AnalogML™: Analog Inferencing for System-Level Power Efficiency

David W. Graham, Ph.D.
Founder & CSO, Aspinity
Professor, Lane Dept. of CSEE, West Virginia University

March 29, 2022
Agenda

01. Analog for battery-powered, always-on inferencing
02. AnalogML overview
03. Audio applications
04. Conclusions
Today’s Sensor Processing at the Edge is Inefficient

41.6B\textsuperscript{1} connected IoT devices by 2025

79.4ZB\textsuperscript{1} of new data from edge sensors

80+\% of the digitized data will be irrelevant

Need to determine data relevance as soon as possible

\textsuperscript{1}International Data Corporation (IDC) \textit{Worldwide Global DataSphere IoT Device and Data Forecast, 2019–2023}, doc #US45066919, June 2019

Copyright 2022 Aspinity, Inc. All rights reserved.
Shifting the ML Workload to Analog

*Inferencing in analog domain at near zero power*

### Traditional Always-On Architecture

- All sensor data processing/analysis handled in digital processor.
- Analog and digital system on 100% of the time.
- Always-on system power draw: **3000-5000μA**

### Aspinity AnalogML™ Architecture

- AnalogML™ chip performs machine learning and other computations in analog, prior to digitization.
- Analog system on 100% of the time.
- Always-on system current draw reduced by > 95%.
- Always-on system power draw: **<100 μA***

*Audio applications*
Efficiency with Analog

- Many operations are more efficient in analog
- Why not do more with analog?
- Historical challenges of analog
  - Versatility
  - Repeatability
  - Ease-of-Use

* Classic studies by Vittoz [1990], Sarapeshkar [1998], etc.
What is AnalogML™?

**Key Features**
- **Sensor interface**: Can be synthesized for multiple sensor types (mic, accelerometer, etc.)
- **Analog feature extraction**: Picks out salient features from raw, analog sensor data, reducing the amount of data going into the neural network
- **Analog neural network**: Efficient, small-footprint analog inferencing block
- **Data compression**: Continuous collection and compression of analog sensor data for low-power data buffering

**Benefits**
- **Software programmable**: Use machine learning models developed using standard training methodologies
- **Flexible**: Leverage reconfigurable concepts to implement a wide variety of applications
- **Smart Wake-Up**: Let ADC and digital sleep until needed
- **Efficient Processing**: Do more at less power with analog processing capabilities
AnalogML™: Configurable Computing Chip

Software configures functionality, parameters, & connections of CABs

Analog NVM stores circuit parameters & NN weights
- Proprietary floating-gate non-volatile memory
- Allows wide assortment of circuit functions & parameters with a minimal set of circuits
- Provides offset removal & trimming
Analog Neural Network

Individual Neuron

- Analog multiplier (custom 4 quadrant multiplier)
- Weights stored locally as analog NVM

Neural Network Capabilities
- 3 configurable NN blocks
- Fully connected & recurrent networks
- Enabled by efficient analog feature extraction

Analog NN Training
- Select size & activations
- Train using standard tools PyTorch
- SDK maps NN settings to weights & connections
Example of a Simple AnalogML™ Audio Chain

SDK provides
• Library of components at different levels of hierarchy
• Language for connecting the components
• Configuration file as a bitstream
AML100: The first AnalogML™ Core
Application: Glass Break Detection

AML100

Spectral Energy
SNR Estimates
Zero Crossing Rate

Feature Extraction
Find “Thuds” & “Shatters”
Ensure Correct Event Sequence

AML100

Sequence-Based Decision
Detection (Interrupt)

Always-On Current Draw

<table>
<thead>
<tr>
<th>Component</th>
<th>Current Draw</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mic</td>
<td>50µA</td>
</tr>
<tr>
<td>AnalogML™</td>
<td>10µA</td>
</tr>
<tr>
<td>ADC</td>
<td>~0µA</td>
</tr>
<tr>
<td>MCU</td>
<td>~0µA</td>
</tr>
<tr>
<td>Total System</td>
<td>60µA</td>
</tr>
</tbody>
</table>
Application: Voice Activity Detection

Low false-alarm rate is critical for low system power
Application: VAD + Preroll for WWEs

- AnalogML™ = 15µA (VAD + Preroll Compression)
- System <100µA in always-on mode
- Preroll capture maintains wake-word detection accuracy

Copyright 2022 Aspinity, Inc. All rights reserved.
Conclusion

• Rethink the standard digital paradigm for ML
• AnalogML™ moves the ML workload to analog → Inferencing before digitization
• Enables the versatility, repeatability, & ease-of-use of digital in the lower-power analog domain
• Opens door to new battery-powered products
• AML100 → The first AnalogML™ core
• Evaluation kits support development of products with AML100
Thank You

David Graham, Founder and CSO, Aspinity
david@aspinity.com

Company Information
Website: https://www.aspinity.com/
Email: info@aspinity.com
Address: 2000 Smallman Street
Suite 201
Pittsburgh, PA 15222
tinyML Summit 2022 Sponsors
Copyright Notice

This presentation in this publication was presented as a tinyML® Summit 2022. 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