

tinyML[®] Foundation

Enabling Ultra-low Power Machine Learning at the Edge

tinyML 2.0:
celebrating 5th Year of tinyML and the road ahead!

Evgeni Gousev, Senior Director, Qualcomm AI Research & tinyML Foundation BoD Chair



www.tinyML.org

Happy 5th Birthday tinyML !



It all started with an **INSPIRATION!**

- *can one run ML on a tiny (MCU-like) device ?*

TensorFlow

Goals: Tiny

- Framework that fits in 5KB of RAM, 20KB of Flash
- Speech demo with 30KB of RAM, 40KB of Flash

tinyML Summit 2019 - Pete Warden : TF-lite for tinyML

Pete Warden, Google

Our always-on vision research and innovation

Integrated vision sensor & processor, independent of main processor

Sensor
Low Power Pixels
Low Power Custom HW
CMOS Image Sensor
Very low power QVGA

Processor
Low Power CPU
Dedicated HW Blocks
Memory
Digital Processor
Trainable algorithms

Metadata output of scene understanding
Commands/Queries
Main Processor
App & OS Software

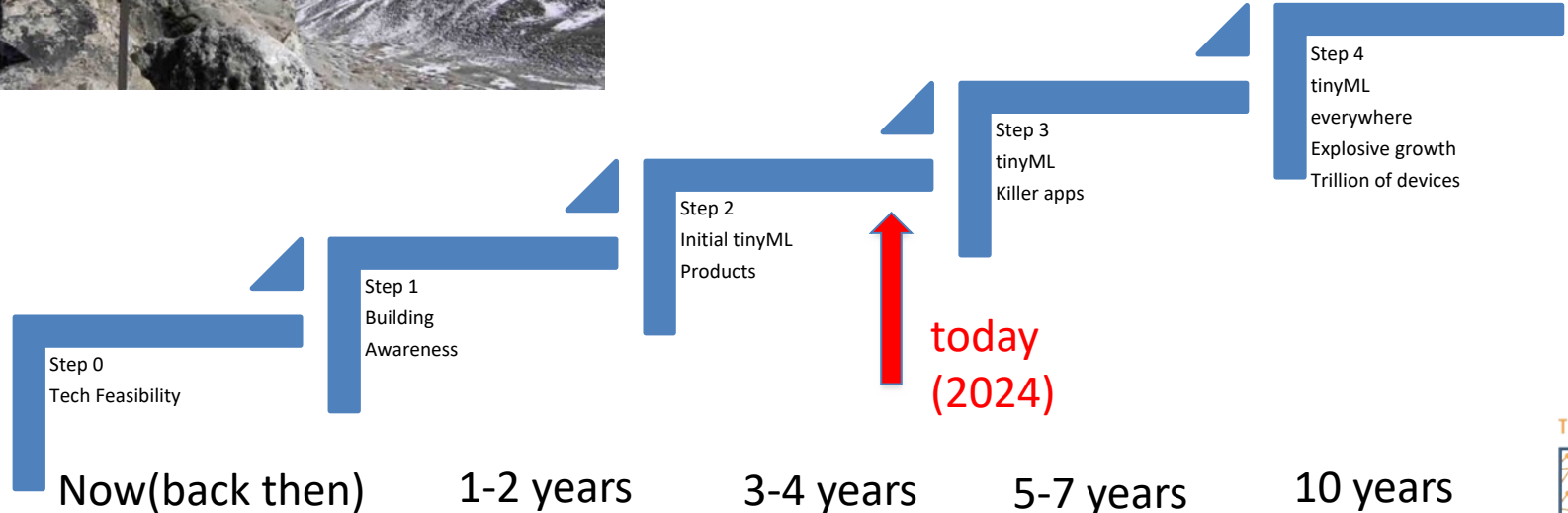
tinyML Summit 2019 - Edwin Park : Ultra-low Power Always-On Computer Vision

Edwin Park, Qualcomm

from 1st tinyML Summit 2019

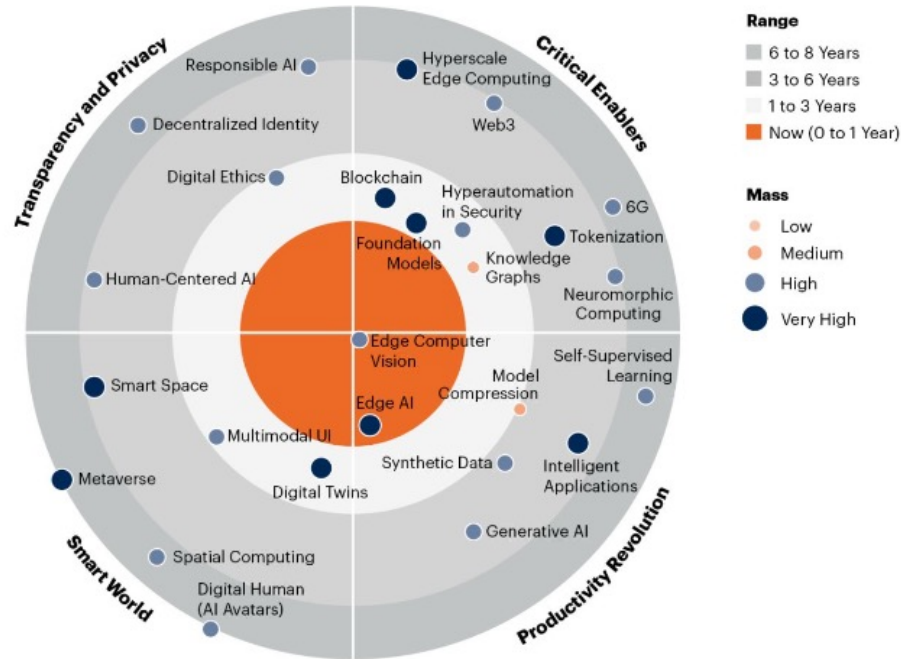


Climbing up tinyML mountain (from 1st, 2019 Summit)



Edge AI is happening now!

2023 Gartner Emerging Technologies and Trends Impact Radar



gartner.com



tinyML Community

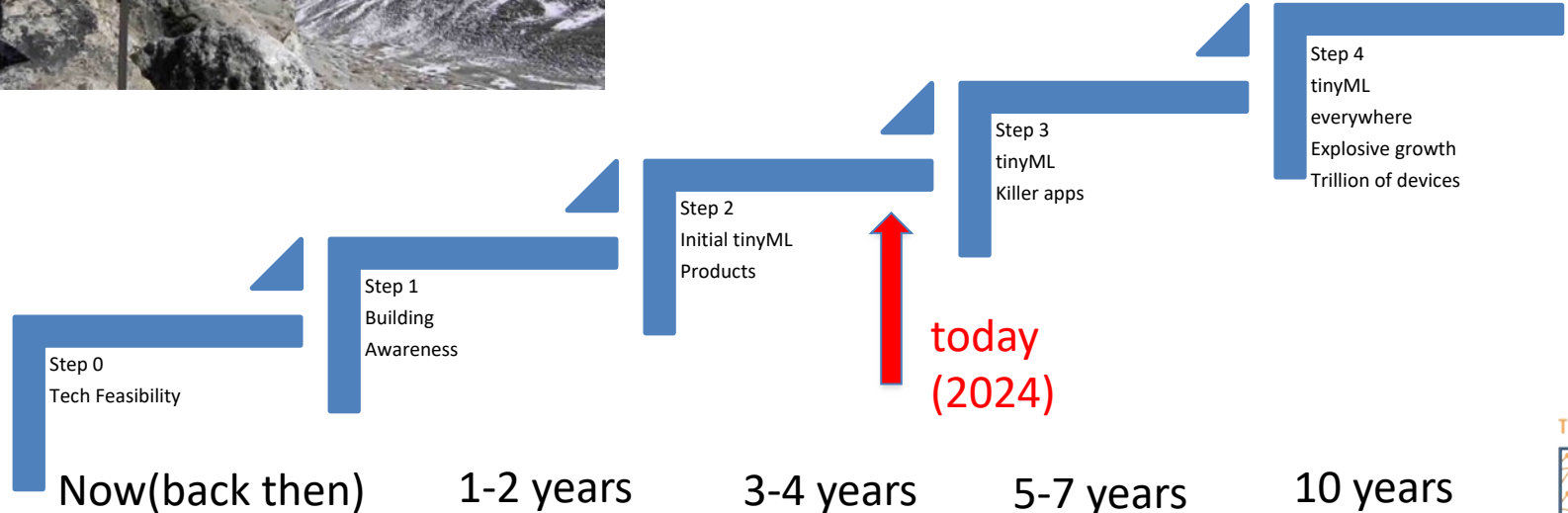
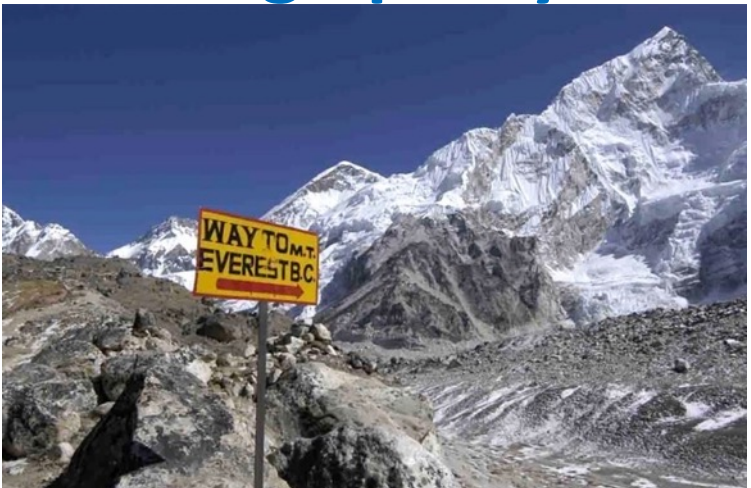


- **21k+ tinyML meetup members in 50 groups in 41 countries**
- **13.2k [youtube.com/tinyML](https://www.youtube.com/tinyML) subscribers, 704 videos, all FREE, 500k views**
- **5k members + 16k followers on LinkedIn**
- **Massive educational initiative, tinyMLedu (e.g. ~100k students in edX)**
- **~ 100 sponsors; amazing diversity**

Sponsors & Partners Make It Possible!



Climbing up tinyML mountain (from 1st, 2019 Summit)



The Future is at hand !



Emergence of GenAI on the Edge!



Released July 18, 2023

=



Released April 18, 2024

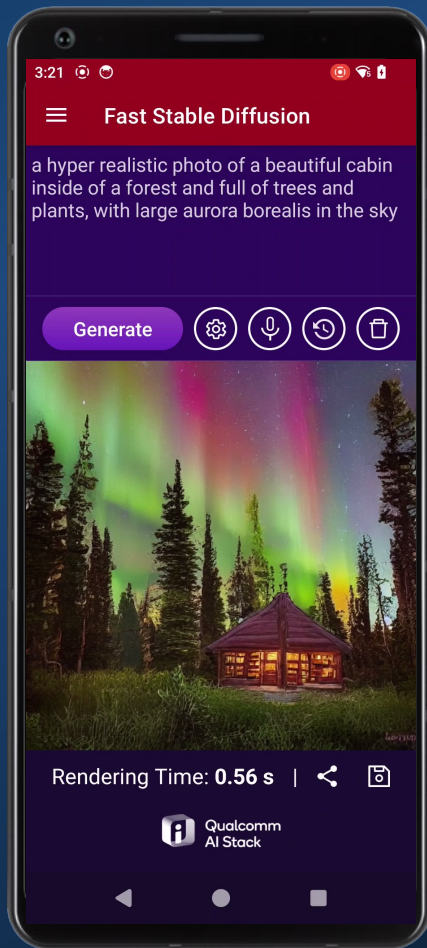
10x model size reduction in 9 months !

and

you can run it on a Raspberry Pi!



World's fastest AI text-to-image generative AI on a phone



Takes less than 0.6 seconds for generating 512x512 images from text prompts

Efficient UNet architecture, guidance conditioning, and step distillation

Full-stack AI optimization to achieve this improvement

SLMs will bring SUPERIOR capabilities to the EDGE

Why small language models are the next big thing in AI

James Thomason

@jathomason

April 12, 2024 12:06 PM

f X in



- “Clem Delangue, CEO of HuggingFace, suggested that
- up to 99% of use cases could be addressed using SLMs, and
 - predicted 2024 will be the year of the SLM.”



Emergence of “agentic” system



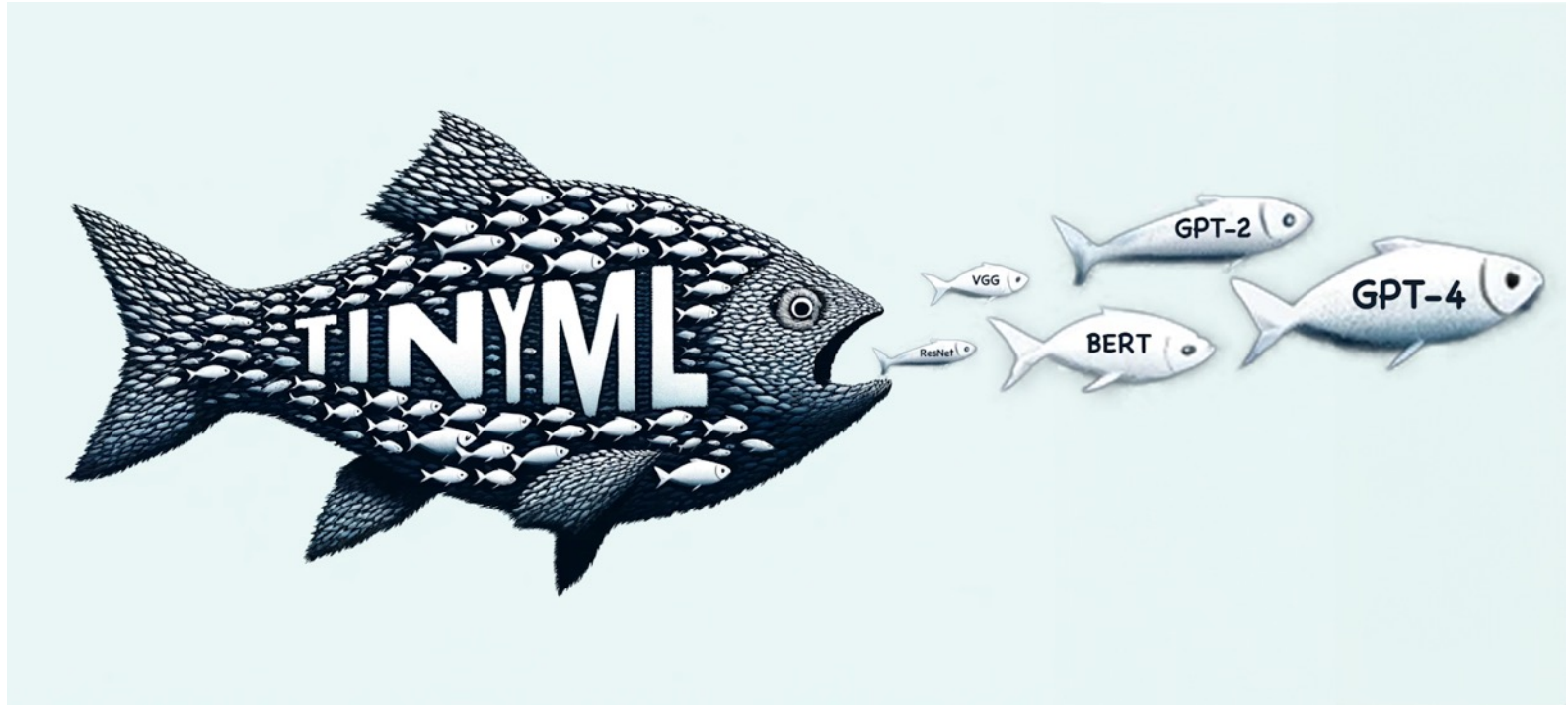
Broadly speaking, “agentic” systems refer to digital systems that can independently interact in a dynamic world

By moving **from information to action**—think virtual coworkers able to complete complex workflows—the technology promises a new wave of productivity and innovation.

Source: www.mckinsey.com (July 24, 2024)



Approaching inflection point for tinyML/edgeAI



Courtesy: Prof. Vijay Reddi and his Harvard team

PETE WARDEN'S BLOG

Ever tried. Ever failed. No matter. Try Again. Fail again. Fail better.

[HOME](#) [ABOUT](#)

Doom, Dark Compute, and AI

JANUARY 5, 2024
By Pete Warden
in [UNCATEGORIZED](#)
Tags: [AI](#),
[ARTIFICIAL-
INTELLIGENCE](#),
[COMPUTER](#),
[PROGRAMMING](#),
[TECHNOLOGY](#)
[LEAVE A COMMENT](#)



Subscribe

Join 1,960 other subscribers

[RSS - Posts](#)

RECENT POSTS

[Understanding the Raspberry Pi Pico's
Memory Layout](#)

[Doom, Dark Compute, and AI](#)

[Why I Love my Chevy Bolt EV](#)

[Stanford's HackLab Course](#)

Rough estimate:

100 B MCU devices @ 100 MHz = 1E22 TOPS (Int8)

2x greater than today's world's GPU/TPU capacity



Distributed Intelligence at the Edge

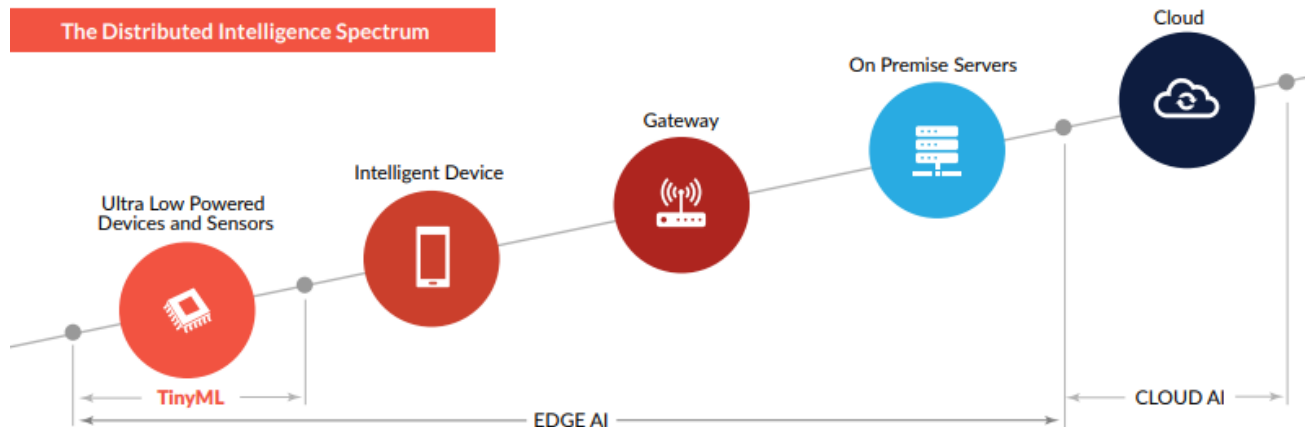
ABIresearch
TRUSTED INTELLIGENCE SINCE 1986



TinyML:
The Next Big
Opportunity
in Tech



The Distributed Intelligence Spectrum



Why tinyML Foundation?

Edge AI is Accelerating Digital Transformations

- Advances in semiconductors, connectivity and software have enabled a wave of AI on devices that securely connect to the cloud
- These capabilities are digitally transforming healthcare, agriculture, manufacturing, retail and many more...
- They take the form of low power sensors, cameras/imaging devices, gateways and robotics platforms – running AI Models with cloud-to-edge management and security

Why tinyML Foundation?

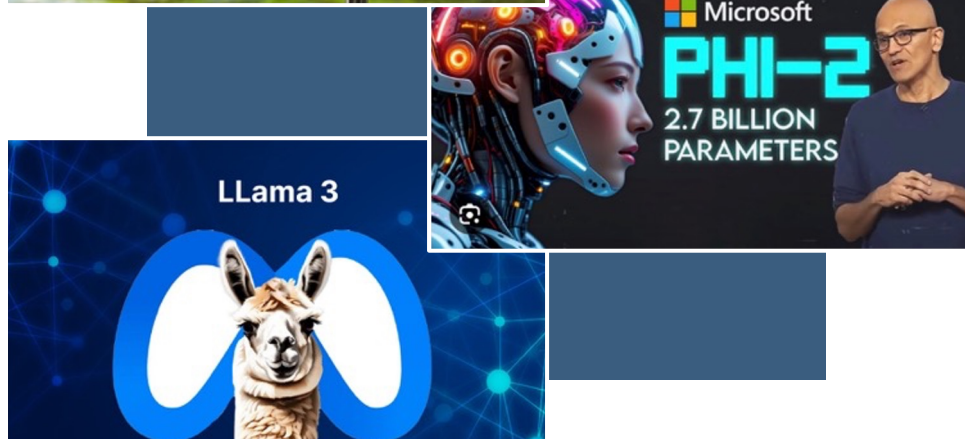
The State-of-the Art is Moving Quickly

- AI models can run as small as 5kb on sensors using only milliwatts – solar and battery powered
- Low power wireless connectivity such as LoRA and 5G Redcap are enabling far-flung deployments
- Generative AI models (and/or “Extractive AI”) coming to the edge – Meta’s LLAMA model shrank 10x in 9 months – for rich sensor descriptions and enhanced telemetry
- “Observational learning” is enabling robotics platforms on the edge with Generative AI



“up to 99% of use cases could be addressed using SLMs . . . 2024 will be the year of the SLM.”

– CEO of Hugging Face



A Broad Community of Expertise

- Community of R&D/academia
- Coursework/training for classes
- Access to technology companies for platforms, tools
- Opportunities for mentorship, internship, career paths
- AI for Good volunteering

- Access to leading edge tech/knowledge
- Access to voice of the customer
- Feedback and input on legislation on behalf of industry



- Knowledge of state-of-the-art
- Partnership and Business Development opps
- Exposure to ecosystem/solutions
- Access to talent
- Responsible AI guidelines/navigation

- Access to leading edge tech
- Exposure to tech suppliers
- Access to commercial customers
- Access to talent
- Responsible guidelines / navigation

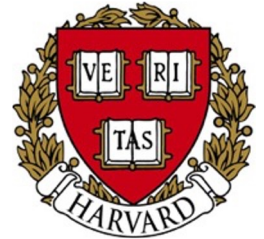
- Provide voice of the customer
- Access to leading edge tech/knowledge
- Provide AI for Good opportunities
- Find solutions with partners
- Access to talent
- Responsible guidelines / navigation

Bridging the Academic and Professional Communities

- tinyML EDU efforts have educated over 100k students worldwide
- tinyML Foundation Symposiums highlight academic research alongside industry events
- Support scholarships for next gen AI leadership
- Opportunities for mentorships, internships and talent pipelines
- Starting IAB (Industry-Academic Board)



The Abdus Salam
International Centre
for Theoretical Physics



MAYO
CLINIC



JOHNS HOPKINS
UNIVERSITY

Scholarship Partner Program

- Provide qualified **value-added services** to the tinyML Foundation Community
- Enable “**most favored nation**” **pricing** to tinyML Foundation Partners
- Drive a virtuous circle with funding for the **Academic Scholarship program**
- First two partners - **5V** and **EDGEIR**

The Tiny ML Foundation's Dedicated Talent Partner

5V Tech is proud to announce our partnership with the TinyML Foundation, solidifying our role as the dedicated talent partner for TinyML teams. We are committed to providing hands-on support for teams who need to scale their engineering and development teams, ensuring they have the right people to drive innovation and success.

Our expertise in talent solutions will help TinyML teams overcome challenges and achieve their goals.

TOM WHITE
CEO | 5V VALUES CONSULTING GROUP

"At 5V Tech, we are always looking for opportunities to foster innovation and support the pioneers who are shaping the future of technology. Partnering with the TinyML Foundation allows us to bring our expertise in scaling tech teams to a community at the forefront of making machine learning more efficient and accessible. We are excited to help drive the growth of TinyML, and work alongside the teams leveraging this technology to create meaningful change. As part of our joining the TinyML Foundation as a dedicated talent partner, please feel free to book a call with me directly to explore how we can help you."

[BOOK A CALL](#)

EDGEIR.com
Industry Review

ANNUAL SUPPORTERS PACKAGE

Become an EdgeIR.com Supporter

COMPLETE PACKAGE

- Enhanced company profile page featuring logo, video, whitepapers/webinars, interviews, articles and more.
- Targeted category directory ads
- Run of site 125x125 spotlight ad & description in rotation
- 728x90 & 300x250 banner ad in rotation & a bonus 468x60 in the newsletter when available.
- Brand focus feature interview
- White paper/webinar promotion and free lead generation
- Editorial preference and more perks!

*6 month test packages also available

Contact Candice Rodriguez - candice@edgeir.com or call 305-926-7751

[@Edge_IR](#) [/company/edgeir](#) [/edgeindustryreview](#)

Navigating Responsible AI and AI for Good Engagements for our Community

- Member of the ITU AI for Good effort
 - tinyML Workshop at the 2024 Summit in Geneva
- World Economic Forum member
 - Developing best practices across industries, educating policy makers
 - Member, AI Governance Council
- Sponsoring Hackathons tackling societal challenges
 - 2023 Pedestrian Safety Hackathon with Infineon, Sony and Brainchip



Strengthening Community with In-Person Events



“It is my first time attending the tinyML Summit, and the experience has been fantastic. I had the opportunity to network with top edge ML researchers from both industry and academia, discussing the current and future challenges of implementing ML on low-power devices...”

– Wilfredo Lugo-Beauchamp,
Assistant Professor, University of Puerto Rico



tinyML Foundation Milan – June 2024



*“The forum was not just about sharing knowledge but also about absorbing the creative energy from many friends and colleagues I met for the first time. The future of innovation is bright, and it's built on the ****dense layers of human connections**** we make...”*

– Nabarun Dasgupta, ST Microelectronics



Sample Topics Addressed by the Community

Symposiums



Online and in-person, presentations and panels with industry and academic researchers that dig into what's next with Edge AI

- AI/ML for Embodied Systems at the Edge: Generative Models, LLMs and Beyond
- Transformer-Based Model Deployment on Edge Devices through MicroNPUs Operator Converter
- On-device Contextual AI: Challenges and Opportunities
- AI models for Medical Devices
- Inspired by 'Her': AI interaction models we'd like to see
- High Efficiency Computer Vision
- Ultra-Efficient On-Device Object Detection on AI-Integrated Smart Glasses
- Tiny Graph Neural Networks for Radio Resource Management
- Comparing Classic Machine Learning Techniques with Deep Learning for TinyML Human Activity Recognition
- Simulating Battery-Powered TinyML Systems Optimized using Reinforcement Learning in Image-Based Anomaly Detection
- SpokeN-100: A Cross-Lingual Benchmarking Dataset for the Classification of Spoken Numbers in Different Languages

RECENT EVENTS

Asia 2023
(Seoul)

Predictive Maintenance & Anomaly Detection

Summit 2024
(Silicon Valley)

GenAI on the Edge

EMEA 2024
(Milan)

Partnership Value

Technology Providers



Understand state-of-the-art, develop cutting-edge partnerships, engage with customers and solution providers, marketing value

Commercial Customers



Discover solution and technology partners, Edge AI technology trends, talent pipeline

Academic Institutions



Collaboration on AI R&D, mentorship/ internship opps for students in technology industry

We are on a mission. . .

- Thought leadership
- Accelerating partnerships
- Voice of the Customer
- Cultivate next-gen leadership

High bandwidth business results

- Regional Community Events
- Working groups
- Partnerships
- Hackathons for Industry Verticals
- Knowledge Certification programs

Accelerate academia

- Symposiums to highlight cutting edge research
- Scholarships and funding
- Talent pipelines, and mentorship/internship programs

Societal impact

- AI for Good via United Nations and World Economic Forum ++
- Responsible AI thru policymaker education
- Security/transparency working groups
- Hackathons for societal solutions

Make Edge AI a Digital AND a Societal Transformation



tinyML Foundation support and organization

FOUNDATION



Pete Bernard
Executive Director
pete@tinyML.org



Ira Feldman
Operations
ira@tinyML.org



Elfego Solares
Program and Event
Manager
elfego@tinyML.org



Rosina Haberl
Event Organizer
rosina@tinyML.org



Olga Goremichina
Meetups and Talks
olga@tinyML.org



Pete Bernard, Executive Director



- EDGECELSIOR - Founder & Principal
- Microsoft - 18 years
 - Senior Director - Azure Edge HW/5G
 - Principal Group Program Manager - WDG Product Engineering
- Silicon Valley - 14 years
 - Insignia Solutions - Chief Product Officer
 - E Color - VP Products
 - Phoenix Technologies - Dr. of Technology
- BS Computer Engineering - Boston University

Join Growing tinyML Communities:



The tinyML Community

<https://www.linkedin.com/groups/13694488/>



The tinyML Channel

<https://www.youtube.com/tinyML>

