

Embedded ML Software





PyTorch



TensorFlow

Caffe



scikit
learn

theano



torch

Canned Estimators

Estimators

Keras Model

Layers

Datasets



Python Frontend

C++

Java

Go

...

TensorFlow Distributed Execution Engine

CPU

GPU

OS

XLA (Accelerated LinAlg Compiler)

CPU

GPU

TPU

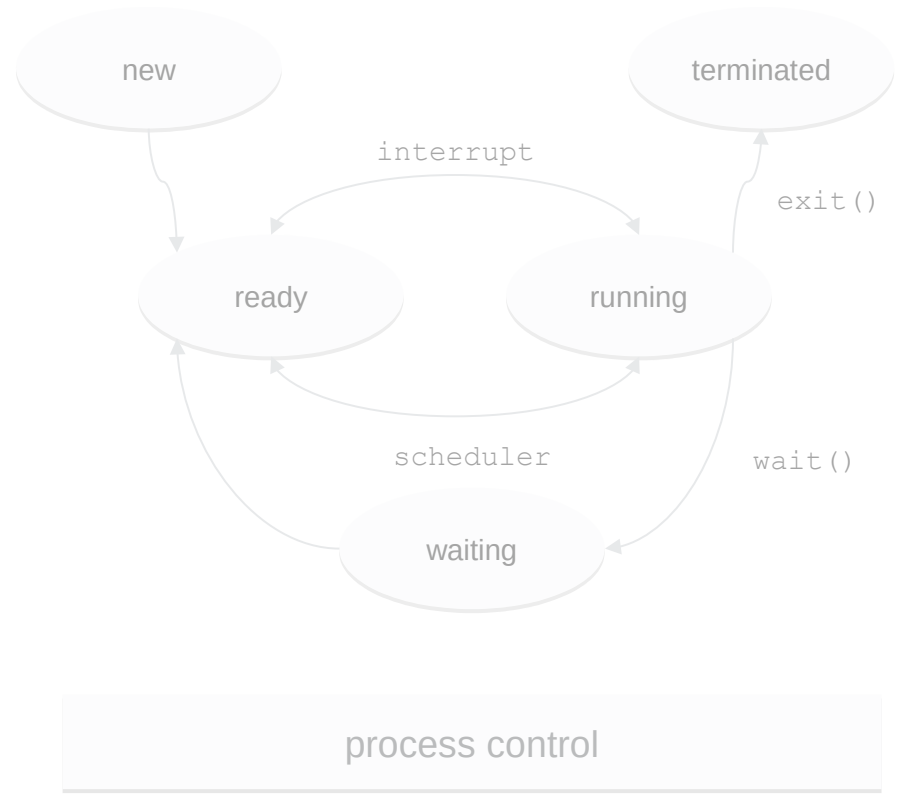
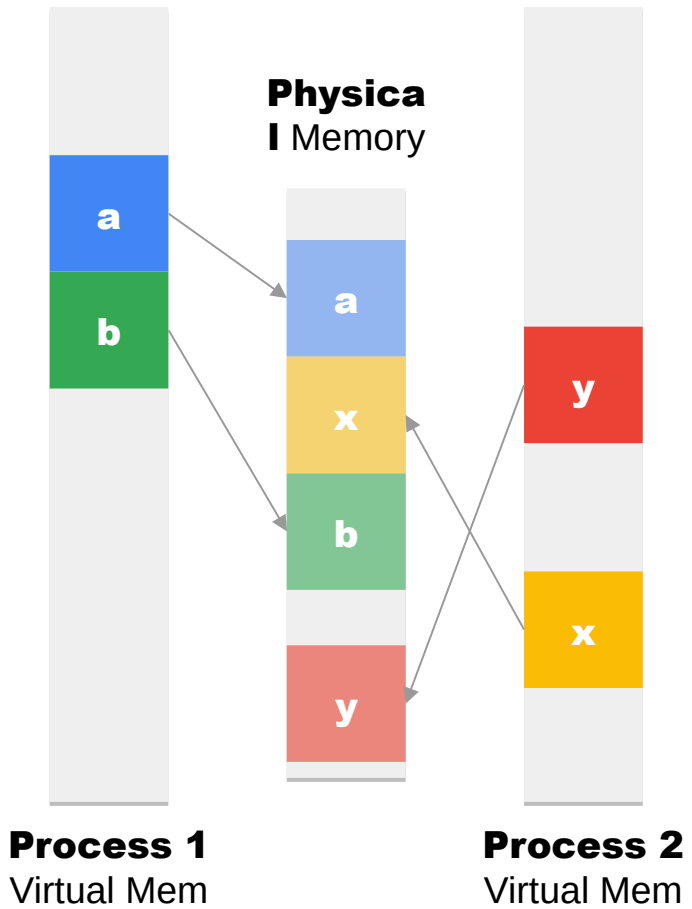
...

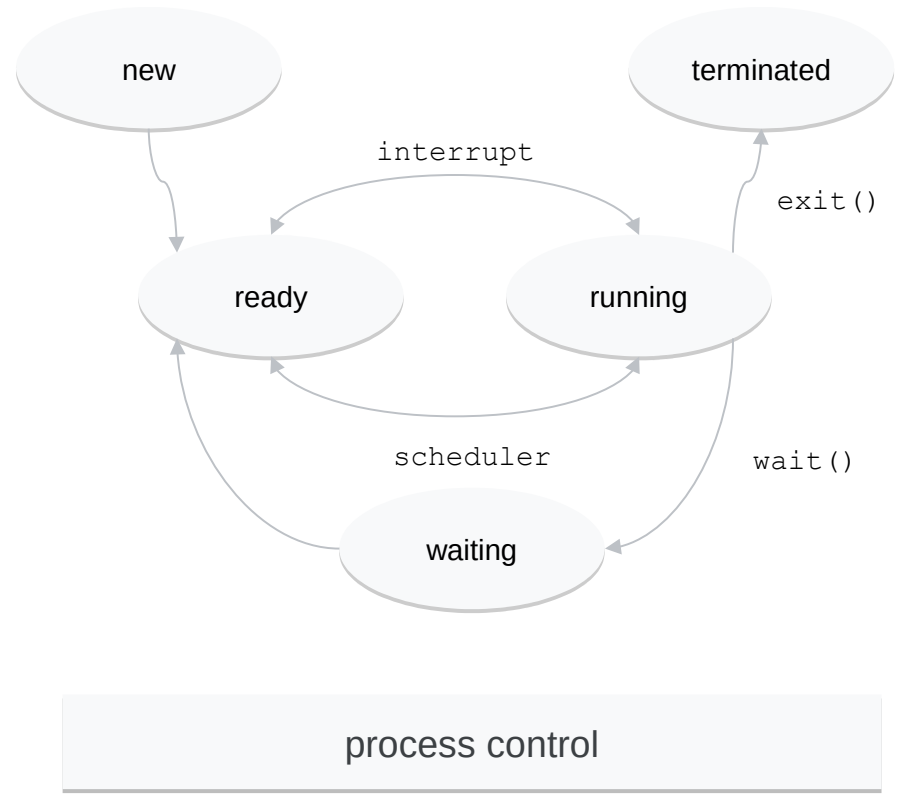
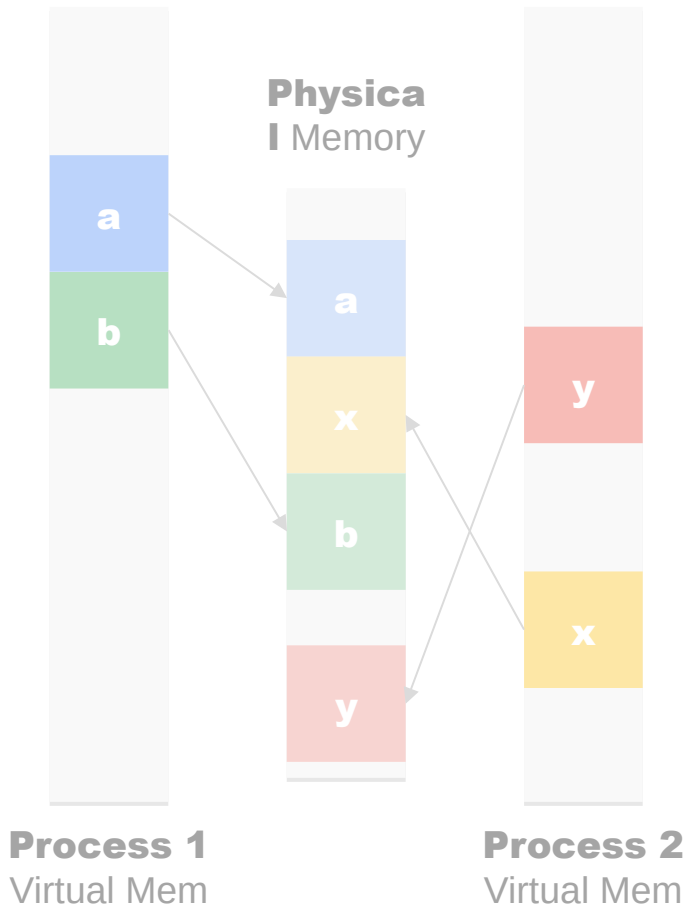


Linux



Mobile OS





Embedded Systems

Arduino
BLE Sense 33

Himax
WE-I Plus EVB

SparkFun
Edge 2

Espressif
EYE

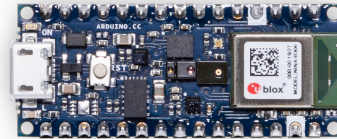
Embedded Systems



Arduino
BLE Sense 33


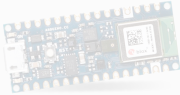
Himax
WE-I Plus EVB

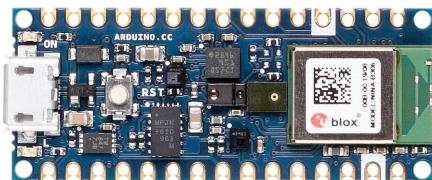
SparkFun
Edge 2

Espressif
EYE



	Microprocessor	>	Microcontroller
Platform			
Compute	1GHz–4GHz	~10X	1MHz–400MHz
Memory	512MB–64GB	~10000X	2KB–512KB
Storage	64GB–4TB	~100000 X	32KB–2MB
Power	30W–300W	~1000X	150μW–23.5mW

	Microprocessor	>	Microcontroller
Platform			
Compute	1GHz–4GHz	~ 10X	1MHz–400MHz
Memory	512MB–64GB	~ 10000X	2KB–512KB
Storage	64GB–4TB	~ 100000X	32KB–2MB
Power	30W–300W	~ 1000X	150μW–23.5mW



Less memory

Limited OS support

Lower compute power

Only focused on *inference*



TinyML Inference Framework

```
graph TD; A[TinyML Inference Framework] --- B[Model]; A --- C[Software]; A --- D[Hardware];
```

Model

Software

Hardware

ML Framework Design Checklist

Model

Support for training?

Yes

No

Support for inference?

Yes

No

How many ops?

few

ML Framework Design Checklist

Model

Support for training?

Yes

No

Support for inference?

Yes

No

How many ops?

few

Software

Need quantization/optimization tools?

Yes

No

Can rely on virtual memory support?

Yes

No

ML Framework Design Checklist

Model

Support for training?	Yes	No
Support for inference?	Yes	No
How many ops?	few	

Software

Need quantization/optimization tools?	Yes	No
Can rely on virtual memory support?	Yes	No

Hardware

Support a diverse range of processor hardware on a device?	Yes	No
Need to support many different platforms and architectures?	Yes	No

“Tiny” Machine Learning Frameworks



tvm



TensorFlow Lite Micro



uTenso



GLOW

r



STM32
Cube.AI

Arduino
BLE Sense 33

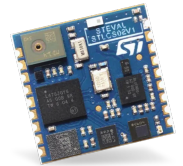
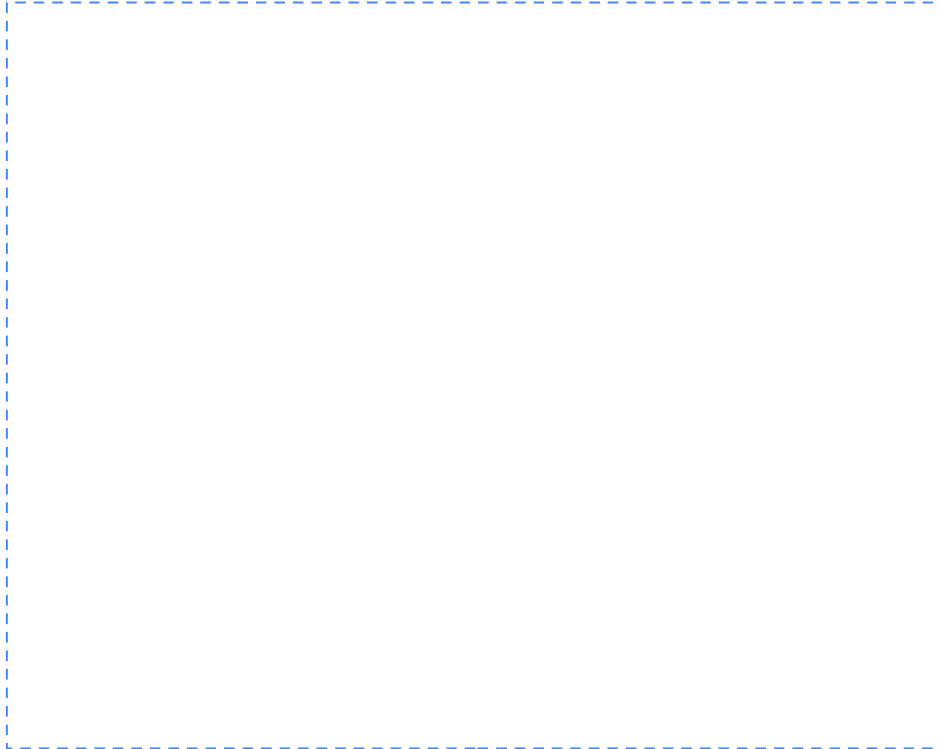
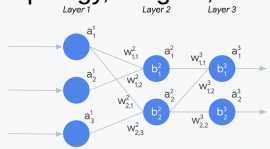
Himax
WE-I Plus EVB

SparkFun
Edge 2

Espressif
EYE

pre-trained NN model

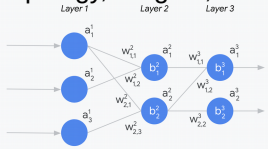
topology, weights, bias



STM32L476
Cortex-M4 **MCU**

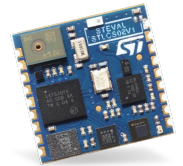
pre-trained NN model

topology, weights, bias



STM32
Cube.AI

command line
interface



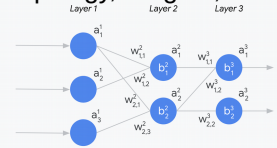
STM32L476
Cortex-M4 **MCU**

parameters:

(1) name, (2) compression factor, (3) targeted STM32

pre-trained NN model

topology, weights, bias

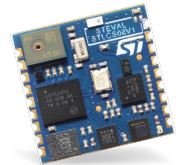


upload

Generate



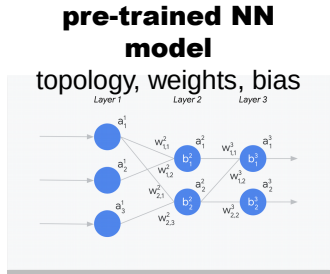
command line interface



STM32L476
Cortex-M4 **MCU**

parameters:

(1) name, (2) compression factor, (3) targeted STM32



upload



PINNR

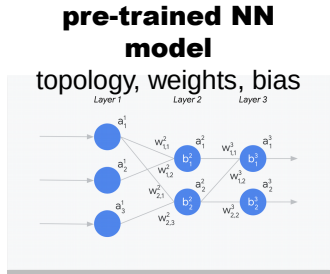
Generate



STM32L476
Cortex-M4 **MCU**

parameters:

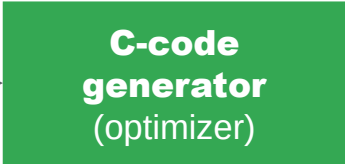
(1) name, (2) compression factor, (3) targeted STM32



upload



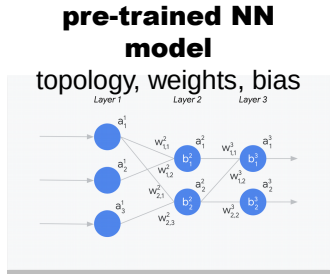
PINNR



Generate



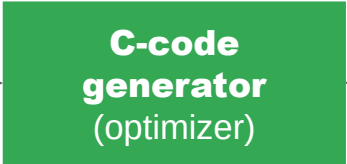
STM32L476
Cortex-M4 **MCU**



upload



PINNR



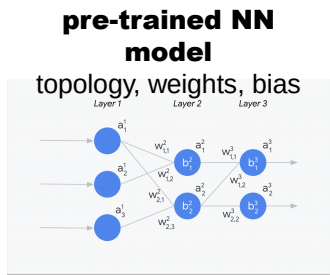
Generate

parameters:

(1) name, (2) compression factor, (3) targeted STM32



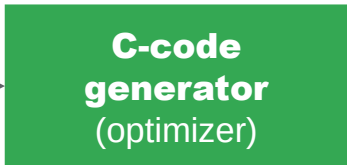
STM32L476
Cortex-M4 **MCU**



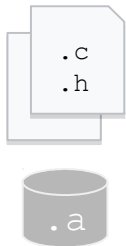
upload



PINNR



Generate



target-dependent
**optimized kernel
runtime libraries**

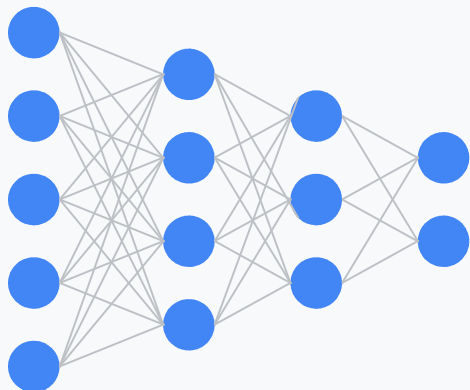


STM32L476
Cortex-M4 **MCU**

parameters:

(1) name, (2) compression factor, (3) targeted STM32

Model



3 dense **layers**

321 **parameters**

No quantization or compression

STM32 
Cube.AI

27kB Flash

5kB RAM

77uS Inference
Time

Closed Source



TensorFlow Lite

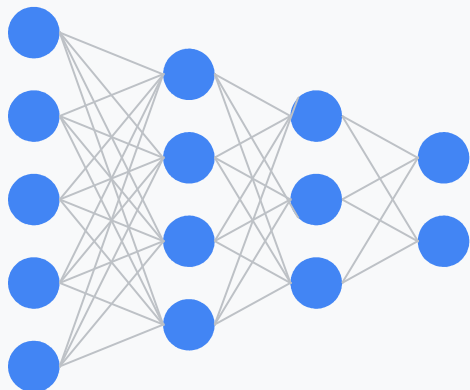
50kB Flash

4.7kB RAM

104uS Inference
Time

Open Source

Model



3 dense **layers**

321 **parameters**

No quantization or compression

STM32 
Cube.AI

27kB Flash
5kB RAM
77uS Inference

Time

Closed Source



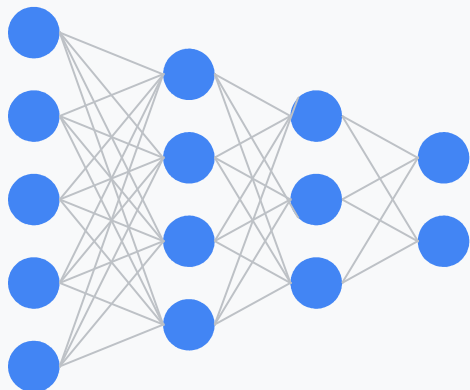
TensorFlow Lite

50kB Flash
4.7kB RAM
104uS Inference

Time

Open Source

Model



3 dense **layers**

321 **parameters**

No quantization or compression

STM32 
Cube.AI

27kB Flash

5kB RAM

77uS Inference

Time

Closed Source



TensorFlow Lite

50kB Flash

4.7kB RAM

104uS Inference

Time

Open Source

Choosing Frameworks

Hardware, IDE

compute
memory constraints

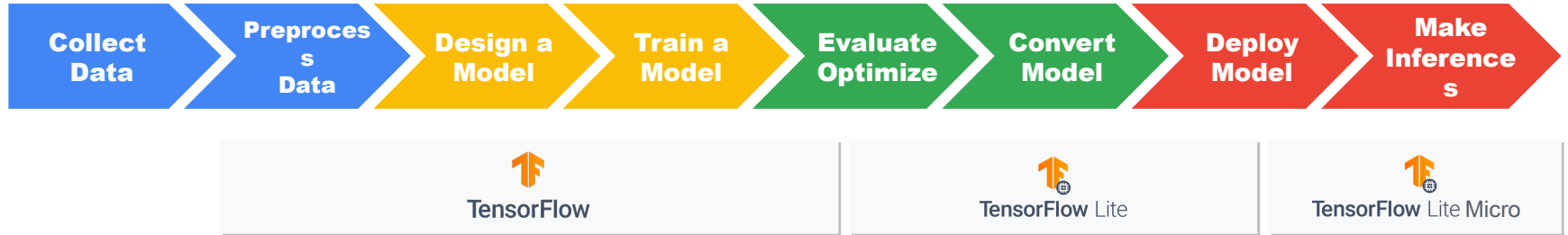
Training/Embedded

training framework
embedded
framework (you'll
export to)

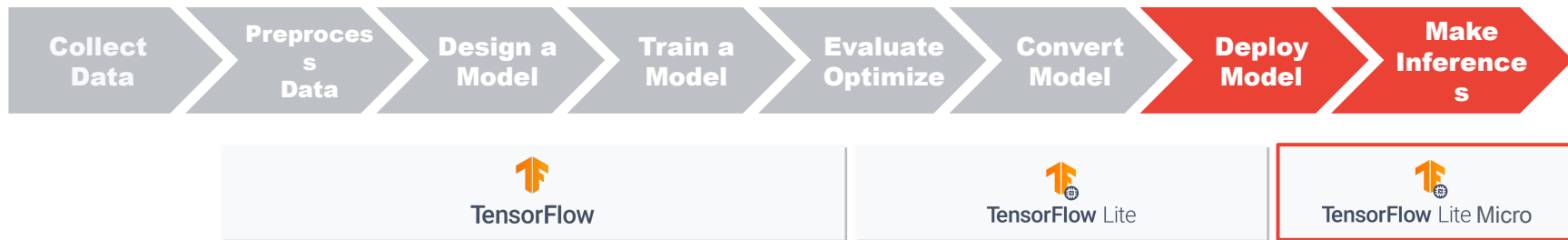
other

documentation
sample code
(for use case)
personal experience

Choosing Frameworks



Choosing Frameworks



Pete Warden, Technical Lead, TensorFlow Mobile and Embedded Team, Google.