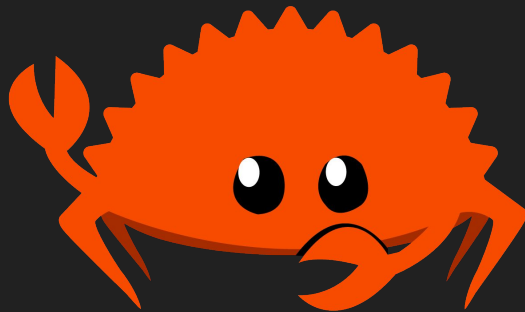


Offline ML with Rust



Daniel Bank

<https://github.com/danielbank/offline-ml>

Why Rust?

- Concurrency, Safety, Performance
- Use Cases in Web Assembly, Embedded, Machine Learning, and more
- Easy Package Manager and Dependency System
- Friendly Ecosystem

Connect with Rust Devs in Phoenix

{az}devs:

- Website: <https://rust.azdevs.org>
- Slack Channel: **#rust**

Meetup:

- Last Wednesday of the Month / HeatSync Labs in Mesa
- Biweekly Booze.rs Drink Up

Offline Voice Recognition

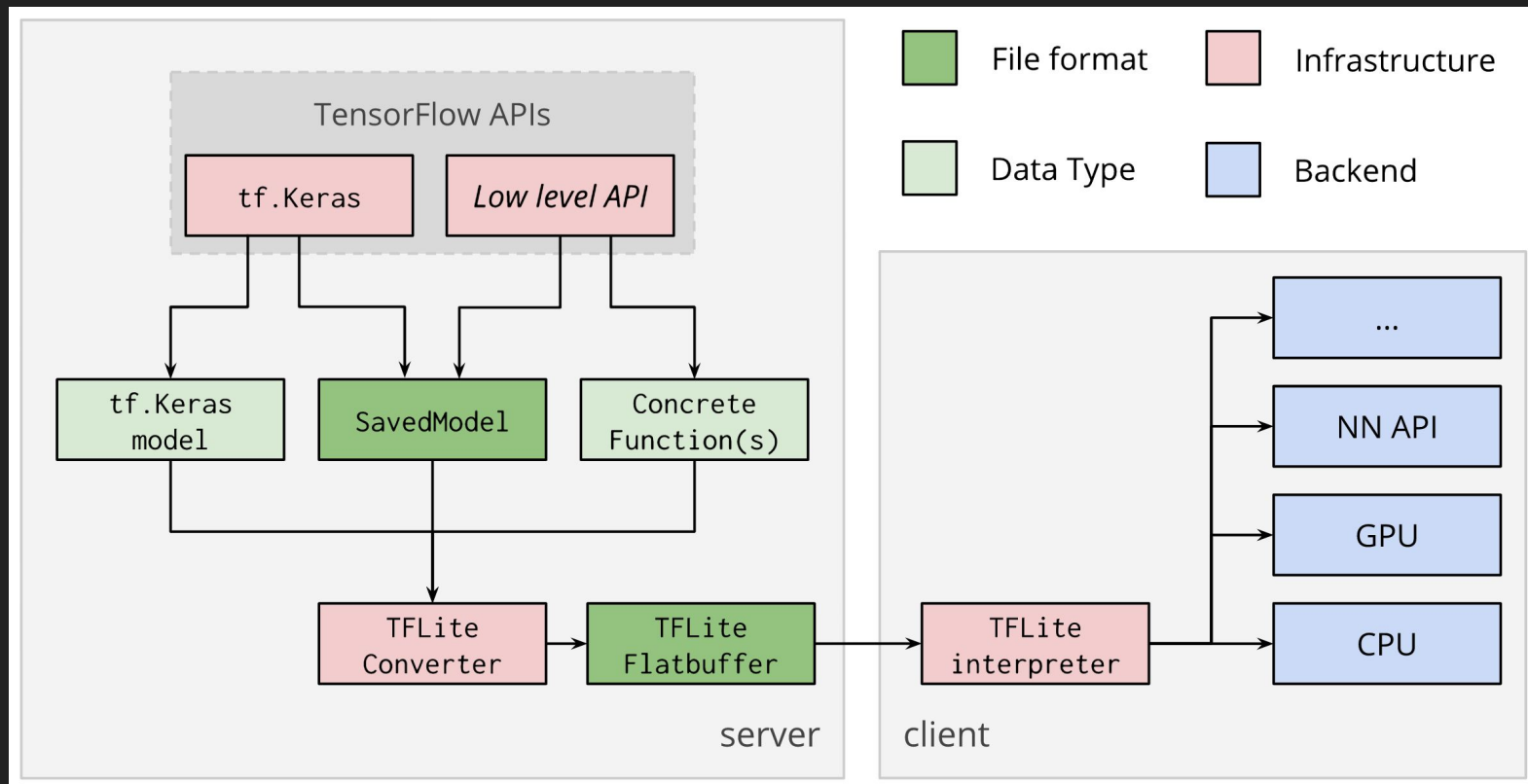
Microsoft ONNX Runtime / PyTorch



TensorFlow Lite

- TensorFlow
 - Open Source library that allows expressing arbitrary computations as a graph of data flow
- TensorFlow Lite
 - Set of tools that let developers run TensorFlow models on embedded systems
 - Only supports a subset of TensorFlow operations
 - Uses a precompiler to convert TensorFlow models to its own format
- Convert TensorFlow models to TensorFlow Lite models using CLI tools
 - TensorFlow models saved in .tflite file (FlatBuffer Schema)

TensorFlow Lite Conversion Workflow



TensorFlow Lite for Microcontrollers Demo

- TensorFlow Lite for Microcontrollers
- Pre-trained model recognizes 10 wanted words
- Model is ~20KB with runtime and operators to run speech detection
- Smallest device I have seen a working demo on (MCU)
- Limited performance (it's not all the crappy microphone's fault)
- Wanted Words Model running on a SparkFun Edge Development Board:

<http://bit.ly/sparkfunedgedemo>

Snips (www.snips.ai)

- Edge-based AI Voice Platform enabling Private by Design voice assistants
 - Targeting any type of hardware: MCU and CPU
 - Simple commands to full NLP
 - On Device / Offline Operability / End User Privacy
- Opensourced Rust Libraries: tract, snips-nlu, and snips-nlu-rs
- Calculator Assistant running on a Snips Seed Voice Interaction Development

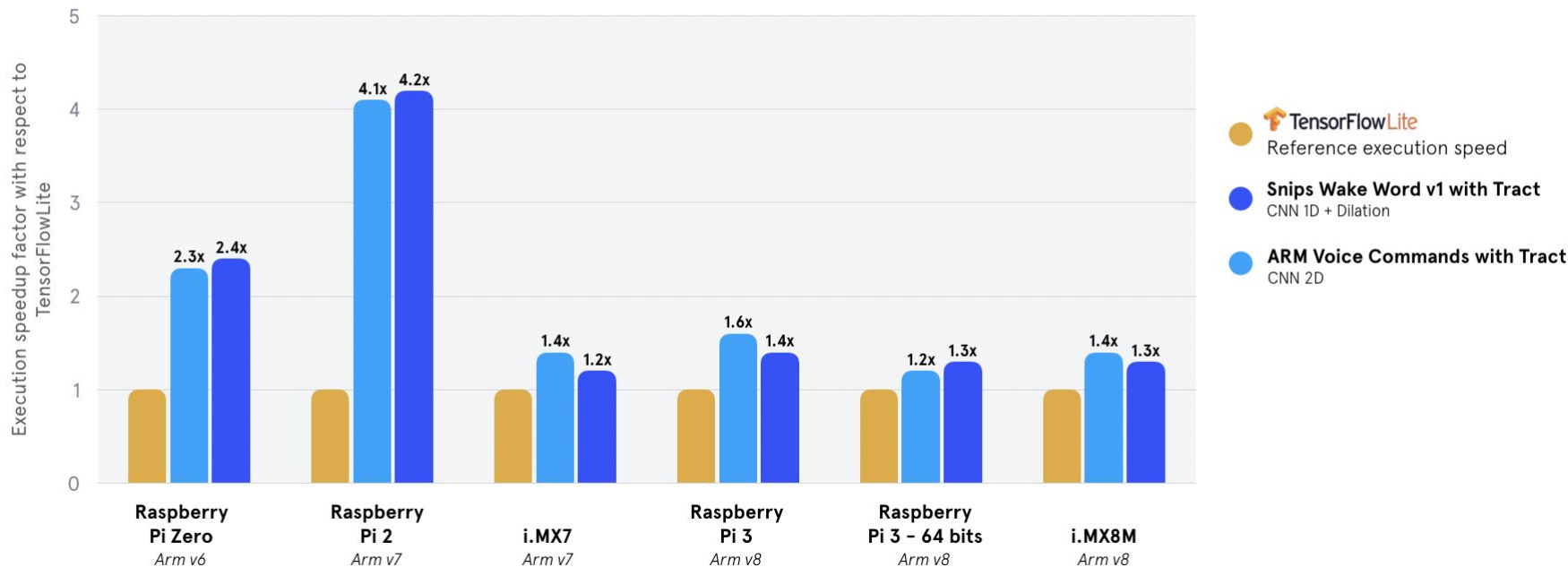
Kit: <http://bit.ly/snipsseed>

Tract

- TensorFlow / ONNX compatible inference library
- Loads frozen model from ProtoBuf format and flows data through it
- Real time streaming support
- Optimizations done before runtime (not network translation phase)

Tract Performance

Tract - up to 4.2x times faster than TensorFlow Lite



Offline ML with Rust

Offline ML with Rust Example

- REST API Server that can receive images via POST and respond with a prediction tuple (score, class) [[all classes](#)]
- Uses MobileNetV2, a lightweight image classification model (23MB)
- Cross Builds down to 15MB Rust binary
- Quasi-Offline: Model resides within the Rust App

Install Rust with Rustup

Go to <https://rustup.rs>

```
rustup update
```

```
rustup show
```

```
rustup toolchain install nightly
```

```
rustup default nightly
```

```
rustup default stable
```

```
rustup target list
```

Code Walkthrough

- Code: <https://github.com/danielbank/offline-ml>
- Cargo.toml
 - Image = Image Encoders and Decoders
 - Hyper = Low Level HTTP Library
 - Gotham = Flexible Web Framework
 - Tract = Neural Network Inference Engine
- Main.rs
 - Main Function
 - Router Function
 - Prediction Handler Function
 - Get Image Function

Cross Compilation

- Instructions: <https://rust.azdevs.org/2019-07-24>
- Rustup Targets
 - `rustup target add armv7-unknown-linux-gnueabihfpwd`
- Cross
 - `cargo install --force --git https://github.com/rust-embedded/cross cross`
 - Docker has to be running!
 - You have to use the cross command instead of cargo (or it will build for your local architecture)
 - `cross build --release --target=armv7-unknown-linux-gnueabihf`

Demo Time

- Curl Command

```
curl -i -X POST -F "image=@<IMAGE_PATH>" http://<IP>
```

- Image Classes

AZ Dev Community Links

- [Desert Rust Meetup](#)
- [AI/ML DevFest](#)
- [IoT DevFest](#) (Coming up in January 2020)
- [{az}devs](#)

Rust ML Links

- [Are We Learning Yet: State of ML in Rust](#)
- [Weld: Parallel Code Generation for Data Analytics Frameworks](#)
- [Rust Crates for Numerical Simulation](#)
- [Tract Repo](#)
- [Snips NLU Rust Repo](#)
- [Rusty Machine: General Purpose ML Library](#)

Snips Links

- [Snips Open Sources Tract Medium Article](#)
- [Deep Dive on Snips at OxidizeConf with Hubert de la Jonquière](#)
- [Snips Uses Rust to Build an Embedded Voice Assistant](#)
- [Tract Repo](#)
- [Snips NLU Rust Repo](#)
- [Snips NLU Repo](#)

TensorFlow Lite Links

- [TensorFlow Lite on SparkFun Edge Codelab](#)
- [TensorFlow Lite Micro Speech Example](#)
- [TensorFlow Lite Guide](#)
- [TensorFlow for Microcontrollers](#)