Industrial View on the Future of Edge AI

Holger Schmidt
25.11.2022

Hardware for Edge AI

- Motivation:
  - Existing edge hardware is not powerful and smart enough to enable the digitalization trends in all application fields and to support the handling of the challenges of the 21st century

- Current state:
  - Overall performance of current edge device is very limited
  - Mature AI acceleration mainly in the cloud
  - Most AI accelerator for edge device on prototype level or only for simple applications

- Way forward:
  - Investigating new technologies for efficiently processing different applications on devices at the edge (e.g. Spiking Neuronal Networks)
  - Developing better periphery components for the accelerators
  - Exploring and developing new components for AI accelerators e.g. new memory technologies
  - Realize better tool chains
  - Reduce size and energy consumption of AI accelerators and improve their performance at the same time

Copyright © Infineon Technologies AG 2022. All rights reserved.

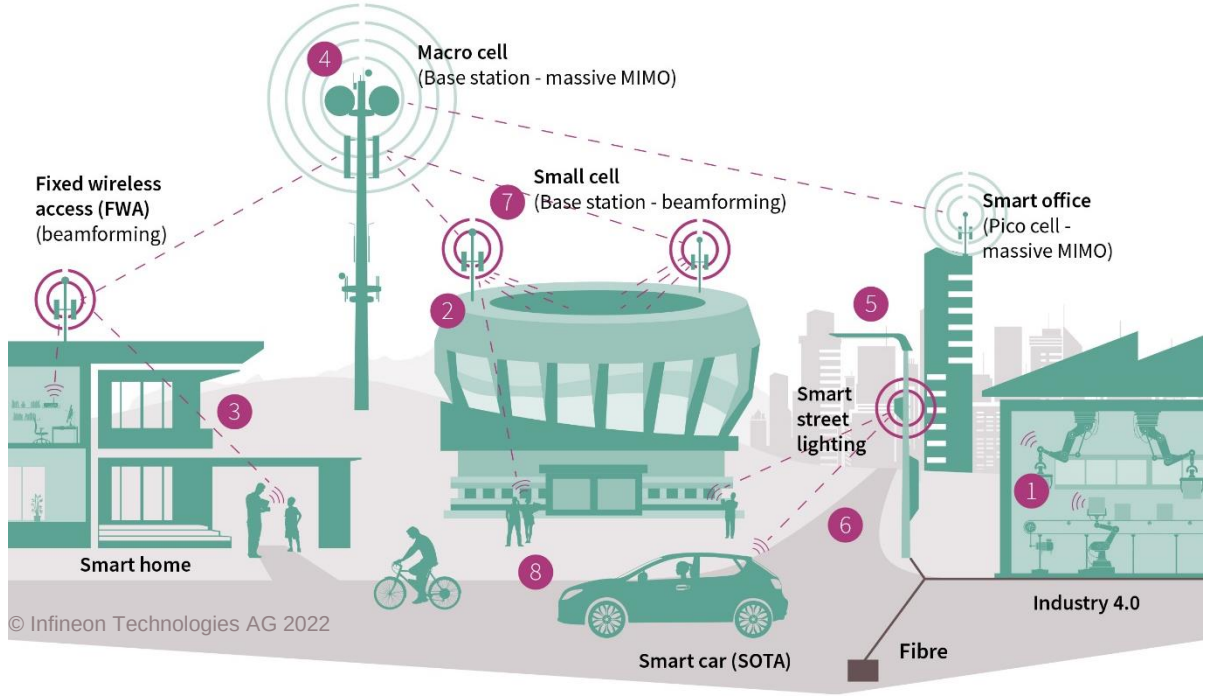
Going to the extreme Edge

- State of art AI is too big for some applications

- Extreme Edge / tinyML advantages:
  - Reduction of the amount of send data leading to the saving of energy
  - Improved real time capabilities
  - Add functionality to devices in an efficient manner

- Further reduction of the size of algorithms and AI accelerator hardware are necessary

- tinyML applications:
  - Basis for an efficient IoT
  - Smart sensors: key-word spotting
  - Hardware monitoring, data filtering
  - And more …



Copyright © Infineon Technologies AG 2022. All rights reserved.

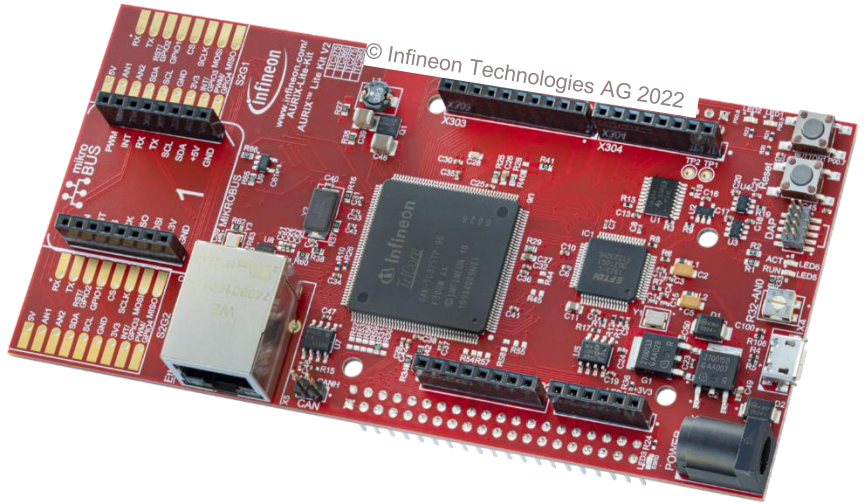
Explainable AI and Hardware for Trustworthy and Safety Critical AI

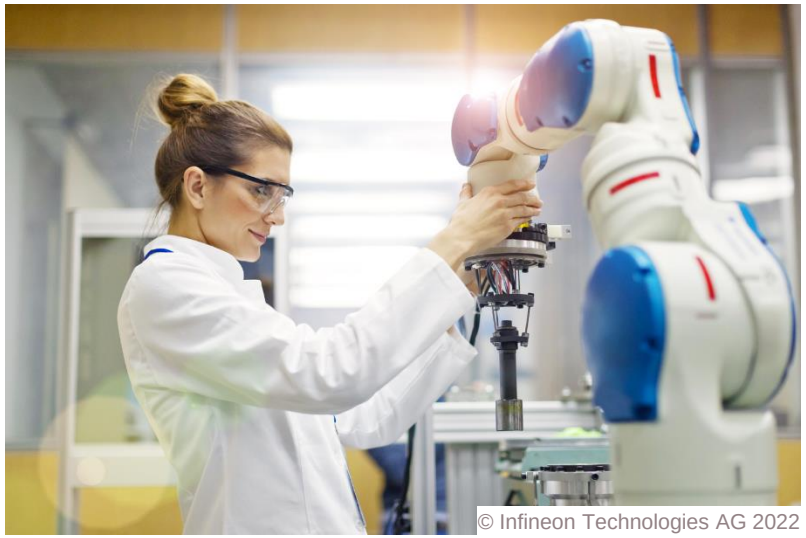
- Motivation:
  - Trust is an important aspect for the adaptation of AI
  - Understanding of how neuronal networks process data is limited

- Methods for interpreting and explaining AI are important for the certification, maintenance, safety and trust of AI-based products

- Specialized hardware is necessary for running safety critical AI similar to IFAG AURIX™ that is used for safety critical general software execution

- Only with such methods and hardware, AI can be integrated into safety sensitive applications such as robots or medical devices

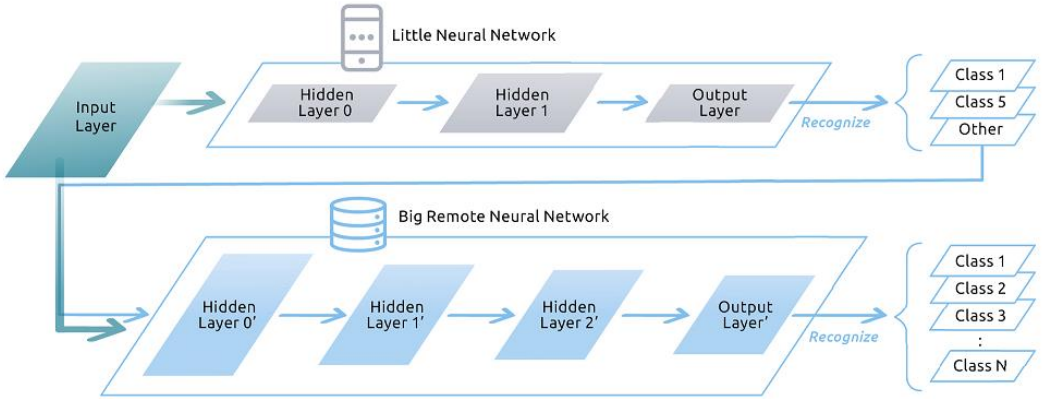




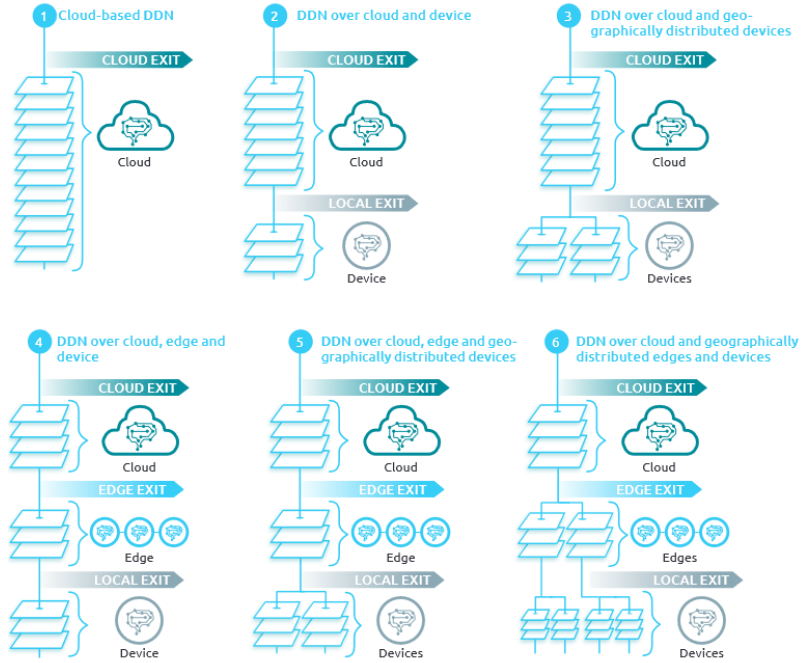
Copyright © Infineon Technologies AG 2022. All rights reserved.

Distributed AI for complex Applications

- Current focus is on single model solutions for applications
- Challenge:
  - Complex applications such as self-driving cars or multi purpose robots cannot be solved by one model
- Methods for efficiently combining different types of AI models at the edge and in the cloud into one system
- These systems require approaches to efficiently train, maintain and monitor the joint AI they are based on
- Edge hardware should include the accelerators for different kinds of AI to support these systems



Source: EPOSS



Source: EPOSS

Copyright © Infineon Technologies AG 2022. All rights reserved.

Other Challenges

- Tracing the history of an AI model will be essential for maintenance and determination of responsibility in case of damage producing events such as car accidents
  - History entails, e.g., data sets used for training, companies involved in the “manufacturing” of the model and updates
  - Blockchain could be a helpful tool in solving this challenge

- Edge AI is prone to attacks especially adversarial attacks
  - e.g. adversarial patches on traffic signs manipulating sign classification algorithms in a car

- Realization of analog AI to replace building blocks like state-machines in hardware to make devices more performant and efficient

- Reliability of edge AI in face of events that were not considered during training needs to be improved to enable the full potential of the IoT

- Many more challenges besides the ones shown need to be solved to solve the challenges of 21st century and to enable applications that will improve the lives of humans

Copyright © Infineon Technologies AG 2022. All rights reserved.