life.augmented

Get to know ST and how it leads AI at the Edge

Danilo PAU

Technical Director

IEEE, ST and AAIA Fellow,

APSIPA Life Member, Sigma XI

System Research and Applications

Top 5 Semiconductor foundries

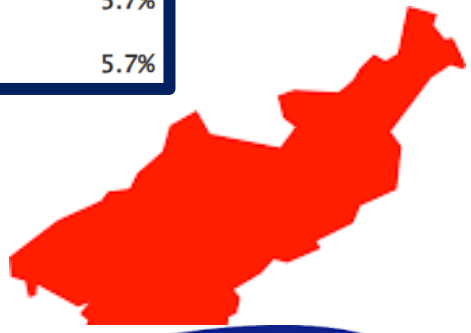
intel

GLOBAL FOUNDRIES

Q1/2024 market share 89.2%.

TSMC	61.7%
Samsung	11%
GlobalFoundries	5.1%
UMC	5.7%
SMIC	5.7%

SMIC



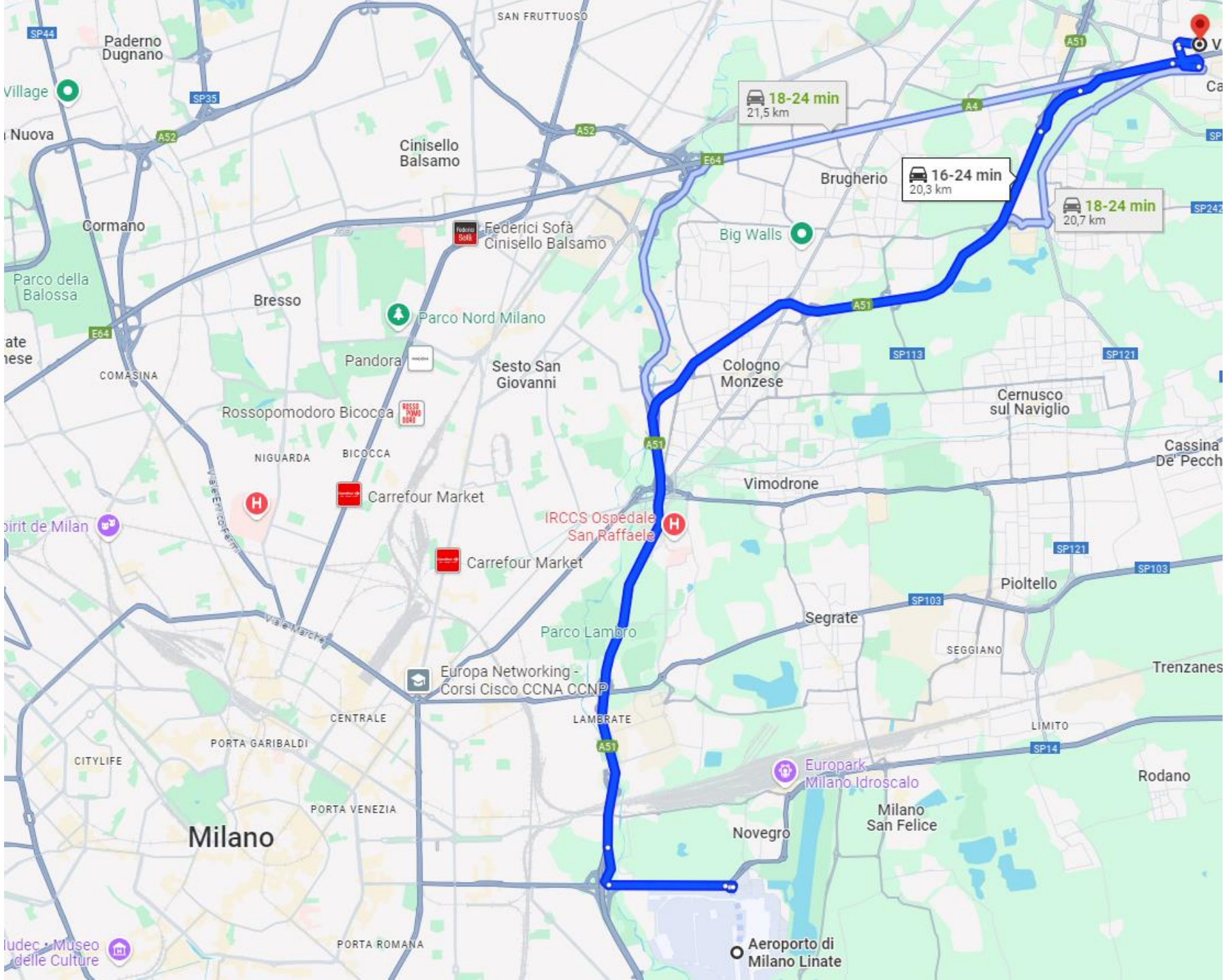
SAMSUNG

SEMICONDUCTORS

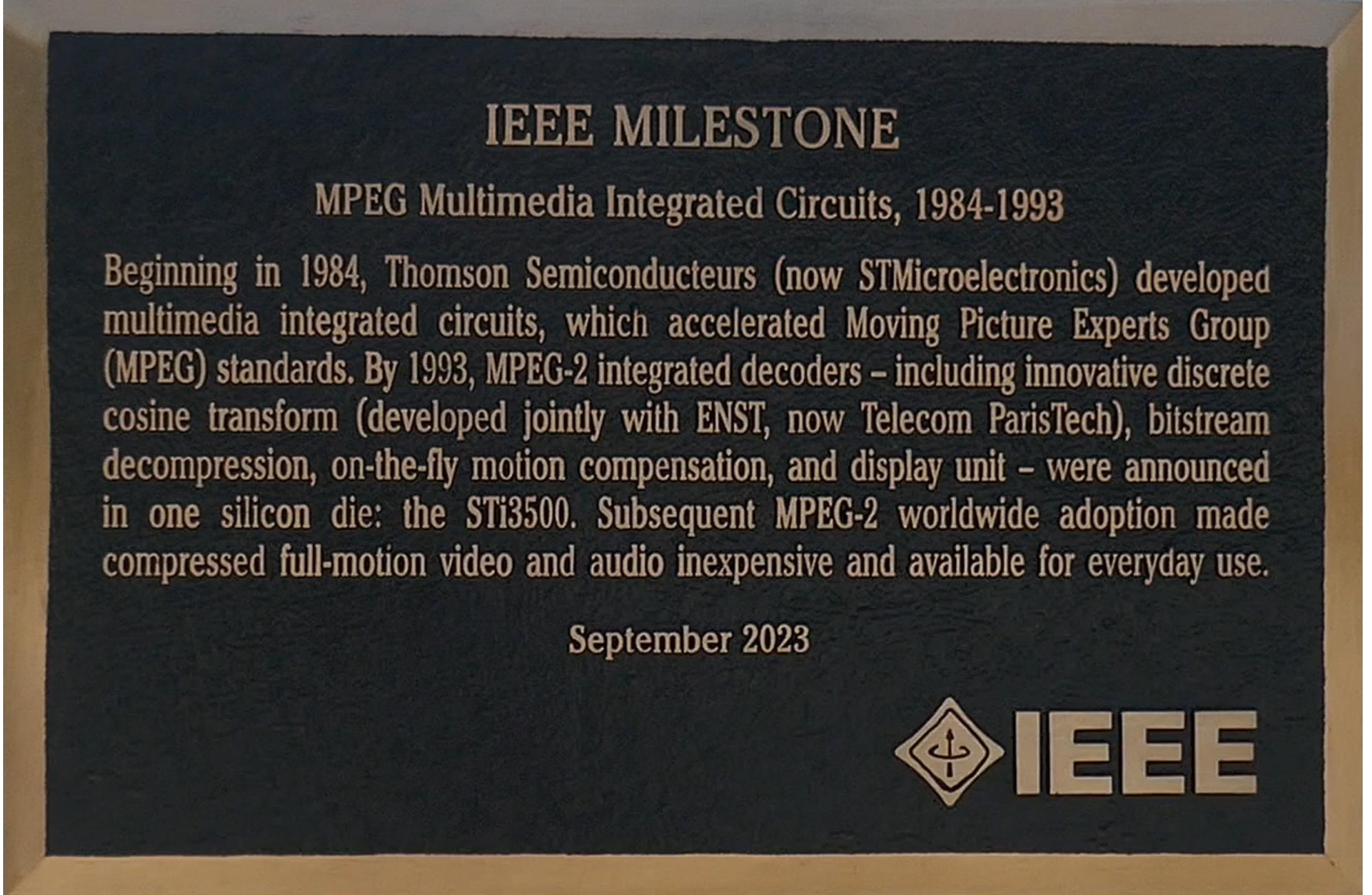
tsmc

UMC

life.augmented


18-24 min
21,5 km
16-24 min
20,3 km
18-24 min
20,7 km
Paderno Dugnano
Cinisello Balsamo
Cormano
Federici Sofà Cinisello Balsamo
Big Walls
Brugherio
Parco della Balossa
Bresso
Parco Nord Milano
Pandora
Sesto San Giovanni
Cologno Monzese
Cernusco sul Naviglio
Rossopomodoro Bicocca
NIGUARDA
BICOCCA
Carrefour Market
IRCCS Ospedale San Raffaele
Vimodrone
Cassina De' Pecch
Carrefour Market
Pioltello
Segrate
Parco Lambro
SEGGIANO
Europa Networking - Corsi Cisco CCNA CCNP
CENTRALE
LAMBRATE
PORTA GARIBALDI
CITYLIFE
LIMITO
Europark Milano Idroscalo
Rodano
PORTA VENEZIA
Milano
Milano San Felice
Novegro
PORTA ROMANA
Aeroporto di Milano Linate
SAN FRUTTUOSO
Cinisello Balsamo
COMASINA
Trenzanes

life.augmented

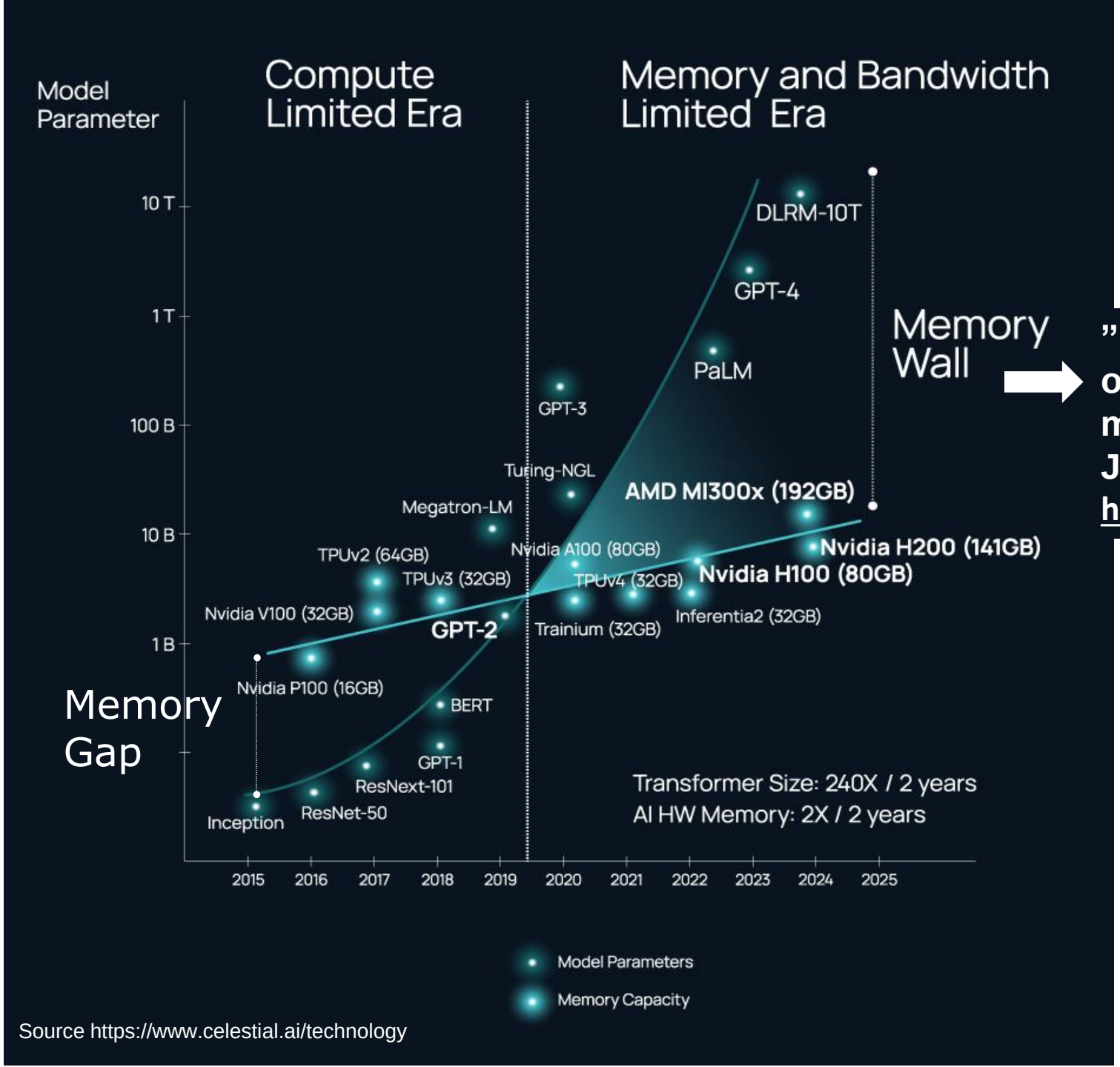


life.augmented

life.augmented

life.augmented

Energy-efficiency edge AI solutions for a more sustainable world



”Process trillions of bytes, billions of times. Distribute workloads on millions of GPUs”
Jensen Huang, CEO nVIDIA
https://youtu.be/MwiM_nPyx5Y

Source https://www.celestial.ai/technology

life.augmented

LLava on STM32MP2

The man is sitting on the side of a road, which is covered in debris. He appears to be in a state of distress. It seems like he is also in a state of pain. The ground he is sitting on is made of concrete, and there are other people standing in the distance. It looks like there are also orange lights visible in the distance, which could be part of a rescue operation. The scene is quite dramatic and tense.

Where to deploy GenAI ?

Data centers ? **OR** **At the Edge ?**

- The average word length in the English language is 4.7 characters [1].
- In most cases, the text content of a Tweet can contain up to 280 characters [2].
- A Tweet can contain maximum 60 tokens

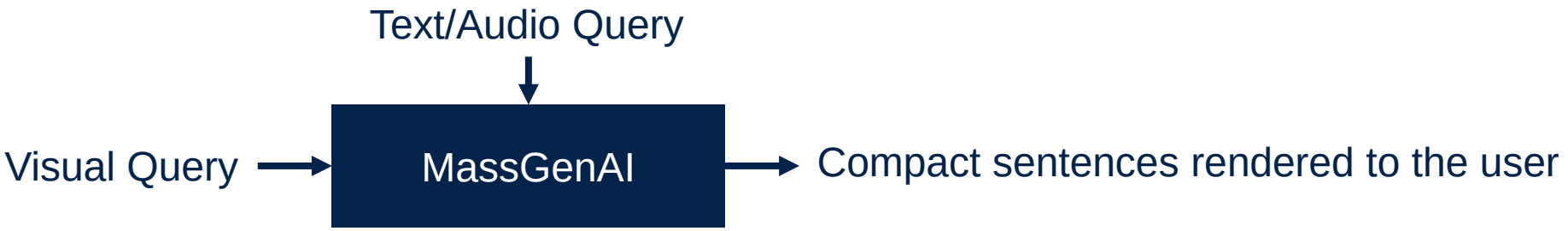
- iPhone users worldwide are expected to rise to 1.56 billion by the end of 2024 [3].
- Android OS users worldwide are expected to rise to 3.6 billion by the end of 2024 [4].
- By 2024, there will be a potential of 5.16 billion total available users (TAU) for **M**ultimodal **Ass**istant powered by **Gen**erative **AI** [**MAssGenAI**]

[1] https://www.wyliecomm.com/2021/11/whats-the-best-length-of-a-word-online/
[2] https://developer.x.com/en/docs/counting-characters#:~:text=In%20most%20cases%2C%20the%20text,as%20more%20than%20one%20character.
[3] https://www.coolest-gadgets.com/iphone-statistics#:~:text=In%20the%20first%20three%20months,of%20the%20global%20smartphone%20market.
[4] https://www.coolest-gadgets.com/android-statistics#:~:text=By%202023%2C%20there%20will%20be,reach%203.6%20billion%20by%202024.

life.augmented

danilo.pau@st.com

MassGenAI Workload Example



- **Qwen2-VL-7B-Instruct-GPTQ-Int8** [5]:
  - state-of-the-art performance on visual understanding benchmarks, including MathVista, DocVQA, RealWorldQA, MTVQA, etc.
  - can understand videos over 20 minutes for high-quality video-based question answering, dialog, content creation, etc.
  - can be integrated with devices like mobile phones, robots, etc., for automatic operation based on visual environment and text instructions.
  - supports the understanding of texts in different languages inside images, including most European languages, Japanese, Korean, Arabic, Vietnamese, etc.
- **Performances** (NVIDIA A100 80GB)
  - Speed (tokens/s) 31.6 (input length 1)
  - GPU memory (GB) 10.11

life.augmented

[5] https://huggingface.co/Qwen/Qwen2-VL-7B-Instruct-GPTQ-Int8

danilo.pau@st.com

How Much is MAssGenAI Feasible on the Cloud ?

- Hypothetical service condition for GenZ.
  - Monthly subscription → 15 $/month
  - Maximum acceptable latency → 5 sec

$$\frac{60\,\frac{tokens}{user}}{31.6\,\frac{tokens}{sec}} \sim 2\,\frac{sec}{user}$$

$$5{,}160{,}000{,}000\ users * 2\ \frac{sec}{user} \sim 10{,}320{,}000{,}000\ \text{sec}\ \boldsymbol{or}\ 119{,}444.444\ days\ \boldsymbol{or}\ 327.3\ years$$

$$level\ of\ acceleration\ required\ to\ support\ 5.16\ B\ users\ simultaneously$$

$$\frac{10{,}320{,}000{,}000\ sec}{5\ sec} \sim \boldsymbol{2{,}064{,}000{,}000\ or\ 2B\ times\ or}$$

$\boldsymbol{41{,}280\ Cortex\ AI\ Superclusters\ (50{,}000\ H100, 130\ MW\ each)}$ [6] **!**

[6] https://technologymagazine.com/ai-and-machine-learning/a-first-look-at-elon-musks-new-cortex-ai-supercluster

life.augmented

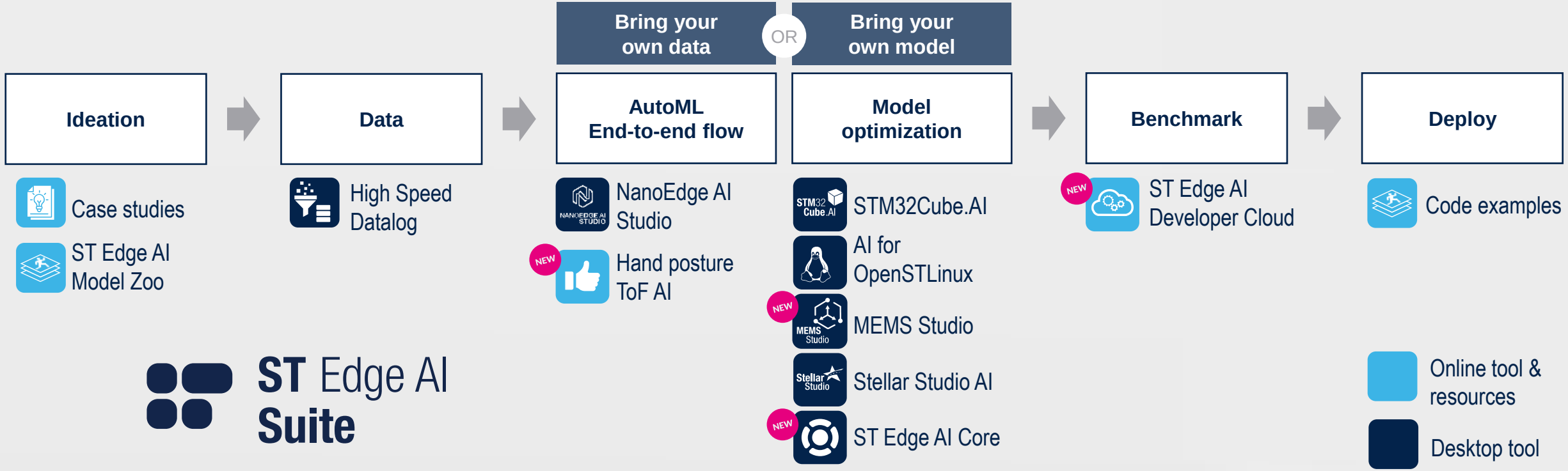
danilo.pau@st.com

Edge AI is enabled by a different category of software and hardware

life.augmented

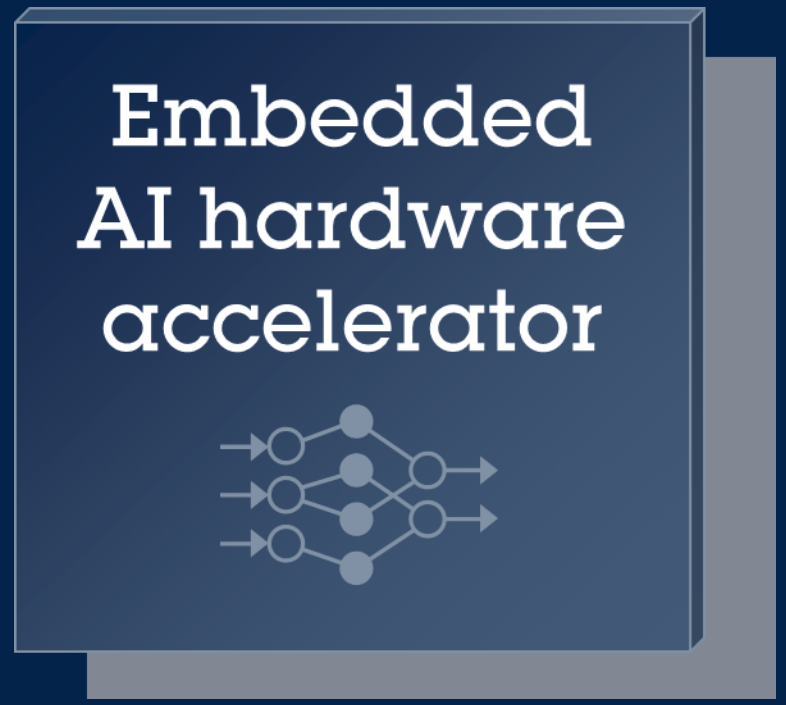
Free tools to run edge AI on MCUs, MPUs, and smart sensors

Specific tools developed for edge AI can greatly help to solve most of the challenges, especially minimizing the risk in term of time and resource investment.



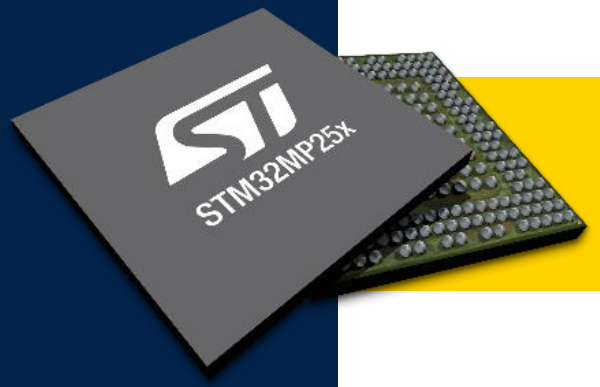
Bringing advanced compute capabilities to enable a new class of embedded machine learning applications

Embedded AI hardware accelerator



Sensors are enabled with an intelligent sensor processing unit (ISPU) for in-sensor edge AI processing.

STM32N6 adding neural acceleration to the STM32 lineup for unmatched levels of performance in an MCU.



STM32MP2 bringing neural acceleration to the embedded Linux world.

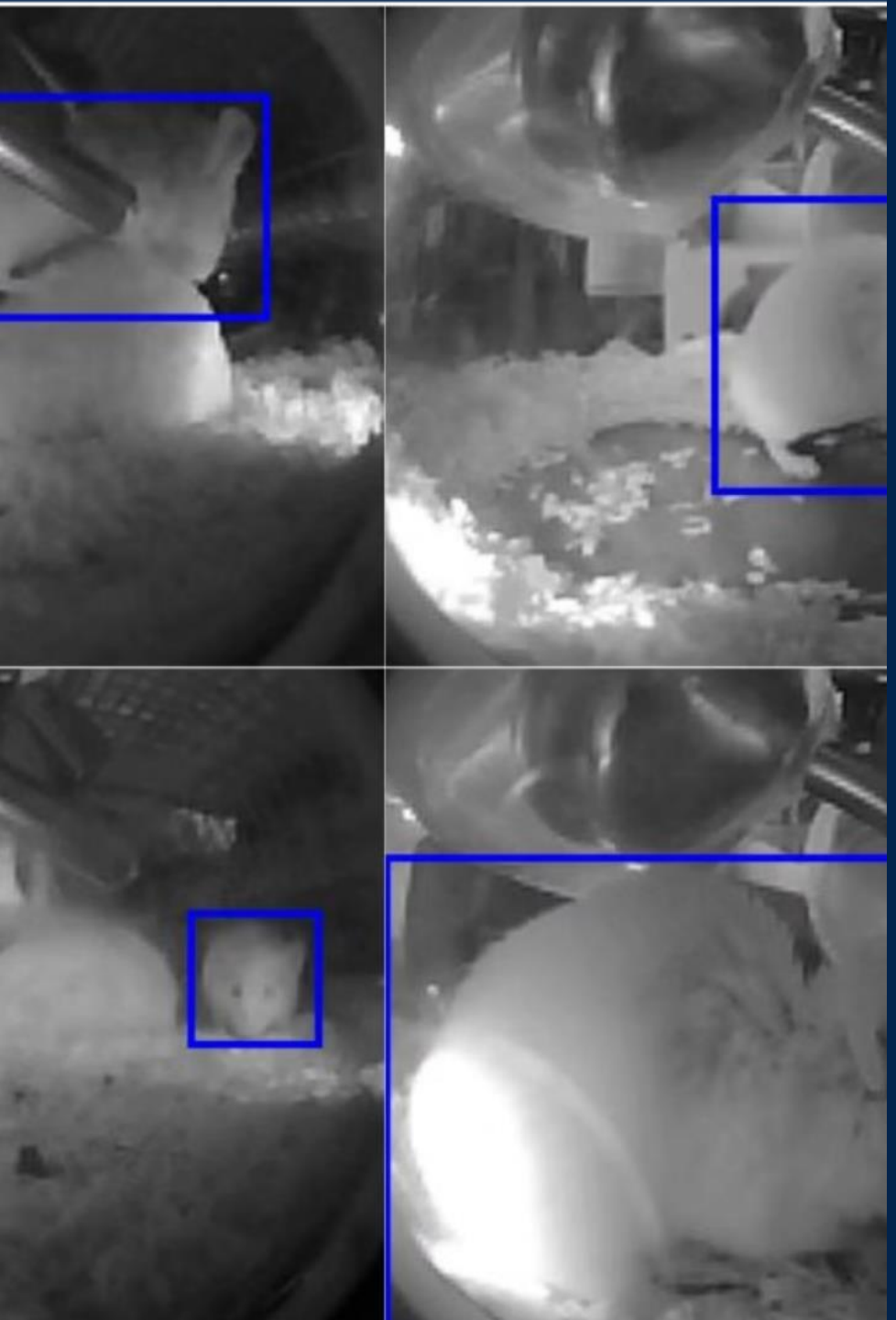
life.augmented

Deep Learning Based Object Detection for Embedded Systems: Exploring Algorithms and Optimization Techniques

by **Beshoy Guirges**

Supervised by **Prof. Claudio Cusano** (thesis and internship supervisor) and by **Danilo Pau** (ST company tutor).

life.augmented



Context: Mice in Pre-clinical Trials

Reason

Ethical & Pragmatic

Mice Tracking

Avoiding human interactions

Goal

Low stress, high welfare

Requirements

1. **Real-time Inference**
2. **MCU Deployability**
3. **High Accuracy**

Results: Accuracy and Performance

Model	Val mAP	Test mAP	**INT8 Test mAP**
YOLOv5-fpn_175K	0.735	0.722	**0.707**
YOLOv5n_264K	0.763	0.752	**0.706**
YOLOv8n_325K	0.753	0.744	**0.623**

Model	**FPS**	RAM (MB)	FLASH (MB)
YOLOv5-fpn_175K	**3.03**	0.74	0.35
YOLOv5n_264K	**2.73**	0.74	0.45
YOLOv8n_325K	**1.50**	1.24	0.51

life.augmented

YOLOv8n (no PTQ) achieves 0.09 FPS

ZINC
not another conference

IEEE

D. Pau, M. Garzola and B. Guirges, "Visual Monitoring of Mice Wellness by Tiny Detectors," 2024 Zooming Innovation in Consumer Technologies Conference (ZINC), Novi Sad, Serbia, 2024, pp. 66-71, doi: 10.1109/ZINC61849.2024.10579289.

https://ieeexplore.ieee.org/document/10579289/authors#authors

life.augmented

Our technology starts with You

life.augmented