

CMP-101

Enabling AI/ML at the Edge – With or Without Connectivity



Sai Bharadwaj

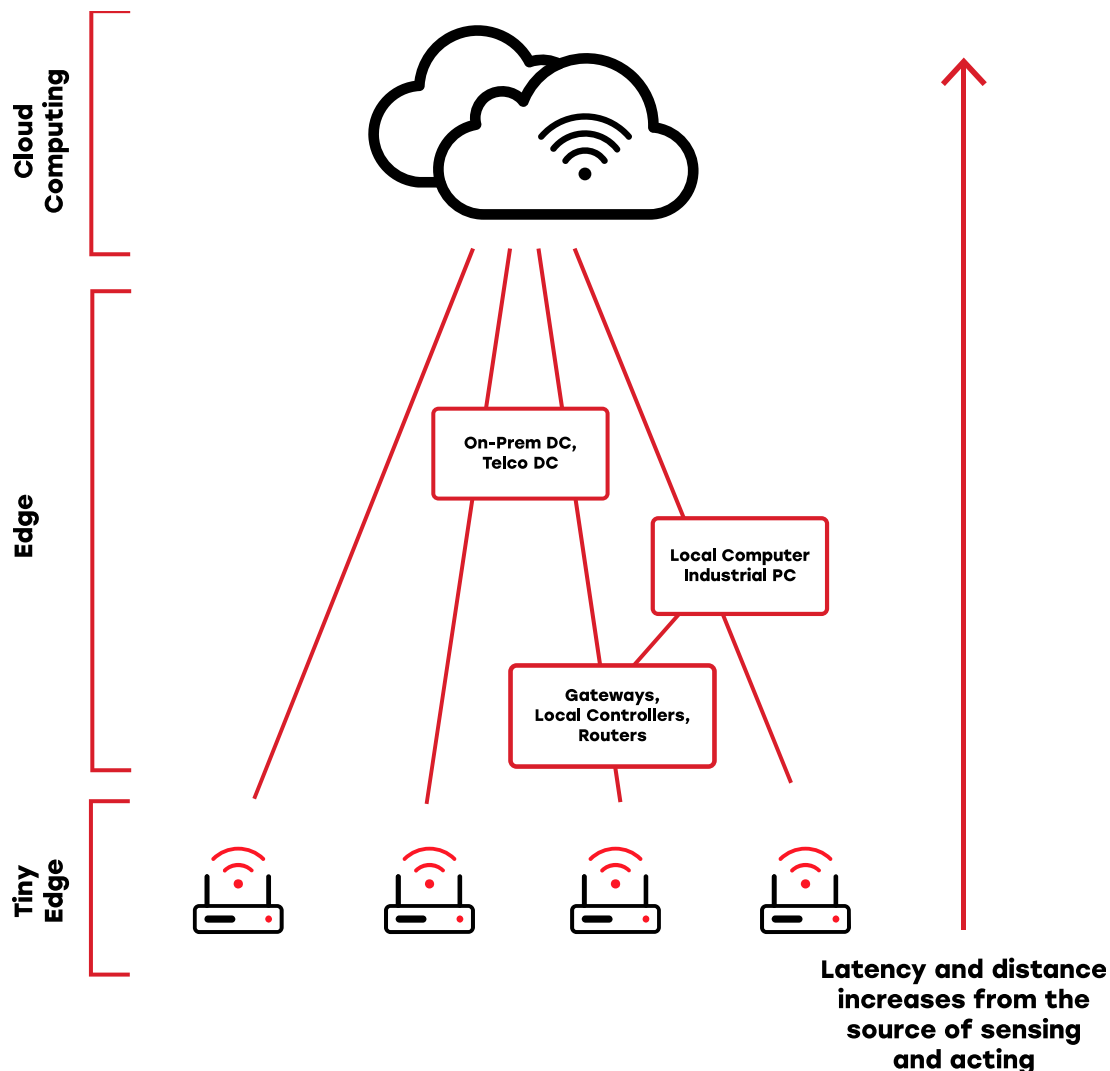
Product Marketing Manager,
Silicon Labs



Jon Gettinger

Head of GTM at ModelCat.AI

Artificial Intelligence(AI) and Machine Learning(ML) at the Tiny Edge



- **Lower latency**
 - Moving decision making closer to where data is collected allows for better real time decision making
- **Privacy, IP Protection, and Security**
 - Can now send anonymous decisions to larger monitoring system rather than sensitive data
- **Removes bandwidth constraints**
 - Decision centric data transmission limits overall amount of bandwidth needed
- **Enables offline mode operation**
 - Eliminates the need for connectivity to make use of critical AI/ML capability
- **Reduces overall system and operational costs**
 - Simplified BOM and data usage lowers overall implementation cost for ML enabled edge devices

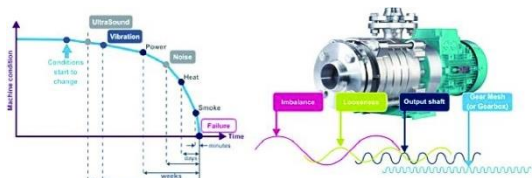
Benefits of non-Wireless TinyML



- **Localized decision making allows for broadened use cases and higher reliability**
 - Removes dependency on cloud or other infrastructure capability
- **Existing wired interfaces can provide backbone to send decisions to central location if needed**
 - Eliminates need for adoption of wireless technologies alongside AI/ML
- **Alerts or notifications can be made via localized interfaces**
 - Can use sounds or LEDs to alert operators, technicians, or consumers when attention is needed
- **AI/ML can be added to existing setups quickly and with minimal disruptions**
 - Add intelligent sensors to equipment or networks without impacting overall system

Machine Learning Applications Supported by Silicon Labs

SENSOR

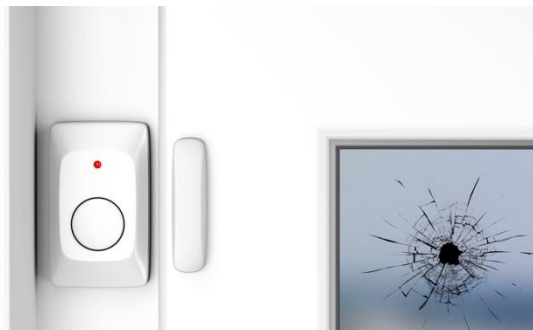


Signal processing (time series low-rate)

- Predictive/Preventative Maintenance
- Bio-signal analysis (healthcare and medical) e.g., pulse detection, EKG
- Cold chain monitoring
- Accelerometer use-cases e.g., fall detection, pedometer, step counting
- Battery monitoring
- Agricultural use-cases e.g., moisture sensing
- Anomaly detection

RAM*: 96kB
Ops/s: 5M

AUDIO



Audio pattern matching

- Security applications e.g., Glass break, scream, shot detection
- Cough detection
- Machine malfunction detection
- Breath monitoring

RAM*: 128kB
Ops/s: 6M

VOICE

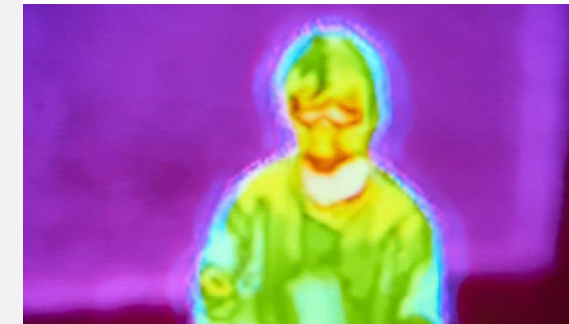


Voice commands

- 10 words command set for smart appliance
- Wake-word detection (Always-On voice)
- Smart device voice control
- Voice assistant

RAM*: 128kB
Ops/s: 40M

VISION



Low-resolution vision

- Wake-up on object detection
- Presence detection
- People counting, people-flow counting
- Movement detection
- Fingerprint

RAM*: 256kB w/hardware accelerator,
Ops/s: 100M

*Suggested minimum chip RAM size

ML Applications at the Tiny Edge with Silicon Labs

AI/ML Enabled MCU Portfolio



LOW POWER, HIGH PERFORMANCE

- Cortex M33 + AI/ML Accelerator
- Up to 256kB RAM and 1024kB Flash
- High performance analog peripherals
- LCD Controller for up to 192 Segments



MORE MEMORY, MORE GPIO

- Cortex M33 + AI/ML Accelerator
- Up to 512kB RAM and 3200kB Flash
- Up to 64 GPIOs
- High performance analog peripherals

Benefits of the MVP ML Hardware Accelerator

Dedicated **ML computing subsystem** next to the CPU: Matrix Vector Processor (MVP)

Optimized MVP to accelerate ML inferencing with a lot of processing power **offloading the CPU**

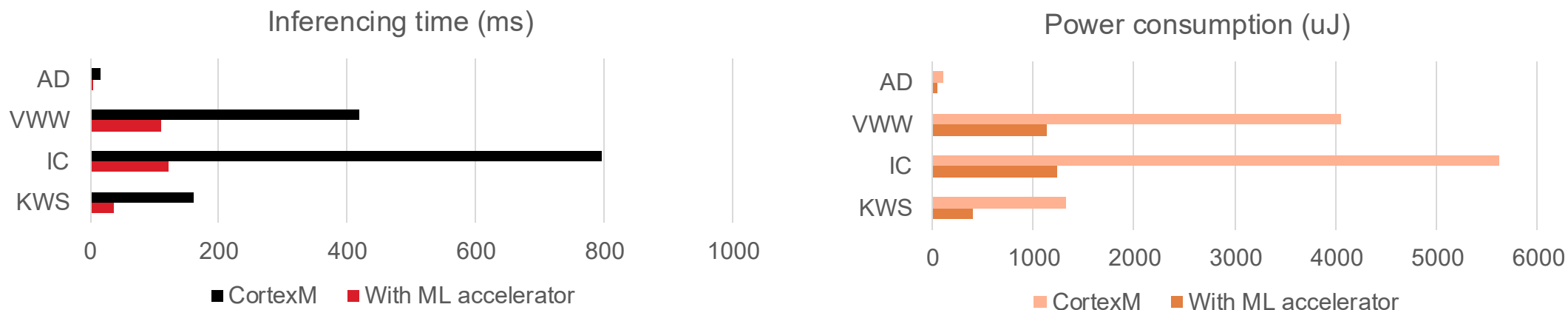
Up to 8x faster inferencing over Cortex-M (see below perf. benchmark)

Up to **6x lower power** for inferencing (see below perf. benchmark)

Dedicated OPNs for MVP accelerated parts → EFR32MG24B[2]... or [3]



Performance Data with ML Hardware Accelerator vs. Pure SW on CortexM*



*Standardized performance benchmark validated by independent benchmarking body **MLCommons.org**. Published in MLPerf Tiny v1.0. Results are for inferencing only (not for the complete application). You can refer to MLCommons as validated results-



New feature in GSDK: MVP Math library

- Accelerate and do more efficient linear algebra operations with internal MVP subsystem
- Math APIs (alternative to CMSIS_DSP) available in GSDK

VECTOR OPERATIONS

- Vector Add
- Vector Absolute Value
- Vector Clip
- Vector Dot Product
- Vector Multiply
- Vector Negate
- Vector Offset
- Vector Scale
- Vector Sub
- Complex Vector Conjugate
- Complex Vector Dot Product
- Complex Vector Magnitude
- Complex Vector Magnitude Squared
- Complex Vector Multiply
- Complex Vector Multiply Real
- Vector Copy
- Vector Fill

MATRIX OPERATIONS

- Matrix Initialize
- Matrix Multiply
- Matrix Scale
- Matrix Sub
- Matrix Transpose
- Matrix Multiply Vector
- Matrix Add
- Complex Matrix Multiply
- Complex Matrix Transpose

- ✓ **Faster and more efficient** execution of many algorithms with large data for example filtering algorithms
- ✓ **Saving CPU cycles, saving power, resulting longer battery life**
- ✓ **Option to win sockets against faster CPUs**

CortexM only

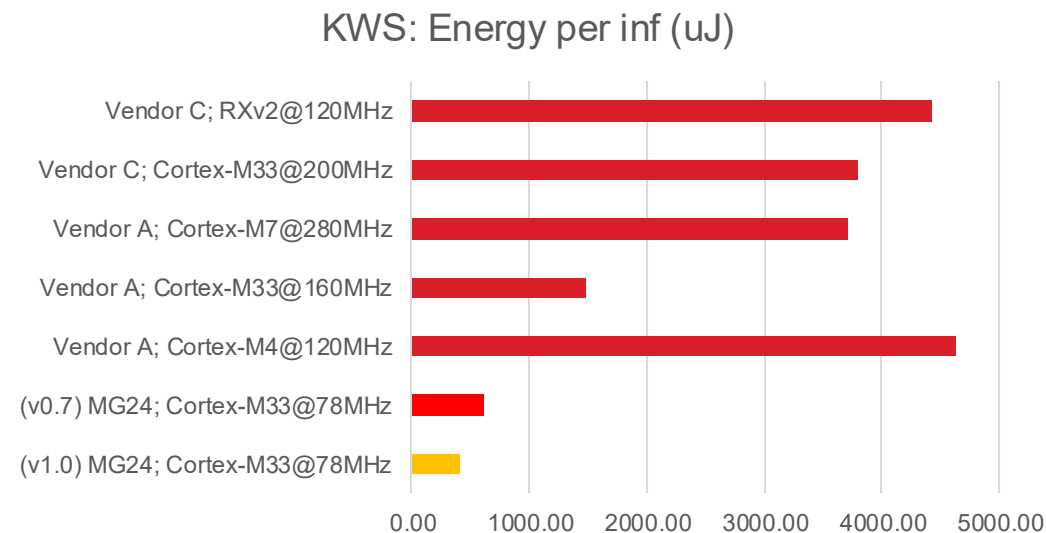
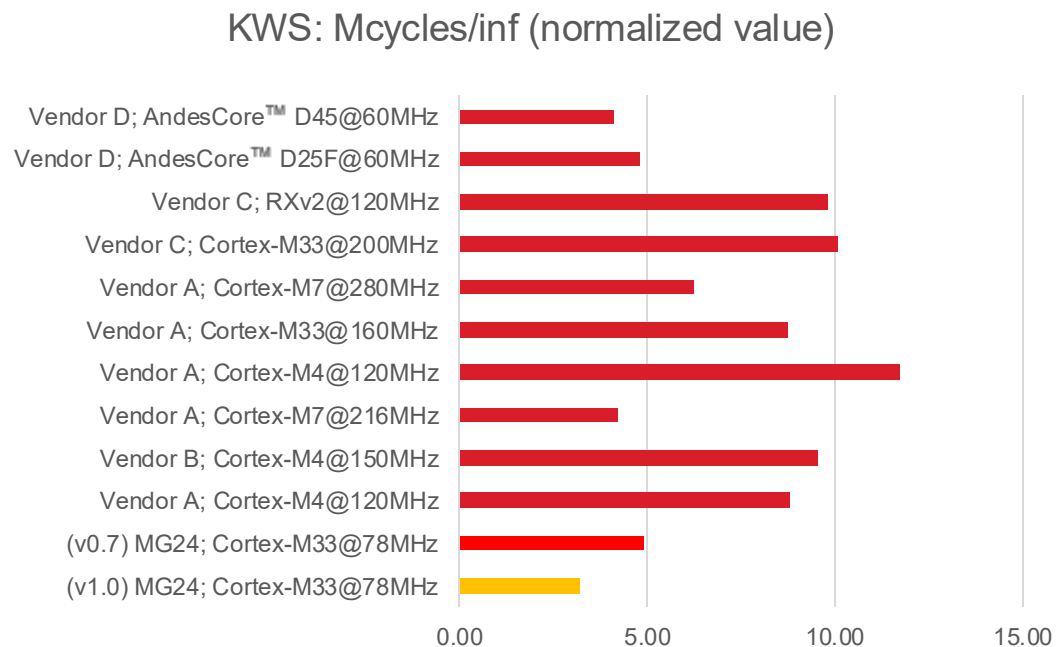


Matrix dims.		CMSIS f32 cpu-cycles	CMSIS f16 cpu-cycles	MVP cpu-cycles	instr	stalls
2x2	2x2	226	304	403	8	0
4x2	2x4	602	913	424	32	0
6x2	2x6	1210	1921	464	72	0
8x2	2x8	2050	3321	516	128	0
10x2	2x10	3122	5113	592	200	0
12x2	2x12	4426	7297	676	288	0
14x2	2x14	5962	9873	784	392	0
16x2	2x16	7730	12841	904	512	0
18x2	2x18	9730	16201	1036	648	0
20x2	2x20	11962	19953	1192	800	0
20x4	4x20	17962	27956	1593	1200	1
20x6	6x20	23742	39956	2193	1600	201
20x8	8x20	27562	47556	2793	2000	400
20x10	10x20	33162	59556	3393	2400	601
20x12	12x20	37162	67156	3993	2800	801
20x14	14x20	42762	79156	4593	3200	1000
20x16	16x20	46762	86756	5193	3600	1201
20x18	18x20	52362	98756	5793	4000	1401
20x20	20x20	56362	106356	6393	4400	1600



~ 9x less cycles

ML_Perf-Tiny v0.7 (and v1.0) Performance Benchmark*

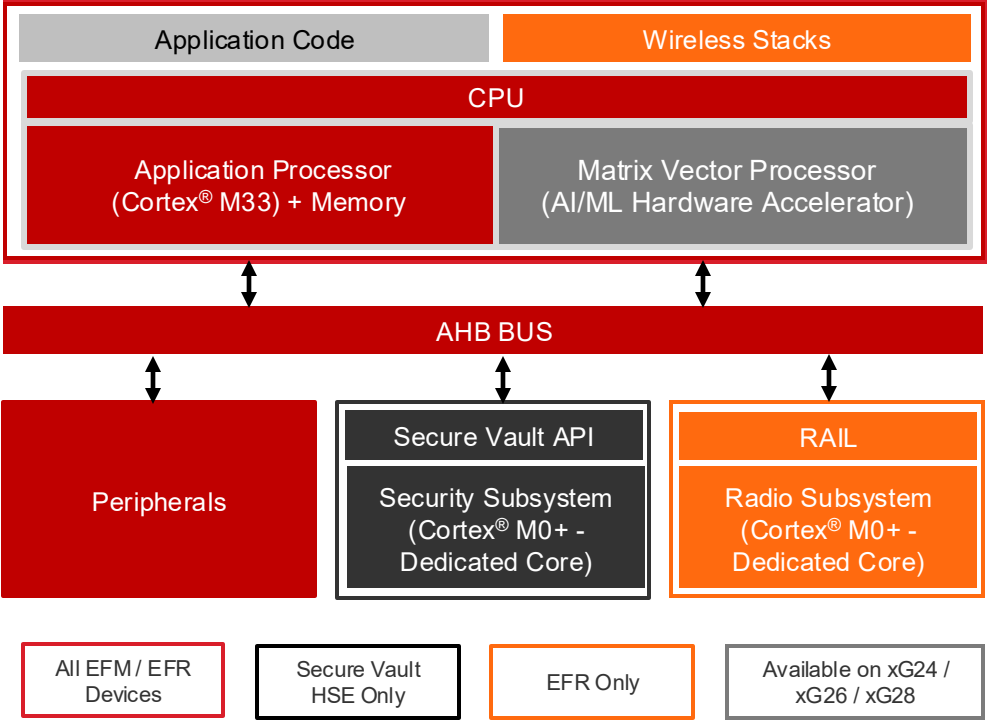


MLPerf Tiny 0.7 benchmark results on xG24-DK2601B board; source: mlcommons.org





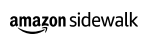



*Standardized performance benchmark validated by independent benchmarking body.
Results are for inferencing only (not the complete application).

Multi-Core and AI/ML Solution



- Multi-core architecture gives design flexibility and optimization across EFM and EFR platforms
 - Dedicated application, radio¹, and security² cores share system burden for better resource utilization
- Common development platform for connected and non-connected products
 - Simplicity Studio gives developers a common development platform for entire product portfolio
- Common Security and AI/ML subsystems
 - Allows for design consistency independent of connectivity needs
- Footprint and firmware compatibility between EFM and EFR families
 - Simplified SKU management and code base development lowers development cost and complexity

	BG 	MG 	FG 	ZG 	SG 	PG 
xG21	✓	✓				
xG22	✓	✓				✓
xG23			✓	✓	✓	✓
xG24	✓	✓				
xG25			✓			
xG26	✓	✓				✓
xG27	✓	✓				
xG28			✓	✓	✓	✓
EFR Device Families						EFM

AI/ML Enabled Wireless SoCs



OPTIMIZED 2.4GHZ AI/ML SOC

- Cortex M33 + AI/ML Accelerator
- Bluetooth, Matter, Proprietary Support
- Up to 256kB RAM and 1536kB Flash
- High performance analog peripherals



SUB-GHZ + BLUETOOTH AND AI/ML

- Cortex M33 + AI/ML Accelerator
- Sub-GHz and Sub-GHz + Bluetooth
- Up to 256kB RAM and 1024kB Flash
- High performance analog peripherals
- LCD Controller for up to 192 Segments



FUTURE-PROOF AI/ML SOLUTION

- Cortex M33 + AI/ML Accelerator
- Bluetooth, Matter, Proprietary Support
- Up to 512kB RAM and 3200kB Flash
- High performance analog peripherals
- LCD Controller for up to 160 Segments

Silicon Labs Machine Learning Solution Benefits

- Industry's widest portfolio of low power solutions combined with ML for Tiny Edge devices
 - Platformed approach to AI/ML for simplified use across connected and non-connected products
 - Options for Bluetooth, 802.15.4/ZigBee/Thread, Matter, Z-Wave, Proprietary, Wi-Sun, Sidewalk
- Integrated ML hardware accelerator provides 8X faster ML inferencing with 1/6th of energy
 - Reduces BOM, footprint and design complexity while minimizing latency
 - Allows for smaller batteries and extended maintenance cycles
- ML development tools and solutions for explorers to experts for faster application development
 - TensorFlow Lite Micro supported in GSDK
 - Partnerships with Eta Compute, Edge Impulse, SensiML, MicroAI to accelerate embedded ML development
 - Silicon Labs' ML Tool Kit on GitHub provides complete control & flexibility for the expert developers
- Wide range of use cases including low data rate sensors, audio/voice and low-res images

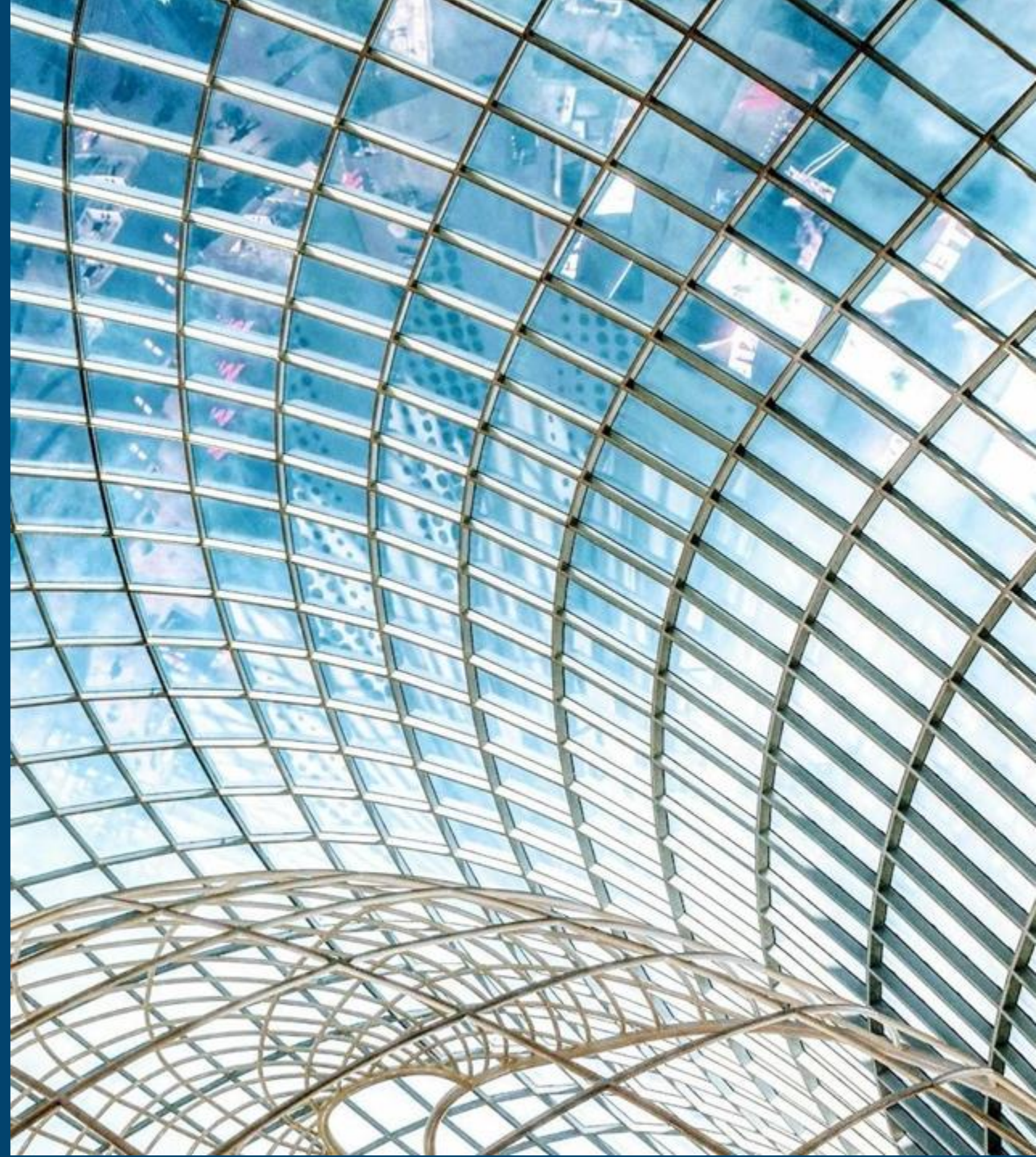
End-to-End Machine Learning Solution for Wireless IoT Edge Devices



Aptos:

Automated ML Model Builder
for Silicon Labs Devices

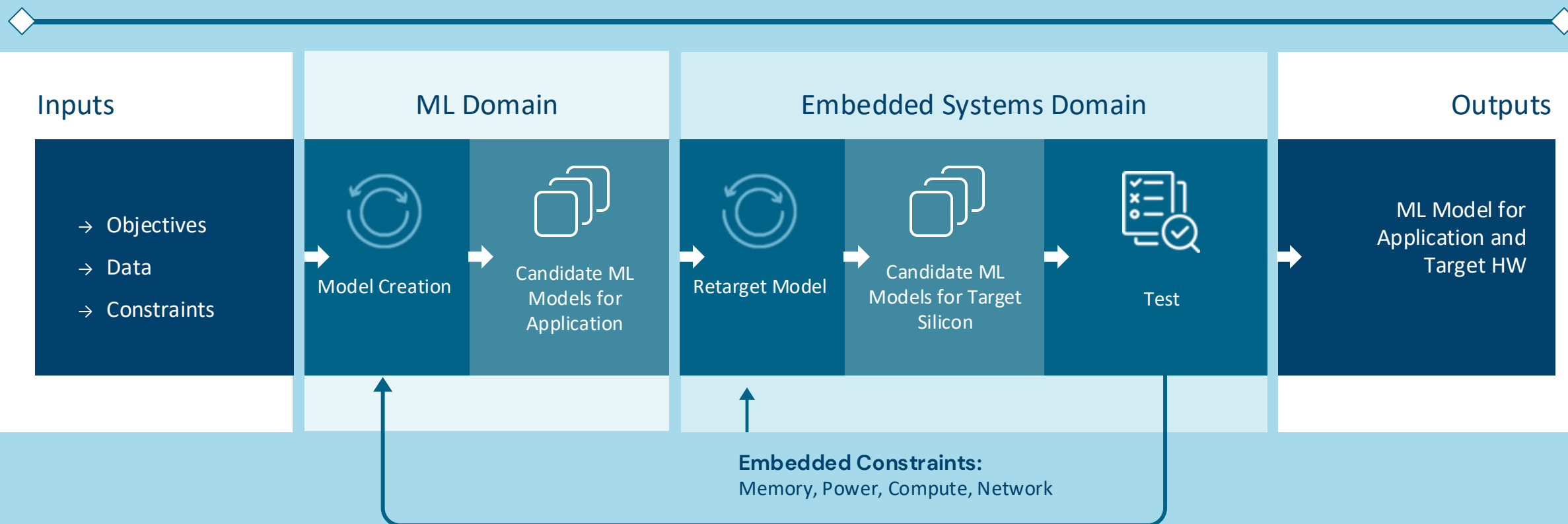
July 10, 2025



Today's Edge ML Development Flow

DEVICE AND EDGE

12 – 18 months



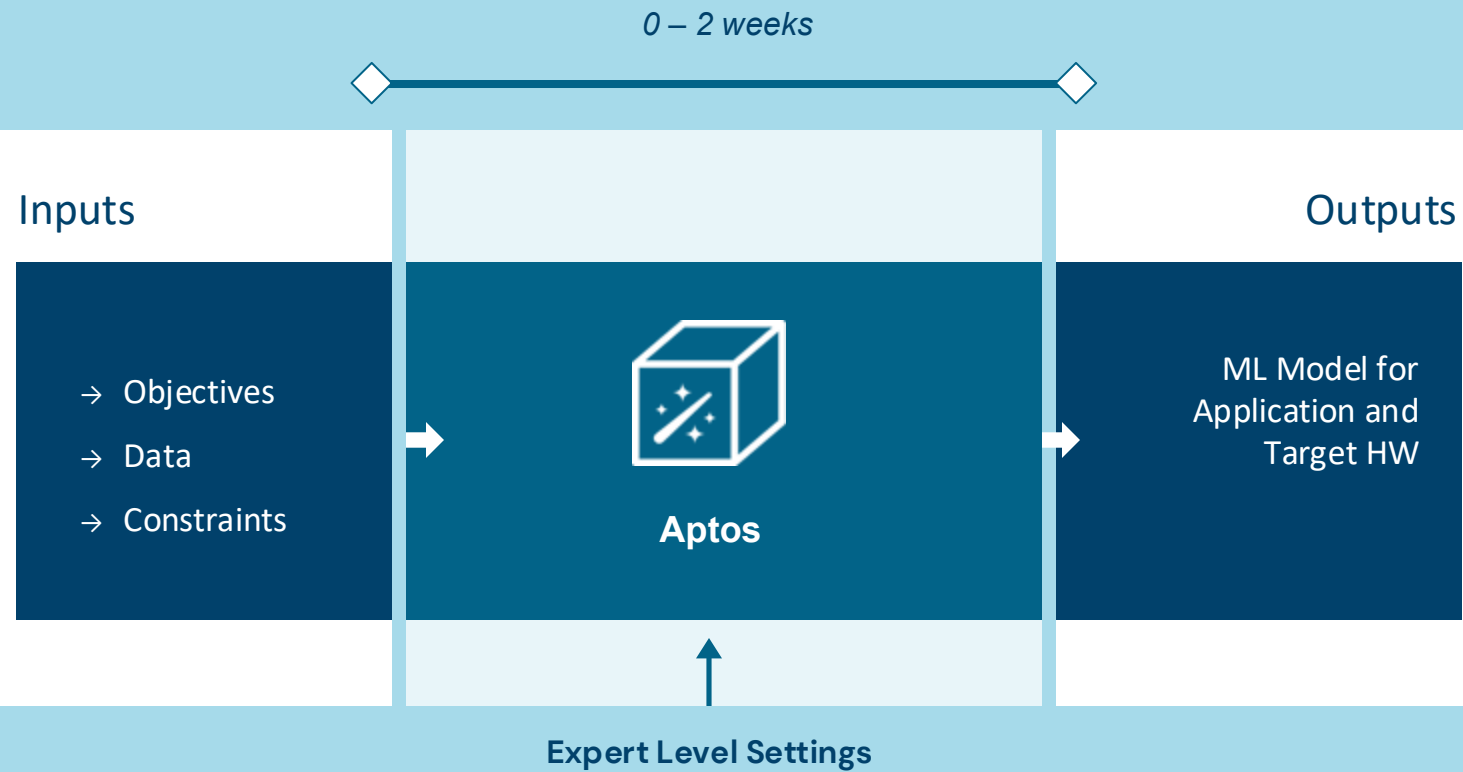
“The majority of edge
ML projects
still *fail to move past the
experimental phase.*”

LIAN JYE SU, RESEARCH DIRECTOR ABI RESEARCH

What's Wrong with this Picture?

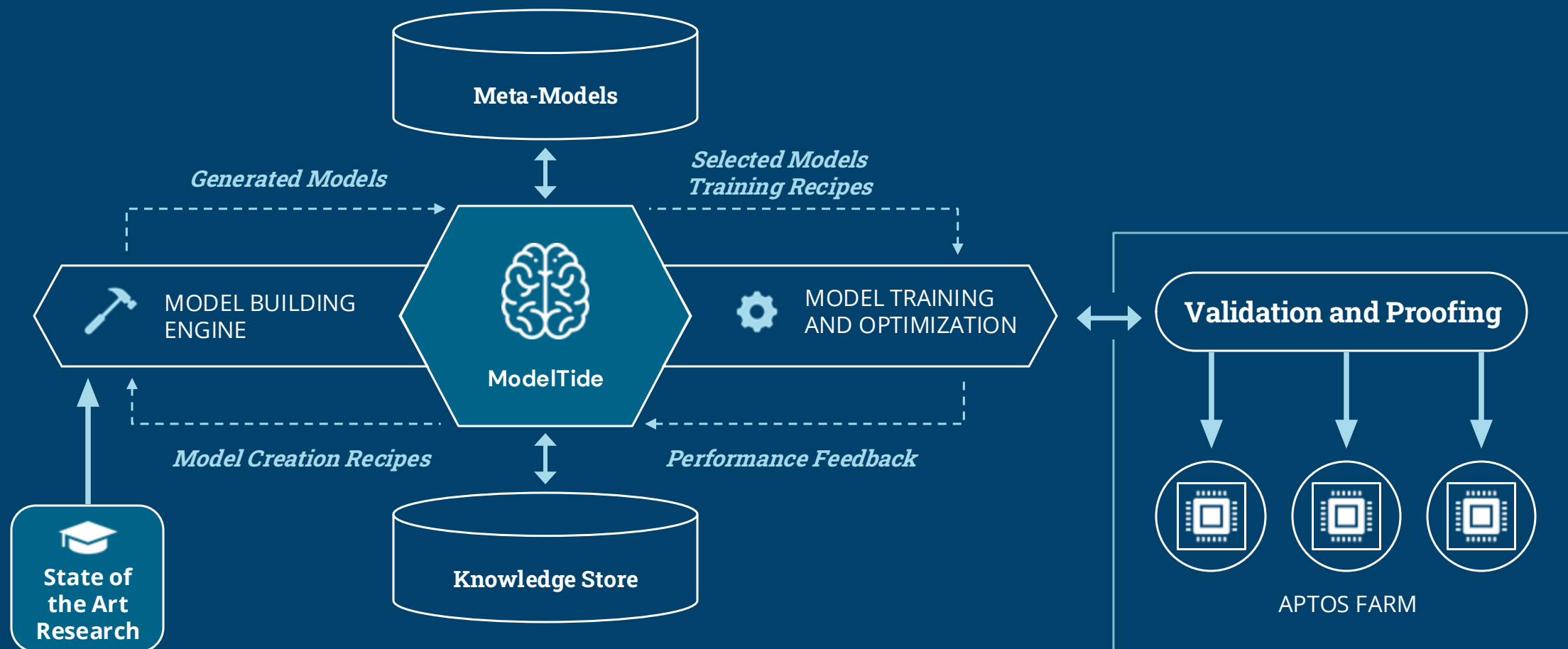
- **It's SLOW:** 12-18 months to develop and train a model and get it onto the device,
- **70% of projects FAIL!** The resulting ML is often ineffective and inefficient
- **It's BROKEN:** The cost and risk are a huge blocker. Data value remains trapped!

The Aptos Approach: *Data In, Model Out*



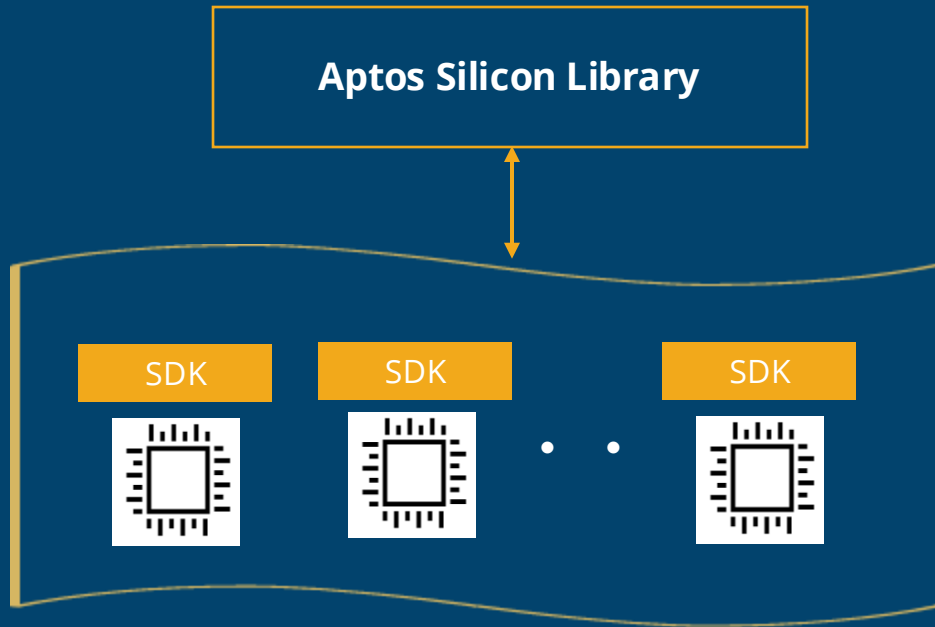
WHAT'S IN THE BOX

AI That Builds AI

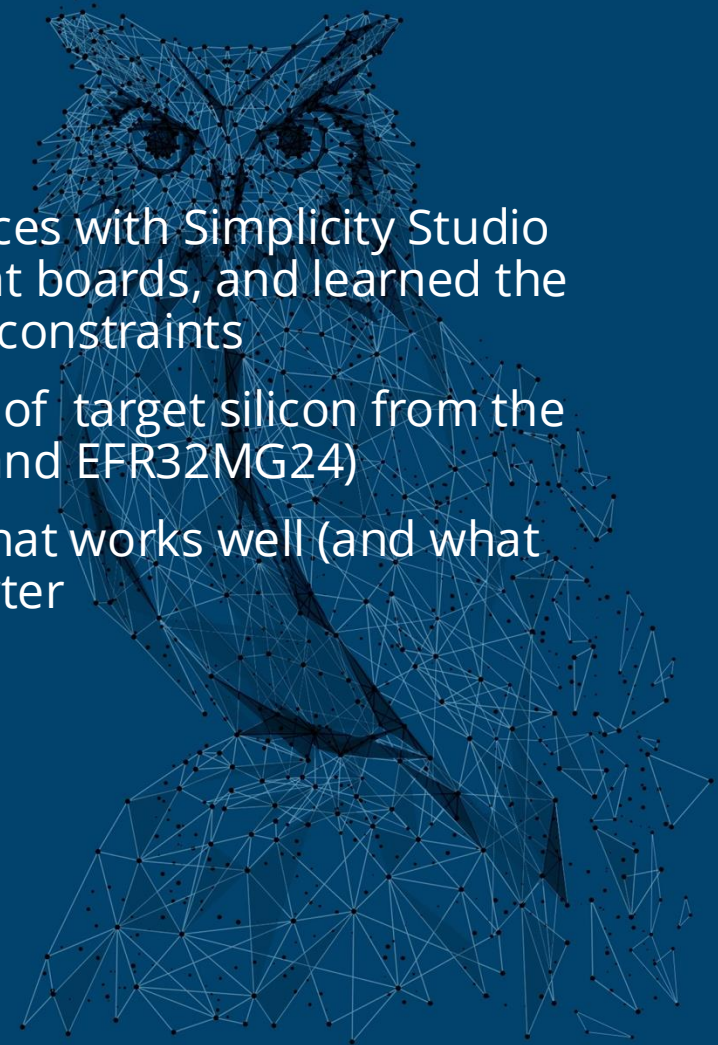


* Protected by issued and pending patents

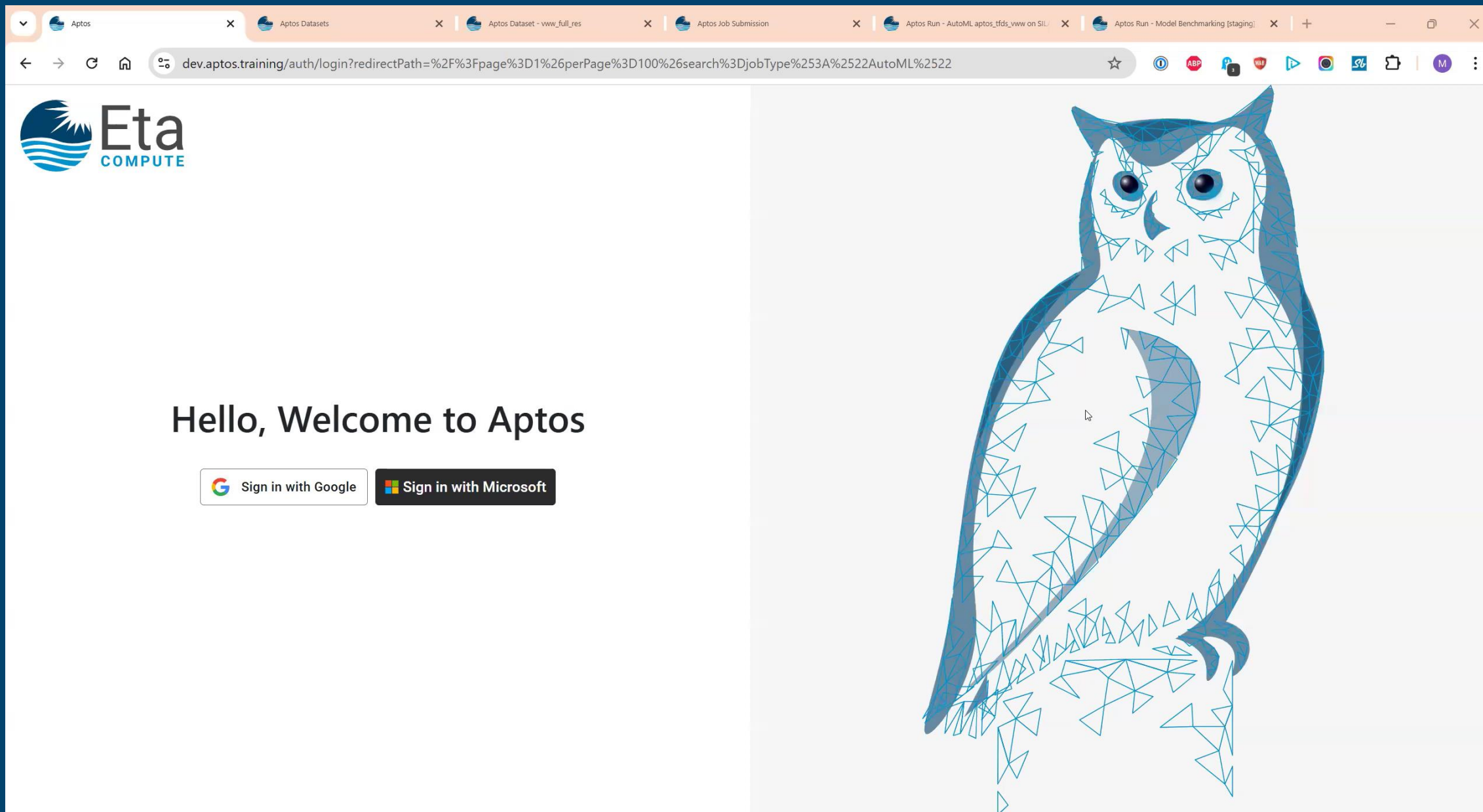
Eta Compute Worked with Silicon Labs to “onboard” Devices into Aptos Silicon Library



- Aptos characterized the devices with Simplicity Studio and Silicon Labs development boards, and learned the range of ML capabilities and constraints
- Users can select their choice of target silicon from the library (initially EFR32MG26 and EFR32MG24)
- Aptos continuously learns what works well (and what doesn't). Each use gets smarter



Visual Wake Demo



The screenshot shows a web browser window with multiple tabs open, including 'Aptos', 'Aptos Datasets', 'Aptos Dataset - vww_full_res', 'Aptos Job Submission', 'Aptos Run - AutoML aptos_tfds_vww on SIL', and 'Aptos Run - Model Benchmarking [staging]'. The address bar shows the URL: `dev.aptos.training/auth/login?redirectPath=%2F%3Fpage%3D1%26perPage%3D100%26search%3DjobType%253A%2522AutoML%2522`. The main content area on the left features the Eta Compute logo and the text 'Hello, Welcome to Aptos'. Below this are two buttons: 'Sign in with Google' and 'Sign in with Microsoft'. On the right side of the page, there is a large, stylized illustration of an owl. The owl is composed of a blue wireframe mesh, with its body and wings filled with a pattern of small blue triangles. The owl is facing forward, with large, dark eyes and a small beak. The background of the owl illustration is a light gray gradient.

AI/ML Demo



Aptos Transforms the Development of Edge ML

Drive more Products into Volume Production

- Overcomes the gap between ML and Embedded Systems through advanced tooling and automation
- One Step Model enables engineers to rapidly and successfully incorporate optimal edge ML models
- Enables ML talent to succeed in an embedded systems environment & leverage a target chip's unique ML capabilities



Free Trial

See for yourself what Aptos can do

Jon Gettinger

jon@etacompute.com

sales@etacompute.com

[Etacompute.com](https://etacompute.com)

Thank You

sales@etacompute.com

www.etacompute.com



SILICON LABS

CONNECTED INTELLIGENCE