“Running and Managing Fleets of Single Board Computers at Scale”

Seth Clark – Co-founder & Head of Product
Modzy

May 2, 2023
Thank you, tinyML Strategic Partners, for committing to take tinyML to the next Level, together.
Executive Strategic Partners
The Leading Development Platform for Edge ML

dgeimpulse.com
Advancing AI research to make efficient AI ubiquitous

Power efficiency
- Model design, compression, quantization, algorithms, efficient hardware, software tool

Personalization
- Continuous learning, contextual, always-on, privacy-preserved, distributed learning

Efficient learning
- Robust learning through minimal data, unsupervised learning, on-device learning

Perception
- Object detection, speech recognition, contextual fusion

Reasoning
- Scene understanding, language understanding, behavior prediction

Action
- Reinforcement learning for decision making

A platform to scale AI across the industry
Accelerate Your Edge Compute

SYNTIANT

Making Edge AI A Reality

www.syntiant.com
Platinum Strategic Partners
Renesas is enabling the next generation of AI-powered solutions that will revolutionize every industry sector.
DEPLOY VISION AI AT THE EDGE AT SCALE
Gold Strategic Partners
Witness potential made possible at analog.com.

Where what if becomes what is.
Easily deploy your tinyML solutions with Arduino Pro

arduino.cc/pro
Arm AI Virtual Tech Talks

The latest in AI trends, technologies & best practices from Arm and our Ecosystem Partners.

Demos, code examples, workshops, panel sessions and much more!

Fortnightly Tuesday @ 4pm GMT/8am PT

Find out more: www.arm.com/techtalks
NEUROMORPHIC INTELLIGENCE FOR THE SENSOR-EDGE
The Right Edge AI Tools Can Make or Break Your Next Smart IoT Product

Analytics Toolkit Suite

AutoML

Version Control and Model Management

Team Collaboration

Data Collection

Code Generation

Test & Validation

Code Editing

sensiml.com/tinyML
STMicroelectronics provides extensive solutions to make tiny Machine Learning easy
ENGINEERING
EXCEPTIONAL
EXPERIENCES

We engineer exceptional experiences for consumers in the home, at work, in the car, or on the go.

www.synaptics.com
Silver Strategic Partners
Join Growing tinyML Communities:

**Meetup**
14.4k members in 47 Groups in 39 Countries

tinyML - Enabling ultra-low Power ML at the Edge

**LinkedIn**
4k members & 11.6k followers

The tinyML Community
https://www.linkedin.com/groups/13694488/
Subscribe to tinyML YouTube Channel for updates and notifications (including this video)

www.youtube.com/tinyML
EMEA 2023
https://www.tinyml.org/event/emea-2023

More sponsorships are available: sponsorships@tinyML.org
Reminders

Slides & Videos will be posted tomorrow

tinyml.org/forums  youtube.com/tinyml

Please use the Q&A window for your questions
Seth Clark

Seth Clark is Head of Product & Co-founder at Modzy. Prior to founding Modzy, Seth served as product manager for a number of successful analytics products. He also served as Principal at Booz Allen Hamilton, where he led complex projects at the intersection of Data Science and Software Development. He has degrees in engineering from the University of Southampton and the Massachusetts Institute of Technology.
Running and Managing Fleets of Single Board Computers (SBCs) at Scale

tinyML Talks
2 May 2023
How I got here
2020s: The Wild West Golden Age of SBCs
SBCs are doing amazing things

ZeroPhone
Pi Zero-based open-source mobile phone (that you can assemble for 50$ in parts)

Jetson Clean Water AI
Using AI object detection to detect water contamination

ClippyGPT
Raspberry Pi 3B powered Clippy bot with ChatGPT
Why we love SBCs

- Many, many ports!
- Powerful hardware options
- Operating Systems with lots of software support
- Small & portable

GPIO
CPU
RAM
Display
Cam
HDMIs
Ethernet
USB for days
SBCs are a convenient alternative to microcontrollers

<table>
<thead>
<tr>
<th>SBCs</th>
<th>Microcontrollers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Operating System with an optional user interface</td>
<td>No Operating System</td>
</tr>
<tr>
<td>Built in networking</td>
<td>No networking</td>
</tr>
<tr>
<td>Lots of software options</td>
<td>Limited to software supported by microcontroller’s IDE</td>
</tr>
</tbody>
</table>
Businesses are increasingly adopting SBCs for serious applications

- 3D cameras
- 5G modules
- Robotics
- Smart cities
- Smart manufacturing
- IoT
- Fleet management
- ...

$2.94 billion market in 2022
SBCs are also great for machine learning

» Write, compile, and run custom software in any language you like
» Download ML frameworks, libraries, and even fully trained models
» Plug in peripherals like cameras and sensors for computer vision and anomaly detection
» Use built in hardware accelerators (e.g. GPUs) or plug one in via USB
» Add an HDMI cable and a monitor and you’ve got a fully-functional app
SBCs are great, and so is machine learning, so…
Setting up an SBC takes time and effort

- **Device Setup**: OS installation, system-level dependency configuration, GPU access
- **Resource Constraints**: RAM allocation, number of CPU cores, shared across all app processes
- **External Software**: Custom data connections, monitoring or logging tools
- **Security**: Device and application security, application accessibility, networking
- **Model Requirements**: Programming language(s), model framework dependencies
Running ML apps brings distinctive challenges

- Minimizing resource needs of models
- Running on multiple chipsets (arm, x86/AMD, 32-bit & 64-bit)
- Managing code for pre-processing data & post-processing predictions
- Working with hardware accelerators (GPUs, TPUs, and FPGAs)
- Connecting to sensors and local data sources
- Gracefully handling intermittent network availability
- Monitoring model performance
- Improving and updating models overtime
Running models by hand on 1000s of SBCs would be madness, so what if you could...

» Send model code and dependencies to an SBC dependably, repeatably, and scalably
» Connect to local data sources and other applications
» Run quickly and use as few resources as possible
» Monitor models while they’re running
» Will keep working even when the wifi is down
Enter the container
The basic idea

- ML model + code
- Dependencies
- A simple API

ML Model A → Docker Image → ML Model B → ML Model C

Docker
Linux
Single Board Computer
### A hypothetical ML container image

<table>
<thead>
<tr>
<th><strong>API</strong></th>
<th>gRPC APIs: Status, Run, Shutdown</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Inference script</strong></td>
<td>inference.py</td>
</tr>
<tr>
<td><strong>Trained model</strong></td>
<td>model-weights.pkl</td>
</tr>
<tr>
<td><strong>Dependencies</strong></td>
<td>pickle, cv2, torch, numpy</td>
</tr>
<tr>
<td><strong>Base Image</strong></td>
<td>armv6 Debian OS</td>
</tr>
</tbody>
</table>
Chassis.ml: An easy way to build model containers

Website: https://chassis.ml
Github: modzy/chassis
A few more ingredients for managing containers on lots of SBCs

» Chassis.ml containers
» Dockerhub
» Docker
» Raspberry Pi, Jetson Nano, Intel Upboard
» Modzy Core
» Nats.io + Jetstream
How it works
Benefits of container-based ML on SBCs

» Supports both ARM and x86/AMD chipsets, 32 or 64-bit
» Built-in support for GPUs
» Models never break due to missing dependencies
» Network connectivity is only needed initially to deploy models
» Models can be monitored at any time
» New models and versions can be deployed centrally to many, many devices at once
Examples
Example: Edge-centric NLP App

Website with Chat Box

Emotion Intent Model

MacBook Pro
Philadelphia, PA

Raspberry Pi 3B+
Fairfax, VA
Example: Industrial Defect Detection
Example: Computer Vision Parking Lots
Final thoughts

» Use an edge-centric architecture anytime you need low-latency, non-stop performance on or offline, or models that need to run in more than one place

» Start small, but plan for massive scale
Thanks for joining our webinar!
Scan the QR code to join our Discord server for more resources.
Copyright Notice

This multimedia file is copyright © 2023 by tinyML Foundation. All rights reserved. It may not be duplicated or distributed in any form without prior written approval.

tinyML® is a registered trademark of the tinyML Foundation.

www.tinyml.org
Copyright Notice

This presentation in this publication was presented as a tinyML® Talks webcast. The content reflects the opinion of the author(s) and their respective companies. The inclusion of presentations in this publication does not constitute an endorsement by tinyML Foundation or the sponsors.

There is no copyright protection claimed by this publication. However, each presentation is the work of the authors and their respective companies and may contain copyrighted material. As such, it is strongly encouraged that any use reflect proper acknowledgement to the appropriate source. Any questions regarding the use of any materials presented should be directed to the author(s) or their companies.

tinyML is a registered trademark of the tinyML Foundation.

www.tinyml.org