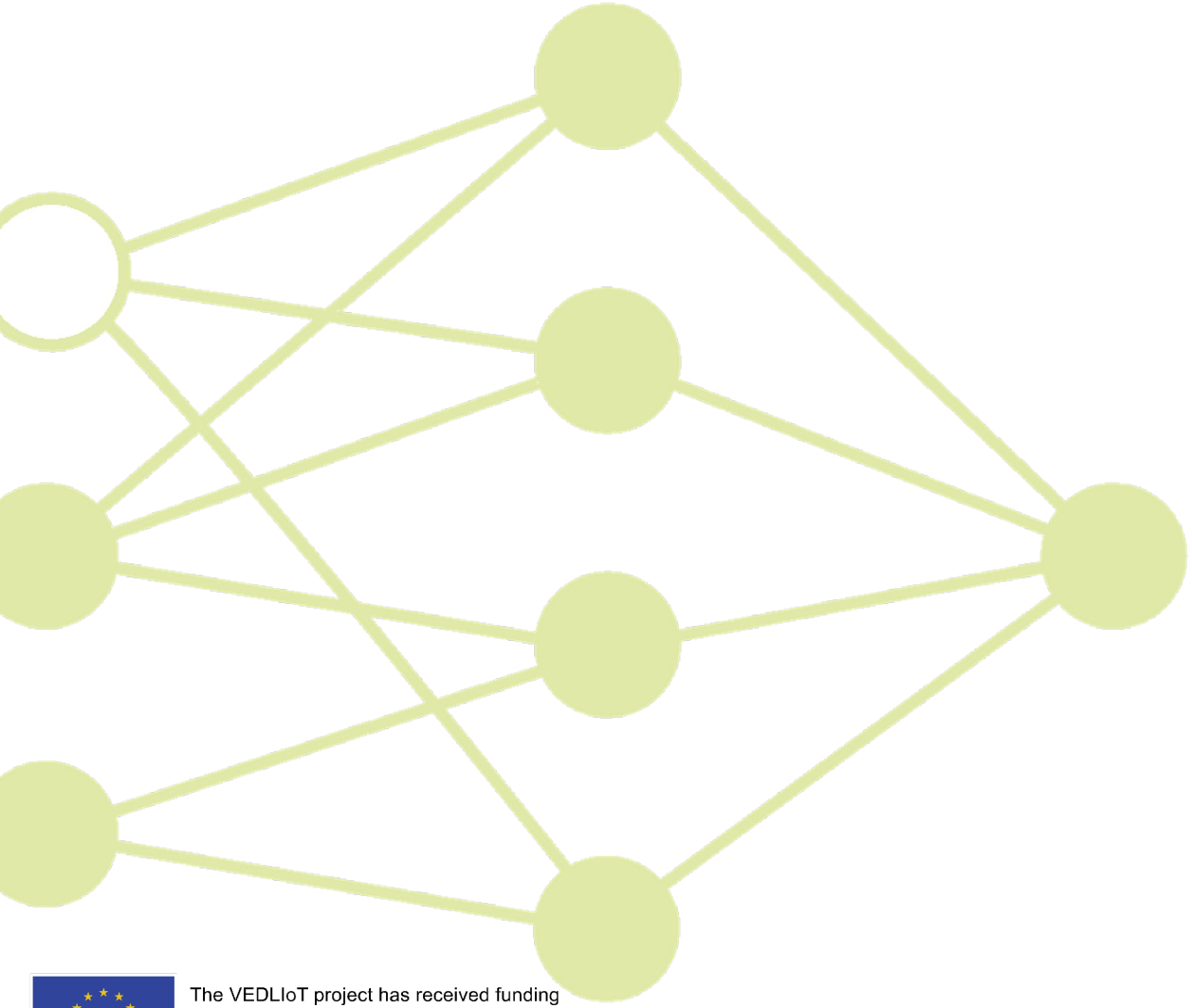


VEDLIoT

Very Efficient Deep Learning in IoT

NG-IoT Workshop on Standardization

VEDLIoT Overview and Standardization activities

Jens Hagemeyer
Bielefeld University

The VEDLIoT project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 957197

Very Efficient Deep Learning for IoT – VEDLIoT

VEDLIoT
Very Efficient Deep Learning in IoT

- **Platform**
  - Hardware: Scalable, heterogeneous, distributed
  - Accelerators: Efficiency boost by FPGA and ASIC technology
  - Toolchain: Optimizing Deep Learning for IoT
- **Use cases**
  - Industrial IoT
  - Automotive
  - Smart Home
- **Open call**
  - 10 projects covering a wide range of AIoT applications
  - Early use and evaluation of VEDLIoT technology

- **Call:** H2020-ICT2020-1
- **Topic:** ICT-56-2020 Next Generation Internet of Things
- **Duration:** 1. November 2020 – 31. Oktober 2023
- **Coordinator:** Bielefeld University (Germany)
- **Overall budget:** 7 996 646.25 €
- **Consortium:** 12 partners from 4 EU countries (Germany, Poland, Portugal and Sweden) and one associated country (Switzerland).

More info:

⇒ https://www.vedliot.eu/
⇒ https://twitter.com/VEDLIoT
⇒ https://www.linkedin.com/company/vedliot/

Big Picture

VEDLIoT
Very Efficient Deep Learning in IoT

Requirements

Smart Home

Industrial IoT

Automotive AI

Security & Safety

Modelling & Verification

Applications

Trusted Execution & Communication

Middleware

Toolchain

embedl

Emulation

RENODE

Benchmarking & Deployment

Kenning

Safety & Robustness

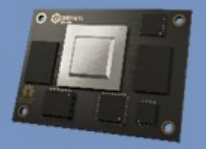
Microserver & Accelerators

Xilinx
Kria

Coral SoM

COM-HPC
Xilinx Zynq
UltraScale+

Jetson AGX
NVIDIA Xavier



RPi CM4
ARVSOM

SMARC
Xilinx Zynq
UltraScale+

Monitoring

Hardware Platforms

Embedded/
Far Edge

u.RECS

Near Edge

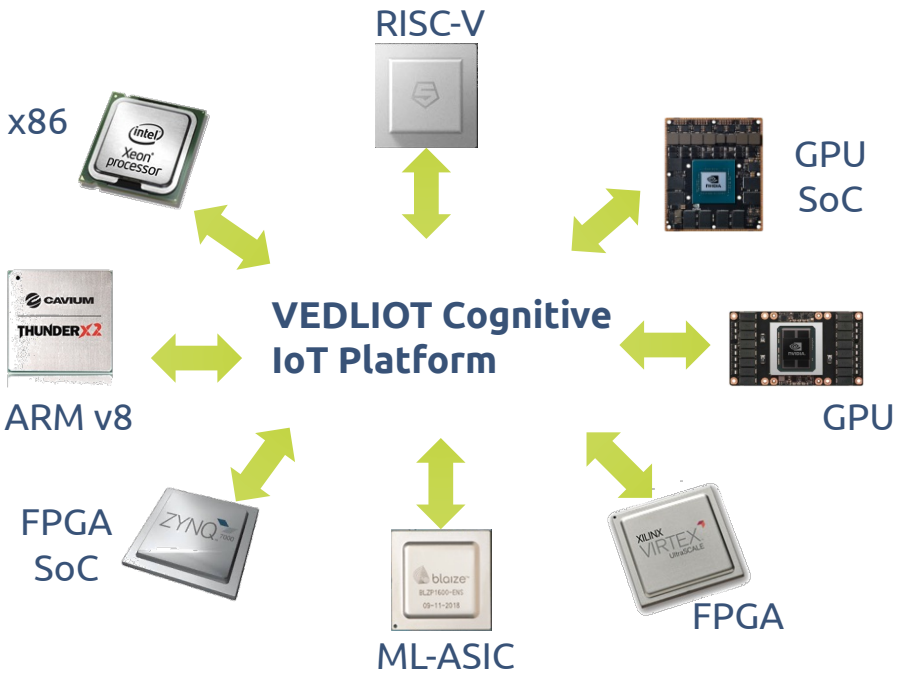
t.RECS

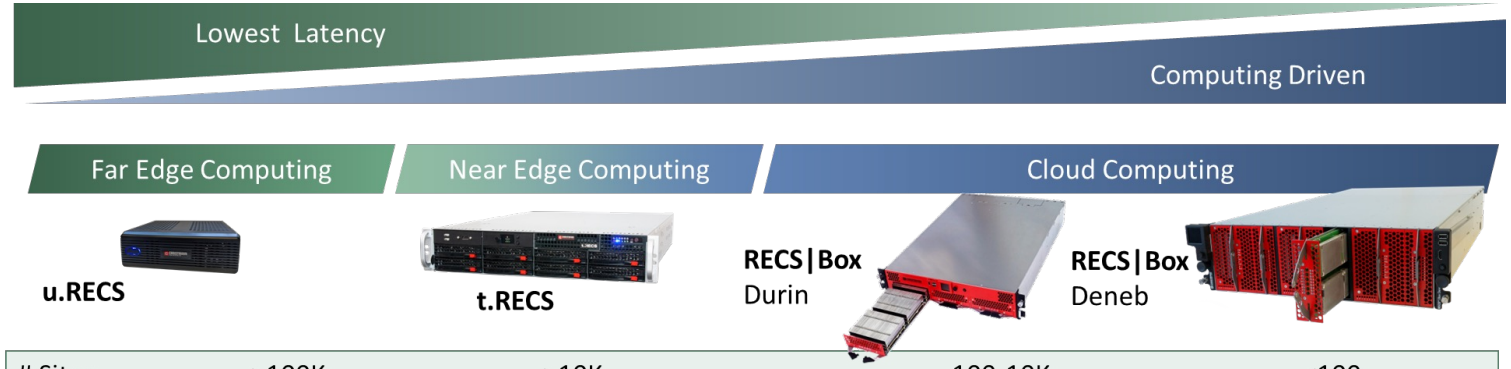
Cloud

RECS|Box

RISC-V extensions

VEDLIoT Hardware Platform

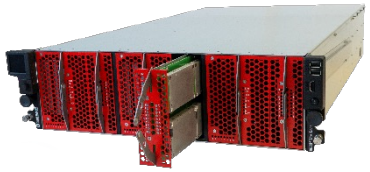




# Sites	>100K	>10K	100-10K	<100
Footprint	Custom	Compact (1RU)	Medium (2RU)	Large (3RU)
Power Budget	<30 W	< 500 W	500 W – 2 KW	> 2 KW
# Microserver	max 2	up to 3	up to 48	up to 144

- Heterogeneous, modular, scalable microserver system
- Supporting the full spectrum of IoT **from embedded over the edge towards the cloud**
- Different technology concepts for improving
  - Performance
  - Cost-effectiveness
  - Maintainability
  - Reliability
  - Energy-Efficiency
  - Safety

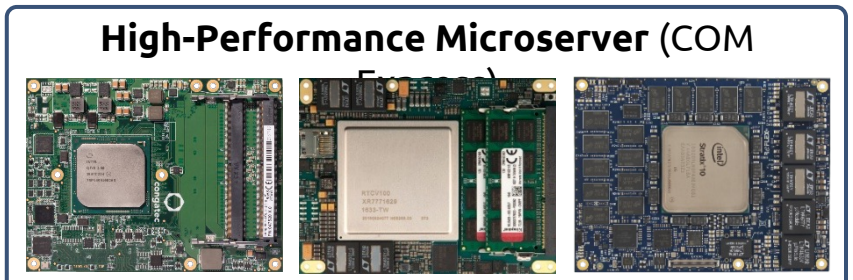
RECS Architecture – RECS|BOX



VEDLIoT
Very Efficient Deep Learning in IoT

High-Performance Carrier
(up to 3 microservers)

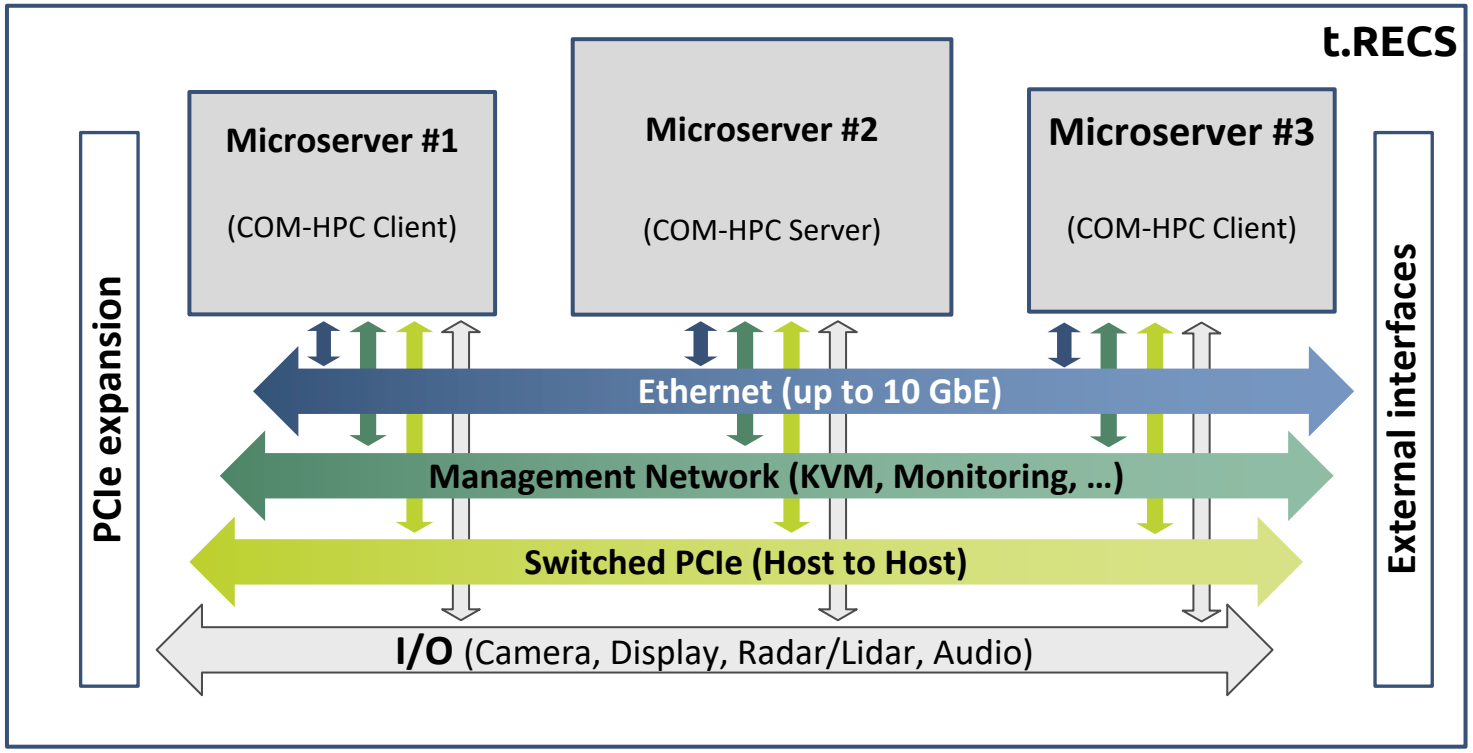
Low-Power Carrier
(up to 16 microservers)



RECS Architecture – t.RECS

t.RECS Edge Server

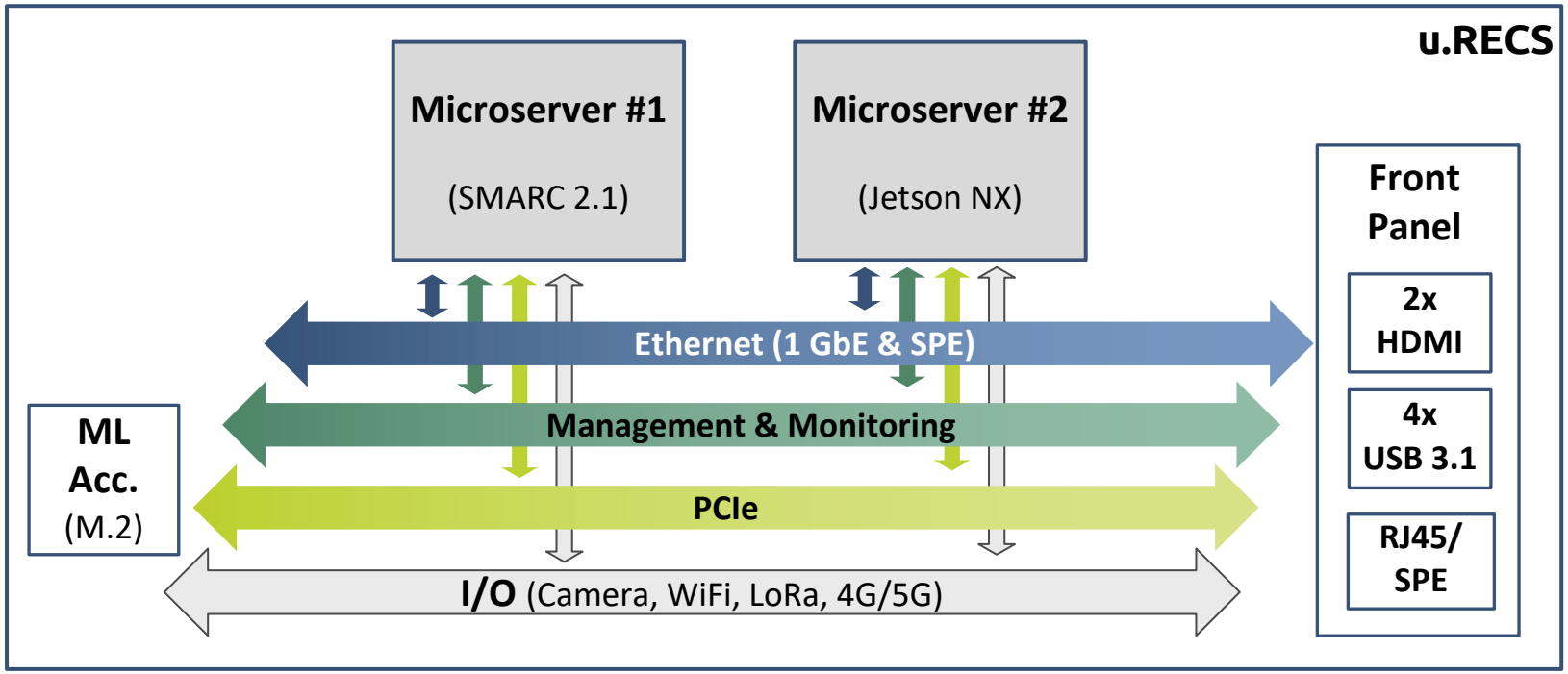
- Optimized platform for local / edge applications
- Provide interfaces for
  - Video
  - Camera
  - Peripheral input (USB)
- Combine FPGA and GPU acceleration
- Compact dimensions
  1 RU, E-ATX form factor
  (2 RU/ 3 RU for special cases)



RECS Architecture – u.RECS

u.RECS AIoT Server

- Supports ML acceleration
  - FPGA
  - ASIC
- Communication interfaces
  - Wired (CAN, Ethernet, CSI)
  - Wireless (WLAN, LoRa, 5G)
- Sensors
  - Camera
  - Environment (Temp./Hum.)
  - Housekeeping
- Embedded Device
  (~ 20x20x6 cm)



Microserver overview

VEDLIoT
Very Efficient Deep Learning in IoT

RECS|BoX

t.RECS

u.RECS

CPU

COM Express
Intel Core i7
8th Gen

Apalis
Exynos (2xARM
Cortex-A15)

COM Express
ARM v8 Server SoC
Hi1616

COM Express
AMD EPYC
3451

COM Express
AMD Ryzen
V1807B

COM-HPC
client size A
Intel Core i7
11th Gen

FPGA SoC

COM Express
Xilinx
Zynq
7045

Apalis
Xilinx Zynq 7020

COM Express
Intel Stratix 10

COM-HPC
client size B
Xilinx Zynq
UltraScale+

COM-HPC
server size D
Intel Agilex

COM-HPC
client size B
Xilinx Versal

GPU SoC

Jetson TX2
NVIDIA
Tegra X2

Jetson nano
NVIDIA
Xavier NX

Jetson AGX
NVIDIA Xavier

COM-HPC Size B
NVIDIA Xavier
AGX

ML SoC

SMARC
Coherent
Logix
HX40416

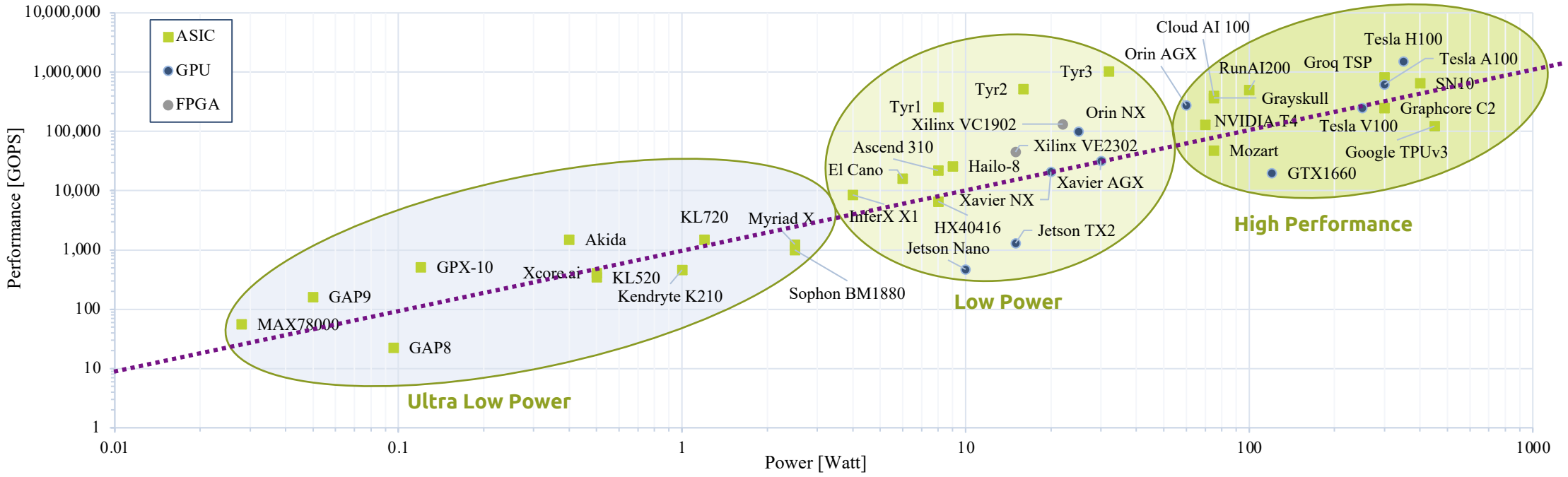
SMARC
Coral SoM

SMARC
Xilinx Zynq
UltraScale

Peak Performance of DL Accelerators

VEDLIoT
Very Efficient Deep Learning in IoT

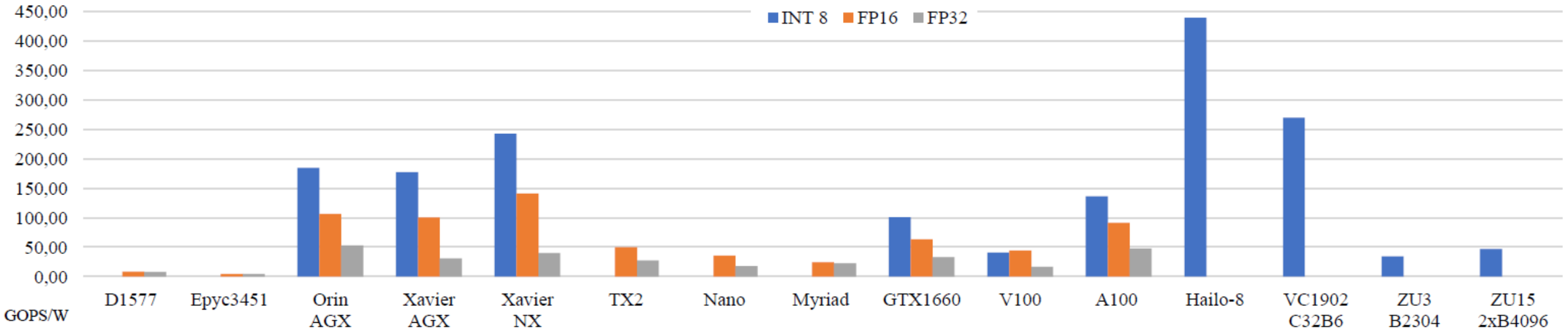
- Peak performance values of specialized accelerators, provided by the vendors (precisions varying from INT8 to FP32)



Average efficiency at 1000 GOPS /W

Yolo v4 accelerator performance

- Performance of Yolo v4 for different hardware platform has been evaluated
- Performance measurement for other networks (Resnet, EfficientNet) available as well



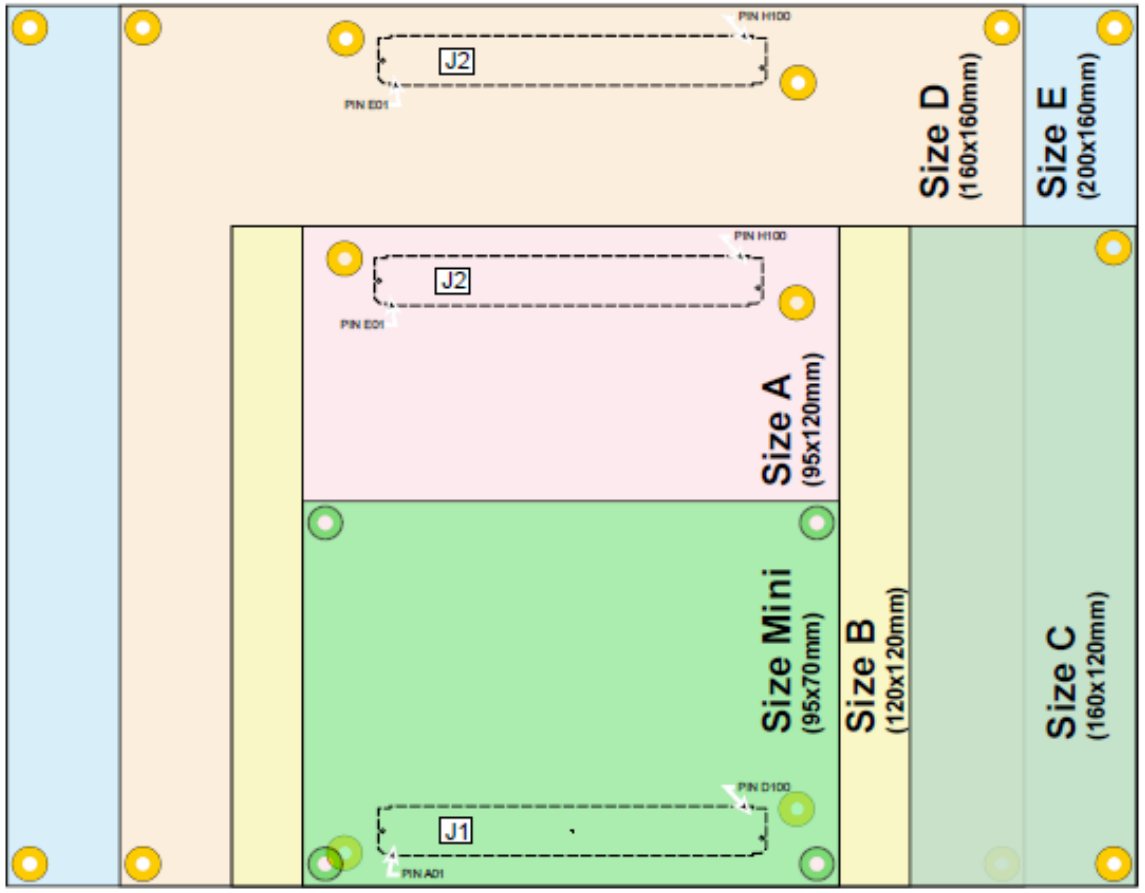
Microserver Standardization – COM-HPC

VEDLIoT
Very Efficient Deep Learning in IoT

COM-HPC™

- Large, open consortium
- Specification final and released
- Driven by industry requirements

COM-HPC Client	COM-HPC Server
49x PCIe	65x PCIe
2x MIPI-CSI	
2x 25GbE KR	8x 25GbE KR
3x DDI	
2x BaseT (up to 10 Gb)	
2x SoundWire, I²S	BaseT (up to 10 Gb)
4x USB4	2x USB4
	2x USB3.2
4x USB2.0	4x USB2.0
2x SATA	2x SATA
eSPI, 2x SPI, SMB	eSPI, 2x SPI, SMB
2x I²C, 2x UART	2x I2C, 2x UART
12x GPIO	12x GPIO



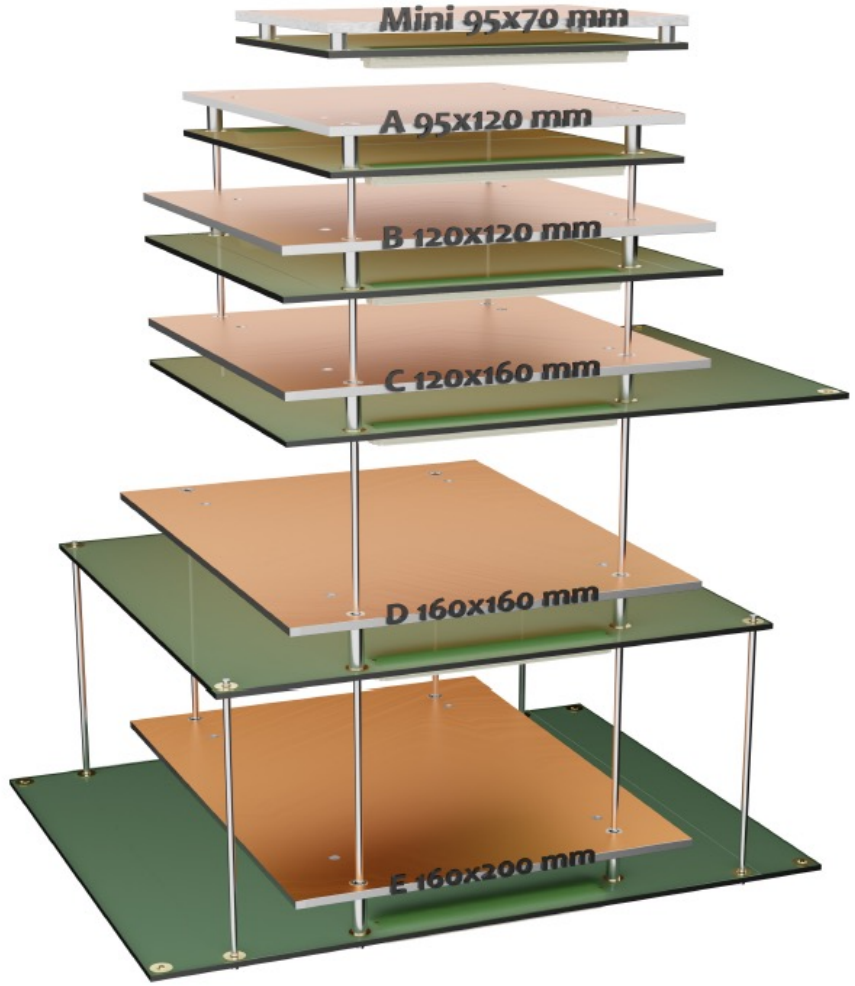
Microserver Standardization – COM-HPC

VEDLIoT
Very Efficient Deep Learning in IoT

COM-HPC™

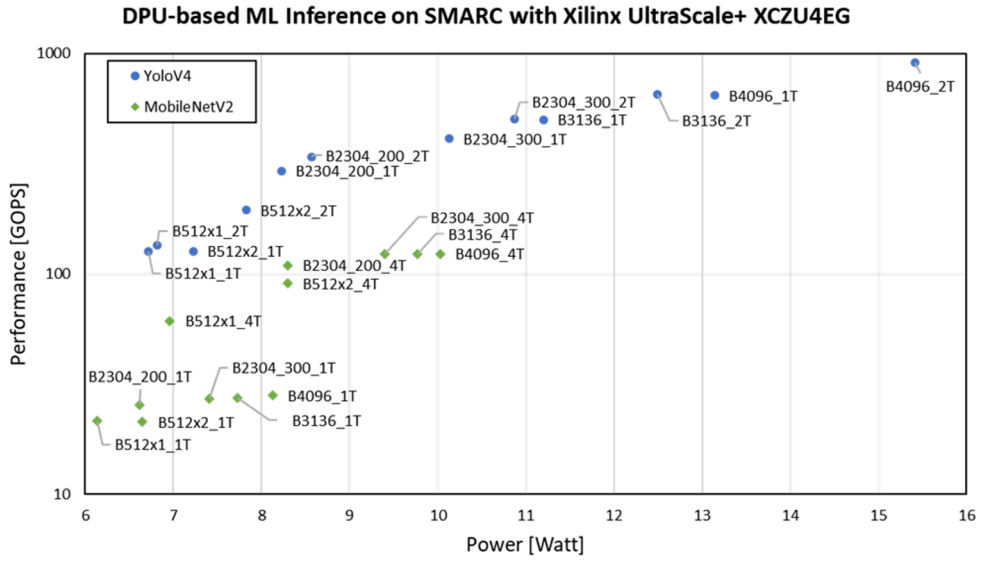
- Large, open consortium
- Specification final and released
- Driven by industry requirements

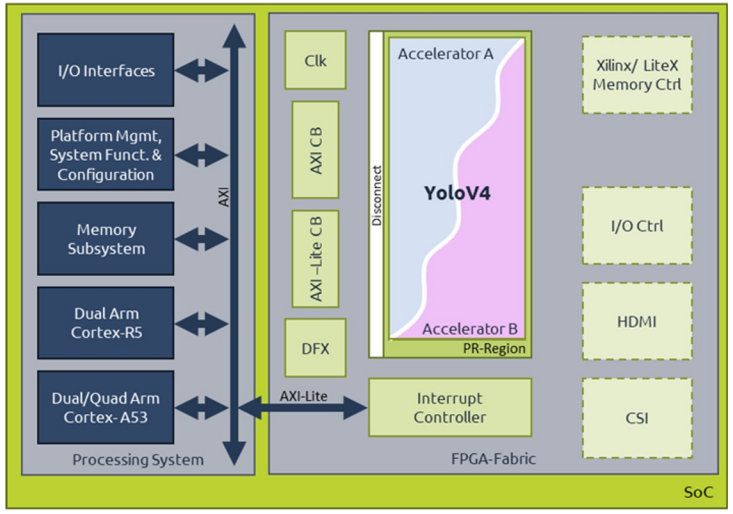
COM-HPC Client	COM-HPC Server
49x PCIe	65x PCIe
2x MIPI-CSI	
2x 25GbE KR	8x 25GbE KR
3x DDI	
2x BaseT (up to 10 Gb)	
2x SoundWire, I²S	BaseT (up to 10 Gb)
4x USB4	2x USB4
	2x USB3.2
4x USB2.0	4x USB2.0
2x SATA	2x SATA
eSPI, 2x SPI, SMB	eSPI, 2x SPI, SMB
2x I²C, 2x UART	2x I2C, 2x UART
12x GPIO	12x GPIO



Reconfigurable DL accelerators

- VEDLIoT accelerators support a large variety of reconfigurable architectures
  - From small embedded FPGAs to large ACAPs
- Large design space for FPGA-based accelerators
- Dynamic hardware reconfiguration
  - Adapt to changing requirements at run-time
  - Change characteristics of DL-accelerator
  - Trade-off between
    power and performance, power and accuracy, etc.
- Inference and training on FPGA
  - Supports quantization from int8 to float32
  - DL and Deep Reinforcement Learning





DL accelerator co-design

VEDLIoT
Very Efficient Deep Learning in IoT

Monolithic design

- One engine computes all the core layers
- E.g. TPU

SEML

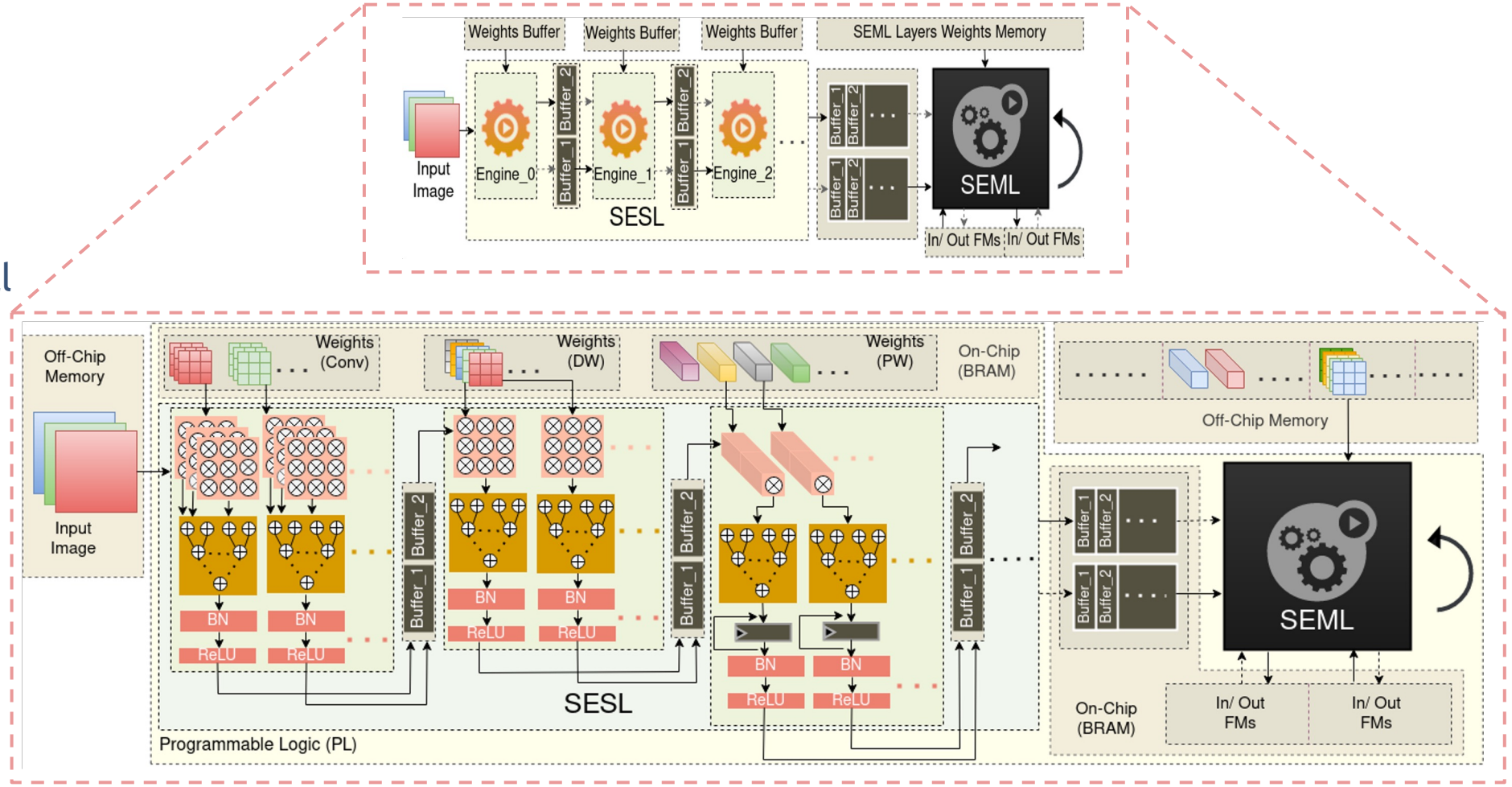
- One engine computes all layers of the same type
- PW engine, DW engine

SESL

- One engine per layer
- E.g. FINN

FiBHA

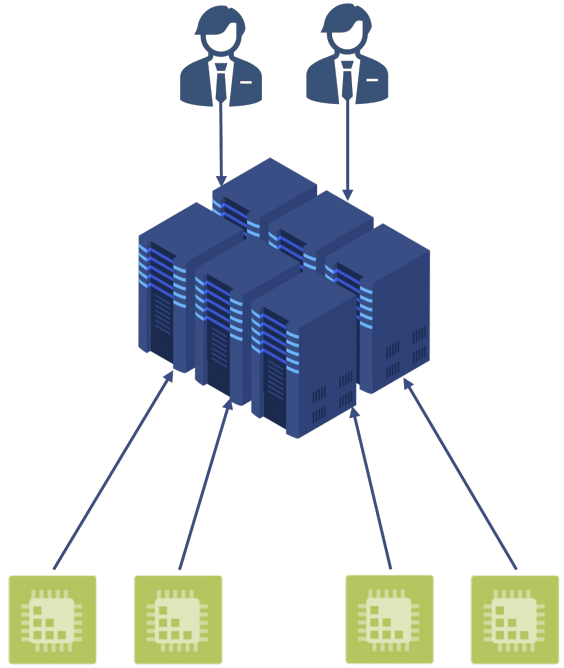
- SESL + SEML

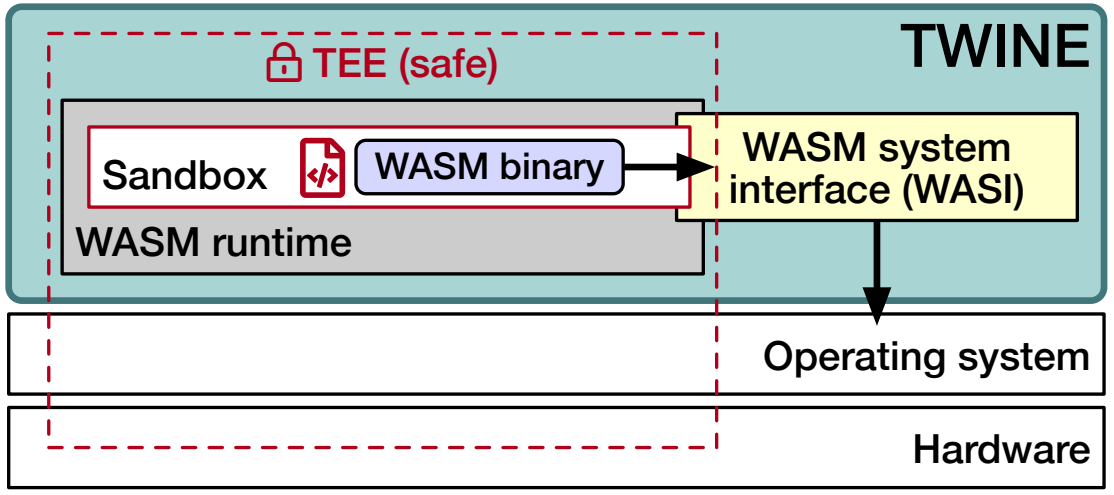


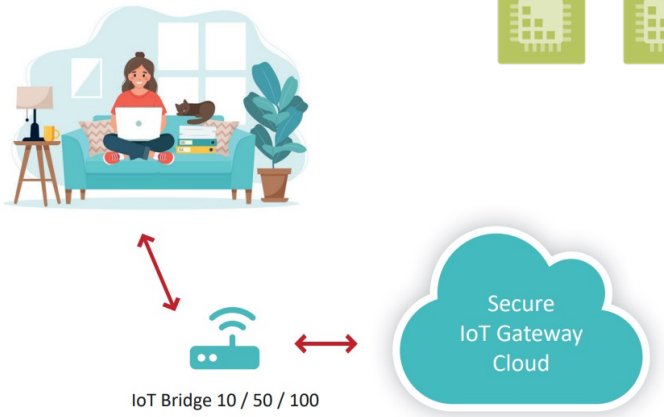
"FiBHA: Fixed Budget Hybrid CNN Accelerator", Fareed Qararyah, Muhammad Waqar Azhar, Pedro Trancoso, IEEE 34th International Symposium on Computer Architecture and High-Performance Computing (SBAC-PAD 2022), Bordeaux, France, November 2–5 2022

Security

- Common environment for running distributed applications
  - WebAssembly runtime + Trusted Execution Environment
  - Security for edge (and cloud) devices
- Advances on attestation
  - Better support for edge devices
  - Distributed (Byzantine fault-tolerant) attestation and configuration service
- Secure IoT Gateway







Safety and Robustness

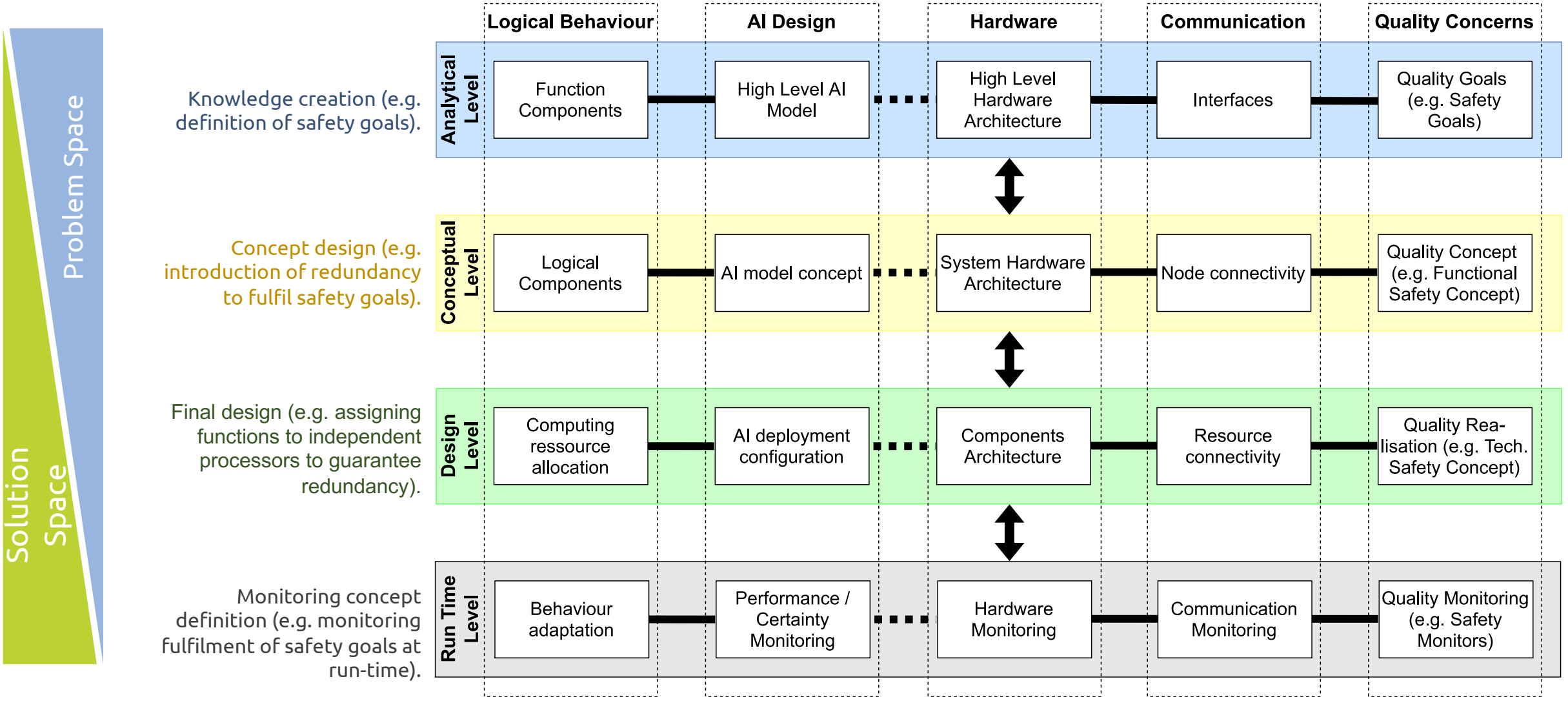
VEDLIoT
Very Efficient Deep Learning in IoT

Simulation platform for ML accelerators

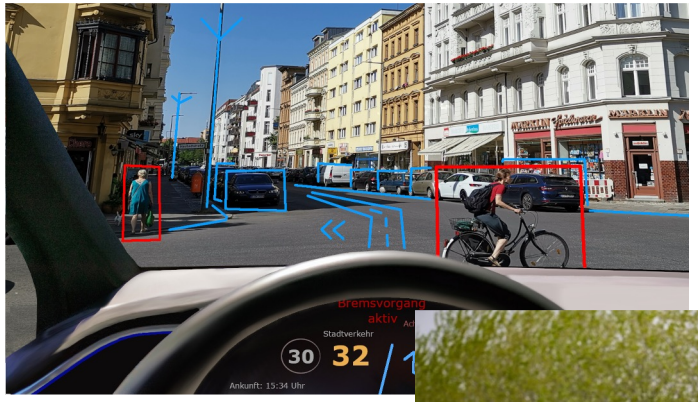
- RISC-V SoCs and Custom Function Units
- Improve test and verification
- Co-simulate Verilog blocks
- Used in Google's CFU Playground
- Continuous integration based in Gitlab and Google Cloud Platform

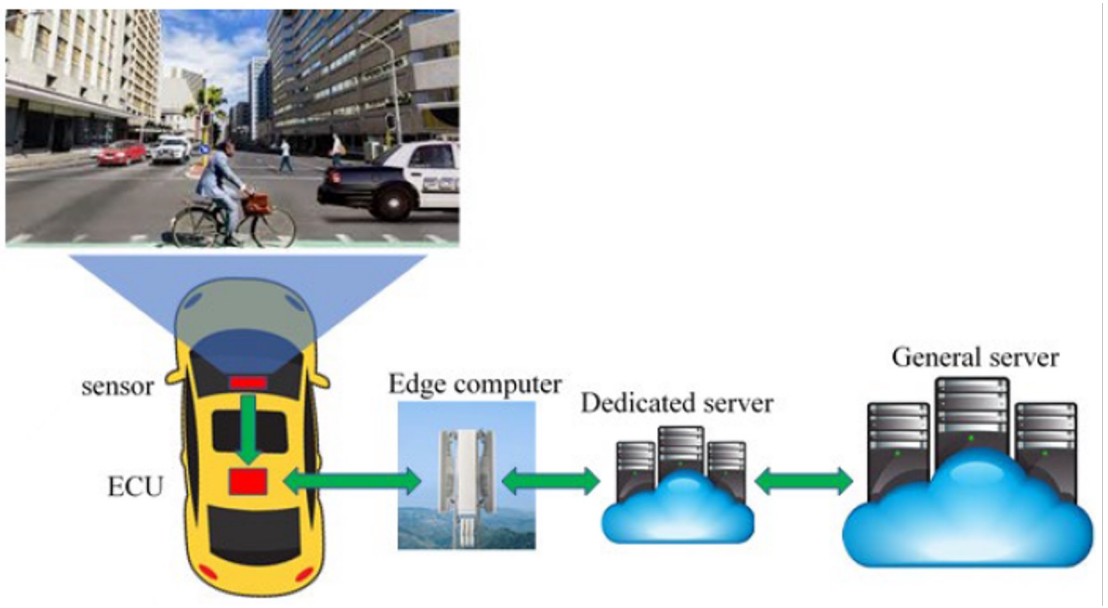
A compositional architecture framework for AIoT

VEDLIoT
Very Efficient Deep Learning in IoT



Use case: Automotive





- Focus on collision detection/avoidance scenario
- Improve performance/cost ratio – AI processing hardware distributed over the entire chain

Challenge: Distribution of work

Use case: Industrial IoT – drive condition classification

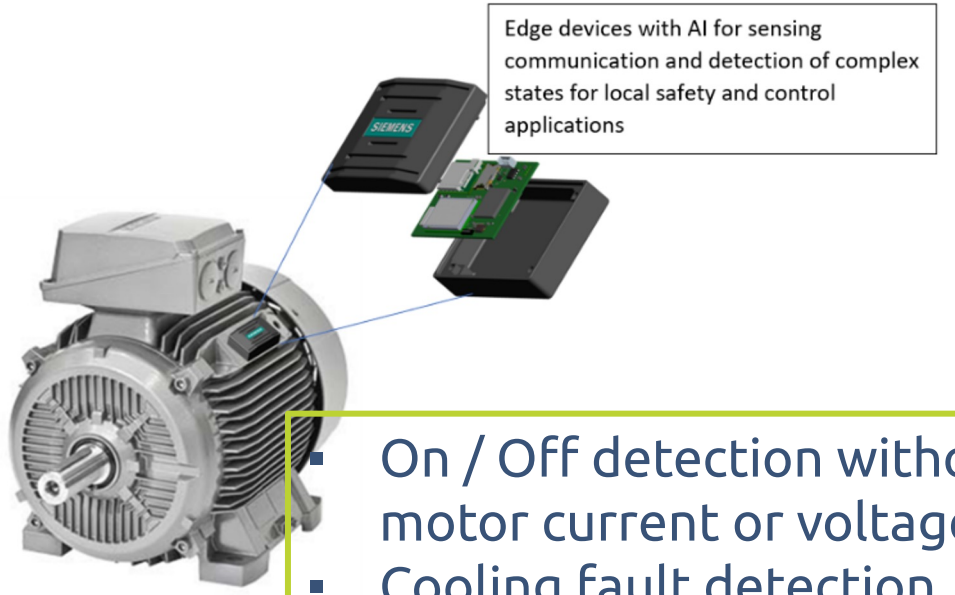
VEDLIoT
Very Efficient Deep Learning in IoT

- Control applications need DL-based condition classification
  - On the edge device for low power consumption
  - Suggestions for control and maintenance

Challenge: Low-power / Efficiency

- DL methods on all communication layers
  - DL in a distributed architecture
  - Dynamically configured systems

- Sensored testbench with 2 motors
  - Acceleration, Magnetic field, Temperature, IR-Cam (temperature), Current-Sensors, Torque

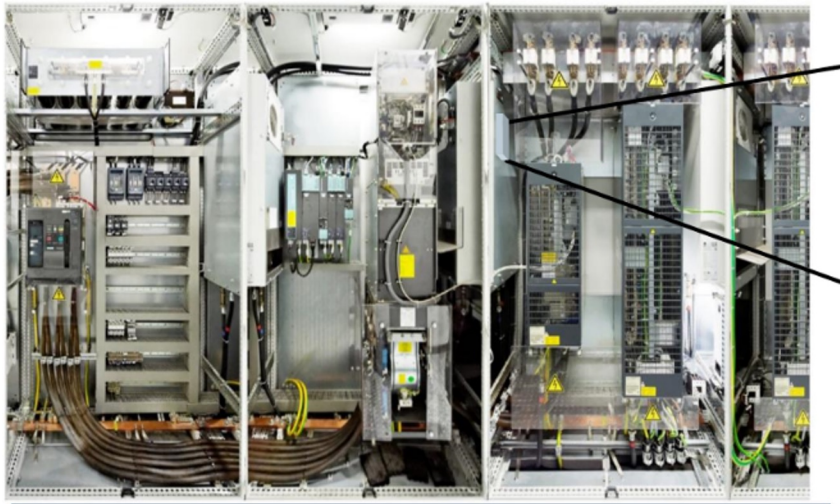


- On / Off detection without motor current or voltage
- Cooling fault detection
- Bearing fault detection

Use case: Industrial IoT – Arc detection

VEDLIoT
Very Efficient Deep Learning in IoT

- AI based pattern recognition for different local sensor data
  - current, magnetic field, vibration, temperature, low resolution infrared picture
- Safety critical nature
  - response time should be <10ms
  - AI based or AI supported decision made by the sensor node itself or by a local part of the sensor network



Combining the information from the IR-Camera and the magnetic field sensor to localize electric faults in power cabinets by deep learning methods

Specifications:
- Industrial temperature range (-20°C ... +85°C)
- Industrial batteries (rechargeable for ID-Tag)
- IP65 protection
- RoHS and IEC 61850-3 complaint
- Pre-certified wireless transceivers
- Target price: 100€ (ID-Tag)
- SIM on Chip*

Challenge: Accuracy

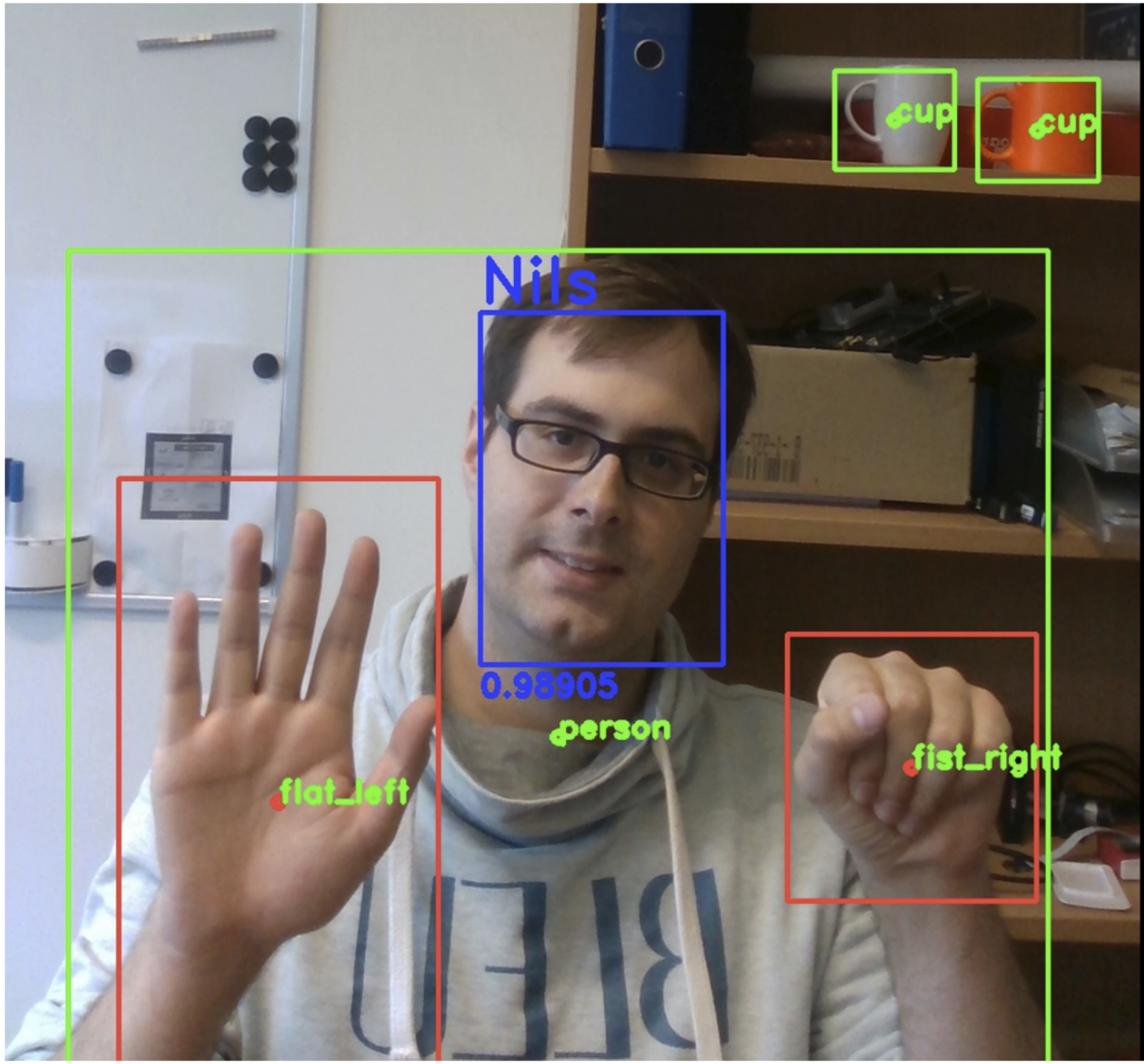
Use case: Smart Mirror – Neural Networks

VEDLIoT
Very Efficient Deep Learning in IoT

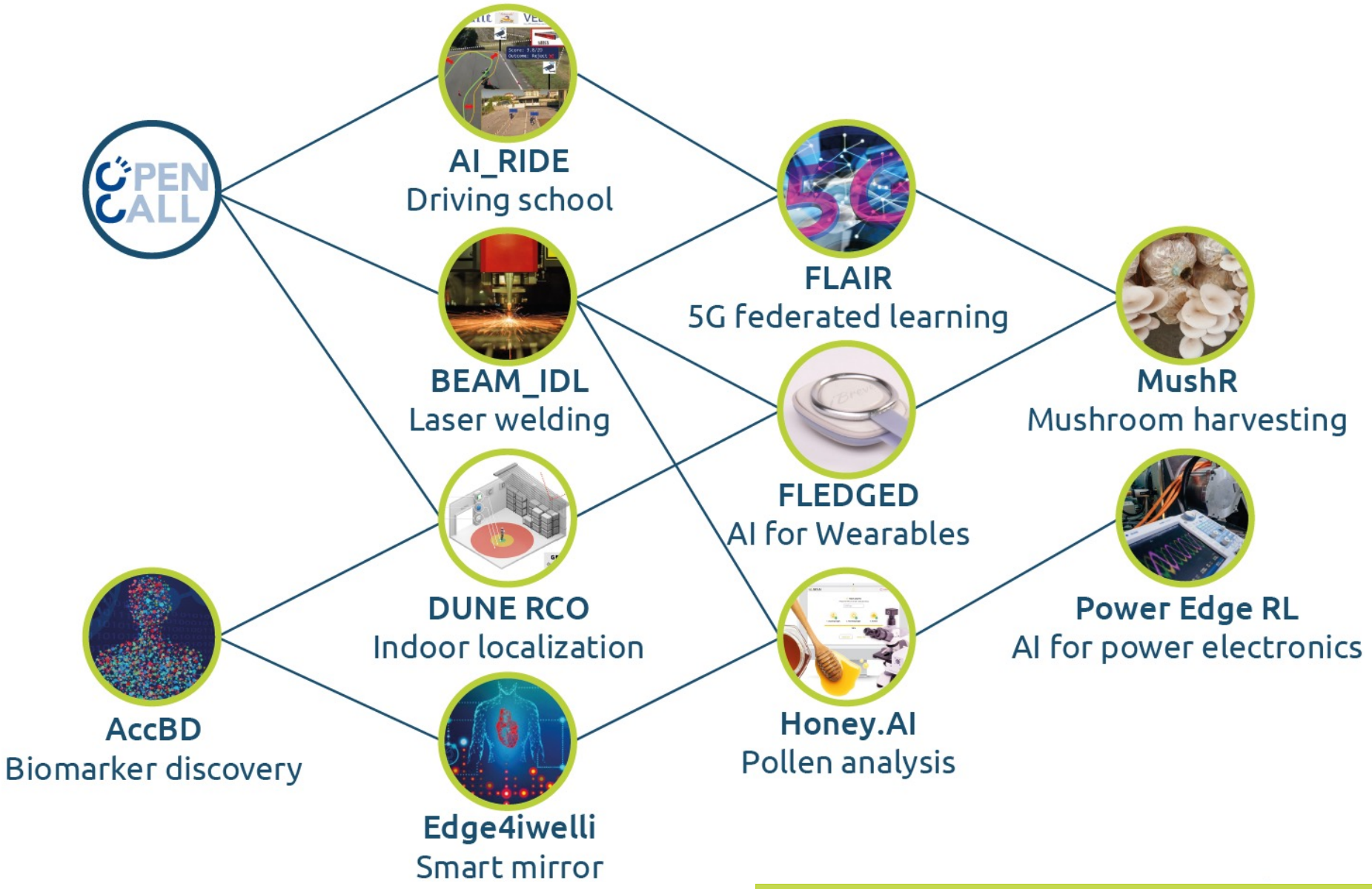
- Face recognition
  - Mobilenet SSD trained on WIDERFACE dataset
- Object detection
  - YoloV3, Efficient-Net, yoloV4-tiny
- Gesture detection
  - YoloV4-tiny with 3 Yolo layers (usually: 2 layers)
- Speech recognition
  - Mozilla DeepSpeech

- AI Art: Style-Gan trained on works of arts
- Collect usage data in situation memory

Challenge:
Data privacy,
Efficiency



Use case: Open calls



Summary – Standardization in VEDLIoT

- Hardware/microserver form factors
  - Active contribution to PICMG Standards COM-HPC and COM Express (https://www.picmg.org/openstandards/com-hpc)
- Several Open Source contributions to large projects (https://vedliot.eu/open-source-software)
  - Renode + Kenning – Emulator and Simulator for distributed IoT, Verilator support
  - Memory Protection for RISC-V: RISC-V PMP
  - TEEs support for WebAssembly: Integation for Trustzone (ARM) and SGX (Intel) into WebAssembly
- Recommendations: Design framework IoT and AI
  - Compositional architecture framework for AIoT developled within VEDLIoT
  - Can help system design to comply with regulatory constraints (e.g. EU AI Act)

VEDLIoT
Very Efficient Deep Learning in IoT

Thank you for your attention.

The VEDLIoT project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 957197

UNIVERSITÄT BIELEFELD
embedl
CHALMERS UNIVERSITY OF TECHNOLOGY
SIEMENS
CHRISTMANN INFORMATIONSTECHNIK + MEDIEN
unine UNIVERSITÉ DE NEUCHÂTEL
UNIVERSITÄT OSNABRÜCK
veoneer
UNIVERSITY OF GOTHENBURG
RI.SE
Ciências ULisboa
FCiências.ID ASSOCIAÇÃO PARA A INVESTIGAÇÃO E DESENVOLVIMENTO DE CIÊNCIAS
antmicro

Contact
Jens Hagemeyer, Carola Haumann
Bielefeld University, Germany
chaumann@cor-lab.uni-bielefeld.de
jhagemey@cit-ec.uni-bielefeld.de