PaddlePaddle

Kubernetes and CUDA

History

- Before open source
  - Baidu internal project
  - developed 4 years ago by Xu Wei
  - 50+ product features in Baidu
  - Twice the Million-Dollar-Prize winner

- After open source
  - open sourced in September 2016
  - new Python API
  - support Jupyter Notebook
  - under a significant rewrite

Design

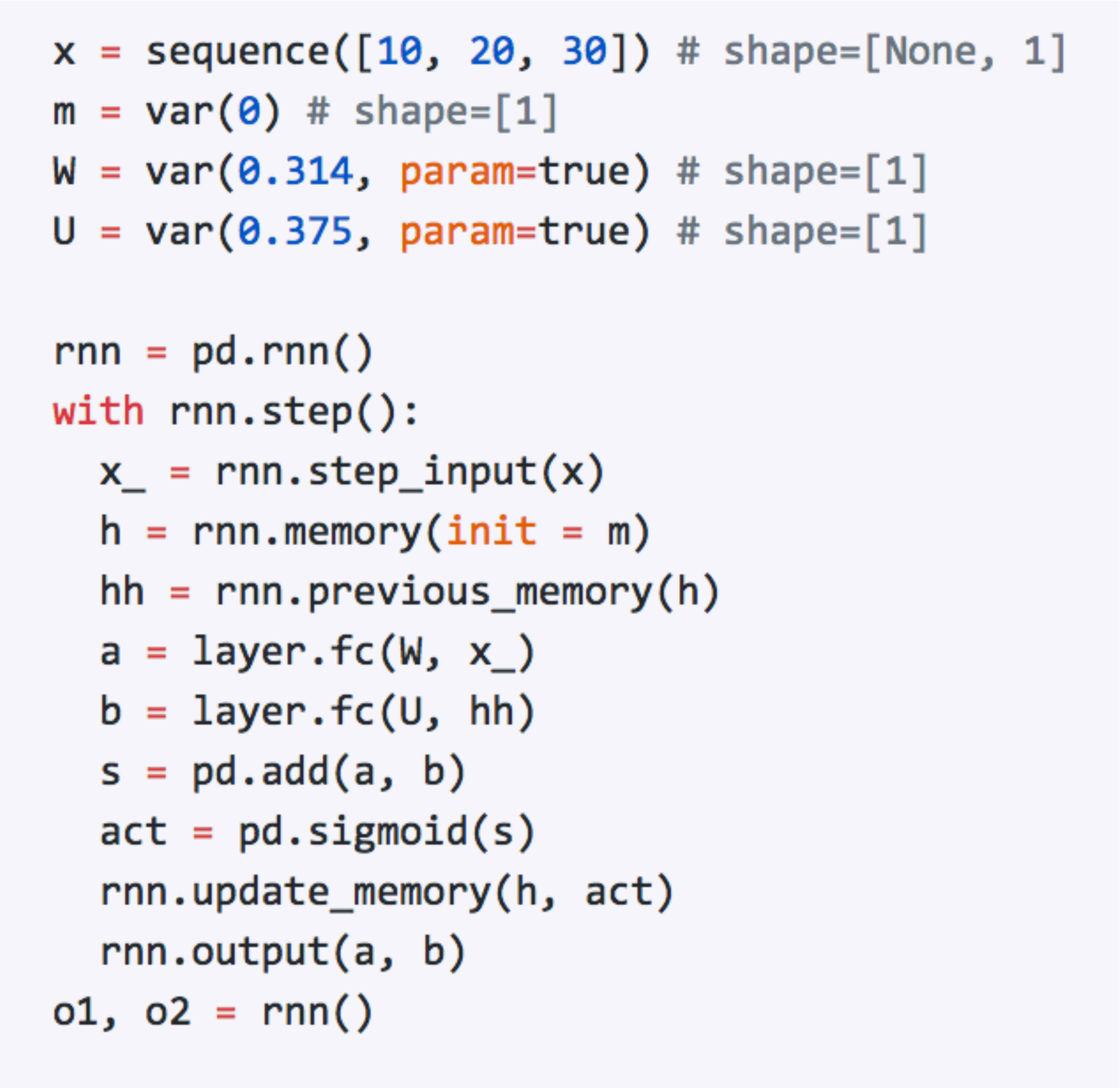
- New representation of deep learning computation
  - Caffe, Torch, Paddle: **sequences of layers**
  - TensorFlow, Caffe2, MxNet: **graphs of operators**
  - PaddlePaddle: **nested blocks**

- Large-scale training and inference
  - auto-scalable deep learning
  - A complete solution — whole business on a private cloud

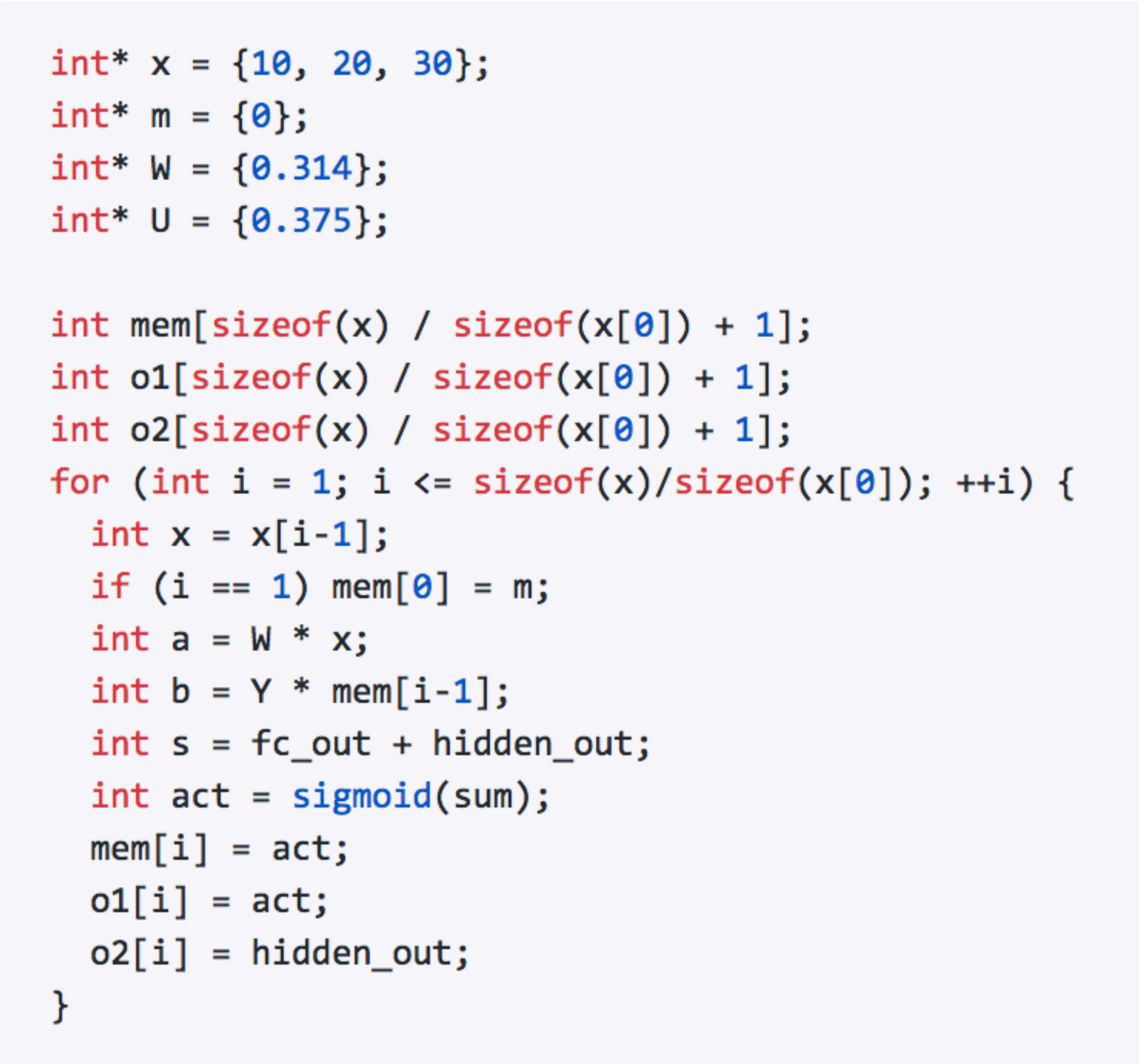
Blocks

programming languages	PaddlePaddle
for, while	RNN, WhileOp
if-else, switch	IfElseOp, SwitchOp
sequential execution	a sequence of layers

RNN / Loop

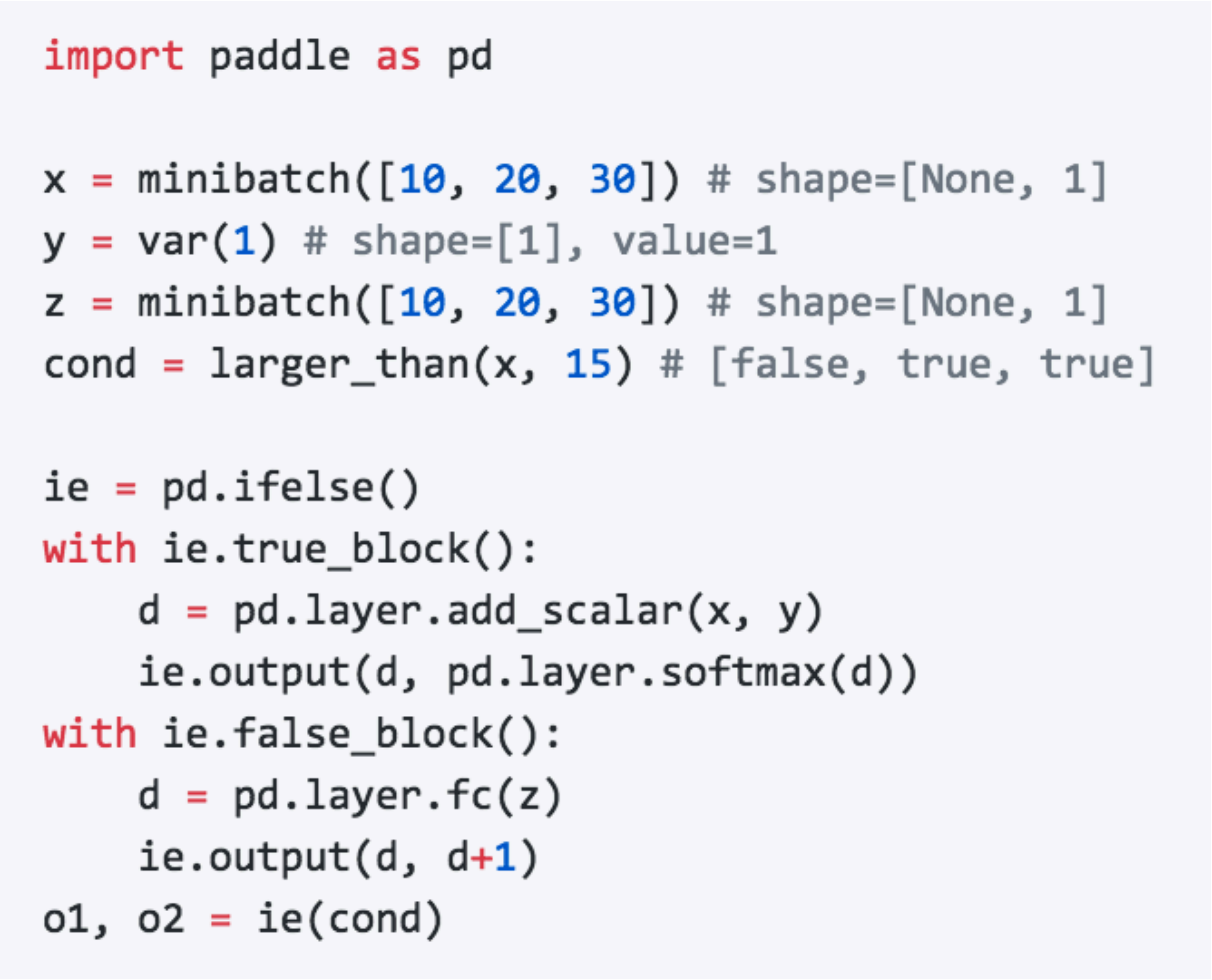


```
x = sequence([10, 20, 30]) # shape=[None, 1]
m = var(0) # shape=[1]
W = var(0.314, param=true) # shape=[1]
U = var(0.375, param=true) # shape=[1]

rnn = pd.rnn()
with rnn.step():
  x_ = rnn.step_input(x)
  h = rnn.memory(init = m)
  hh = rnn.previous_memory(h)
  a = layer.fc(W, x_)
  b = layer.fc(U, hh)
  s = pd.add(a, b)
  act = pd.sigmoid(s)
  rnn.update_memory(h, act)
  rnn.output(a, b)
o1, o2 = rnn()
```

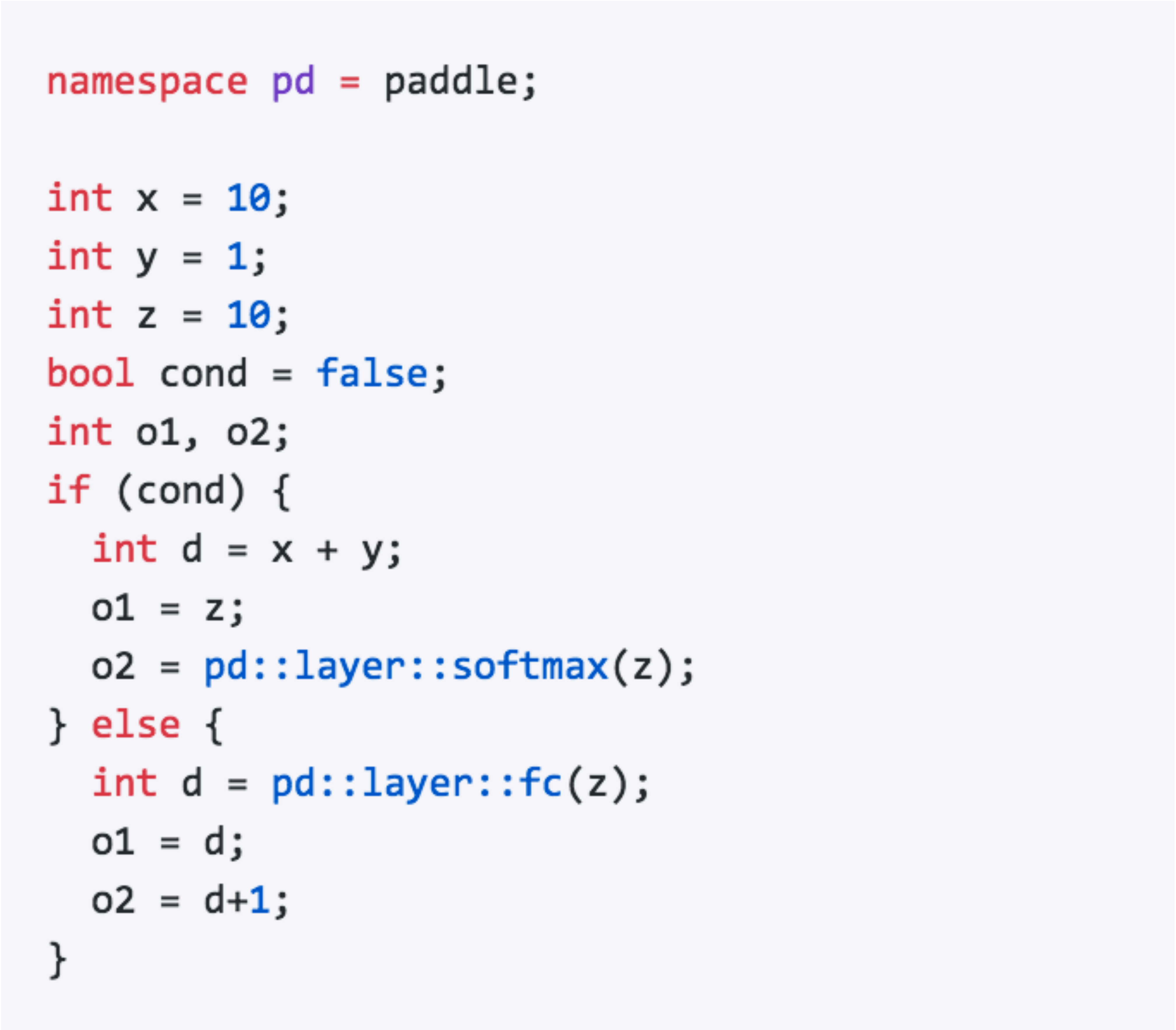


```
int* x = {10, 20, 30};
int* m = {0};
int* W = {0.314};
int* U = {0.375};

int mem[sizeof(x) / sizeof(x[0]) + 1];
int o1[sizeof(x) / sizeof(x[0]) + 1];
int o2[sizeof(x) / sizeof(x[0]) + 1];
for (int i = 1; i <= sizeof(x)/sizeof(x[0]); ++i) {
  int x = x[i-1];
  if (i == 1) mem[0] = m;
  int a = W * x;
  int b = Y * mem[i-1];
  int s = fc_out + hidden_out;
  int act = sigmoid(sum);
  mem[i] = act;
  o1[i] = act;
  o2[i] = hidden_out;
}
```

If-else / IfElseOp



```python
import paddle as pd

x = minibatch([10, 20, 30]) # shape=[None, 1]
y = var(1) # shape=[1], value=1
z = minibatch([10, 20, 30]) # shape=[None, 1]
cond = larger_than(x, 15) # [false, true, true]

ie = pd.ifelse()
with ie.true_block():
    d = pd.layer.add_scalar(x, y)
    ie.output(d, pd.layer.softmax(d))
with ie.false_block():
    d = pd.layer.fc(z)
    ie.output(d, d+1)
o1, o2 = ie(cond)
```



```cpp
namespace pd = paddle;

int x = 10;
int y = 1;
int z = 10;
bool cond = false;
int o1, o2;
if (cond) {
  int d = x + y;
  o1 = z;
  o2 = pd::layer::softmax(z);
} else {
  int d = pd::layer::fc(z);
  o1 = d;
  o2 = d+1;
}
```

Execution

- Programming languages
  - stack push when entering block
  - stack pop when leaving block

- PaddlePaddle
  - stack push when entering block
  - no pop when leaving
  - destroy after a mini-batch

Acceleration

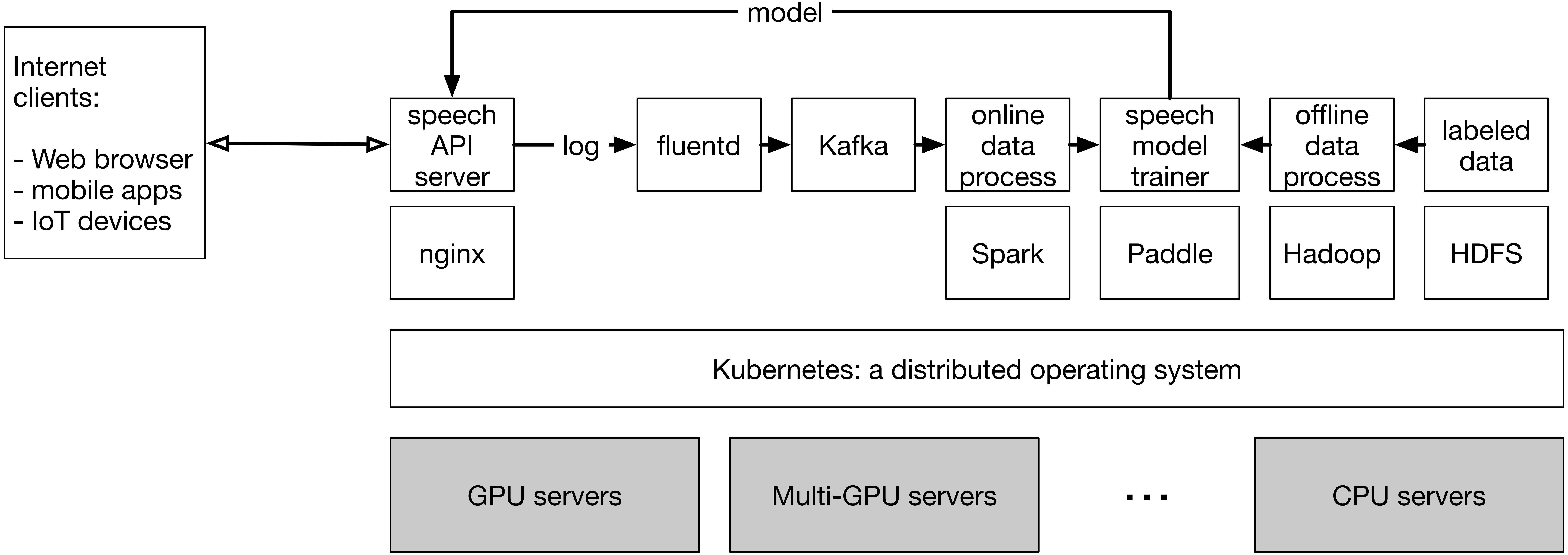
- A block
    - local variables
    - a sequence of instructions

- Instruction types
    - computational
        - fully-connected, CNN
    - control flow
        - IfElse, RNN, While
    - I/O
        - rend/recv, rendezvous

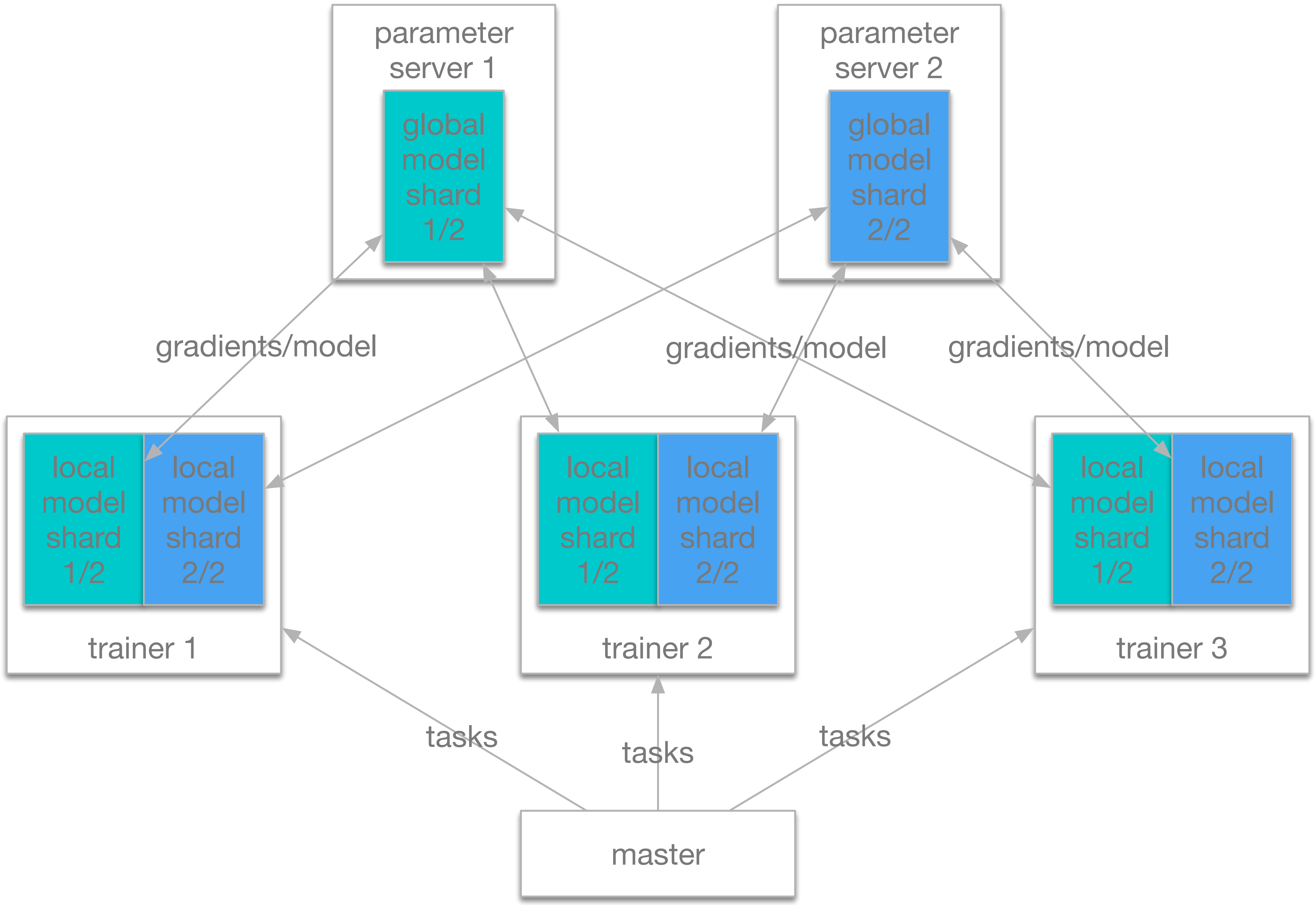
Industrial Solutions

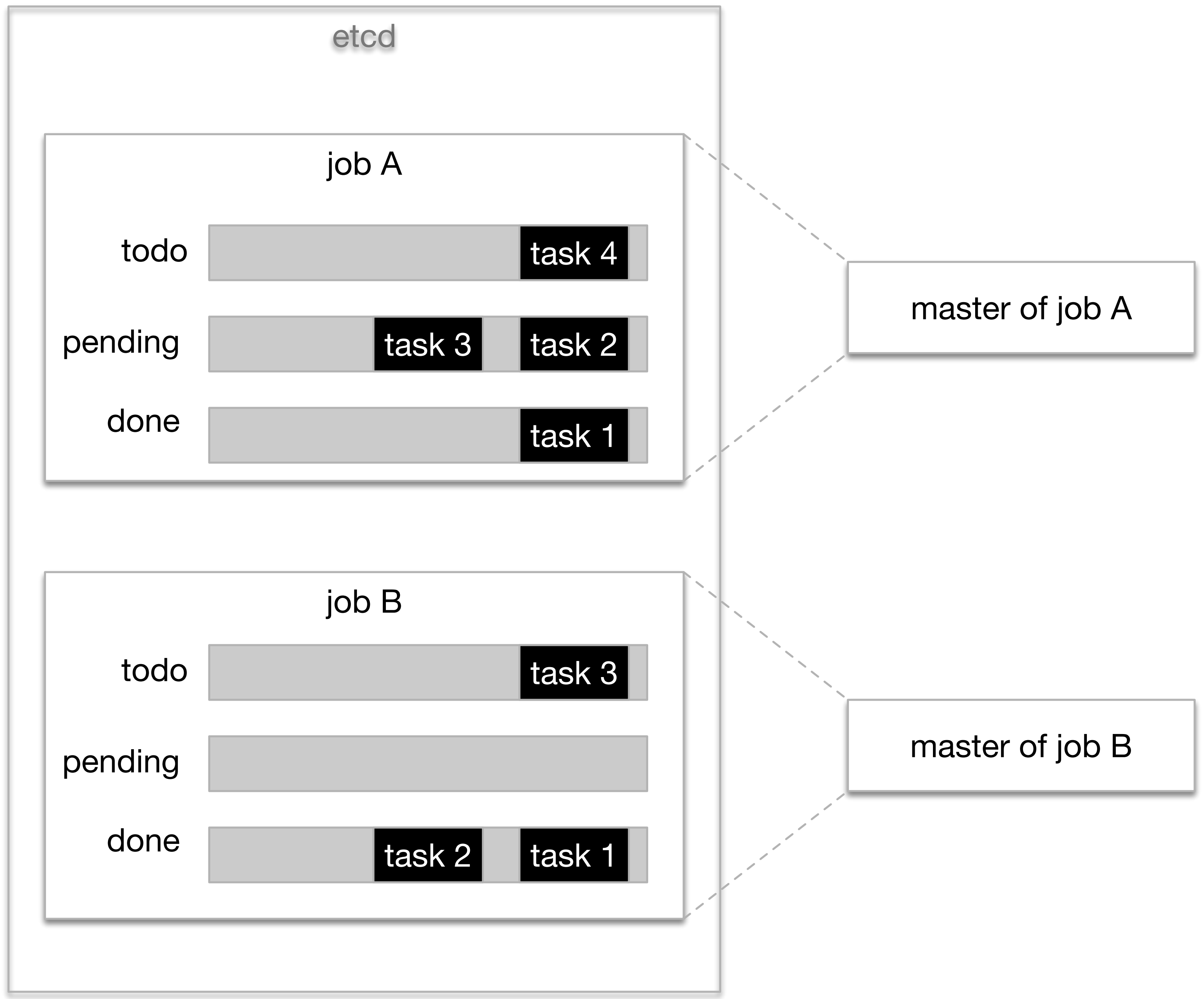
	Internet	traditional
big companies	on-premises cluster	on-premises cluster
small companies	cloud	on-premises cluster

General-purpose Cluster



Fault Recovery





Project Info

- Main repo: https://github.com/PaddlePaddle/paddle
- Paddle Book: http://book.paddlepaddle.org

- Model Bank: https://github.com/paddlepaddle/models
- Paddle Cloud: https://github.com/PaddlePaddle/cloud

Thank you!

Baidu百度