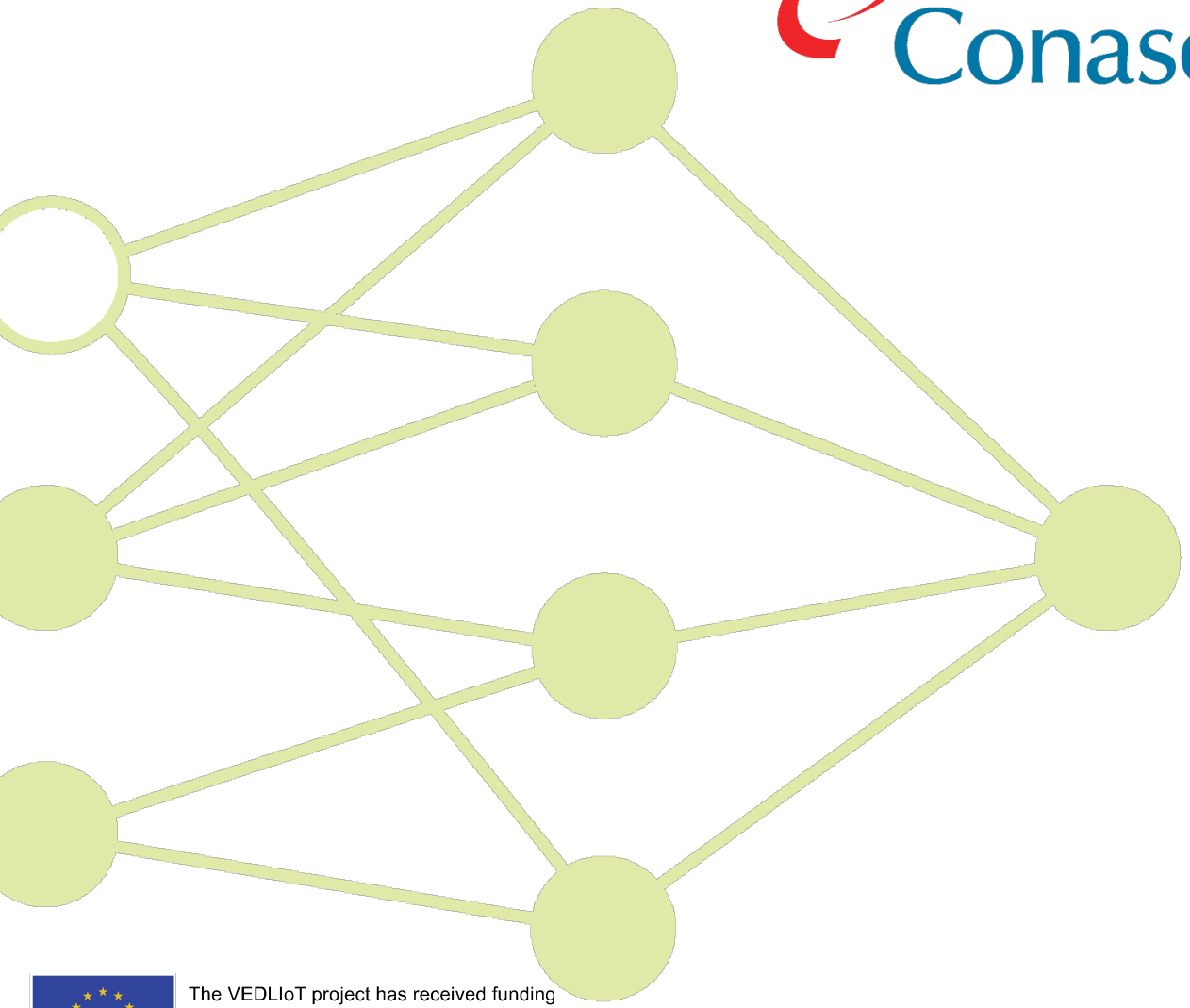


Conasense

VEDLIoT
Very Efficient Deep Learning in IoT

Very Efficient Deep-Learning in IoT (VEDLIoT) – Overview

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

- **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**
  - At project mid-term
  - 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/

Partners

VEDLIoT
Very Efficient Deep Learning in IoT

- Bielefeld University (UNIBI) - Coordinator
- Christmann (CHR)
- University of Osnabrück (UOS)
- Siemens (SIEMENS)
- University of Neuchâtel (UNINE)
- University of Lisbon (FC.ID)
- Chalmers (CHALMERS)
- University of Gothenburg (UGOT)
- RISE (RISE)
- EmbeDL (EMBEDL)
- Veoneer (VEONEER)
- Antmicro (ANT)

Big Picture

VEDLIoT
Very Efficient Deep Learning in IoT

Requirements

Smart Home

Industrial IoT

Automotive AI

Security & Safety

Modelling & Verification

Safety & Robustness

Applications

Open Call

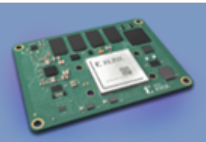
Middleware

Toolchain
embedl

Emulation
RENODE

Benchmarking & Deployment
Kenning

Microserver & Accelerators

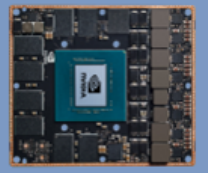


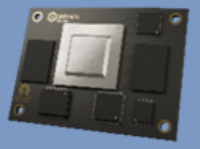
Xilinx Kria

Coral SoM

COM-HPC
Xilinx Zynq UltraScale+

Jetson AGX
NVIDIA Xavier





RPi CM4
ARVSOM

SMARC
Xilinx Zynq UltraScale

Hardware Plattforms

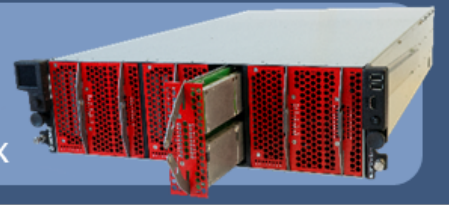
Embedded/ Far Edge

uRECS

Near Edge

t.RECS

Cloud



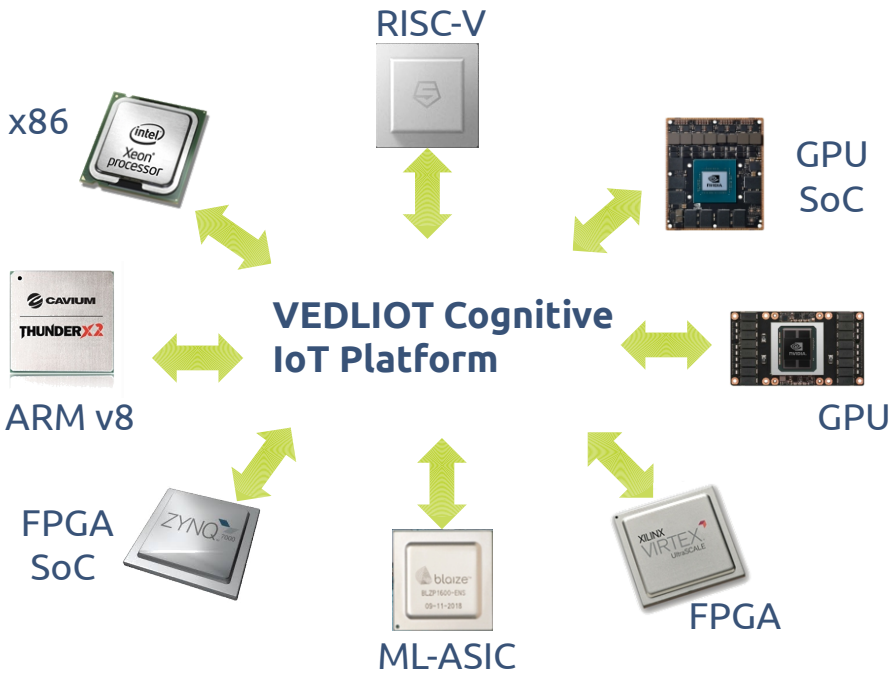
RECS|Box

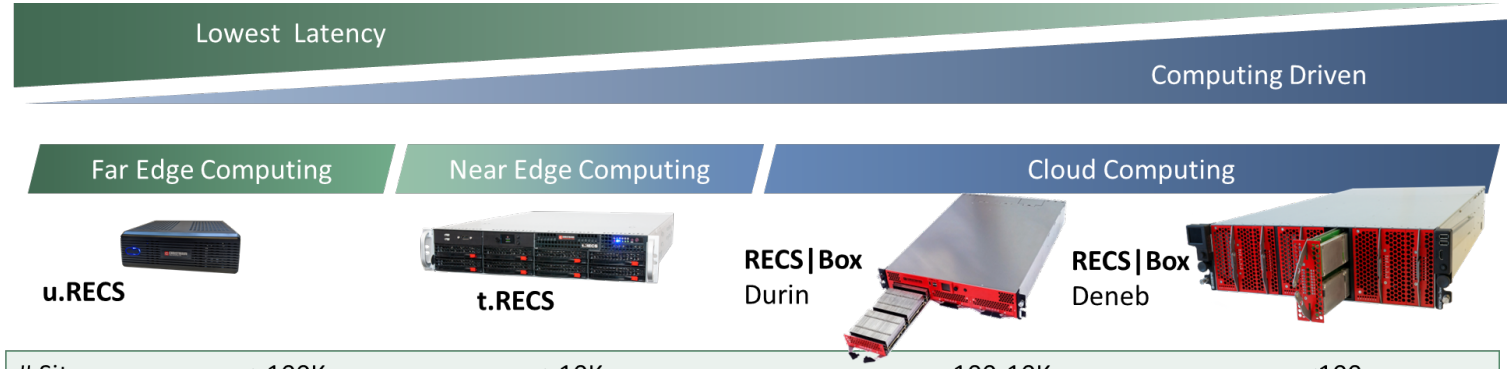
Trusted Execution & Communication

Monitoring

RISC-V extensions

VEDLIoT Hardware Platform

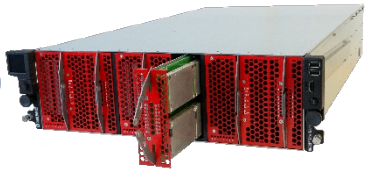




# Sites	>100K	>10K	100-10K	<100
Footprint	Custom	Compact (1RU)	Medium (2RU)	Large (3RU)
Power Budget	<30 W	< 500 W	500 W – 2 KW	> 2 KW
# Microserver	max 2	up to 3	up to 48	up to 144

- 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)

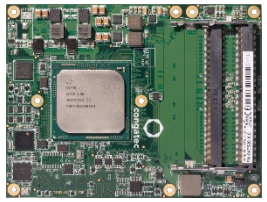


VEDLIoT
Very Efficient Deep Learning in IoT

High-Performance Carrier
(up to 3 microservers)

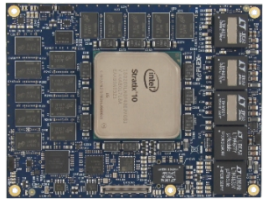
Low-Power Carrier
(up to 16 microservers)

High-Performance Microserver (COM Express)



x86

ARM v8



FPGA SoC

Low-Power Microserver (Apalis/Jetson)

FPGA SoC

ARM Soc

GPU SoC

RECS Server Backplane (up to 15 Carriers)

Carrier (High Performance)

Microserver (High Performance) #1 #2 #3

Carrier (Low Power)

Microserver (Low Power) #1 #2 #3 … **Microserver** (Low Power) #16

Carrier (PCIe Expansion)

e.g. GPU-Accelerator

Compute Network (up to 40 GbE)

Management Network (KVM, Monitoring, …)

High-Speed Low-Latency Network (PCIe, High-Speed Serial)

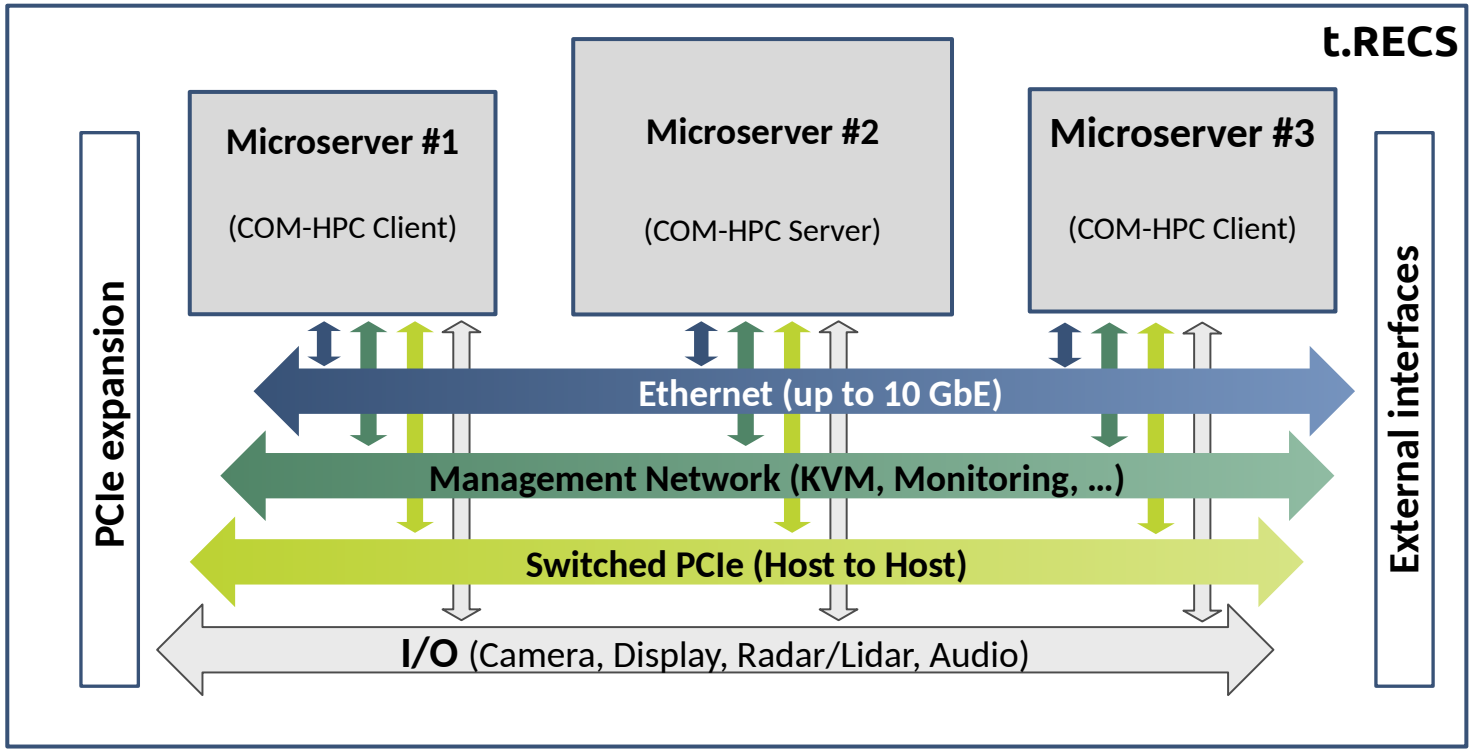
Ext. Connectors
- QSFP+
- HDMI/USB
- RJ45
- iPass+ HD

RECS Architecture (t.RECS)

VEDLIoT
Very Efficient Deep Learning in IoT

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)

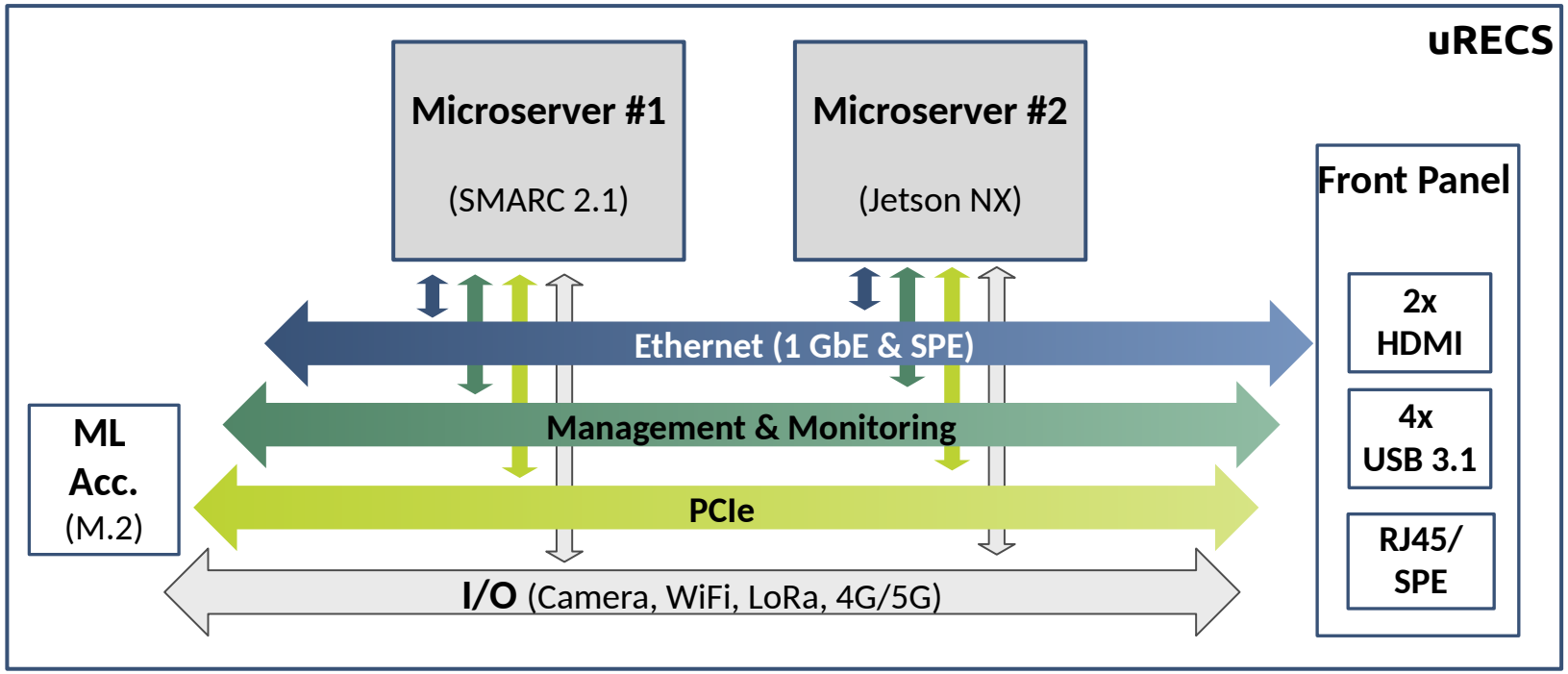


RECS Architecture (u.RECS)

VEDLIoT
Very Efficient Deep Learning in IoT

u.RECS AIoT Server

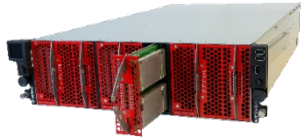
- Supports ML acceleration
  - FPGA
  - ASIC
- Communication interfaces
  - Wired (CAN, Ethernet, CSI)
  - Wireless (WLAN, LoRa, 5G)
- Sensors
  - Camera
  - Environment (Temp./Hum.)
  - Housekeeping
- Embedded Device
  (~ 20x20x6 cm)



Microserver overview

VEDLIoT
Very Efficient Deep Learning in IoT

RECS|BoX

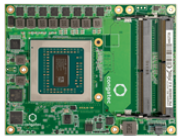


CPU

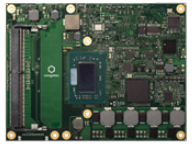
COM Express
Intel Core i7
8th Gen

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

t.RECS

FPGA SoC

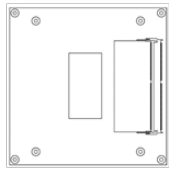
COM Express
Xilinx
Zynq
7045

Apalis
Xilinx Zynq 7020

COM Express
Intel Stratix 10

COM-HPC
client size B
Xilinx Zynq
UltraScale+

COM-HPC
server size D
Intel Agilex



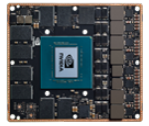
COM-HPC
client size B
Xilinx Versal

uRECS

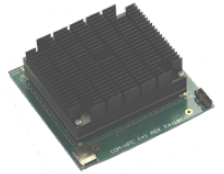
GPU SoC

Jetson TX2
NVIDIA
Tegra X2

Jetson nano
NVIDIA
Xavier NX



Jetson AGX
NVIDIA Xavier



COM-HPC Size B
NVIDIA Xavier
AGX

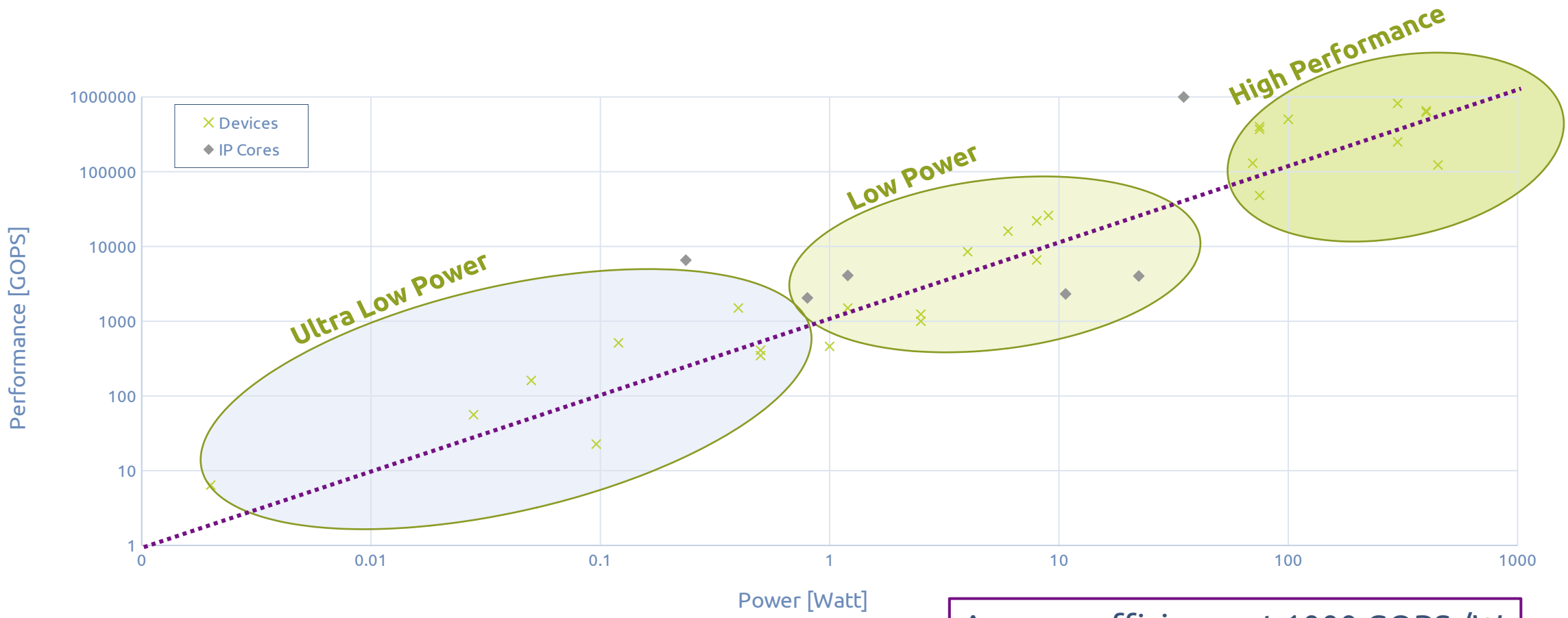
ML SoC

SMARC
Coherent
Logix
HX40416

SMARC
Coral SoM

SMARC
Xilinx Zynq
UltraScale

Peak Performance of DL Accelerators



Average efficiency at 1000 GOPS /W

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

Flexible Accelerators for Deep Learning

VEDLIoT
Very Efficient Deep Learning in IoT

- End of Moore's law & dark silicon
  => Domain Specific Architectures (DSA)
- Efficient, flexible, scalable accelerators for the compute continuum

Algotecture
- Optimized DL algorithms
- Optimized toolchain
- Optimized computer architecture

Heterogeneous DL Accelerator

DL Model → Compiler → CPU, GPU-SoC, ML-SoC

Algotecture/ Co-Designed DL Accelerator

DL Model ↔ Co-Design → FPGA-SoC

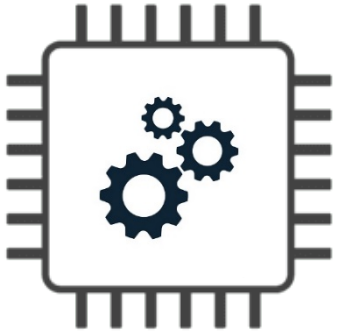
VEDLIoT's Deep Learning Toolchain

VEDLIoT
Very Efficient Deep Learning in IoT

Model Zoo

- Image Classification
- Object Detection
- Semantic Segmentation
- Instance Segmentation
- Extractive Question Answering

Optimization Engine



Compilers & Runtime APIs

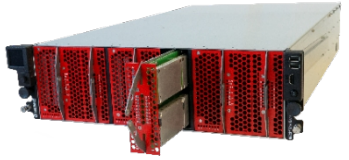
tvm

TensorRT

Arm NN

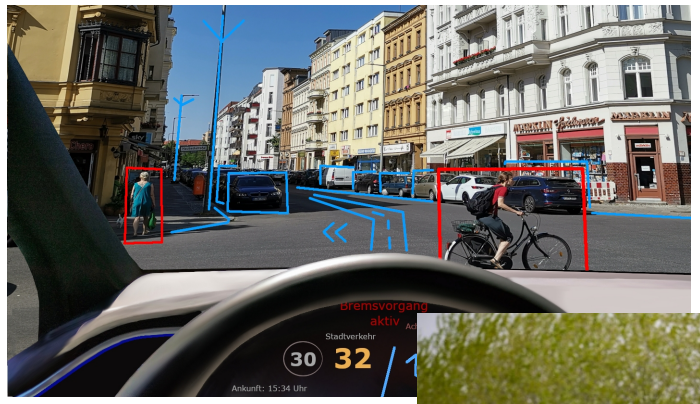
OpenVINO™

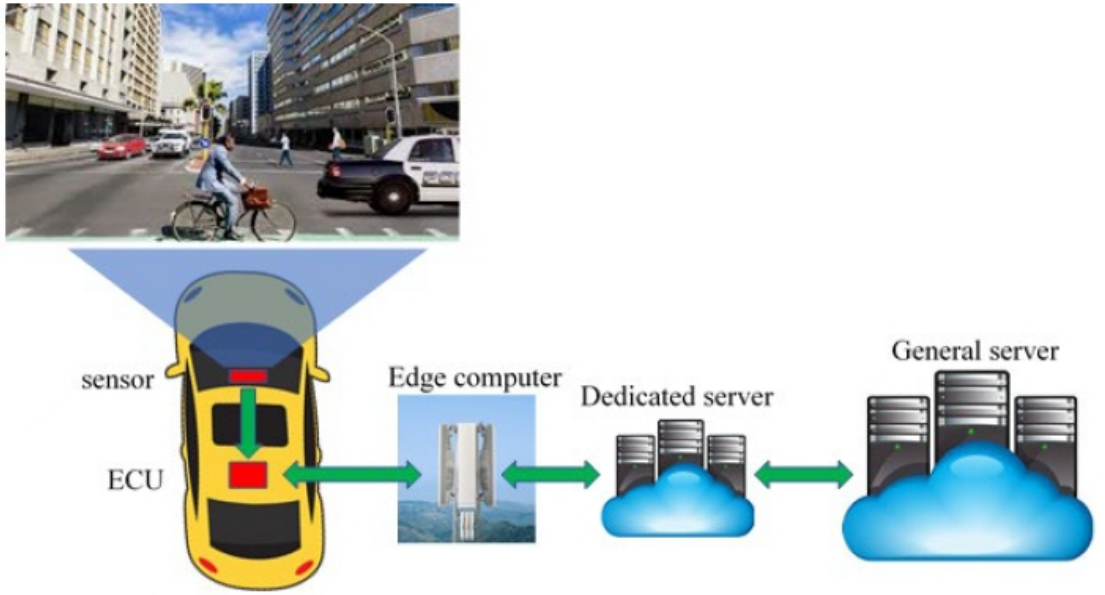
Heterogeneous Hardware Platforms



Use case: Automotive

VEDLIoT
Very Efficient Deep Learning in IoT





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

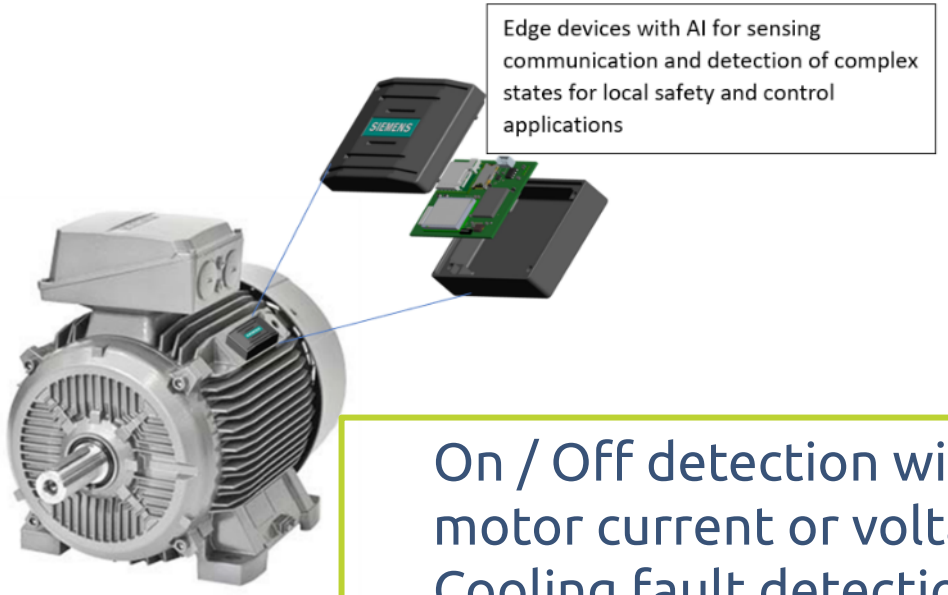
VEDLIoT
Very Efficient Deep Learning in IoT

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

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

Challenge: Low-power / Efficiency



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

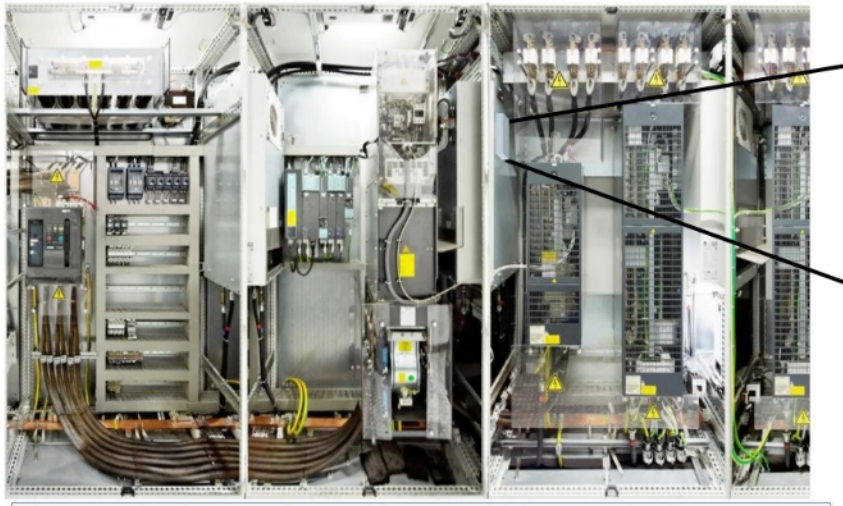
AI based pattern recognition for different local sensor data

current, magnetic field, vibration, temperature, low resolution infrared picture

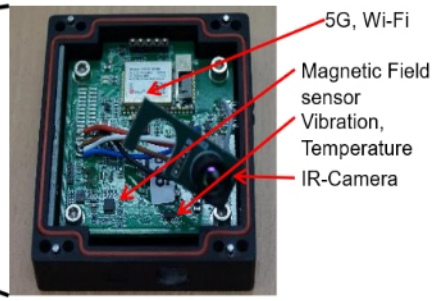
Safety critical nature

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



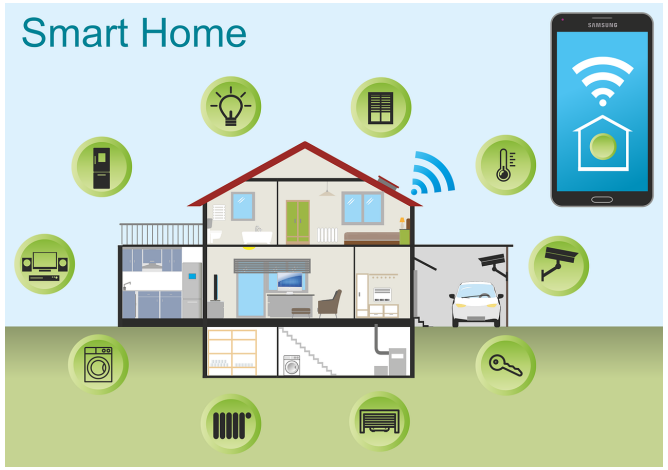
Combining the information from the IR-Camera and the magnetic field sensor to localize electric faults in power cabinets by deep learning methods

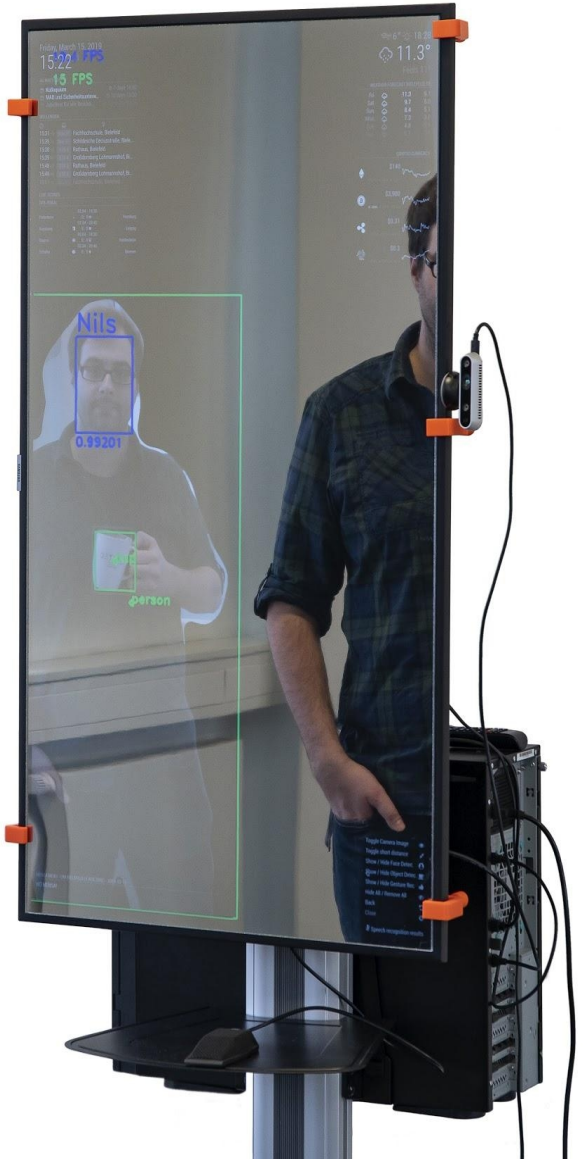


Specifications:
- Industrial temperature range (-20°C ... +85°C)
- Industrial batteries (rechargeable for ID-Tag)
- IP65 protection
- RoHS and IEC 61850-3 complaint
- Pre-certified wireless transceivers
- Target price: 100€ (ID-Tag)
- SIM on Chip*

Challenge: Accuracy

Use case: Smart Home / Assisted Living





Increase safety, health and well being of residents – acceleration of AI methods for demand-oriented user-home interaction

- Smart Mirror as central user interface
  - Own mirror image can be seen normally
  - Intuitive control over gesture and voice
  - Shows personalized information
- Data privacy as the highest priority
  - Edge computation of many neural networks

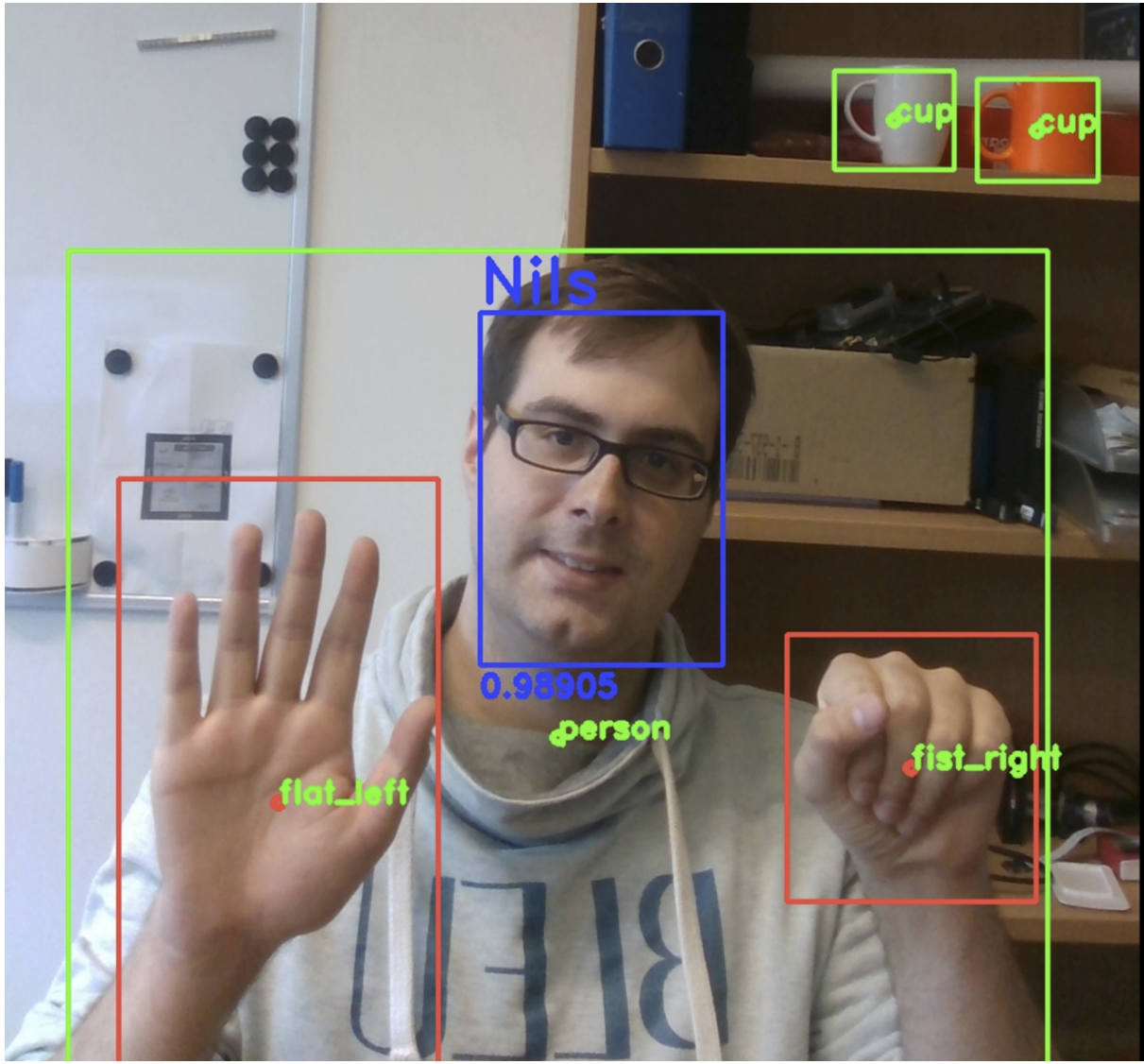
Use case: Smart Mirror – Neural Networks

VEDLIoT
Very Efficient Deep Learning in IoT

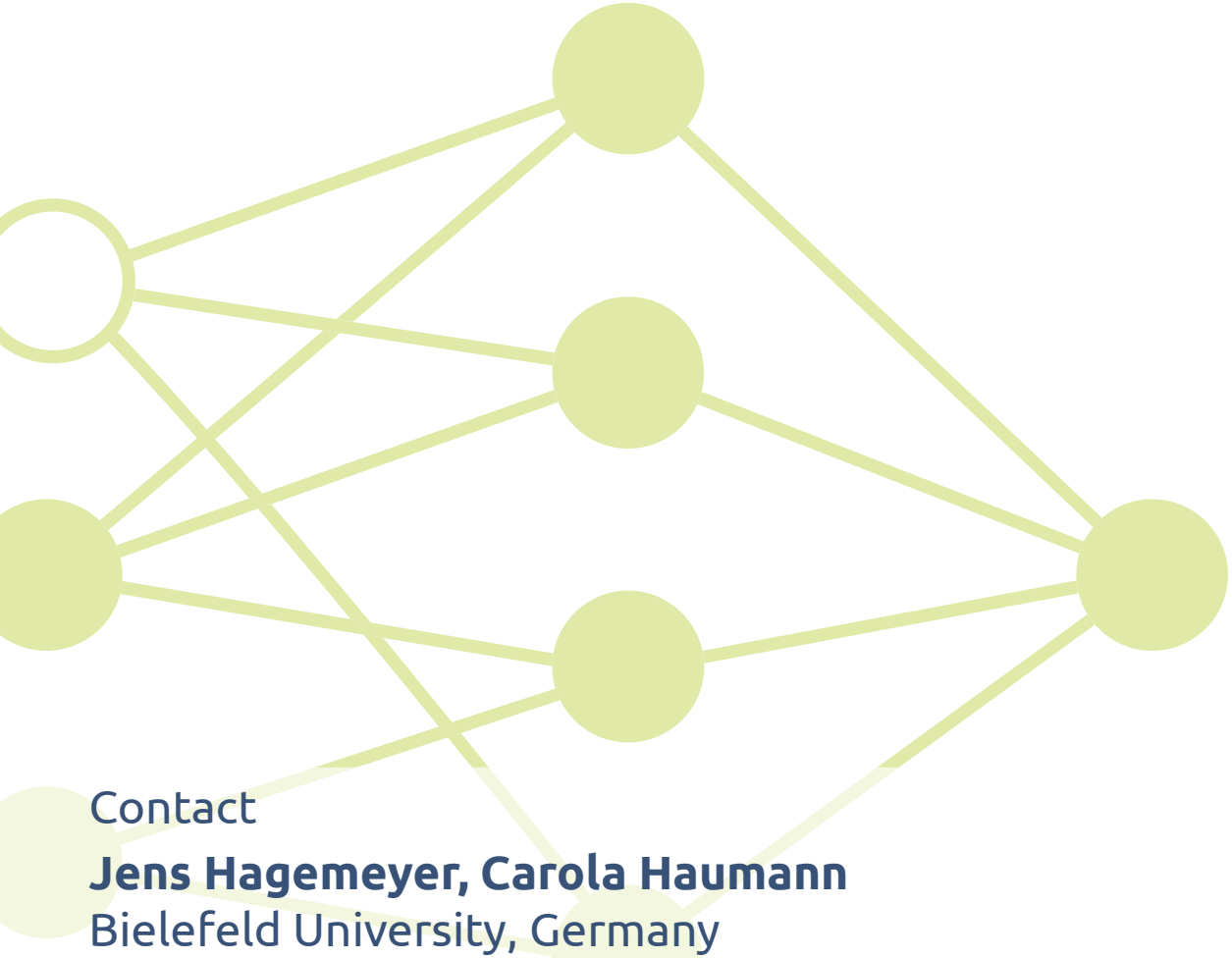
- Face recognition
  - Mobilenet SSD trained on WIDERFACE dataset
- Object detection
  - YoloV3, Efficient-Net, yoloV4-tiny
- Gesture detection
  - YoloV4-tiny with 3 Yolo layers (usually: 2 layers)
- Speech recognition
  - Mozilla DeepSpeech

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

Challenge:
Data privacy,
Efficiency



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

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

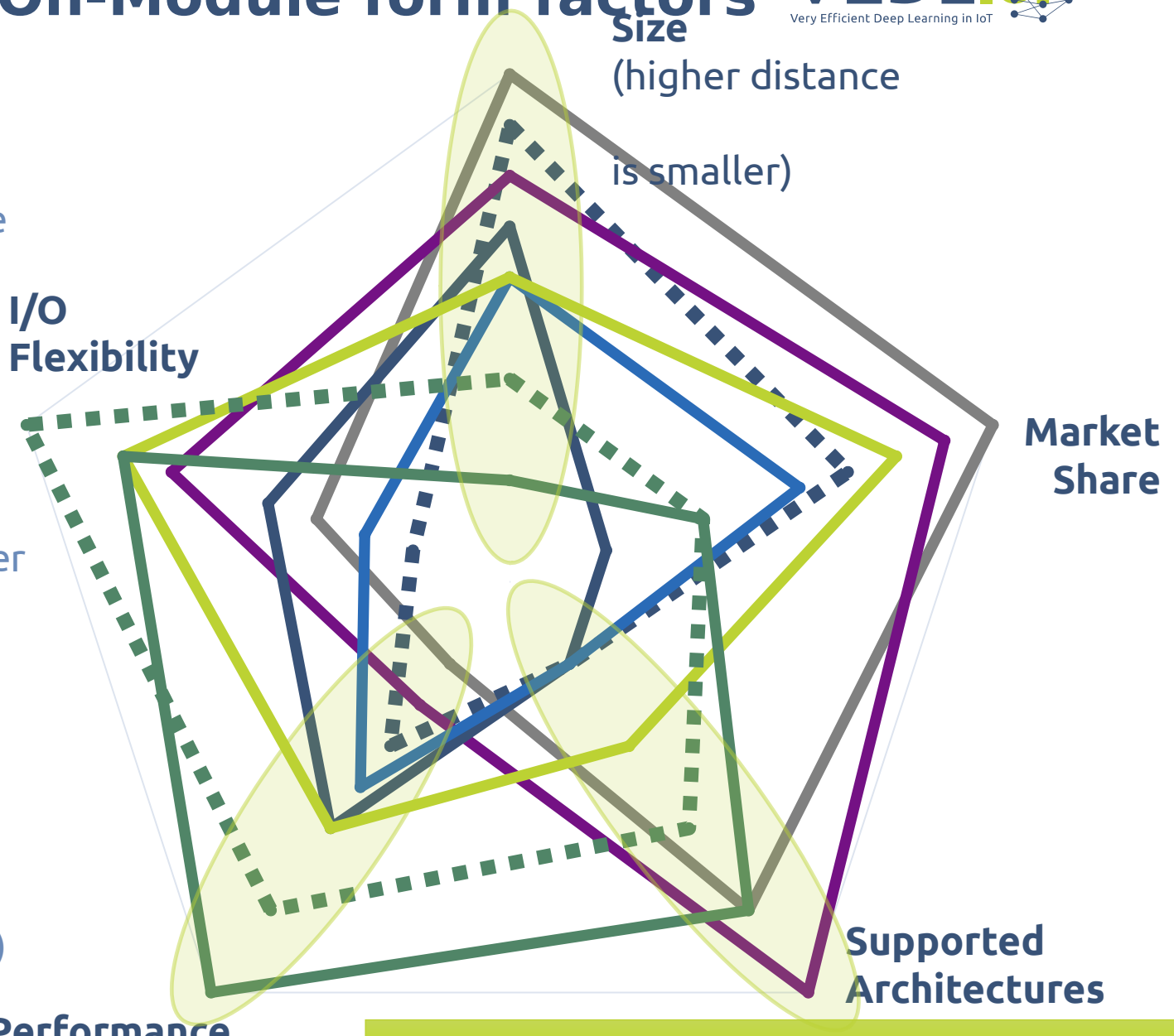
Supported Computer-On-Module form factors

VEDLIoT
Very Efficient Deep Learning in IoT

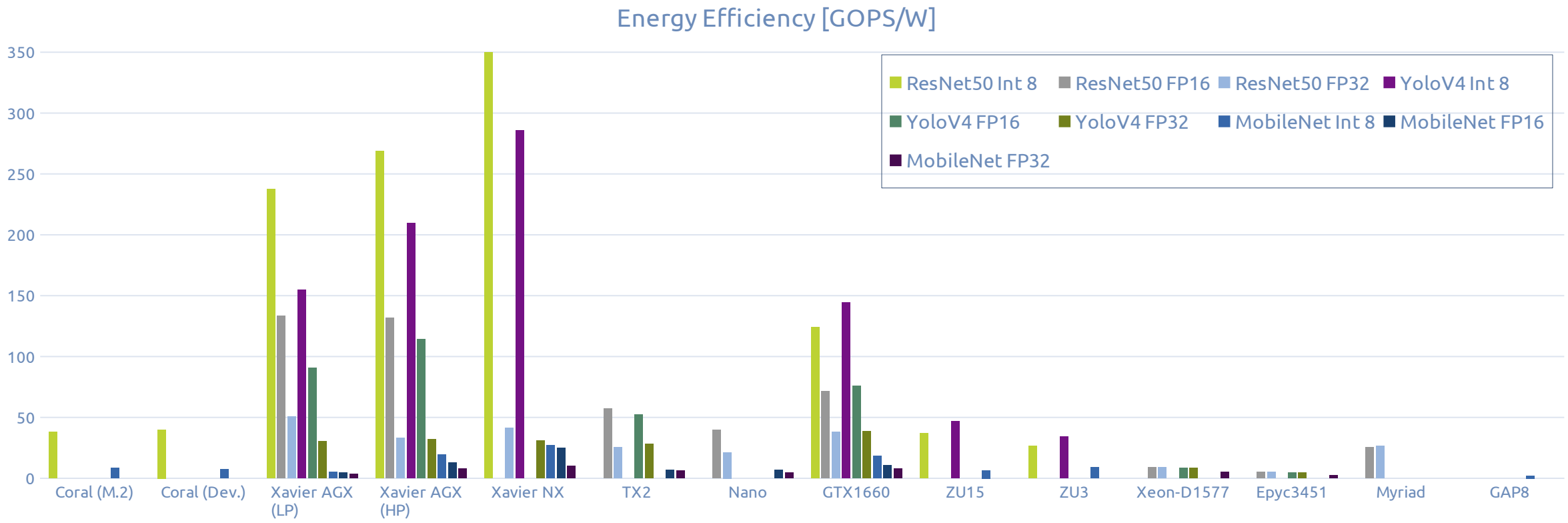
uRECS

RECS|Box & t.RECS

- Raspberry Pi Compute Module 4
- Jetson Xavier NX
- SMARC
- Xilinx Kria
- Jetson AGX Xavier
- COM Express (Type 6/7)
- COM-HPC Client (Type A-C)
- COM-HPC Server (Type D/E)

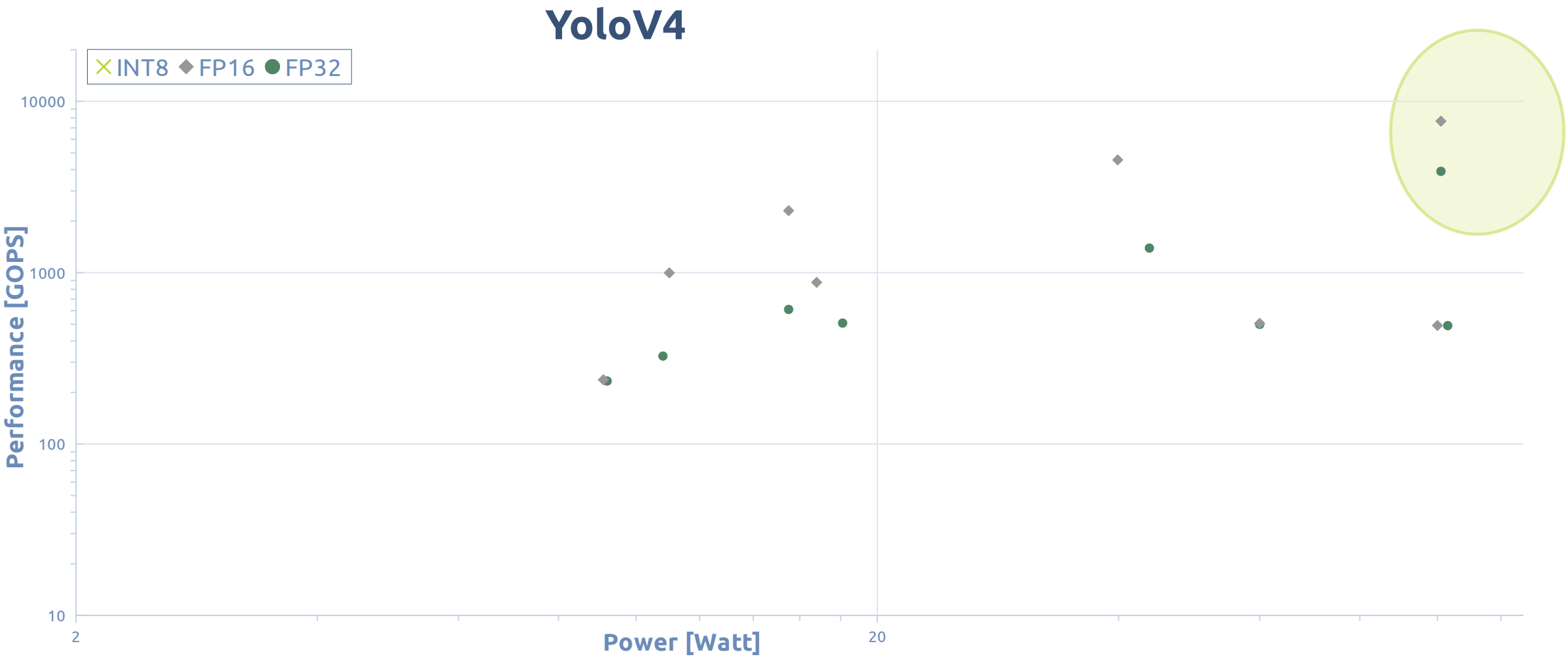


Benchmark performance of DL accelerators



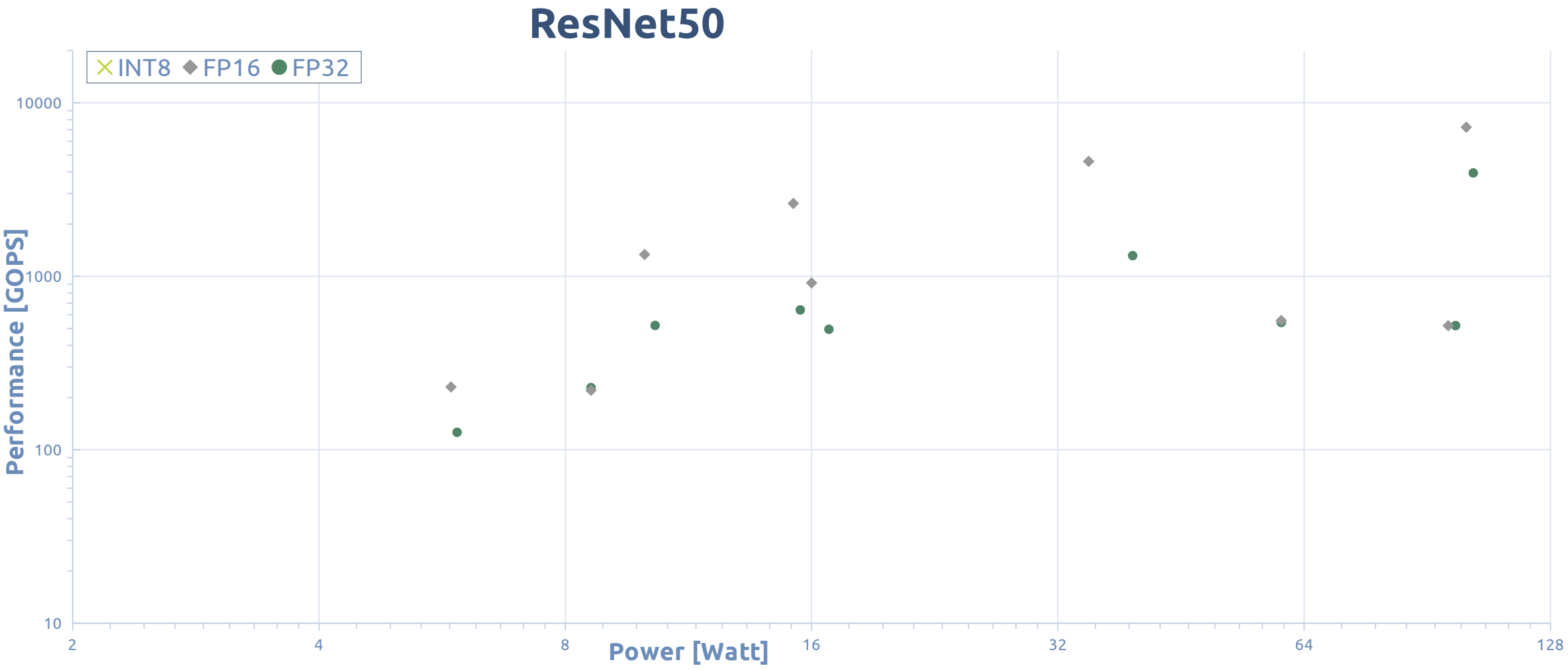
- Comparison based on currently available architectures
- VEDLIoT will include new specialized accelerators

Benchmark performance of DL accelerators



Benchmark performance of DL accelerators

VEDLIoT
Very Efficient Deep Learning in IoT



Benchmark performance of DL accelerators