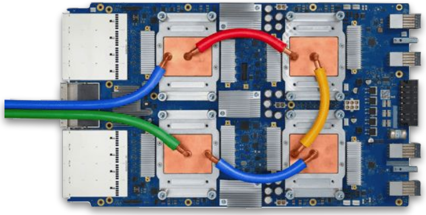
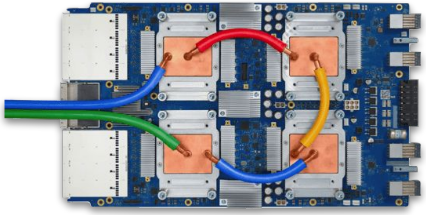
TF vs. TFLite vs. TFLite Micro

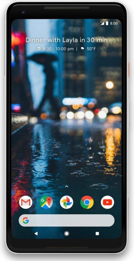
TensorFlow



TensorFlow

TensorFlow Lite

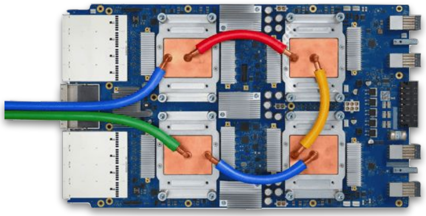


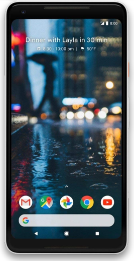


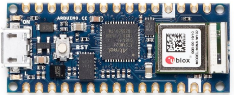
TensorFlow

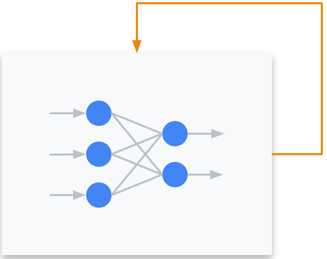
TensorFlow Lite

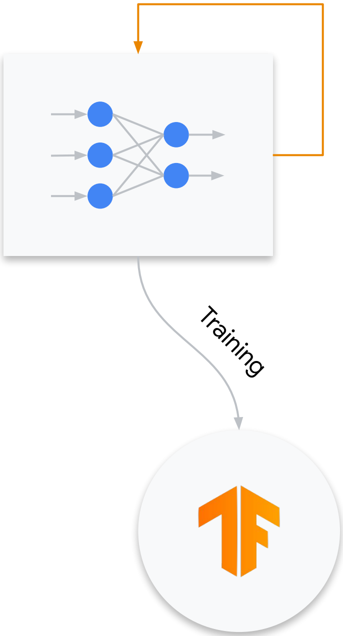
TensorFlow Lite Micro

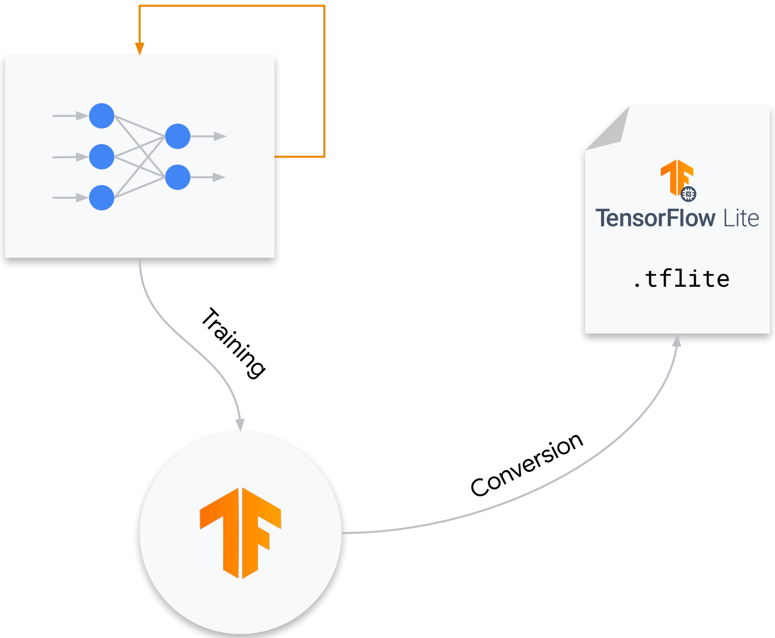


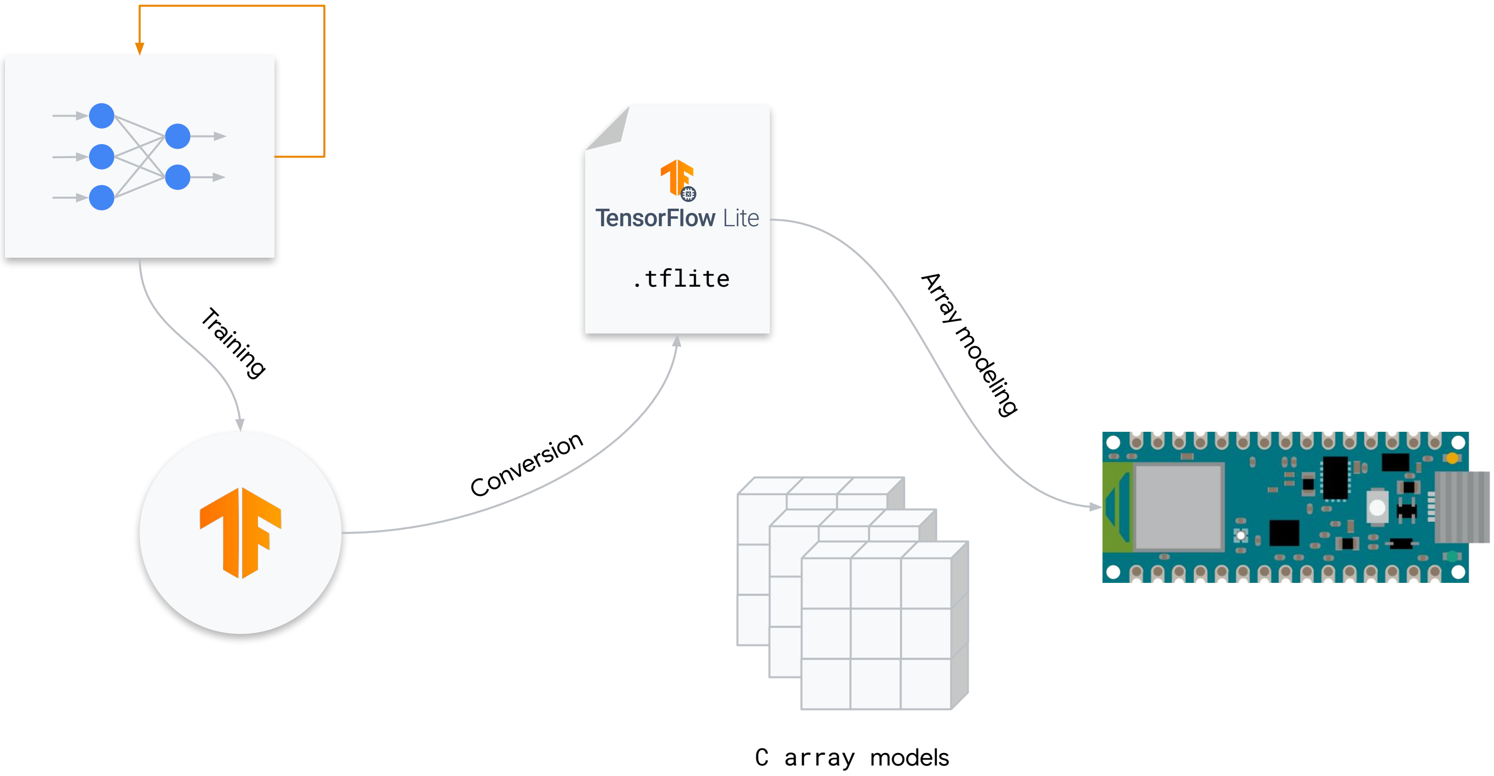


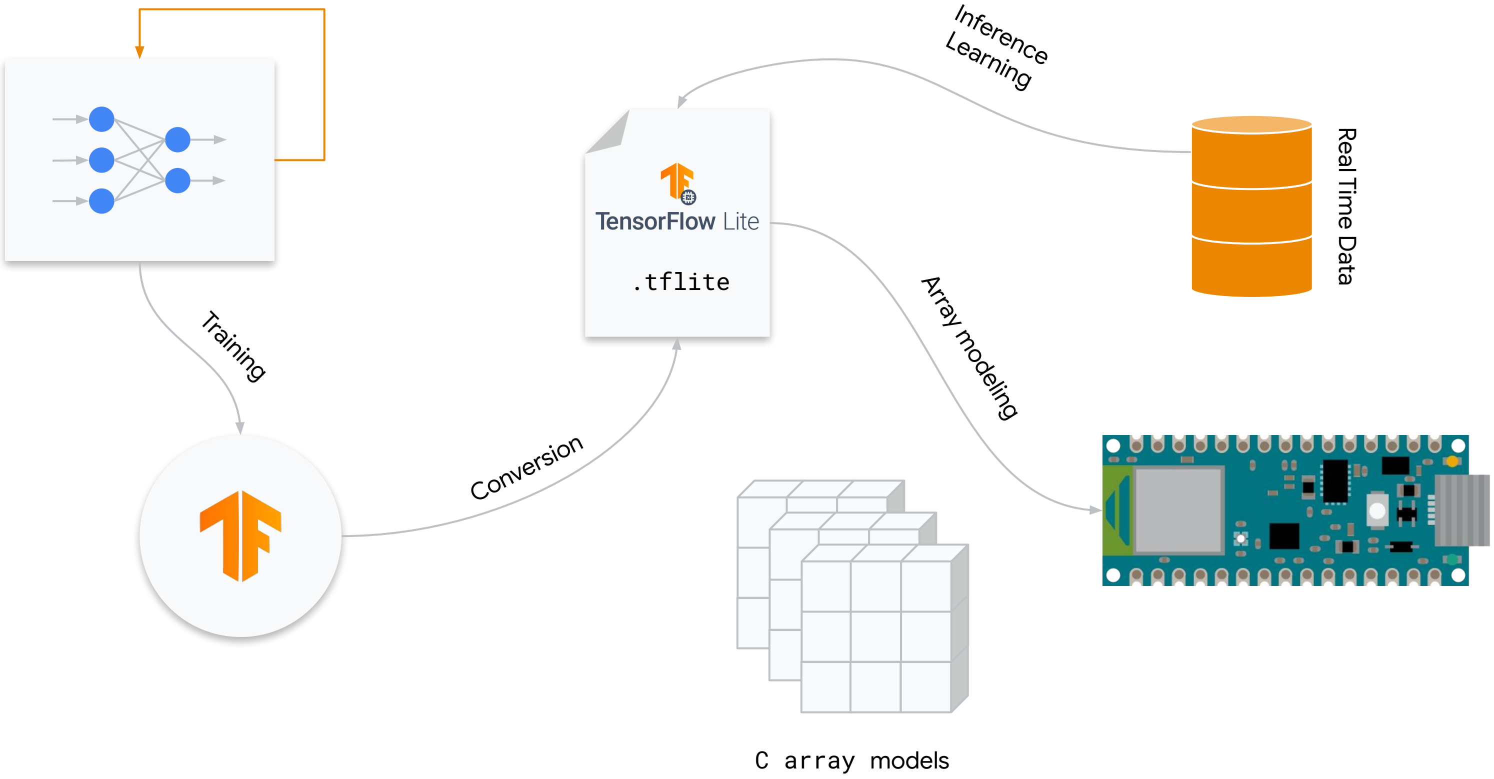


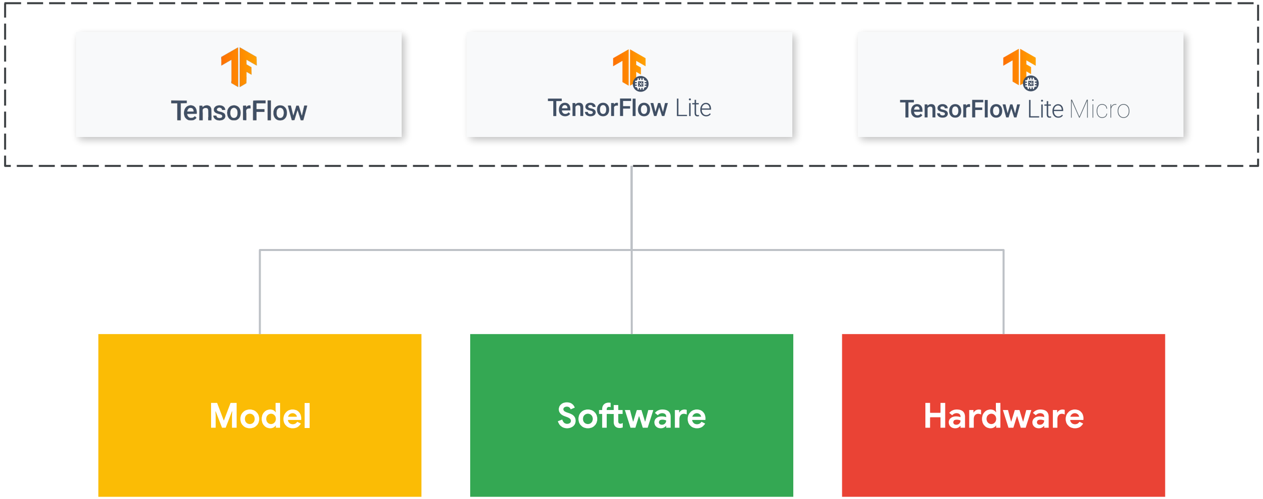



Training


TensorFlow Lite
.tflite
Training
Conversion


TensorFlow Lite
.tflite
Training
Conversion
Array modeling
C array models


Inference
Learning
TensorFlow Lite
.tflite
Real Time Data
Training
Array modeling
Conversion
C array models



TensorFlow

TensorFlow Lite

TensorFlow Lite Micro

Model

Software

Hardware

Model	TensorFlow	TensorFlow Lite	TensorFlow Lite Micro
Training	Yes	No	No

Model	TensorFlow	TensorFlow Lite	TensorFlow Lite Micro
Training	Yes	No	No
Inference	Yes *(but inefficient on edge)*	Yes *(and efficient)*	Yes *(and even* ***more*** *efficient)*

Model	TensorFlow	TensorFlow Lite	TensorFlow Lite Micro
Training	Yes	No	No
Inference	Yes *(but inefficient on edge)*	Yes *(and efficient)*	Yes *(and even **more** efficient)*
How Many Ops	~1400	~130	~50

Model	TensorFlow	TensorFlow Lite	TensorFlow Lite Micro
Training	Yes	No	No
Inference	Yes *(but inefficient on edge)*	Yes *(and efficient)*	Yes *(and even **more** efficient)*
How Many Ops	~1400	~130	~50
Native Quantization Tooling + Support	No	Yes	Yes

Model	TensorFlow	TensorFlow Lite	TensorFlow Lite Micro
Training	Yes	No	No
Inference	Yes *(but inefficient on edge)*	Yes *(and efficient)*	Yes *(and even* ***more*** *efficient)*
How Many Ops	~1400	~130	~50
Native Quantization Tooling + Support	No	Yes	Yes

Software	TensorFlow	TensorFlow Lite	TensorFlow Lite Micro
Needs an OS	Yes	Yes	No

Software	TensorFlow	TensorFlow Lite	TensorFlow Lite Micro
Needs an OS	Yes	Yes	No
Memory Mapping of Models	No	Yes	Yes

Software	TensorFlow	TensorFlow Lite	TensorFlow Lite Micro
Needs an OS	Yes	Yes	No
Memory Mapping of Models	No	Yes	Yes
Delegation to accelerators	Yes	Yes	No

Hardware	TensorFlow	TensorFlow Lite	TensorFlow Lite Micro
Base Binary Size	3MB+	100KB	~10 KB

Hardware	TensorFlow	TensorFlow Lite	TensorFlow Lite Micro
Base Binary Size	3MB+	100KB	~10 KB
Base Memory Footprint	~5MB	300KB	20KB

Hardware	TensorFlow	TensorFlow Lite	TensorFlow Lite Micro
Base Binary Size	3MB+	100KB	~10 KB
Base Memory Footprint	~5MB	300KB	20KB
Optimized Architectures	X86, TPUs, GPUs	Arm Cortex A, x86	Arm Cortex M, DSPs, MCUs

Conclusion

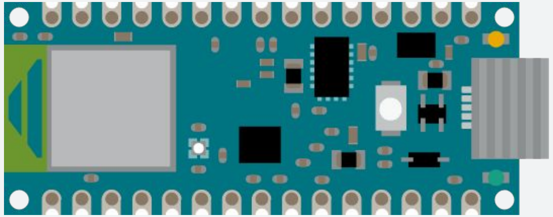
- **Many** different training and inference **frameworks**

Conclusion

- **Many** different training and inference **frameworks**
- **Major differences** between various deployment approaches even **within a single framework**

Conclusion

- **Many** different training and inference **frameworks**
- **Major differences** between various deployment approaches even **within a single framework**
- Open standards are **not** a panacea for platform **neutrality** and **portability**



downstream deployment problems