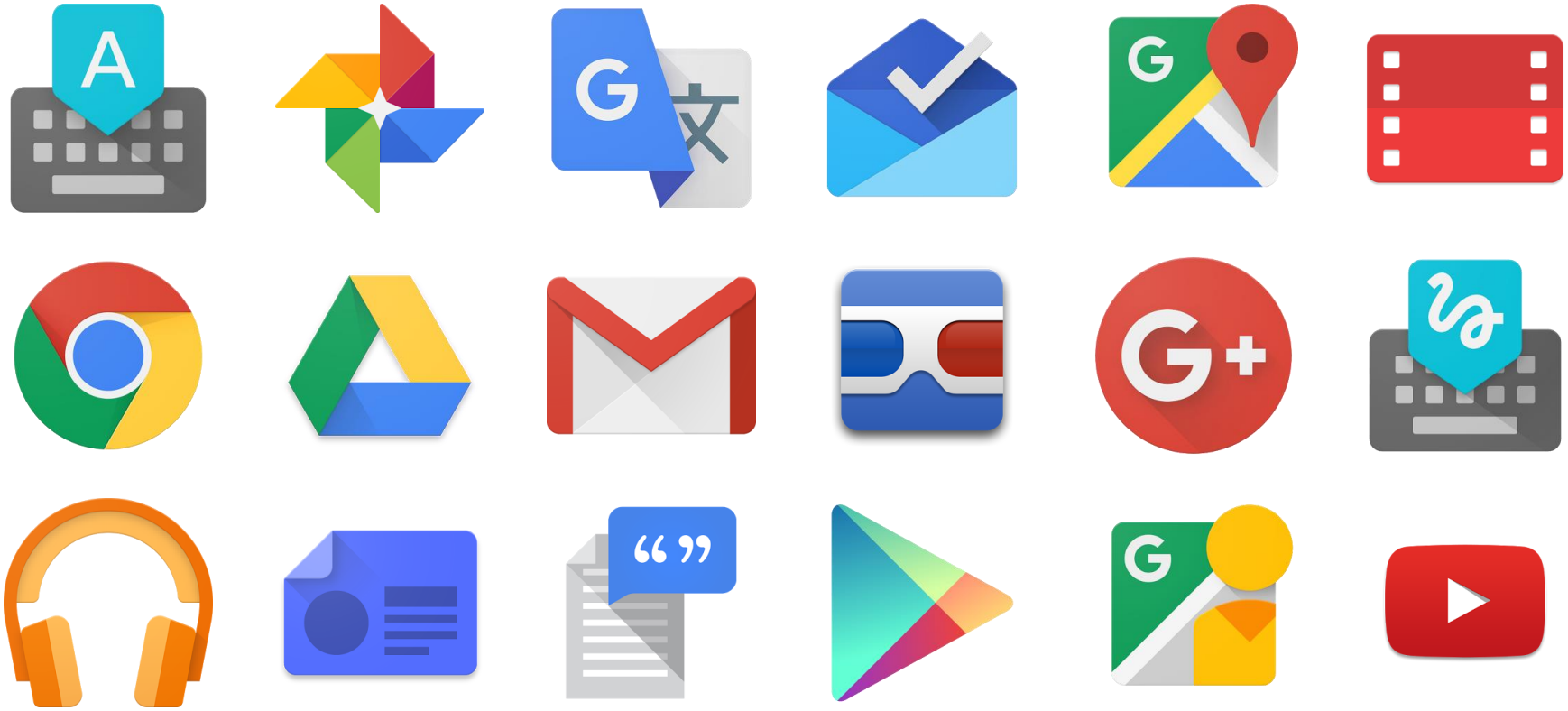
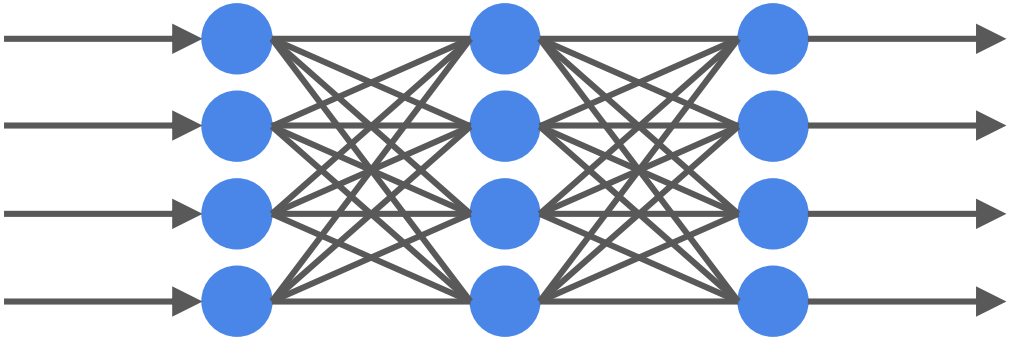
Deep Learning with TensorFlow

Rajat Monga
Engineering Director, TensorFlow

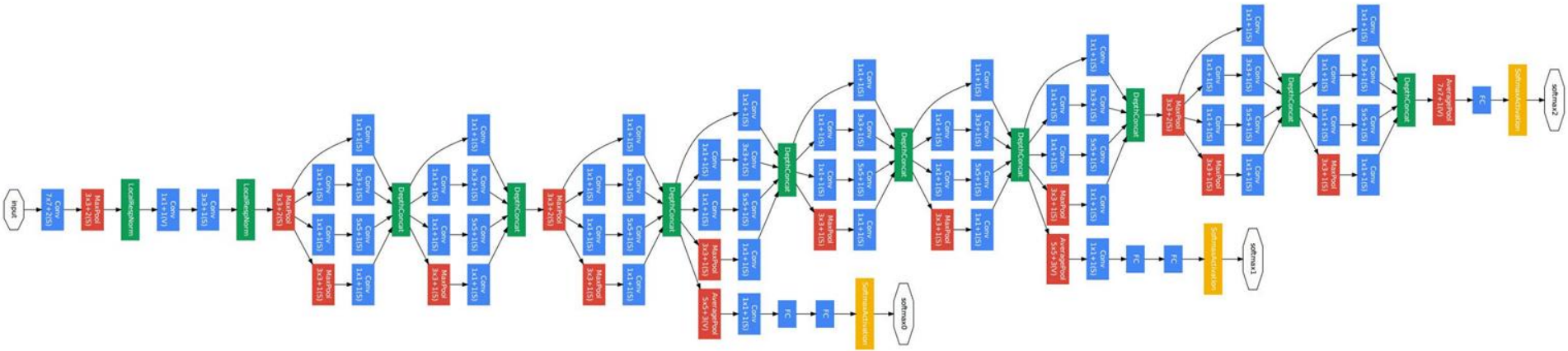


How complex are Deep Learning Systems?

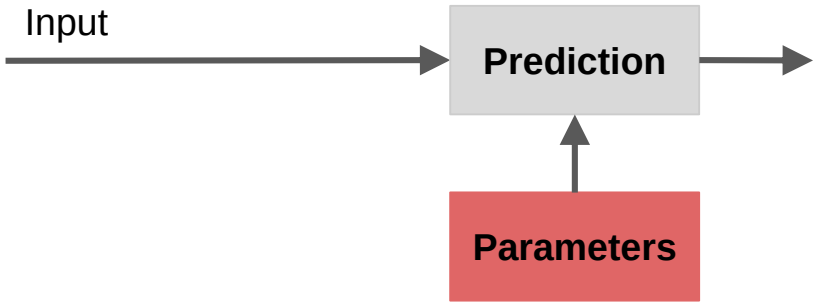
Neural networks



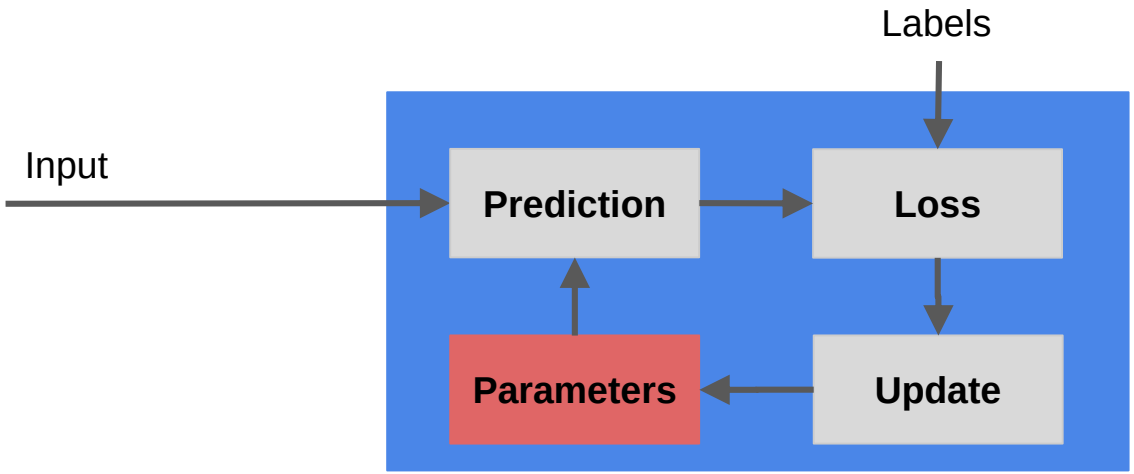
Inception (2015)



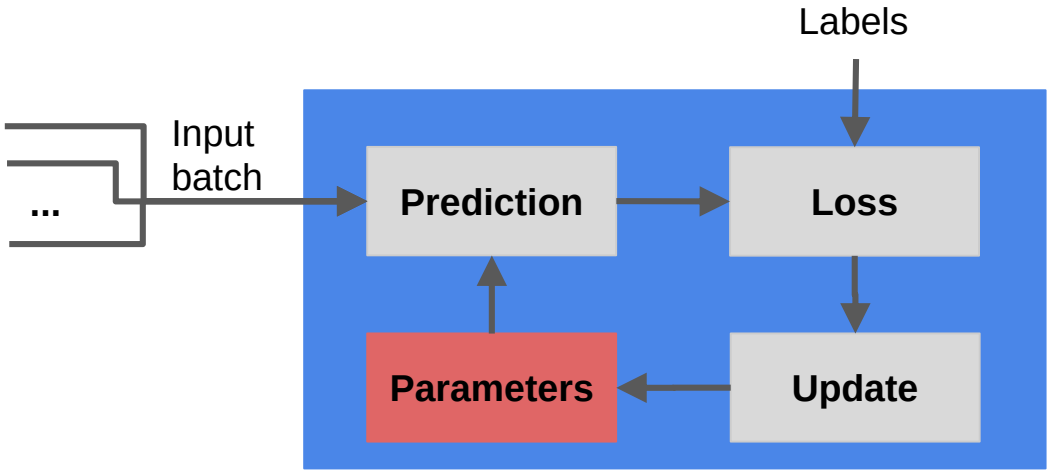
That’s not all...



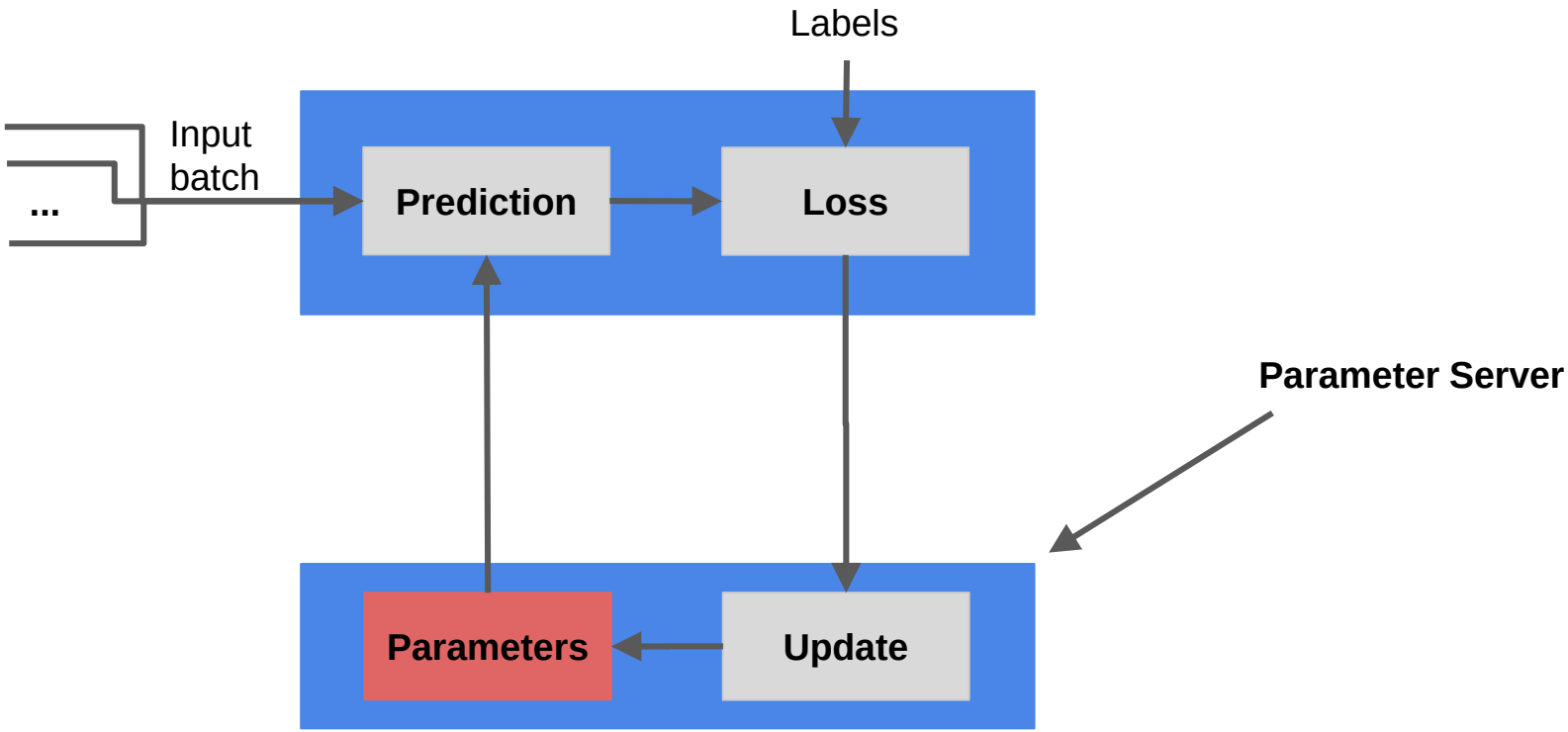
That’s not all...



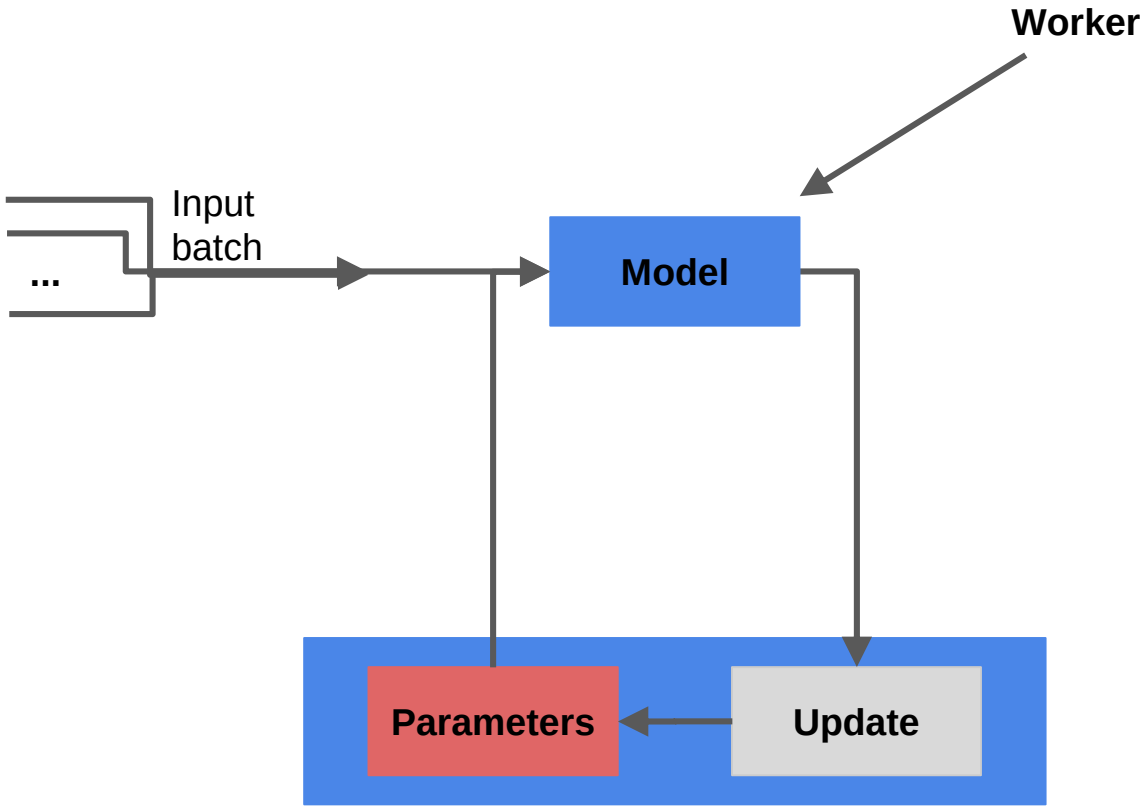
That’s not all...



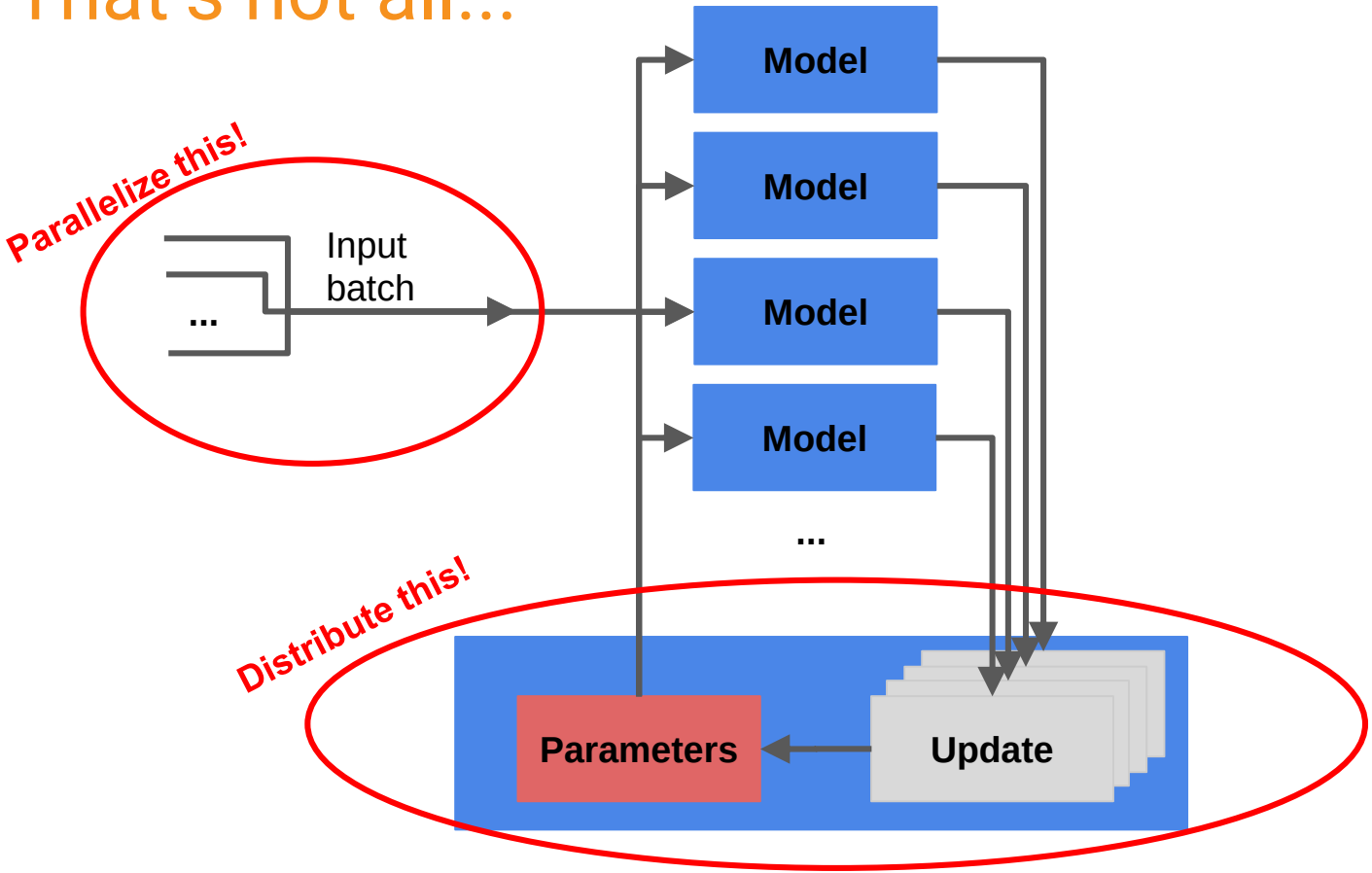
That’s not all...



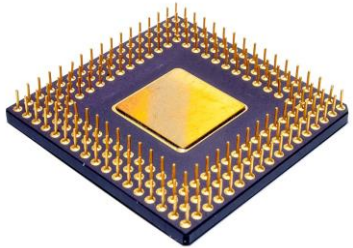
That’s not all...



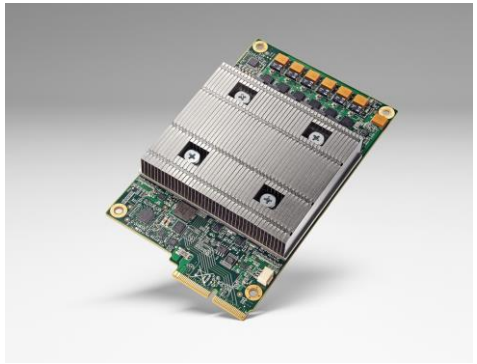
That’s not all...



Portability is a requirement



CPU



TPU

Android

iOS

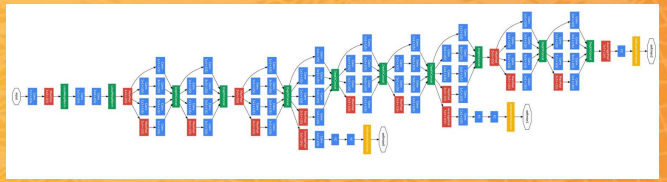
Raspberry Pi

GPU

machine learning gets complex quickly

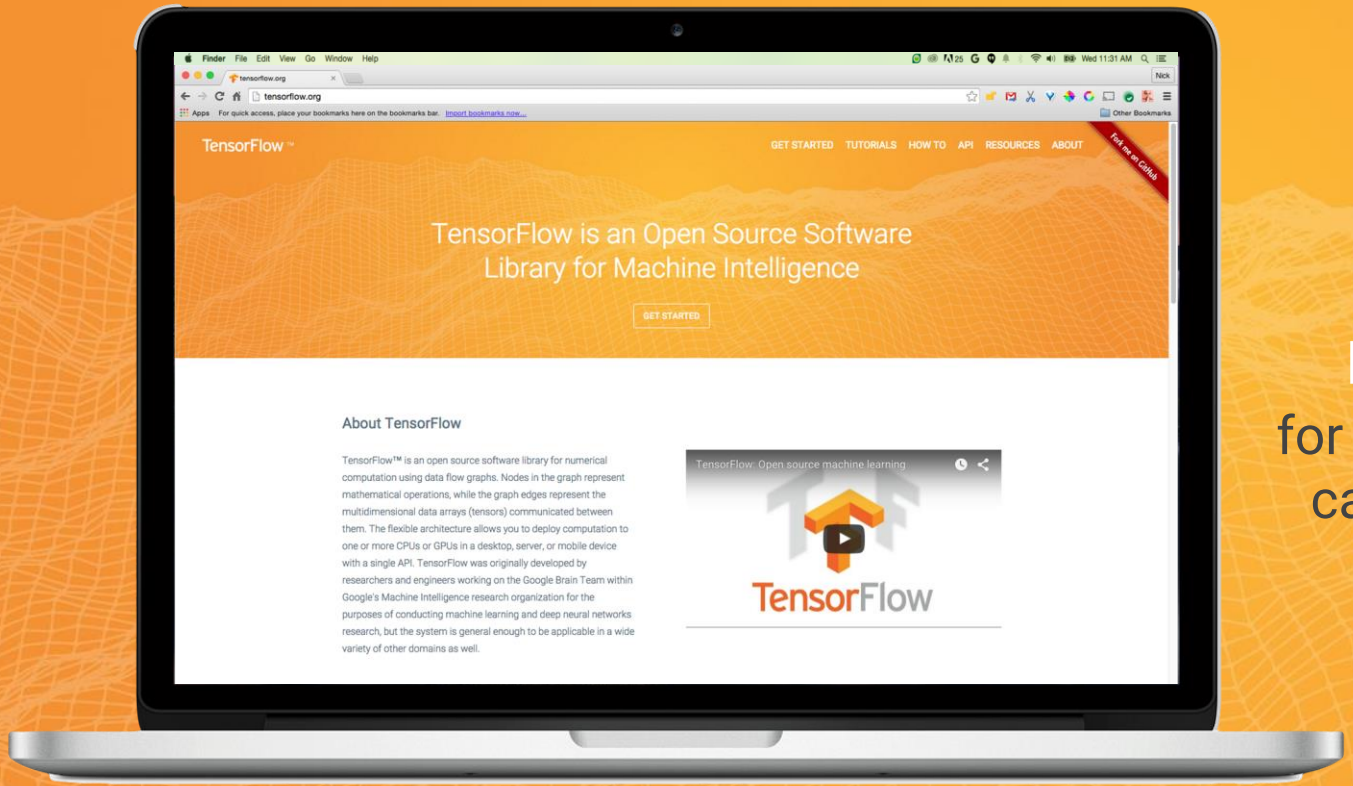
Heterogenous System

Distributed System



Modeling complexity

TensorFlow



#1

repository

for “machine learning” category on GitHub

Some Stats

12,000+ commits since Nov, 2015

570+ contributors

1M+ binary downloads

5000+ TensorFlow related repositories on GitHub

#15 most popular repository on GitHub by stars - across all categories

Used in ML classes at many universities:
Toronto, Berkeley, Stanford, ...


ARM


quanliphi

AIRBUS
DEFENCE & SPACE

CI&T

CEVA

Google

Movidius

UBER

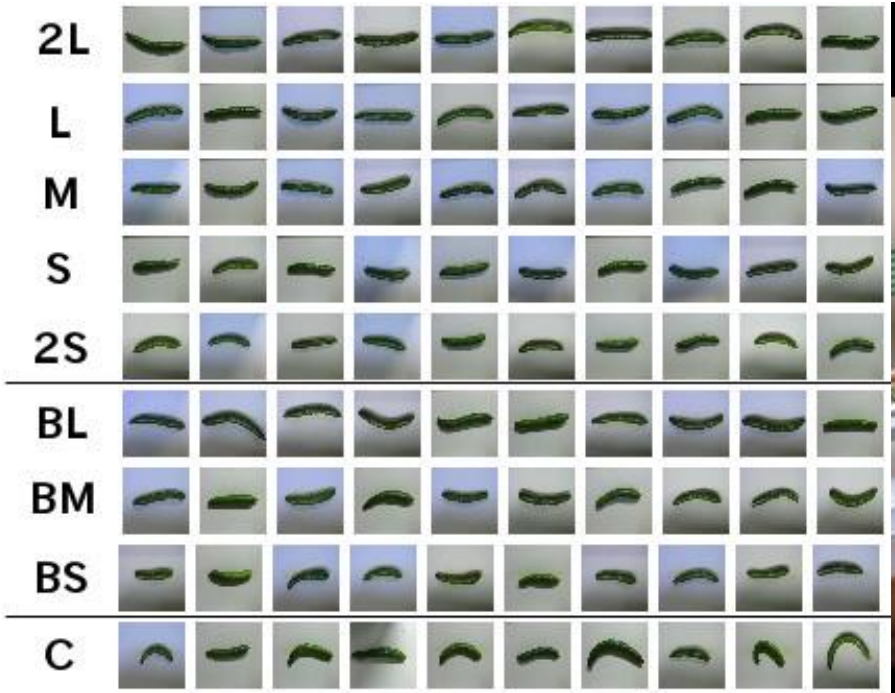
JD.COM 京东


DeepMind

ebay


Dropbox

TensorFlow powered Cucumber Sorter



From: http://workpiles.com/2016/02/tensorflow-cnn-cucumber/

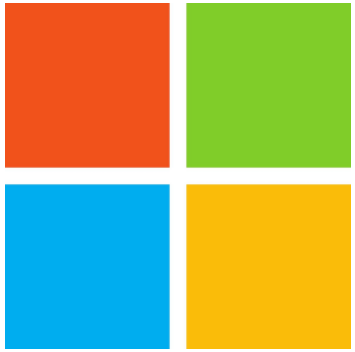
machine learning gets complex quickly

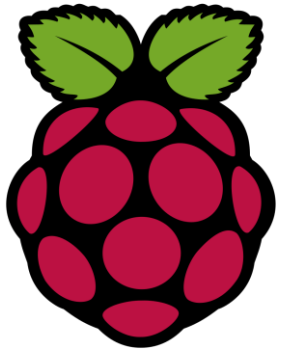
Heterogenous System

Distributed System

Modeling complexity

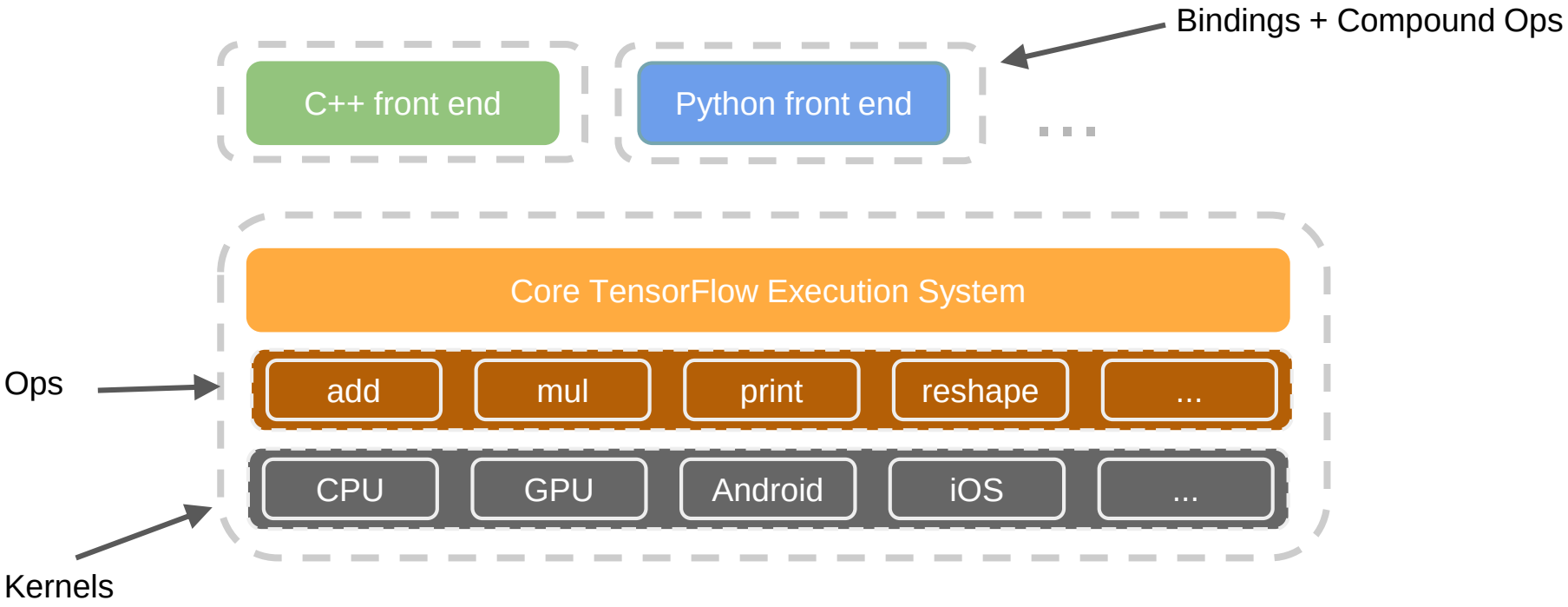
Platforms





Languages

TensorFlow Architecture

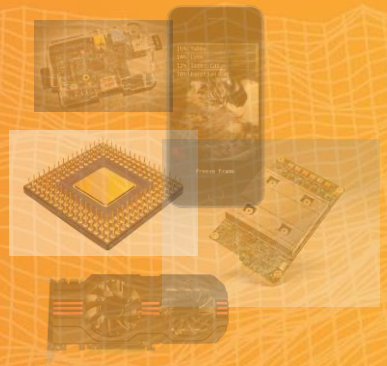


Ecosystem

MESOS

kubernetes

machine learning gets complex quickly

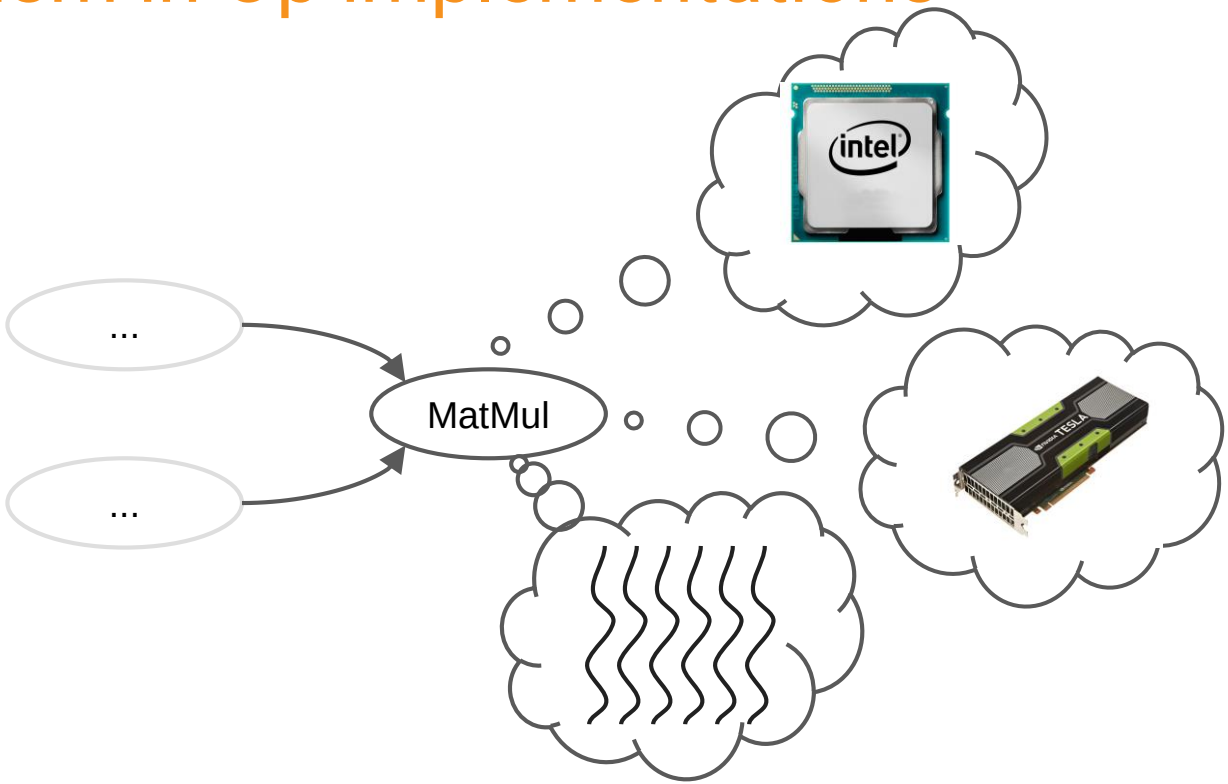


Heterogenous System

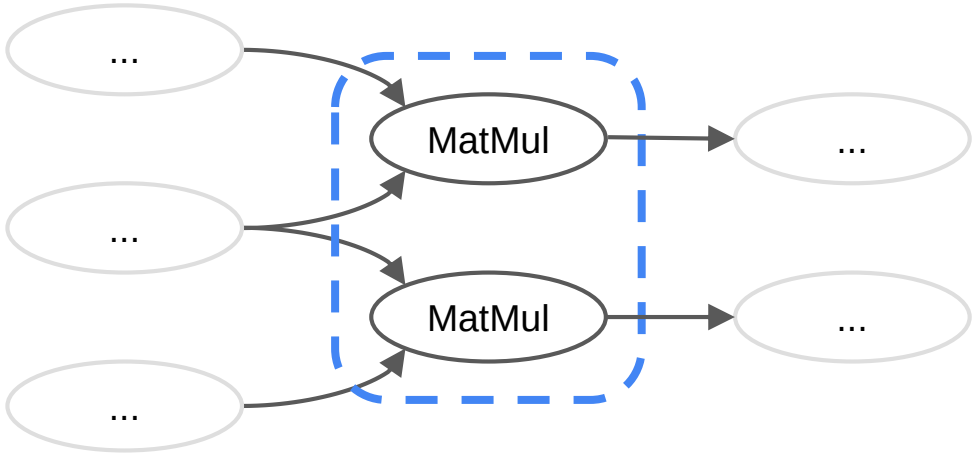
Distributed System

Modeling complexity

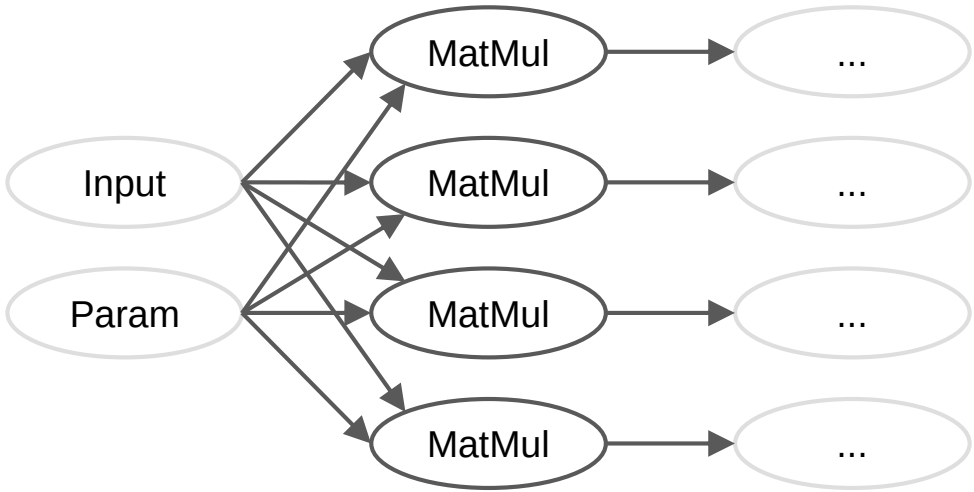
Parallelism in Op implementations



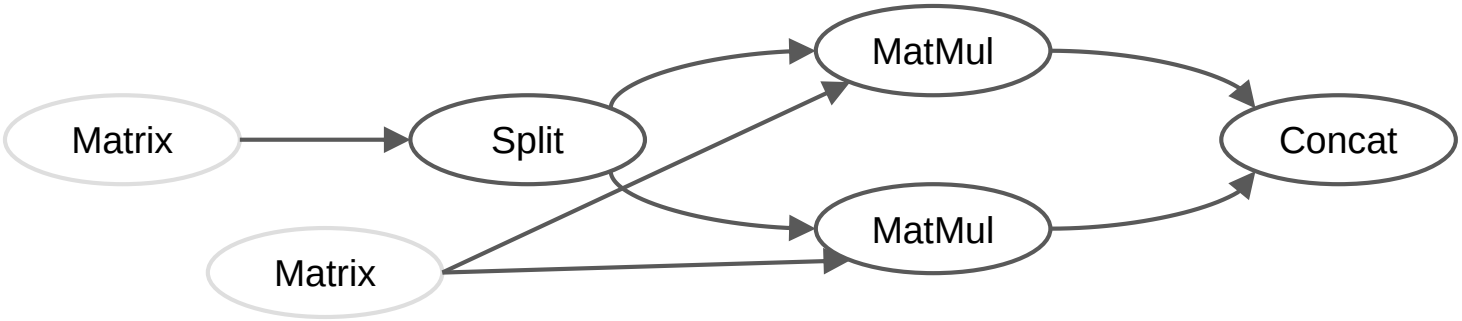
Task Parallelism in DataFlow graph



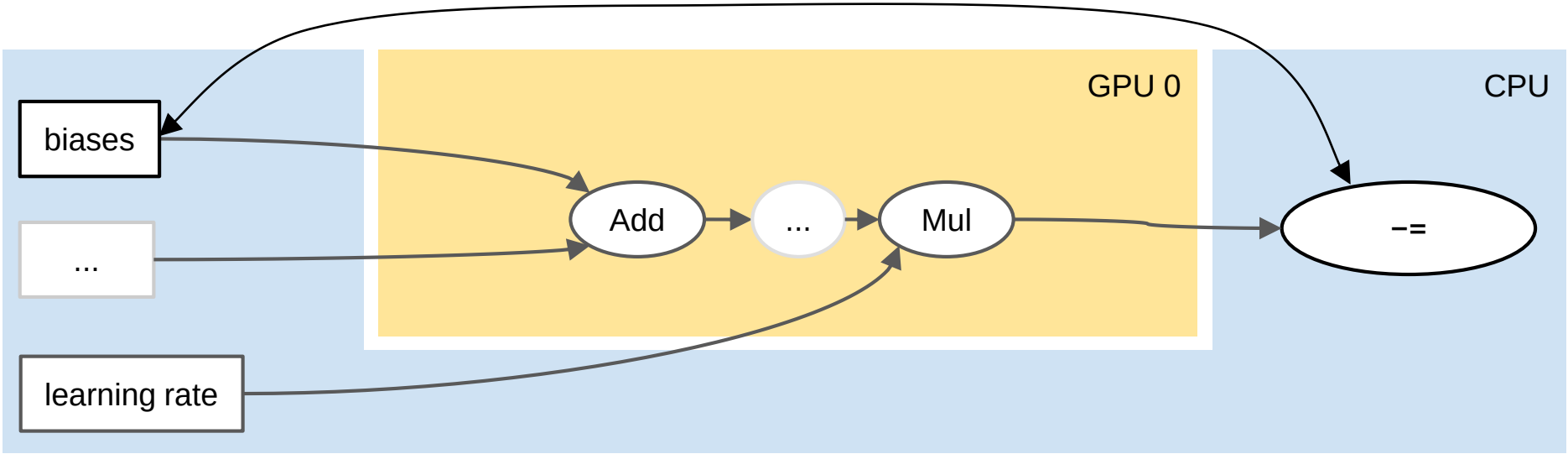
Data Parallelism



Model Parallelism

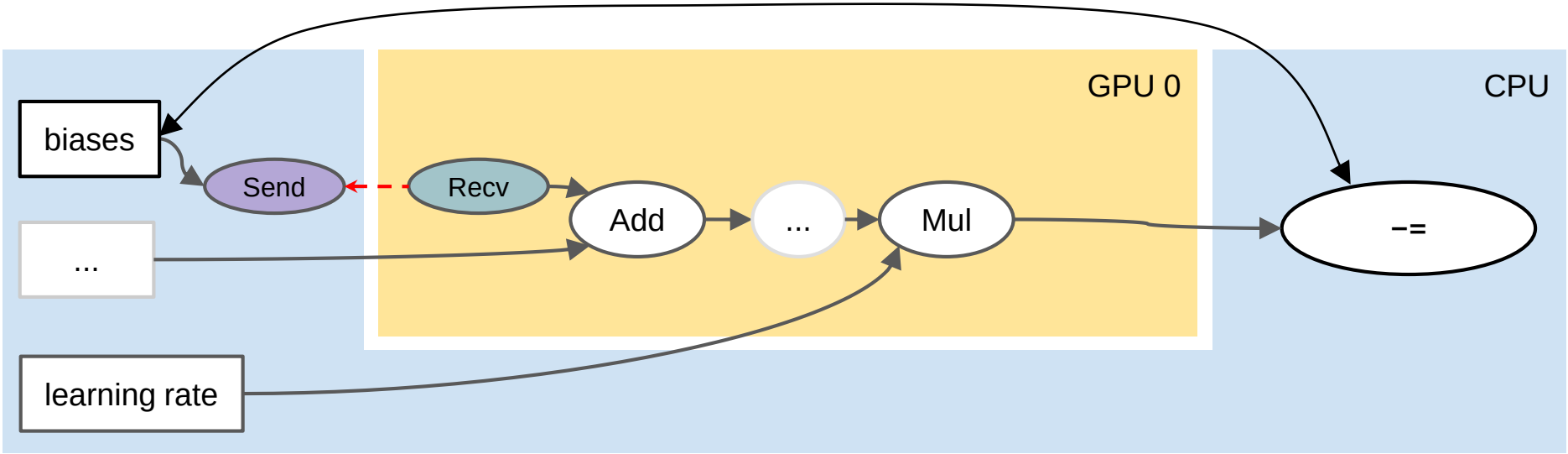


Distribution across Devices



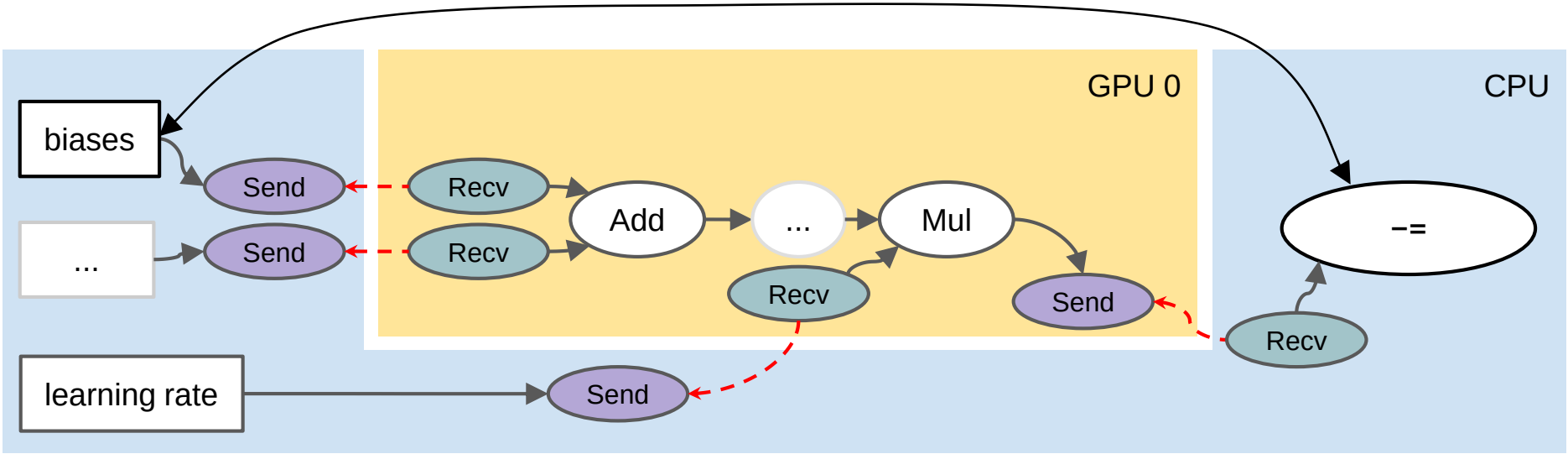
Distribution

- TensorFlow inserts *Send/Recv* Ops to transport tensors across devices
- *Recv* ops pull data from *Send* ops



Distribution

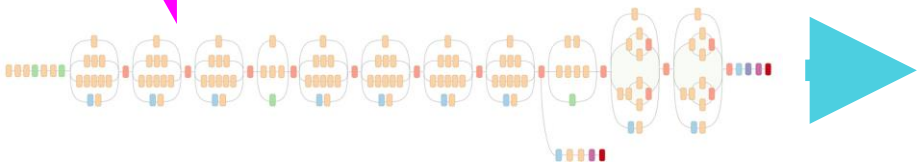
- TensorFlow inserts *Send/Recv* Ops to transport tensors across devices
- *Recv* ops pull data from *Send* ops



Just-In-Time Compilation

XLA: “Accelerated Linear Algebra” Compiler

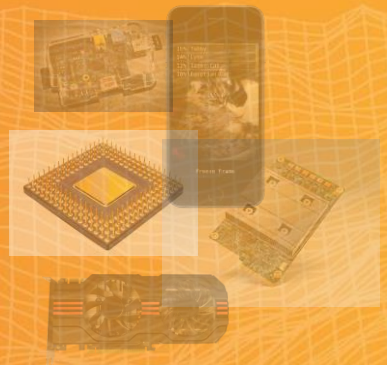
TF graphs go in,



Optimized & specialized assembly comes out.

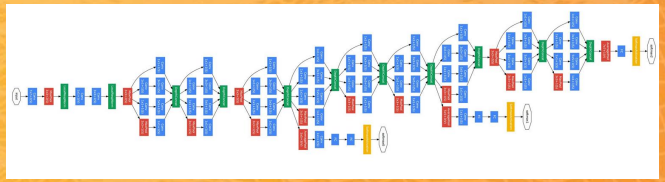
```
0x00000000      movq    (%rdx), %rax
0x00000003      vmovaps (%rax), %xmm0
0x00000007      vmulps  %xmm0, %xmm0, %xmm0
0x0000000b      vmovaps %xmm0, (%rdi)
                        ...
```

machine learning gets complex quickly

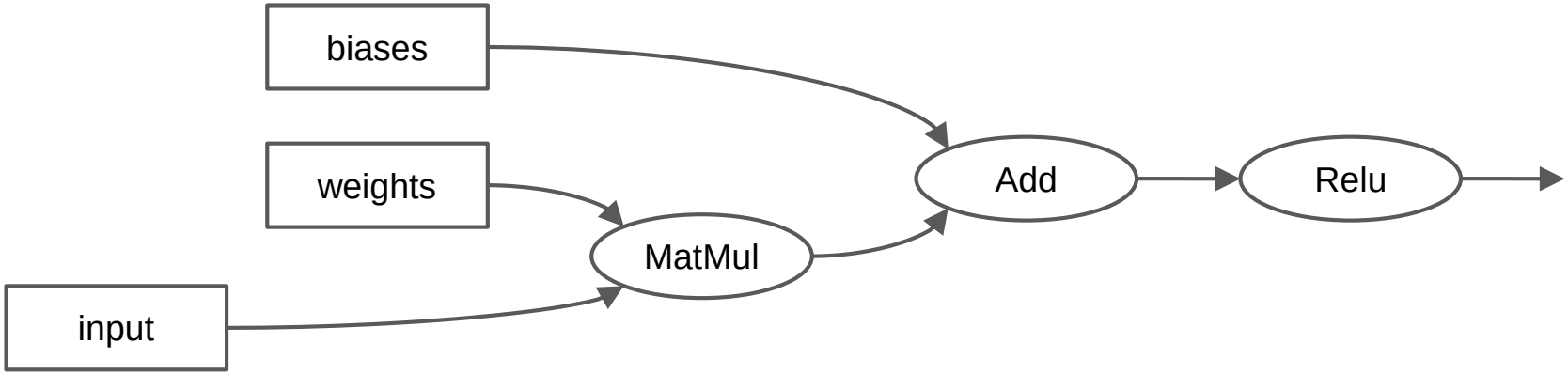


Heterogenous System

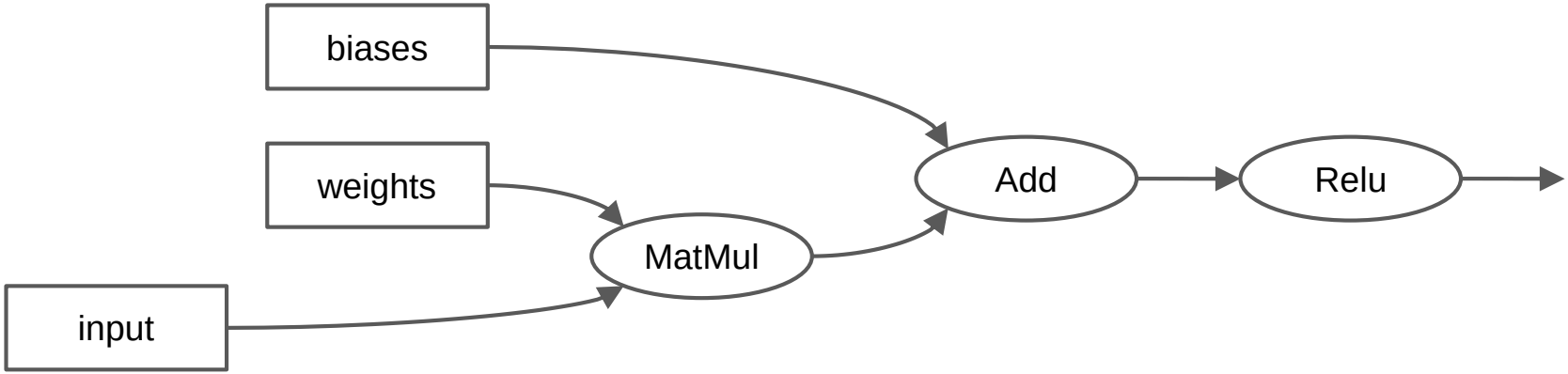
Distributed System



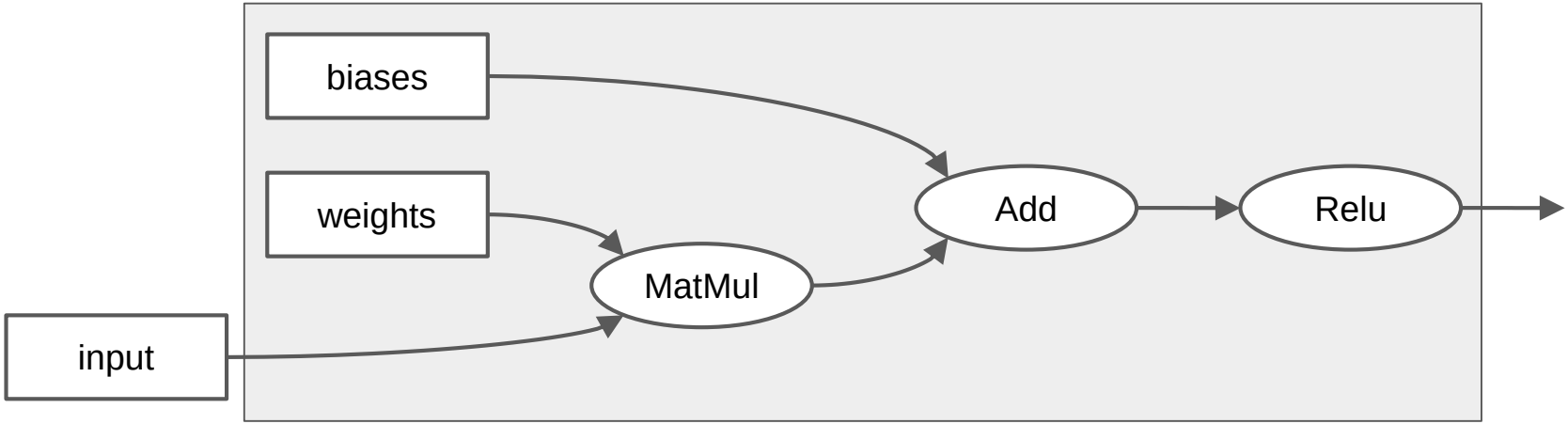
Modeling complexity




```
input = ...
biases = tf.get_variable('biases', ...)
weights = tf.get_variable('weights', ...)
out = tf.matmul(input, weights)
out = tf.add(out, biases)
out = tf.nn.relu(out)
```



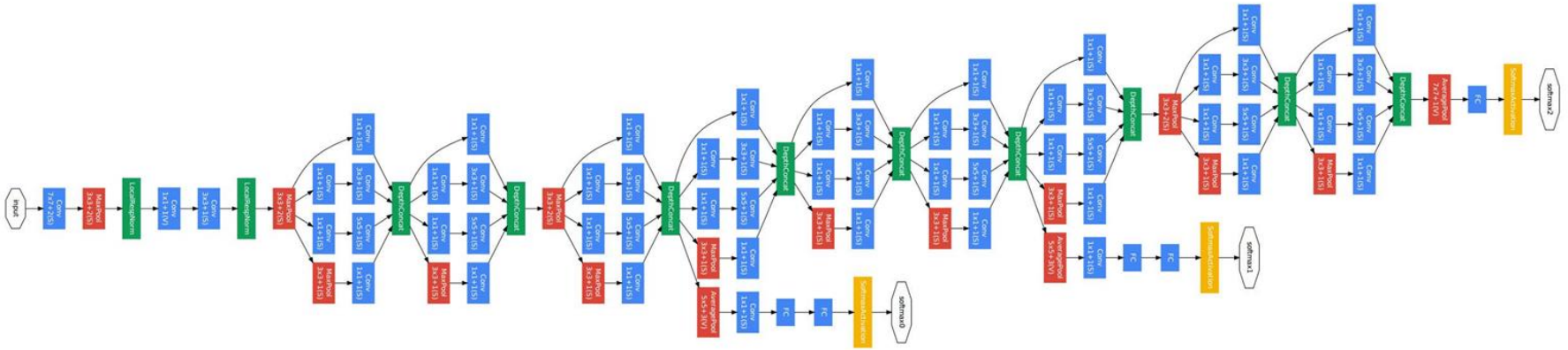

```
input = ...
output = tf.contrib.layers.fully_connected(input, ...)
```




```
input = ...
output = tf.contrib.layers.fully_connected(input, ...)
```

Align cognitive model with programming model

Each box is one line of code!



TensorFlow contains complete algorithms

Linear{Classifier,Regressor}

DNN{Classifier,Regressor}

DNNLinearCombined{Classifier,Regressor}

SVM

KMeansClustering

GMM

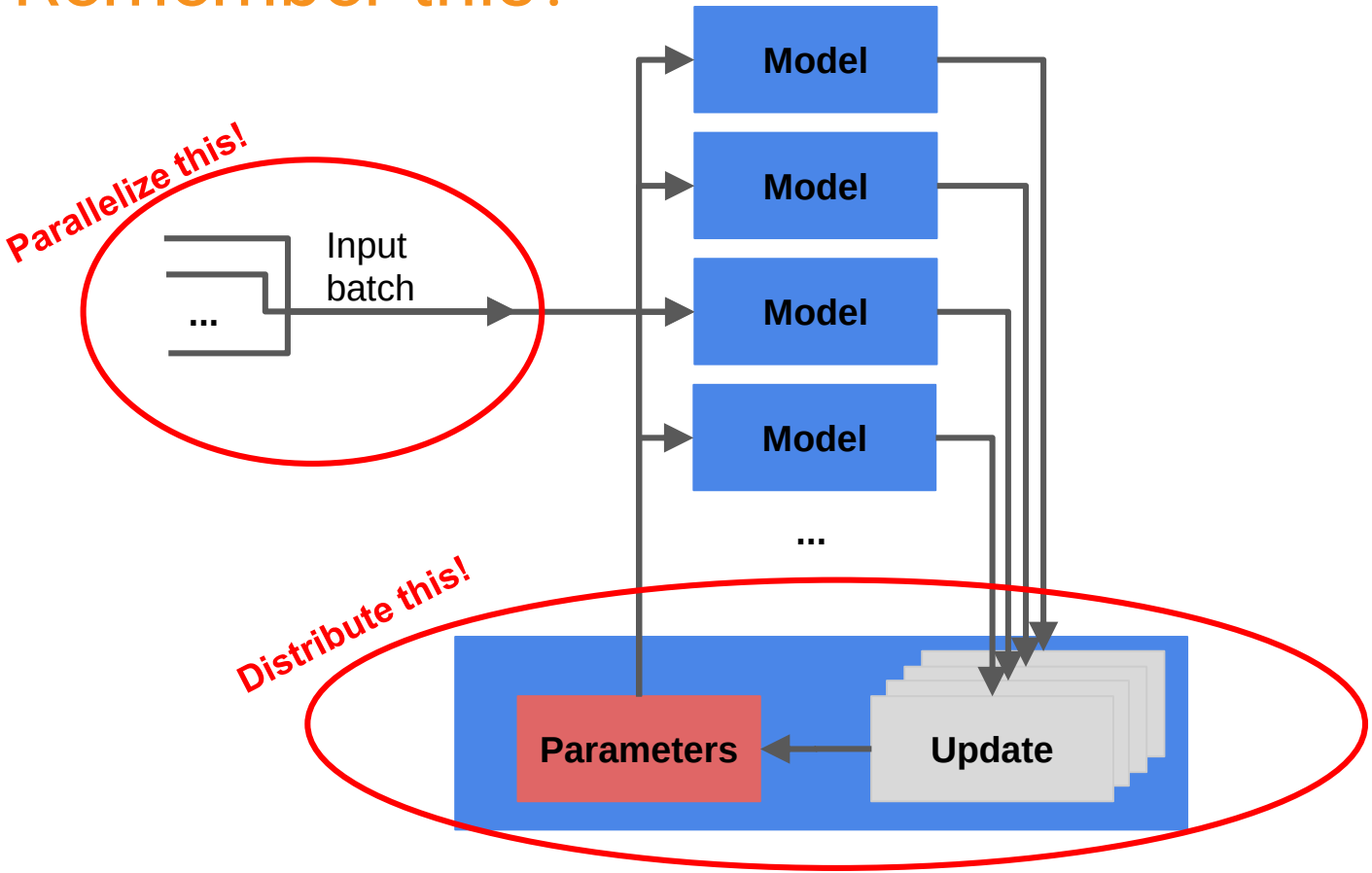
...

Simple machine learning

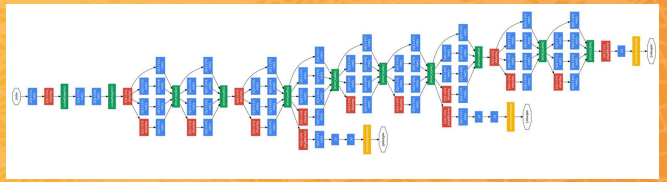
```
classifier = learn.LinearClassifier(feature_columns=feature_columns,
                                    n_classes=10)

classifier.fit(data, labels, batch_size=100, steps=1000)

classifier.evaluate(eval_data, eval_labels)
```

Tooling provided for distributed training and evaluation, graphical debugging, and export to production server (tensorflow/serving).

Remember this?



TensorFlow handles complexity for you...



...so you can focus on your ideas

Thank You!

Rajat Monga
@rajatmonga