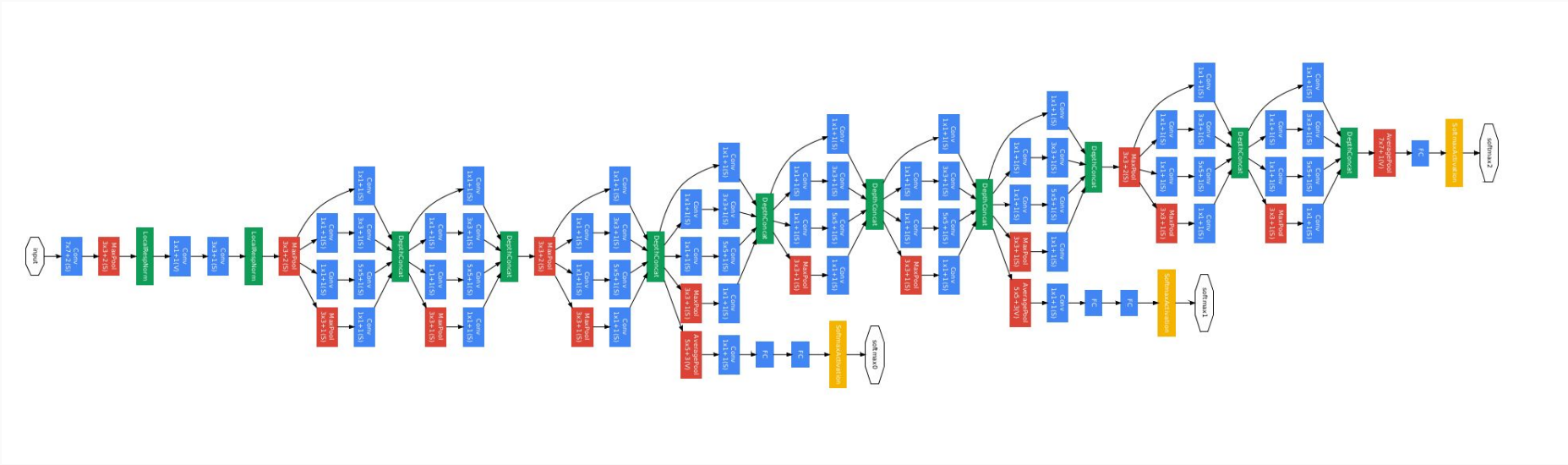
Community contributions

- DQN: github.com/nivwusquorum/tensorflow-deepq
- NeuralArt: github.com/woodrush/neural-art-tf
- Char RNN: github.com/sherjilozair/char-rnn-tensorflow
- Keras ported to TensorFlow: github.com/fchollet/keras
- Show and Tell: github.com/jazzsaxmafia/show_and_tell.tensorflow
- Mandarin translation: github.com/jikexueyuanwiki/tensorflow-zh

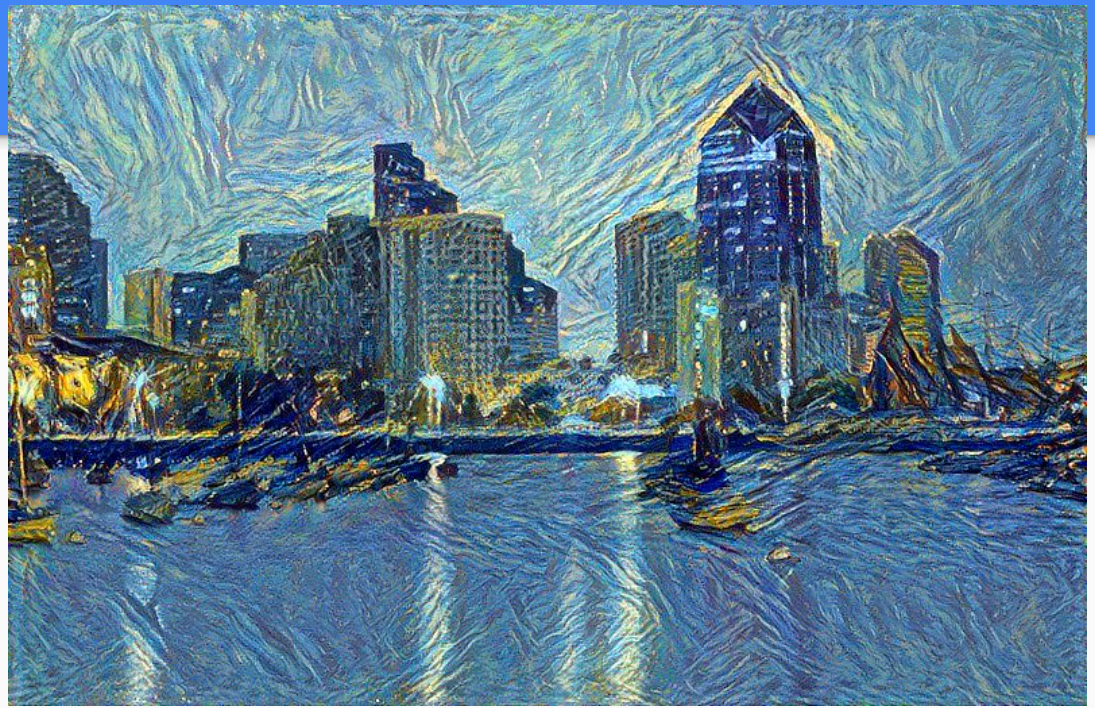
Companies onboard




ImageNet model



NeuralArt



github.com/woodrush/neural-art-tf

TensorFlow at Google

- Text to speech, speech to text
- Translation
- Search ranking
- Ads ranking
- Photo classification
- … and hundreds more