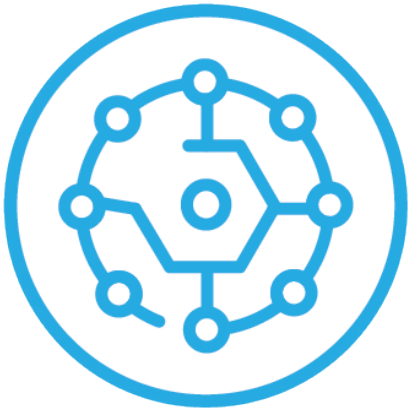
EPoSS.
European Association on
Smart Systems Integration



Joint workshop by EPoSS and
INSIDE Industry Associations

The Future of Innovation in Edge AI

Online, April 04, 2025

Notes from the Workshop
Part III: Bridging the Digital Divide
Caused by Generative AI
Presented by Danilo Pau, STMicroelectronics

1. Understanding the Divide: A Two-Speed World

The emergence of **Generative AI (GenAI)** has created one of the most significant technological divides in modern computing. On one side are a handful of companies – primarily in the U.S. and China – with near-unlimited compute, financial resources, and domain expertise. On the other side is "everyone else" – not just smaller tech companies and startups, but also small and medium-sized enterprises (SMEs), public institutions, and the wider scientific research community.

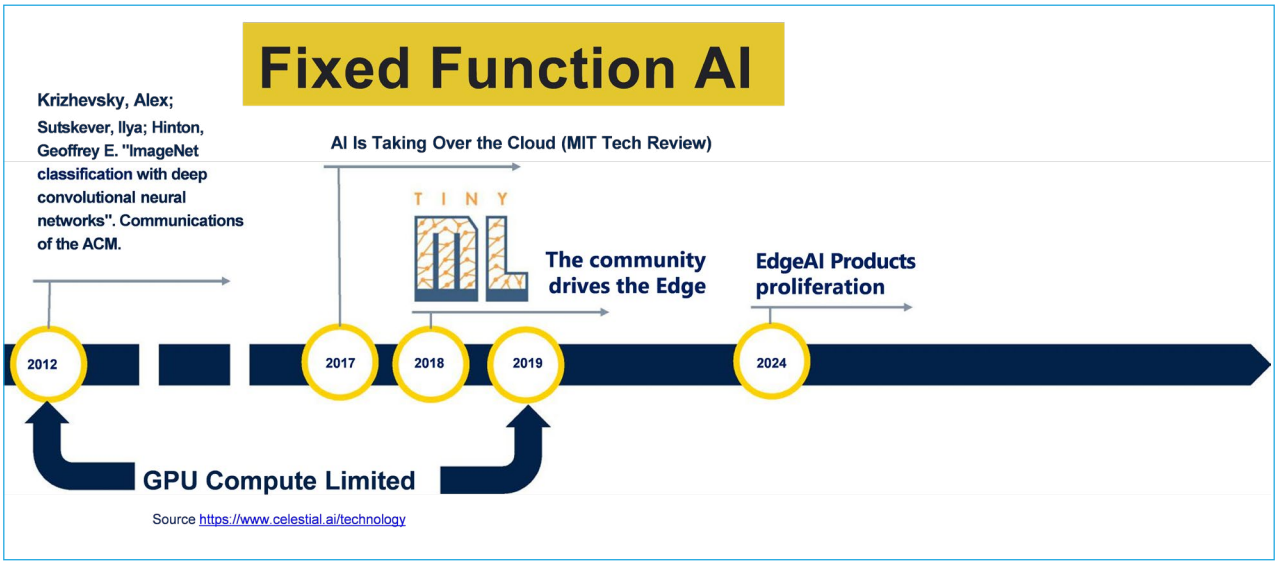
Copyright: STMicroelectronics

This disparity, referred to as the **Digital Divide**, has grown as large-scale GenAI models become increasingly central to innovation – but also increasingly out of reach for those without deep pockets and massive infrastructure.

2. From GPU-Limited to Edge-Capable: The Evolution of AI

The shift began in 2012 with the introduction of **AlexNet**, sparking the **GPU compute-limited era**. Until about 2019, it was still feasible for researchers and developers to train and deploy deep learning models on off-the-shelf GPUs. In parallel, cloud computing services emerged, offering nearly limitless compute power. However, these cloud solutions posed cost barriers that especially burdened SMEs and academic users.

In response, the **TinyML community** emerged in 2019, advocating for efficient, low-power AI at the edge. By 2024, a new era of **Fixed Function AI** had matured, supporting tasks like classification, detection, and segmentation – all within milliwatt power budgets and deployed across heterogeneous embedded devices.



3. Tackling Heterogeneity with ST Edge AI Core Technology

Deploying AI on edge devices presents significant technical challenges, largely due to **heterogeneity** – the variety of hardware platforms, real-time operating systems, sensor types, and AI workloads. STMicroelectronics addressed this challenge by developing the **ST Edge AI Core technology**.

This solution includes:

- **Model importers and analyzers**
- **Optimizers and code generators**
- **Validation tools ensuring lossless deployment**

Crucially, the technology enables developers to train and deploy AI models without needing to install anything locally, using the **STM32Cube.AI Developer Cloud**. This cloud-based service interfaces directly with real hardware – no simulators – allowing developers to test and validate their models on actual microcontrollers and sensors.

4. The Rise of Generative AI: A New Divide Emerges

While fixed-function AI is now effectively deployed at the edge, **Generative AI** has introduced new complexities. Starting around 2014 with **Generative Adversarial Networks (GANs)** and popularized by breakthroughs like **Transformers** (2017), GenAI workloads have become increasingly hyperparameterized and resource-intensive.

This shift created what Danilo Pau recalls be the **Memory Wall** – a significant barrier where even single GPUs are no longer sufficient to handle GenAI training and inference. Solutions proposed by industry leaders involve building compute superclusters with hundreds of thousands of GPUs, often consuming vast amounts of energy and water, and generating significant e-waste and CO_2 emissions.

5. Mass Deployment? Not So Fast

A case study using **Qwen2-VL-7B-Instruct**, a state-of-the-art multimodal GenAI model, illustrates the limitations. Assuming modest usage (60 tokens per user per query with a five-second latency requirement), scaling this model to serve all **5.16 billion smartphone users** would require:

- Over **40,000 AI superclusters** (each the size of NVIDIA's Cortex AI cluster)
- Massive power infrastructure (up to 130 MW per cluster)
- Unfeasible levels of acceleration and cost

In short, deploying GenAI at scale via the cloud is **unsustainable** – economically and ecologically.

6. Toward a U-Turn: GenAI at the Edge

A strategic pivot is needed: **bringing Generative AI to the edge**. In 2024, the **TinyML Foundation** rebranded as the **Edge AI Foundation**, expanding its scope to include GenAI. The goal is to democratize access to generative technologies by enabling them to run efficiently on resource-constrained devices.

Recent studies identified **135 papers** on GenAI edge deployment (2022–2024), of which **66 were deemed in scope** – indicating early but promising traction. Most deployments occurred on:

- **Smartphone hardware** (e.g., Qualcomm Snapdragon, Apple A-series, Samsung Exynos)
- **Developer boards** (e.g., NVIDIA Jetson, Raspberry Pi)
- A marginal share (<10%) on **microcontrollers**

7. Europe's Opportunity in Edge GenAI

Despite lagging in cloud-based GenAI infrastructure, **Europe holds strong potential at the microcontroller level**. STMicroelectronics has released new platforms:

- **STM32MP2 series** – Successfully deployed models like OLlama Q1, achieving draft performances of 2.42 tokens/sec.
- **STM32N6 series** – Used for style transfer and image anonymization at 10 frames/sec in partnership with Fondazione Bruno Kessler (FBK).

These examples demonstrate that **GenAI workloads can be run on the edge** with acceptable speed and power consumption – even with constrained hardware.

8. The Energy Efficiency Race

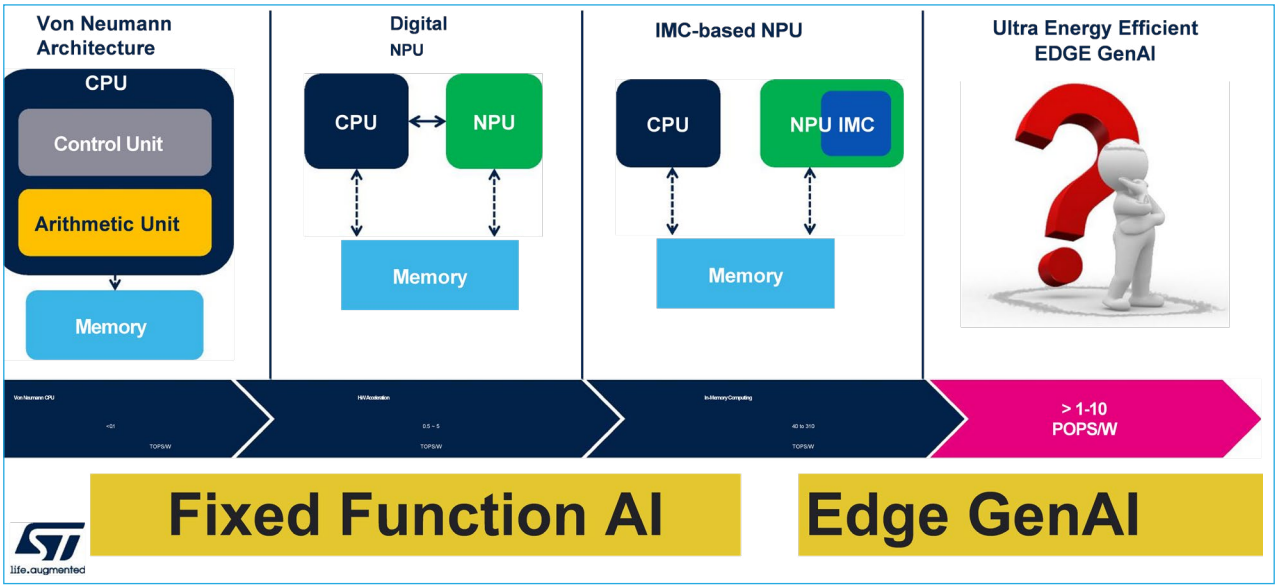
To make edge GenAI truly viable, **energy efficiency must drastically improve**. Current benchmarks show:

- **Fixed function AI** on STM32 achieving 3 TOPS/W
- **In-memory computing (IMC)** reaching up to 300 TOPS/W with binarized workloads

The long-term goal is to push energy efficiency toward **1–10 POPS/W**, unlocking edge GenAI applications that are not only fast and affordable but also environmentally sustainable.

9. Conclusion: Building a Community for Sustainable AI

The only way to bridge the divide is to collaborate across industry, academia, and public stakeholders. The Edge AI Working Group under the auspices of **EdgeAI Foundation** – co-chaired by Danilo Pau and Prof. Hajar Moussanif – has taken a leadership role in this effort, organizing forums like the **Generative AI on the Edge** virtual event in March and October 2024 and a planned one on May 2025.



Copyright: STMicroelectronics

As GenAI continues to shape global technology trends, the future will not be defined by how large a model can grow – but by how efficiently and responsibly it can be deployed. **Edge GenAI** represents not just a technical shift, but a necessary realignment of strategy, sustainability, and inclusion. An opportunity that the research and industry European Community shall not miss to keep worldwide leadership role on silicon and services in the next decade.