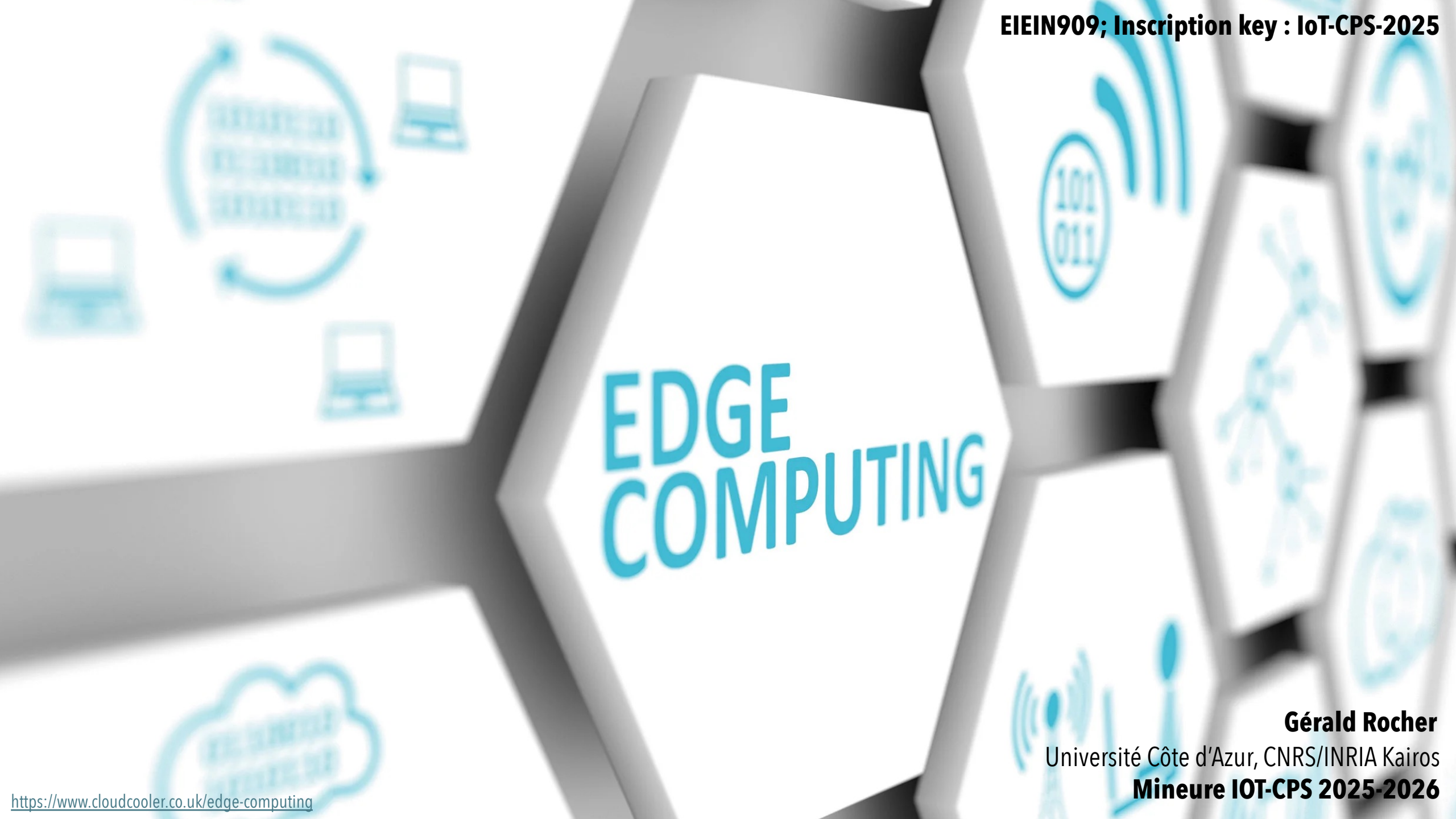
EIEIN909; Inscription key : IoT-CPS-2025



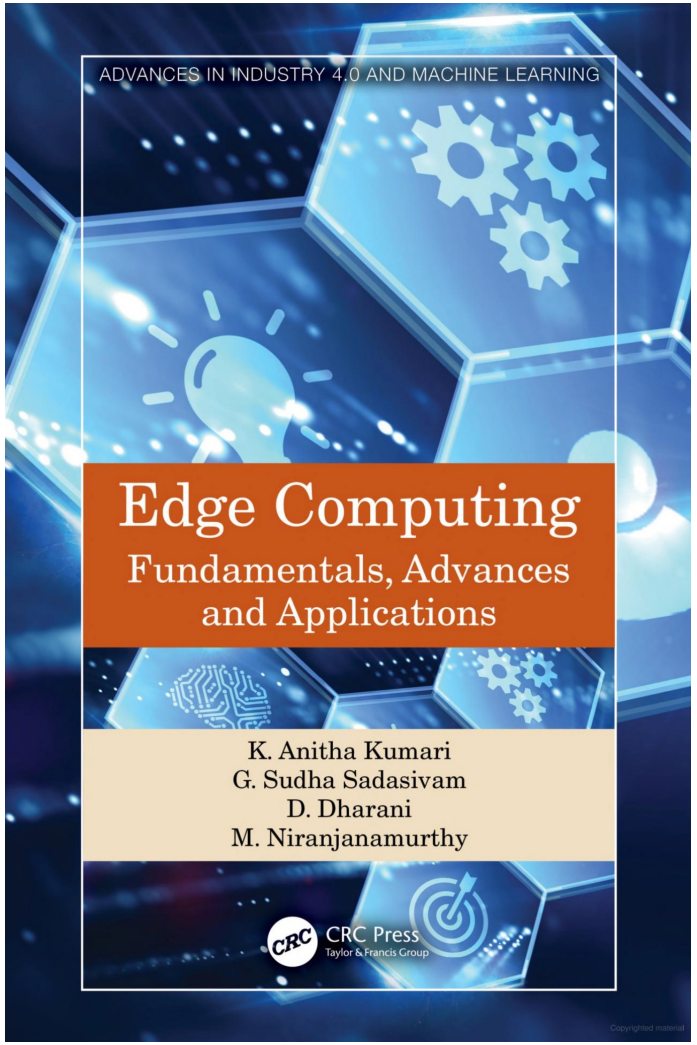
https://www.cloudcooler.co.uk/edge-computing

Gérald Rocher
Université Côte d'Azur, CNRS/INRIA Kairos
Mineure IOT-CPS 2025-2026

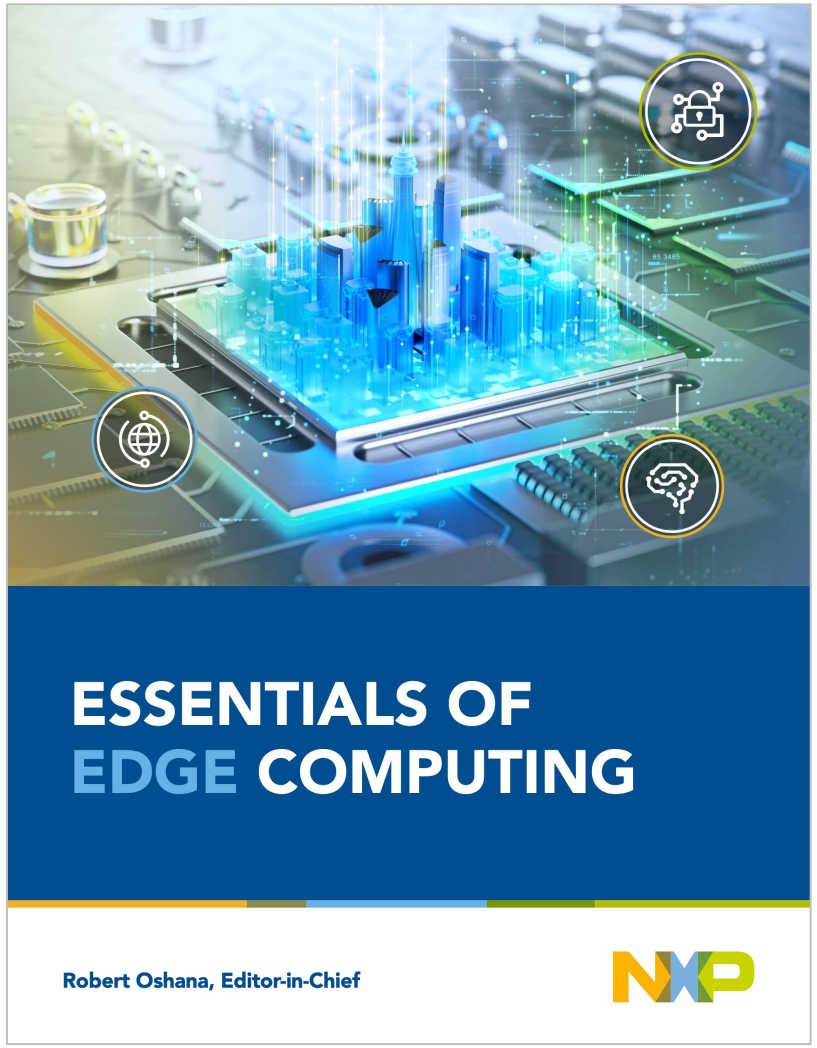
Content and Planning

Courses	Lecturer	Date	Comments
Introduction	Gérald ROCHER	19/11/2025	Project definition
TinyML	Gérald ROCHER	26/11/2025	Project review
NXP (NPU + TFLite)	Gaetan Bahl Chiraz Rayene Harkati	03/12/2025	TP on NXP board
ONNX Runtime	Gérald ROCHER	10/12/2025	Project review & advancement
Apache TVM	Gérald ROCHER	17/12/2025	Project review & advancement
Project Development	Gérald ROCHER	07/01/2025	Project review & advancement
Project Development	Gérald ROCHER	14/01/2025	Project review & advancement
Final Project Review	Gérald ROCHER	21/01/2025	POC presentation + report

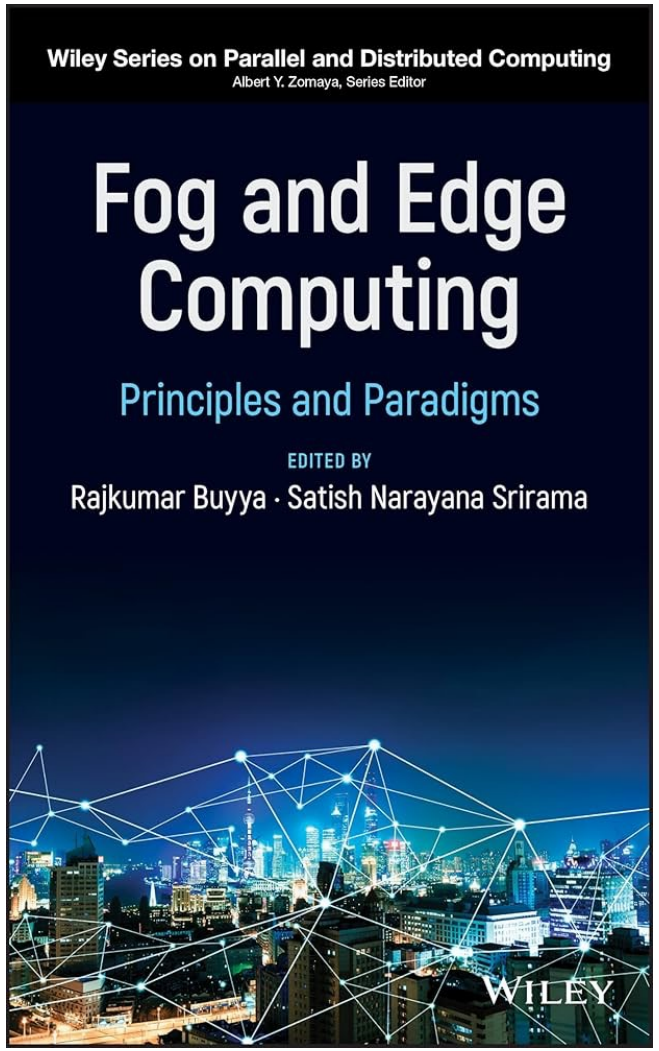
References



Link



Link

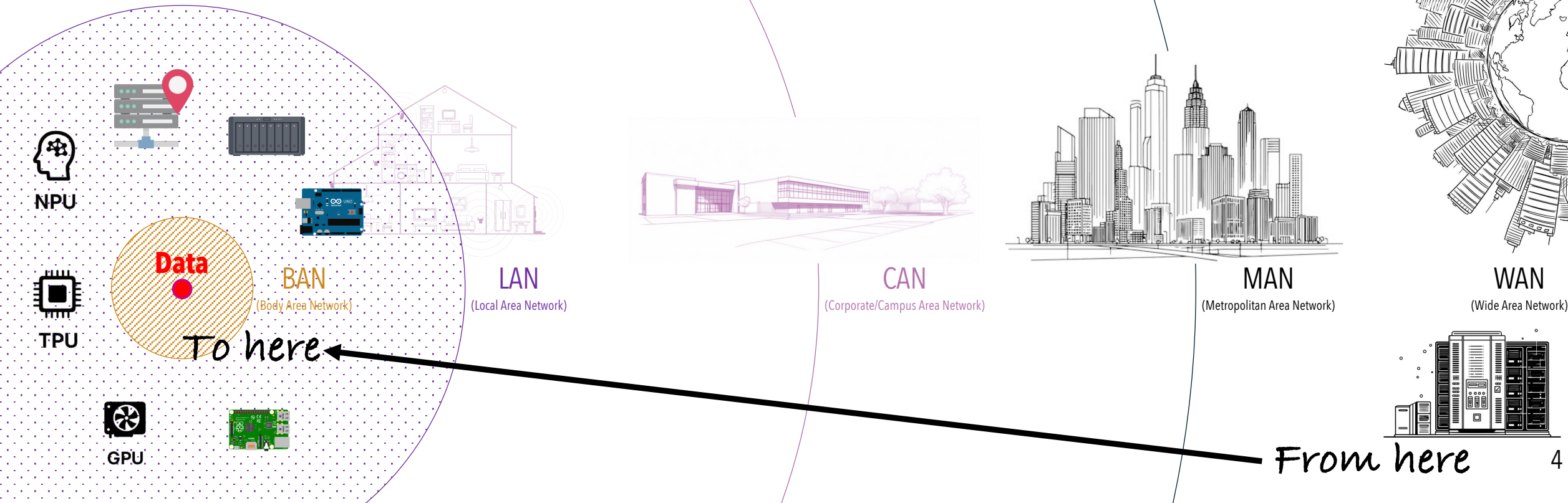


Link

What is Edge Computing ?

Edge computing is a distributed computing model that brings computation and data storage closer to the sources of data.

Wikipedia

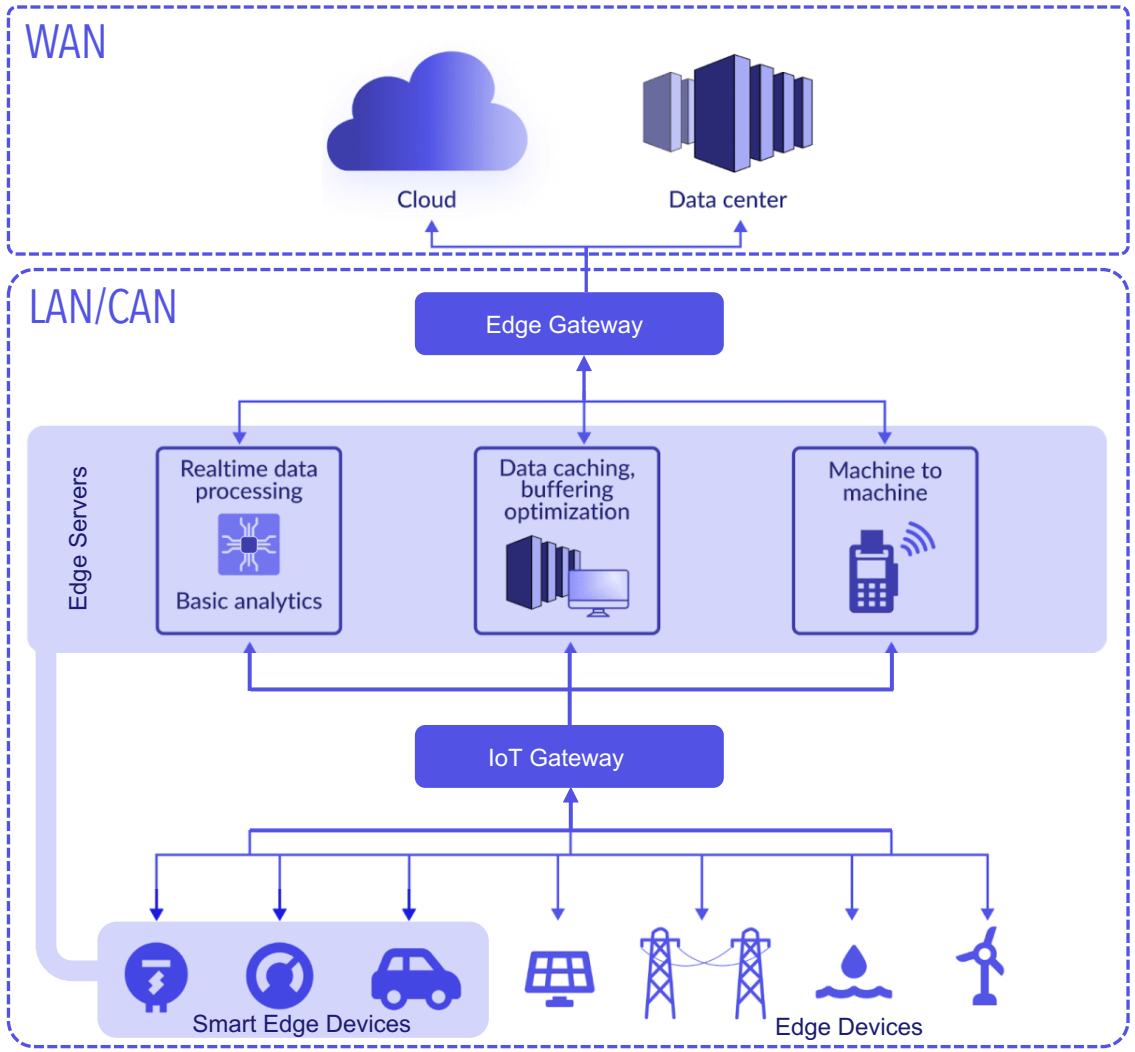


Why Does Edge Computing Matter?

By storing and processing data closer to their producers, Edge Computing helps

- To improve network connectivity and WAN congestion,
- To reduce latency (near real-time response),
- To strengthen security,
- To enhance user experiences,
- To reduce costs.

Edge Computing Architecture



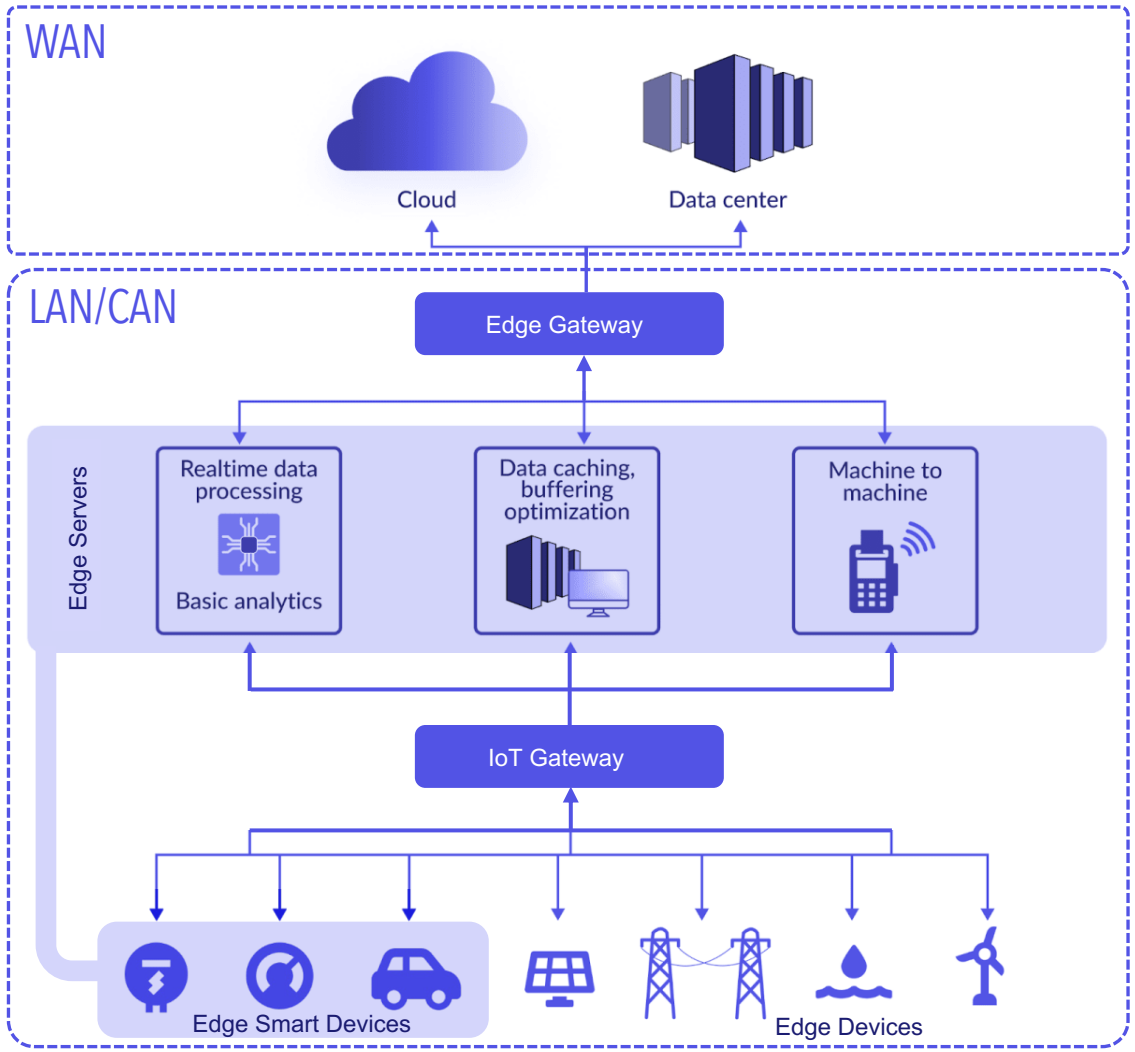
Edge Devices

- Physical devices (sensors, actuators, or embedded systems) that interact with the environment,
- Typically, resource-constrained (limited compute, memory, or power).
- Purposes
  - **Data Collection**: Measure physical parameters (temperature, vibration, etc.),
  - **Basic Control**: Execute simple, real-time actions (e.g., turning a motor on/off),
  - **Local Processing**: Some may run lightweight algorithms (e.g., filtering noisy sensor data).

Smart Edge Devices

- More advanced than basic edge devices, with embedded compute capabilities (e.g., MCU with ML accelerators (ASIC/NPU/TPU/DSP), smart cameras),
- Often run firmware/OS (e.g., Linux, FreeRTOS)
- Purposes
  - **Local Processing:** Sensor Fusion, Lightweight ML Inferences,
  - **Local Analytics**: Pre-process data (e.g., object detection on a camera stream),
  - **Autonomy**: Operate independently for critical tasks (e.g., predictive maintenance at the machine level),
  - **Protocol Translation**: Convert field protocols (Modbus, CAN bus) to IP-based ones (MQTT, HTTP)

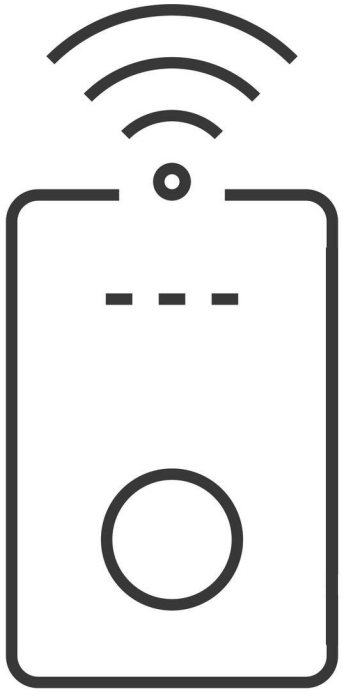
Edge Computing Architecture



Edge Servers

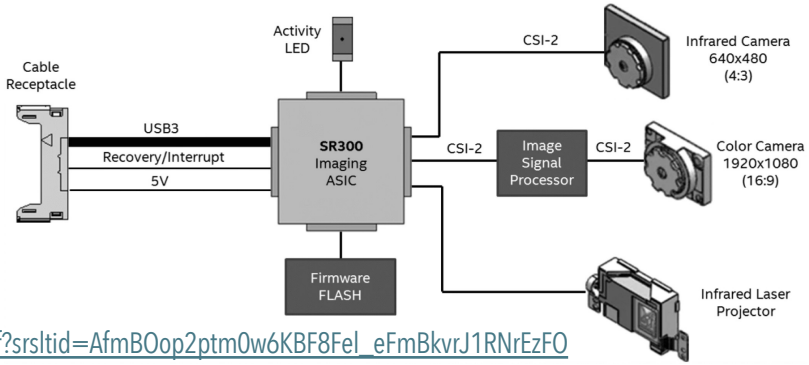
- More powerful compute nodes (physical/virtual) deployed in the LAN (CPU/GPU/TPU, RAM), close to the IoT gateway.
- Handle heavier workloads than edge devices/gateways.
- Purposes
  - **Basic Real-time Analytics Server**
    - Run rule-based alerts (e.g., "temperature exceeds threshold"),
    - ML inference (e.g., anomaly detection),
    - Sensor Fusion,
    - Tools : Time-series DB (influxDB), scripting (Python), low-code (Node-RED).
  - **Data Caching & Buffering**
    - Temporarily store data during network interruptions,
    - Optimize bandwidth by batching data before WAN transmission,
    - Tools : Redis, Apache Kafka (for buffering), local SQL/NoSQL databases.
  - **Data Optimization**
    - Compress data (e.g., from raw video to metadata),
    - Filter redundant data (e.g., remove duplicate sensor readings).
    - Tools : encoding, deduplication, downsampling, etc.
  - **M2M (Machine-to-Machine) Coordination**
    - Orchestrate workflows between devices (e.g., synchronize robots in a factory),
    - Enable pub/sub messaging (e.g., MQTT broker for local device communication).

Smart Edge Devices



Industry / Robotics

Intel SR300



https://www.mouser.com/pdfdocs/intel_realsense_camera_sr300.pdf?srsltid=AfmBOop2ptm0w6KBF8Fel_eFmBkvrJ1RNrEzFOsby5N5zFEofZEM0-zV

Oak-D PRO W

4 TOPS of processing power (1.4 TOPS for AI)

Encoding: H.264, H.265, MJPEG - 4K/30FPS, 1080P/60FPS

Computer vision: warp/dewarp, resize, crop via ImageManip node, edge detection, feature tracking.

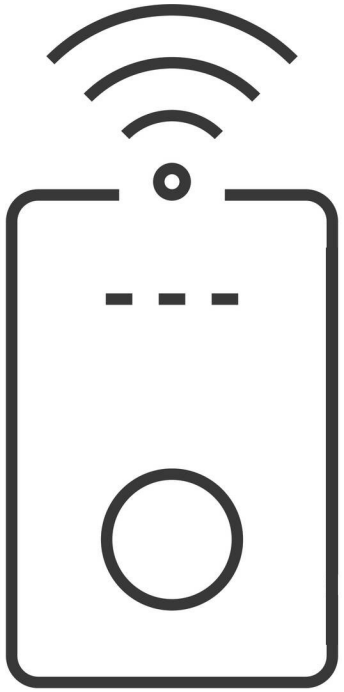
Stereo depth perception with filtering, post-processing, RGB-depth alignment, and high configurability

Object tracking: 2D and 3D tracking with ObjectTracker node

On-device programming: Run custom logic/tasks on-device.

https://github.com/luxonis/depthai-hardware/blob/master/DM9098_OAK-D-S2/Datasheet/OAK-D-S2_Datasheet.pdf

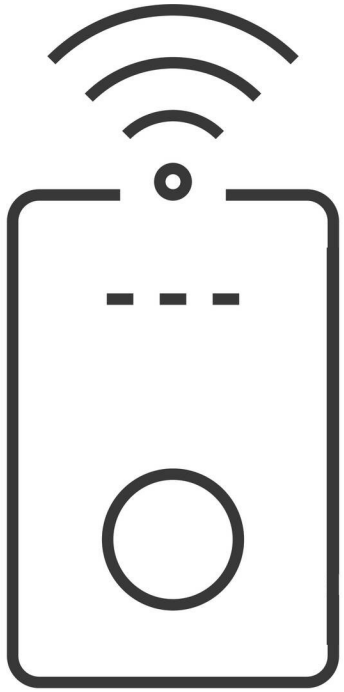
Smart Edge Devices



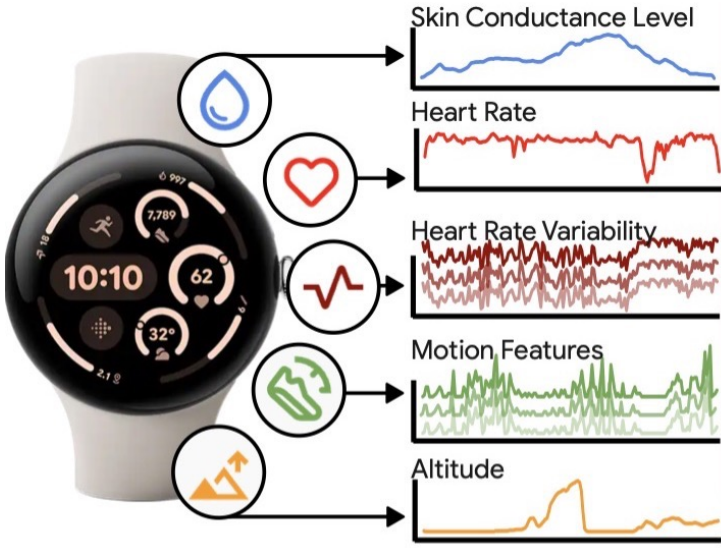
Smart City

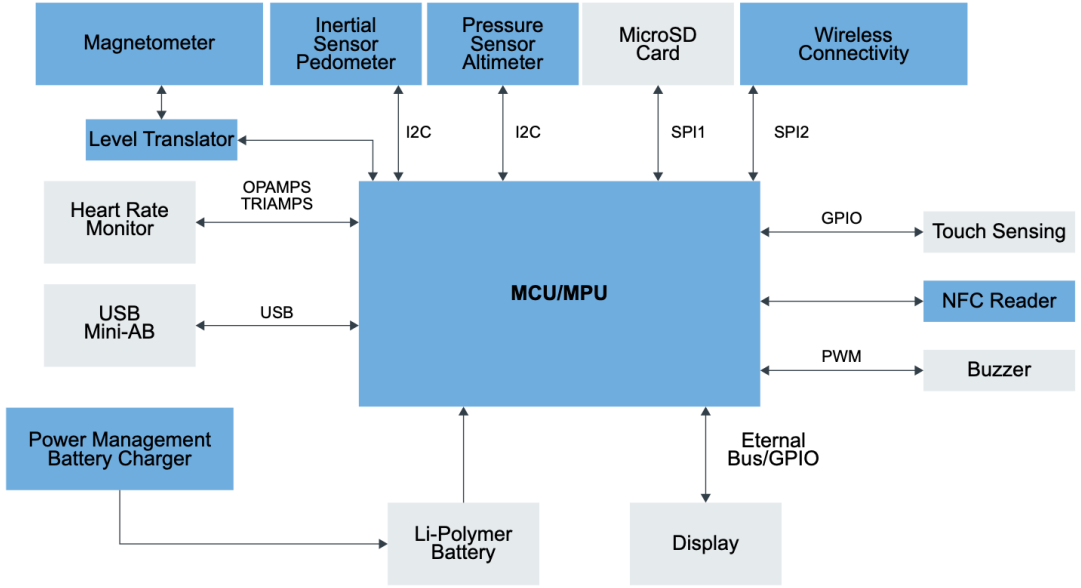
https://intelilight.eu/smart-streetlight-controllers/zhaga-luminaire-controller/
https://intelilight.eu/smart-streetlight-controllers/nema-luminaire-controller/

Smart Edge Devices



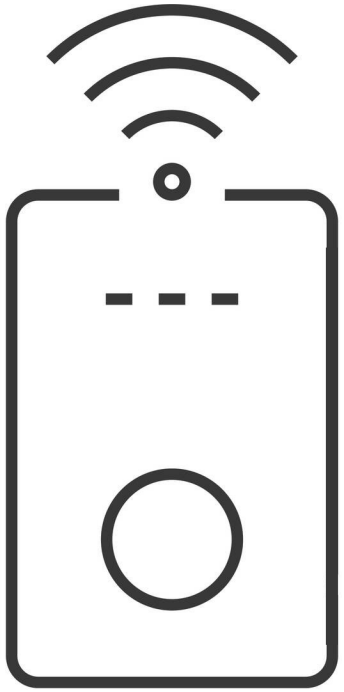
Smart Health



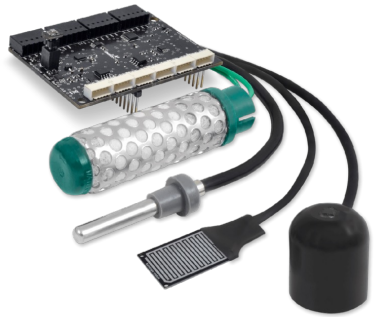


https://www.nxp.com/assets/block-diagram/en/SmartWatchandWristband_SMARTWATCH.pdf

Smart Edge Devices



Smart Agriculture



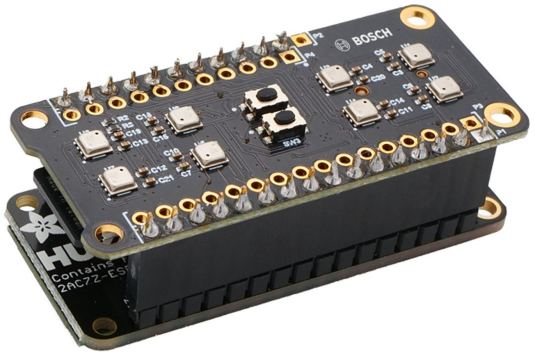
Predictive models powered by Artificial Intelligence, that are calibrated automatically based on your field conditions.

https://www.libelium.com/iot-solutions/smart-agriculture/

Environment

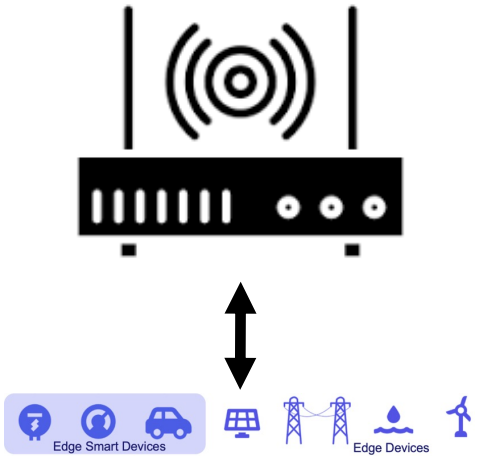
BME688

Digital low power gas, pressure, temperature & humidity sensor with AI.



https://www.bosch-sensortec.com/media/boschsensortec/downloads/datasheets/bst-bme688-ds000.pdf

IoT Gateways



NXP Zigbee Gateway (QN9080DK)
Protocols : Zigbee 3.0, BLE 5.0

RAK7289 WisGate Edge
Protocols : LoRaWAN + LTE backup

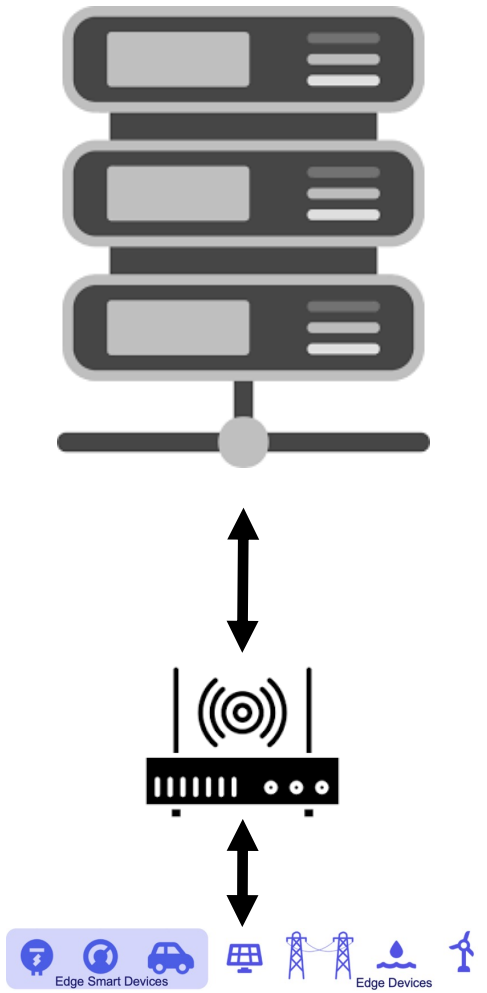
RAK7289 WisGate Edge
Protocols : LoRaWAN

RAK7289 WisGate Edge Pro
Protocols : LoRaWAN, LTE Cat-4 backup, Wi-Fi, BLE (via optional module)

Eurotech ReliaGATE 10-14
Protocols : Wi-Fi, Bluetooth, BLE5, GNSS/LTE (optional)

Custom (SBC + Hats/Dongles)
Protocols : Almost all protocols!

Smart Edge Industrial Servers



Axial AX300 Industrial Edge Server

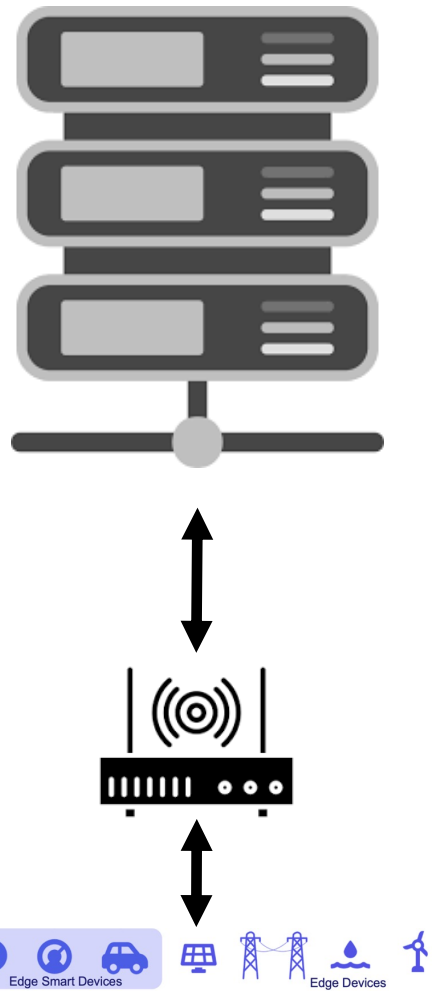
https://static.onlogic.com/resources/manuals/OnLogic-AX300-Product-Manual.pdf

- Up to 7x single slot or 4x dual slot NVIDIA A6000 GPUs
- Dual socket Intel Xeon processing
- Up to 5864 TOPS of AI performance

Performs heavy-duty processing

- AI/ML inference,
- Local database storage,
- Custom business logic.

Smart Edge AI Servers

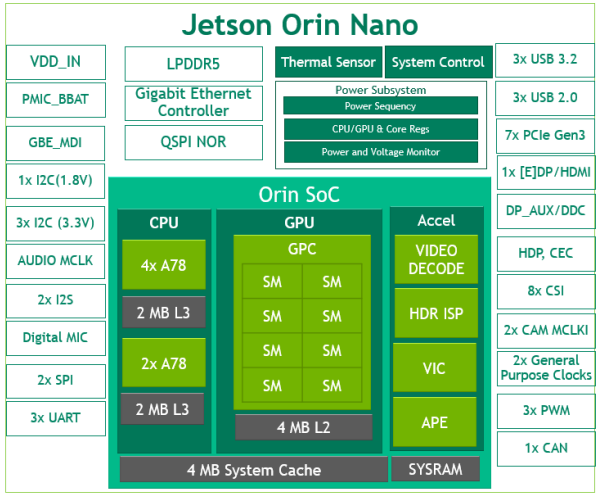


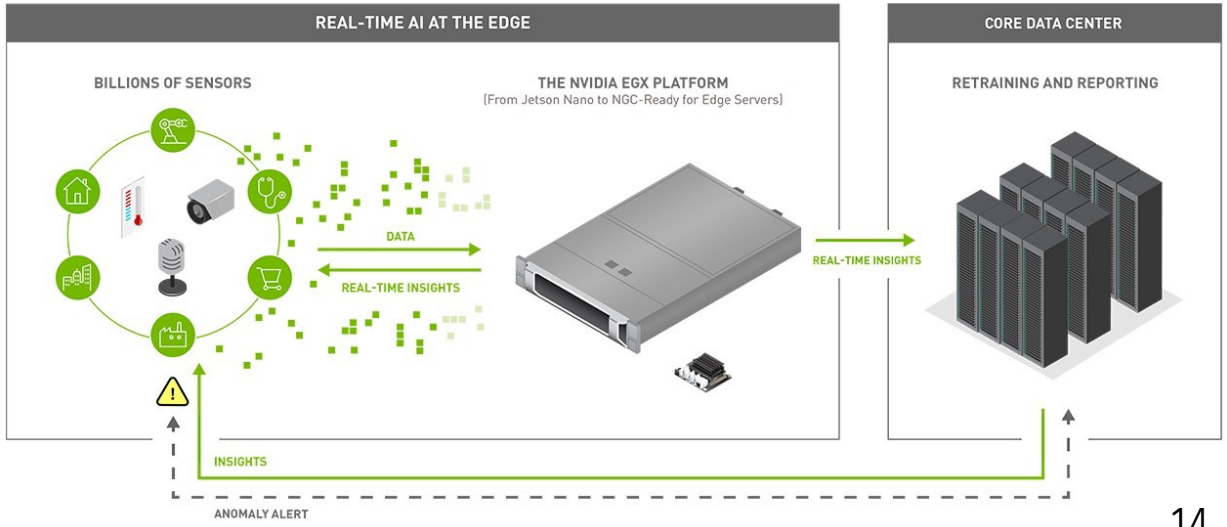
NVidia® Real-time Edge AI Processors and Servers

Nvidia Jetson ORIN Nano

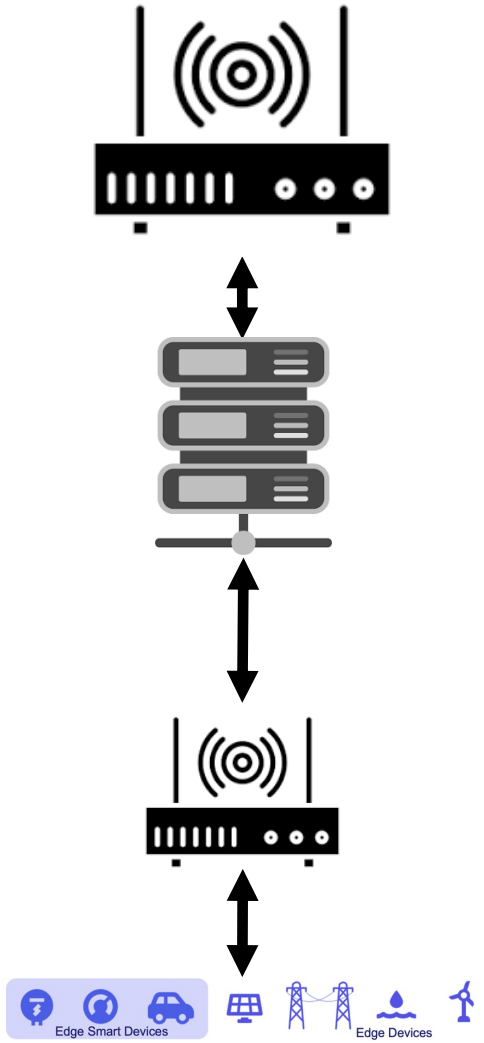
Nvidia Jetson AGX Xavier

https://info.nvidia.com/rs/156-OFN-742/images/Jetson_AGX_Xavier_New_Era_Autonomous_Machines.pdf





Smart Edge Gateways



Intel® Processor-Based Edge Gateways

Features

- Intel Atom® and Core™ processors
- SO-DIMM for DDR4 memory
- Rich I/O: DP++, DVI, VGA, GbE, COM, USB, DI/O
- Security: TPM2.0
- Rich storage: 2.5" SATA / M.2
- Embedded expansion: Mini PCIe/uFM/M.2/USIM
- (EGW-3200) Optional sensor suite: accelerometer, humidity, pressure, temperature
- WiFi-6E/Bluetooth 5.3

Software Support

- Windows 10 IoT Enterprise LTSC 2019
- Linux Ubuntu Server 20.04 LTS
- Dell NativeEdge

Optional Accessories – Qualified and Certified

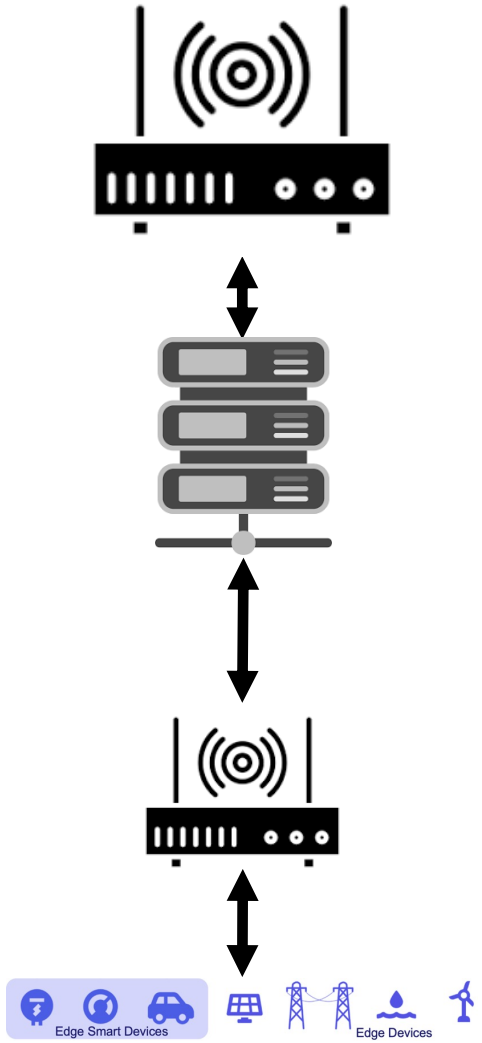
- Expansion modules (mPCIe or uFM) for Isolated COM (RS-232 or RS-422/485), GbE with PoE, GbE LAN, Canbus
- 4G and 5G modules
- AC-to-DC adapter

https://i.dell.com/sites/csdocuments/Products_Documents/en/us/dell-edge-gateway-3200-5200-spec-sheet.pdf

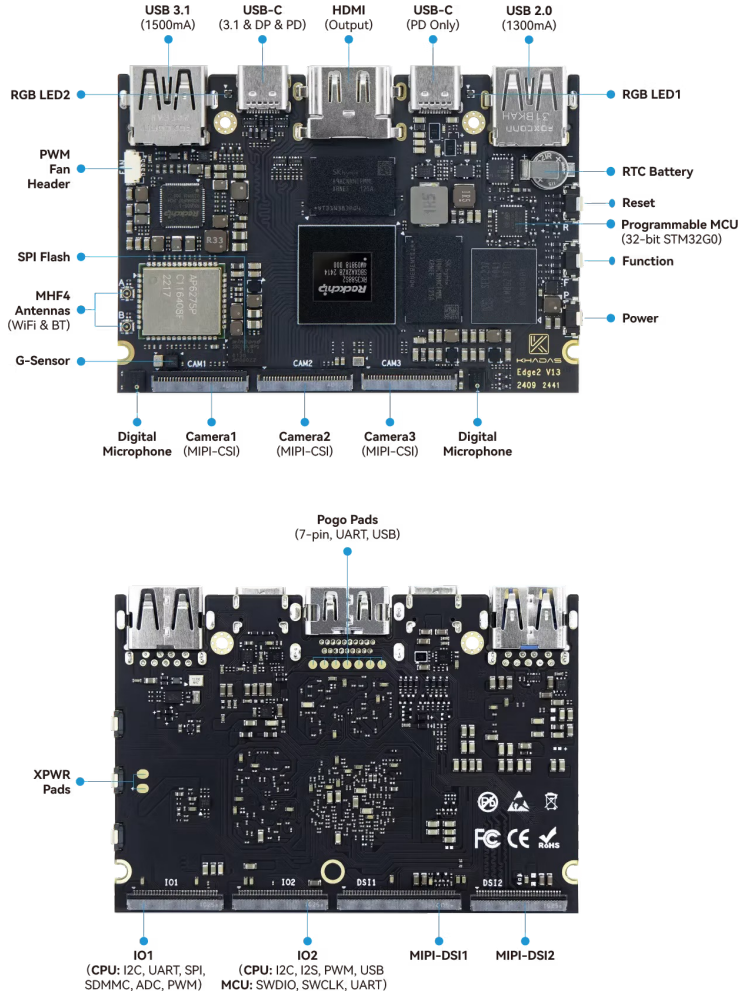
Performs

- WAN Optimization (compression, caching),
- Security (VPN, firewall, etc.),
- Cellular failover.

Smart Edge Lightweight Servers

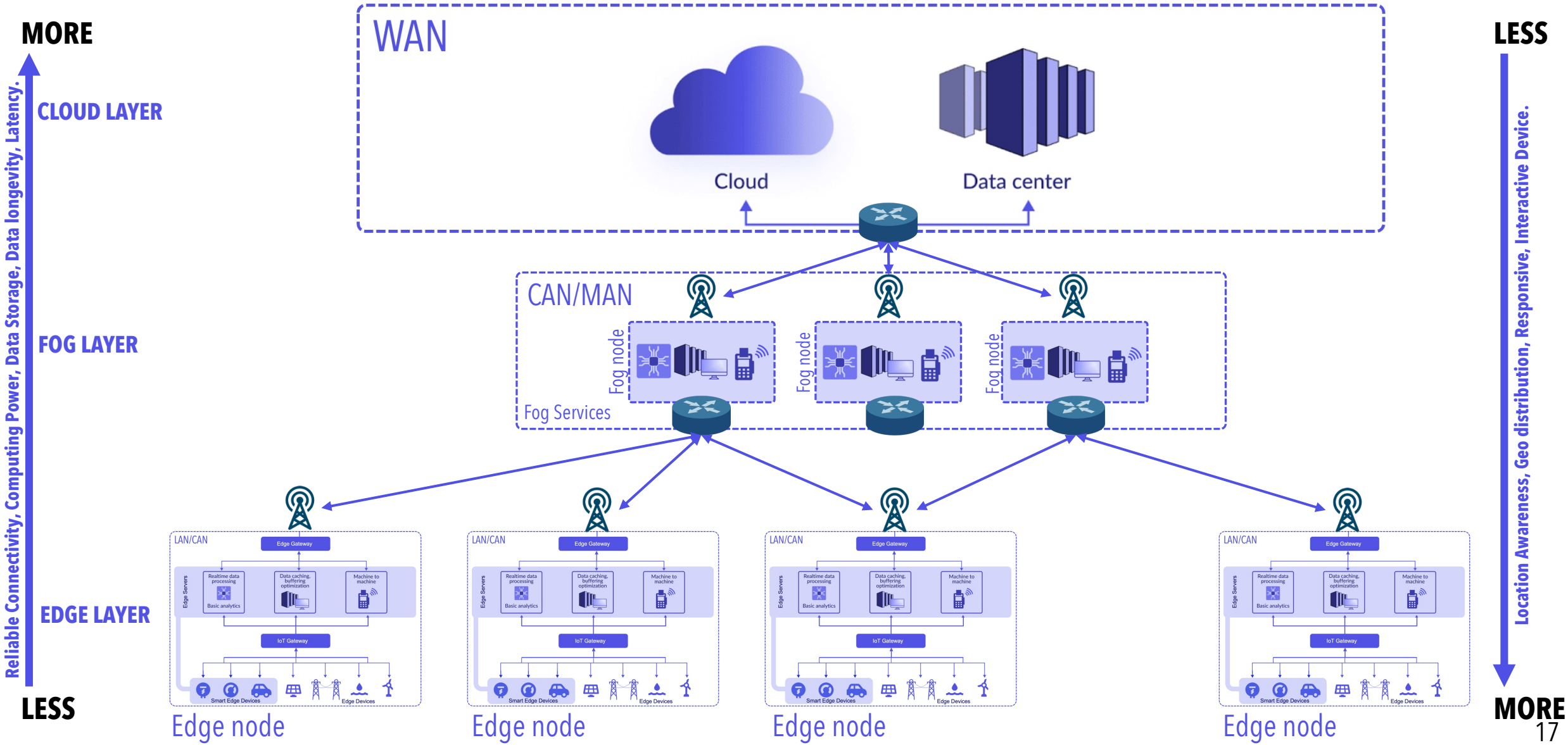


Khadas® Edge2

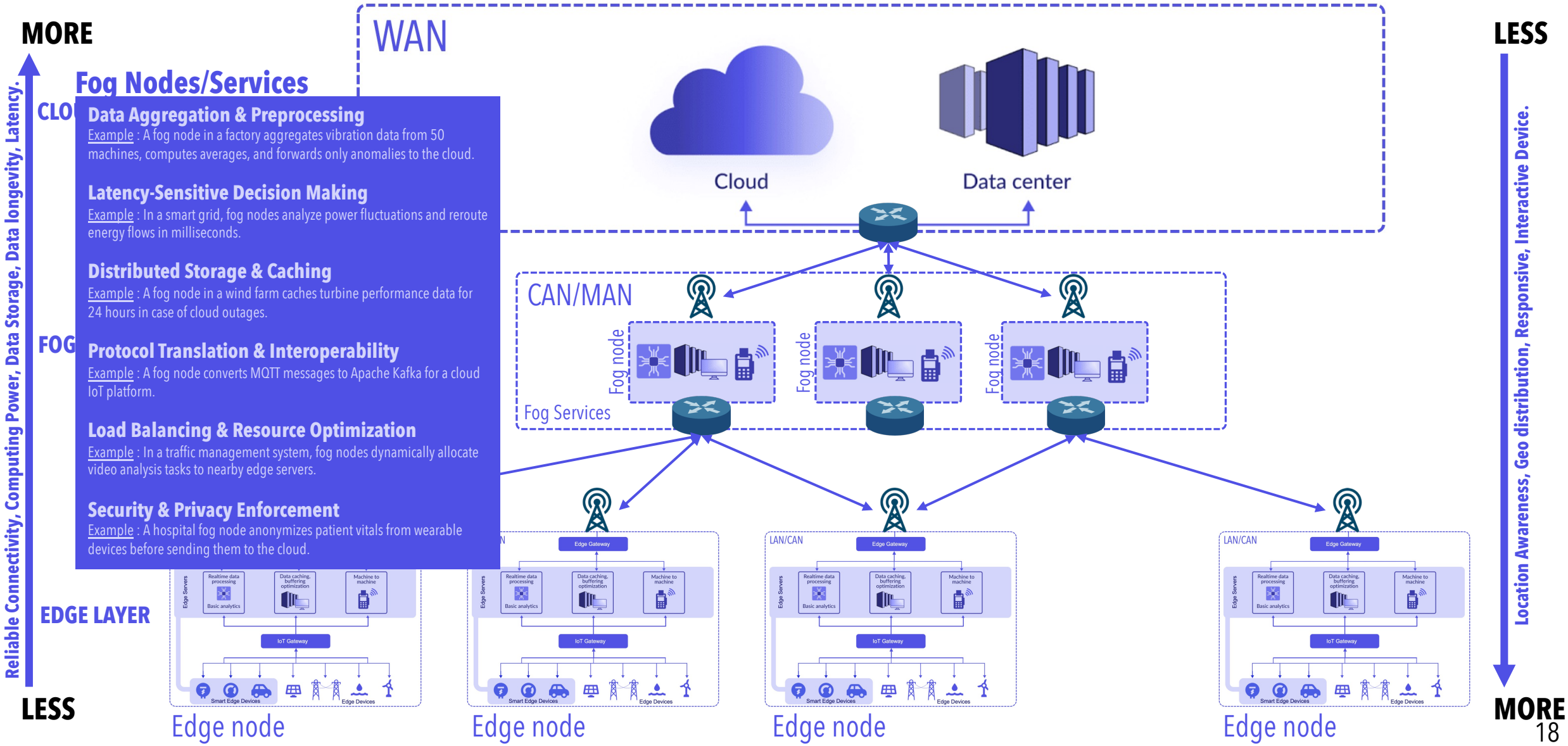


Model	Basic	Pro
SoC	Rockchip RK3588S 2.25GHz Quad Core ARM Cortex-A76 + 1.8GHz Quad Core Cortex-A55 CPU ARM Mali-G610 MP4 GPU up to 1GHz Build-in 6 TOPS Performance NPU 4K@60fps AV1, 8K@60fps H.265 Decoding 8K@30fps H.264/H.265 Encoding HDR, HDR10, HLG Video Processing	
Coprocessor [1]	STM32G031K6	
SPI Flash	32MB	
RAM	8GB LPDDR4X 2112MHz, 64-bit	16GB LPDDR4X 2112MHz, 64-bit
eMMC 5.1	32GB	64GB
Wi-Fi	Ampak AP6275P 2T2R Wi-Fi 6, IEEE 802.11 ax/ac/a/b/g/n	
Bluetooth	Bluetooth 5.0	
USB HOST	x1 USB 3.1 + x1 USB 2.0	
USB-C	x1 PD (Power Deliver) Only x1 USB 3.1 + PD + DP 1.4, up to 4K@60fps [2]	
HDMI	Type-A Female, 8K@60fps HDMI2.1, Dynamic HDR, CEC, DSC 1.2a and HDCP 2.3	
MIPI Display	x1 30-pin 0.5mm FPC Connector 4-lane MIPI-DSI Interface, Resolution up to 4K@60Hz	
Touch Display	x1 40-pin 0.5mm FPC Connector 4-lane MIPI-DSI Interface, Resolution up to 4K@60Hz I2C and GPIO for Touch Panel	
Cameras	x3 30-pin 0.5mm FPC Connectors 4-lane MIPI-CSI Interface per Connector ISP Resolution up to 48MP	
Expand IO	x2 30-pin 0.5mm FPC Connector CPU: I2C, UART, SPI, SDMMC, I2S, ADC, PWM, USB MCU: SWDIO, SWCLK, UART	
Pogo Pads	7-pin, USB, UART, 5V	
Cooling Fan Header	4-pin 0.8mm Header PWM Speed Control	
DMIC	Stereo Digital Microphones	
Sensor	KXTJ3-1057 Tri-axis Digital Accelerometer	
RTC Battery	3V 3mAh, Lithium Rechargeable Battery	
LEDs	x2 RGB LED	
Buttons	x3 (Reset / Func / Power)	
XPWR Pads	For External Power Button	
Mounting Holes	Size M2 x4	
Board Dimensions	82.0 x 57.5 x 5.7 mm	
Board Weight	25g	

Distributed Edge Computing Architecture



Distributed Edge Computing Architecture



Fog Nodes/Services

Data Aggregation & Preprocessing
Example : A fog node in a factory aggregates vibration data from 50 machines, computes averages, and forwards only anomalies to the cloud.

Latency-Sensitive Decision Making
Example : In a smart grid, fog nodes analyze power fluctuations and reroute energy flows in milliseconds.

Distributed Storage & Caching
Example : A fog node in a wind farm caches turbine performance data for 24 hours in case of cloud outages.

Protocol Translation & Interoperability
Example : A fog node converts MQTT messages to Apache Kafka for a cloud IoT platform.

Load Balancing & Resource Optimization
Example : In a traffic management system, fog nodes dynamically allocate video analysis tasks to nearby edge servers.

Security & Privacy Enforcement
Example : A hospital fog node anonymizes patient vitals from wearable devices before sending them to the cloud.

Some notes/recalls on Hardware accelerators @Edge

Edge Computing relies on several Hardware solutions to speed-up data treatment

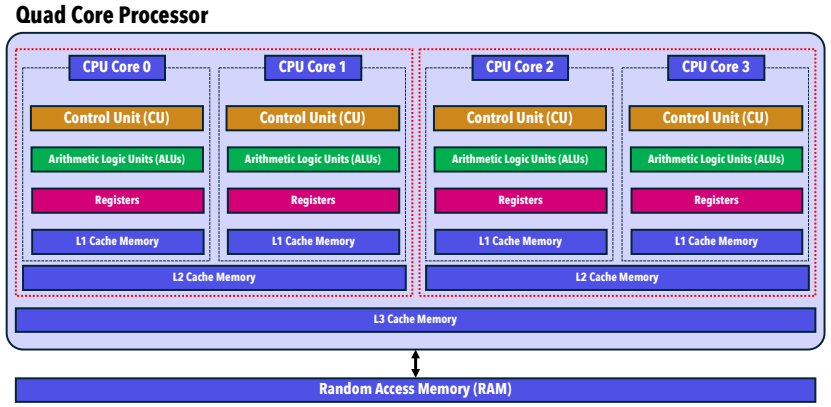
- **CPU** (Central Processing Unit),
- **MCU** (Micro Controller Unit),
- **GPU** (Graphic Processing Unit),
- **TPU** (Tensor Processing Unit),
- **NPU** (Neural Processing Unit),
- **LPU** (Language Processing Unit),
- **DSP** (Digital Signal Processor),

… and System On Chips (**SoCs**) embedding MCU, GPU, NPU, DSP…

Central Processing Unit (CPU)

A CPU is a **general-purpose** processor based on the von Neumann architecture.
The greatest benefit of CPUs is their **flexibility**. One can load any kind of software on a CPU for many different types of applications.
Are Optimized for low latency.

MIMD (Multiple Instruction, Multiple Data)
Each CPU core executes different instructions on different data streams simultaneously



- ALU allows arithmetic (add, subtract, etc.) and logic (AND, OR, NOT, etc.) operations to be carried out,
- Registers are tiny (few bytes), ultra-fast storage locations, used to hold the data the ALU is processing,
- Lx (L1,L2,L3), are small (~32–64 KB (L1), ~256 KB–1 MB (L2) and, ~2–64 MB (L3)), fast memory layers between registers and RAM, storing frequently accessed data.
  - L1 (1-4 cycles) closest to the core, split into L1i (instructions) and L1d (data),
  - L2 (10-20 cycles) shared between cores,
  - L3 (100+ cycles) reduce RAM access (100+ cycles) by caching frequent data.
- The control unit controls the ALU, memory, and input/output (IO) functions, which tells them how to respond to the program that's just been read from the memory.

CPU Operations

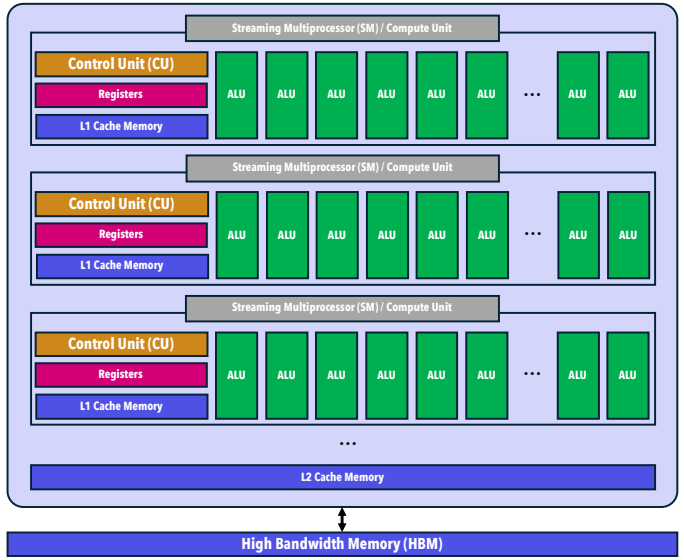
1. FETCH (fetches instruction from memory),
2. DECODE (converts into machine readable format + branching/prediction),
3. EXECUTE (executes the instruction and performs calculations),
4. STORE (stores the result into the memory).

A CPU loads values from memory, performs a calculation on the values and stores the result back in memory for every calculation (sequential). Memory access is slow when compared to the calculation speed and can limit the total throughput of CPUs. This is often referred to as the **von Neumann bottleneck**.

Graphical Processing Unit (GPU)

GPUs integrate thousands of streamlined ALUs (Arithmetic Logic Units) **optimized for parallel workloads**. While originally designed for graphics rendering, their massively parallel architecture makes them general-purpose processors exceptionally efficient for compute-intensive tasks like AI, scientific simulations, and data analytics (GPGPU).

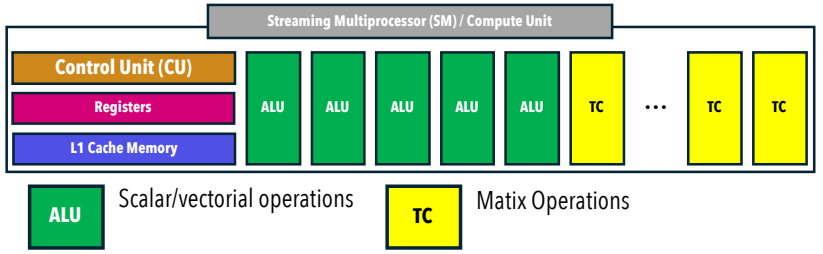
Are optimized for high throughput.



SIMD (Single Instruction, Multiple Data)
All GPU cores execute the same instruction on multiple data points in parallel.

ALUs are **CUDA Cores** in the Nvidia ecosystem.

Tensor Cores (TC) are specialized computation processors optimized for linear algebra computations (e.g., matrix multiplications).



Unit Type	Native Formats	Emulated Formats
CPU ALU	INT8/16/32/64, FP16/32/64	INT4, FP8
GPU ALU	INT8/16, FP16/32	INT4, FP8
Tensor Core	INT4/8, FP8/16	–

GPU Operations

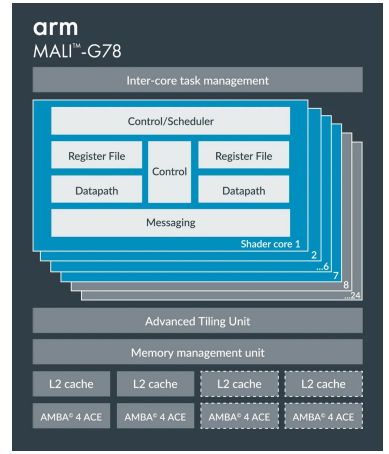
1. FETCH (fetches one instruction into warps of 32 threads. E.g., 10k data → 313 warps spread out to different SMs),
2. DECODE (converts into machine readable format; NO branching/prediction),
3. EXECUTE (32 identical ops /cycle/warp (SIMD)),
4. STORE (stores the result into the memory).

Graphical Processing Unit (GPU) - Examples

Nvidia Edge server Professional GPUs

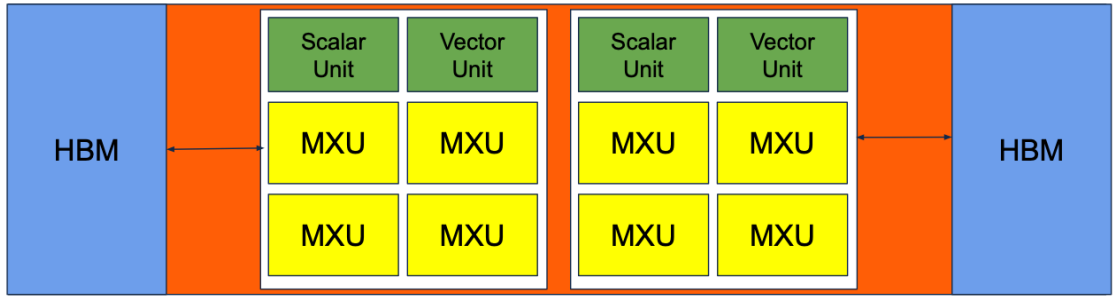
GPU Name	GPU SMs	GPU Cores	VRAM Capacity (GB)	Memory Bus Width (bit)	Memory Speed (Gbps)	Memory Bandwidth (GB/s)	TBP (W)	Launch Date
RTX A4000	48	6,144	16 (GDDR6)	256	14	448	140	Apr 2021
RTX A4500	56	7,168	20 (GDDR6)	320	16	640	200	Nov 2021
RTX A5000	64	8,192	24 (GDDR6)	384	16	768	230	Apr 2021
RTX A5500	80	10,240	24 (GDDR6)	384	16	768	230	Mar 2022
RTX A6000	84	10,752	48 (GDDR6)	384	16	768	300	Oct 2020
RTX A6000 Ada	142	18,176	48 (GDDR6)	384	20	960	300	Dec 2022

In SOCs…



Tensor Processing Unit (TPU)

TPUs are Google's **matrix processors** specialized for **neural network workloads in the Cloud**. This is an Application-Specific Integrated Circuit (ASIC), unlike CPUs and GPUs, TPUs are not general-purpose processors, they are optimized for massive matrix operations used in neural networks at fast speeds.



Google TPU v4 (https://courses.grainger.illinois.edu/cs433/fa2022/projects/Google-TPU.pdf)

TPU Operations

1. HBM feeds weights/activations to the MXU for bulk computations.
2. Scalar Unit manages control flow (e.g., loop over layers).
3. Vector Unit processes element-wise ops (e.g., ReLU, norm).
4. Results stream back to HBM or are reused for next layers.

Matrix Multiply Unit (MXU)

- Is the core of TPU, it replace ALUs/Tensor Cores,
- Built around a massive systolic array (e.g., 128×128 or 256×256) **optimized for INT8/FP16/BF16/FP32 matrix multiplications**,
- Unlike GPU SMs (which handle 32-thread warps), TPUs process entire matrices at once in a deterministic, grid-like dataflow.

Scalar Unit

- Handles control operations (No branch prediction, out-of-order execution, or thread scheduling), address calculations, and non-matrix computations,
- May manage synchronization between different units (MXU, Vector Unit).

Vector Unit

- Handle SIMD operations that are not matrix multiplications (e.g., normalizations, element-wise additions, ReLU activation, pooling, etc.).

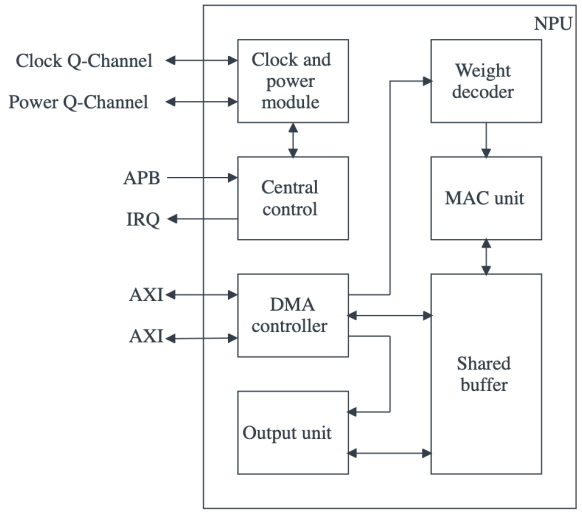
High Bandwidth Memory (HBM)

- Serves as the high-capacity, high-bandwidth memory for the TPU, storing inputs, weights, and intermediate activations (stacked on die).

https://cloud.google.com/tpu/docs/intro-to-tpu
In-Datacenter Performance Analysis of a Tensor Processing Unit, Norman P. Jouppi et al., 2017 https://indico.fccn.pt/event/29/contributions/747/attachments/560/1081/Apresentação%20Google%20Cloud.pdf
https://cloud.google.com/tpu/docs/system-architecture-tpu-vm

Neural Processing Unit (NPU)

NPUs are optimized to perform the mathematical operations that are involved in **neural network processing**, such as matrix multiplication and convolution. NPUs can be integrated into CPUs/MCUs, GPUs, or ASICs, or they can be standalone chips (**edge devices/servers**).



https://developer.arm.com/documentation/102420/0200/Neural-processing-unit-introduction/Description-of-the-neural-processing-unit

NPU Operations

1. DMA fetches weights/inputs ⟶ Shared Buffer.
2. Weight Decoder decompresses weights ⟶ feeds MAC Unit.
3. MAC Unit computations ⟶ outputs to Shared Buffer.
4. Central Control coordinates the next layer.

Central Control

- Manages task scheduling and coordination between all units,
- Decodes NPU instructions (from the host CPU),
- Handles synchronization and error handling.

DMA (Direct Memory Access)

- Transfers weights/activations to/from DRAM (e.g., LPDDR) without CPU intervention,
- Maximizes throughput by prefetching data into the Shared Buffer.

Weight Decoder

- Decodes compressed weights (e.g., sparse/pruned models),
- Reorganizes weights for efficient feeding into the MAC Unit.

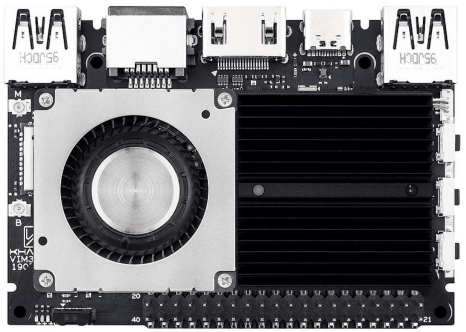
MAC unit (Multiply/Accumulate)

- Performs INT8/INT16/FP16 matrix multiplications and convolutions,
- Processes simple element-wise ops (ReLU, convolutions, pooling, etc.).

Shared Buffer

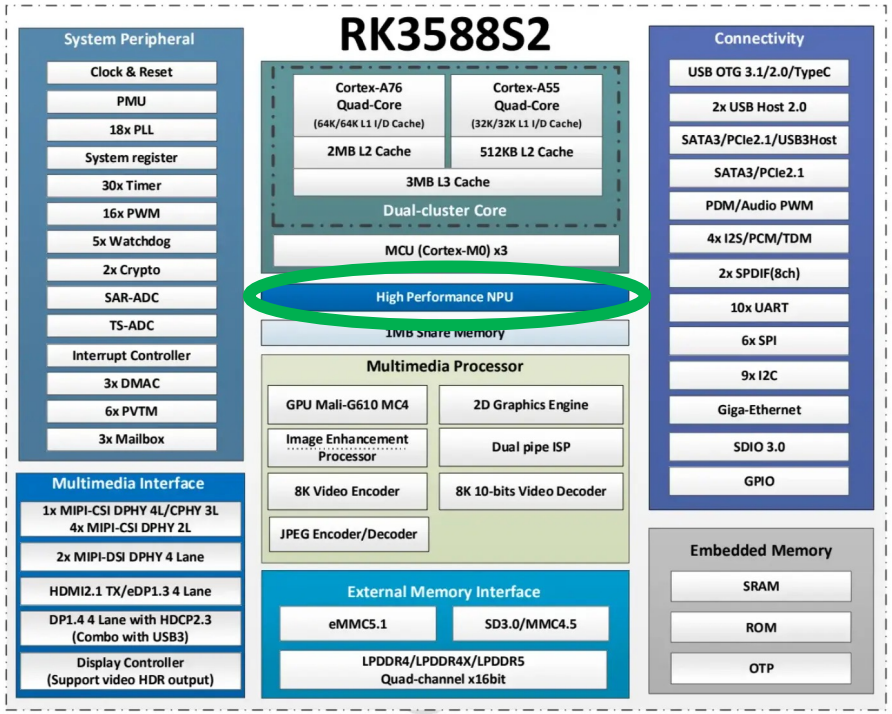
- Stores inputs, weights, and intermediate results for low-latency access,
- Shared across MAC units for data reuse.

Neural Processing Unit (NPU) – SOCs Examples

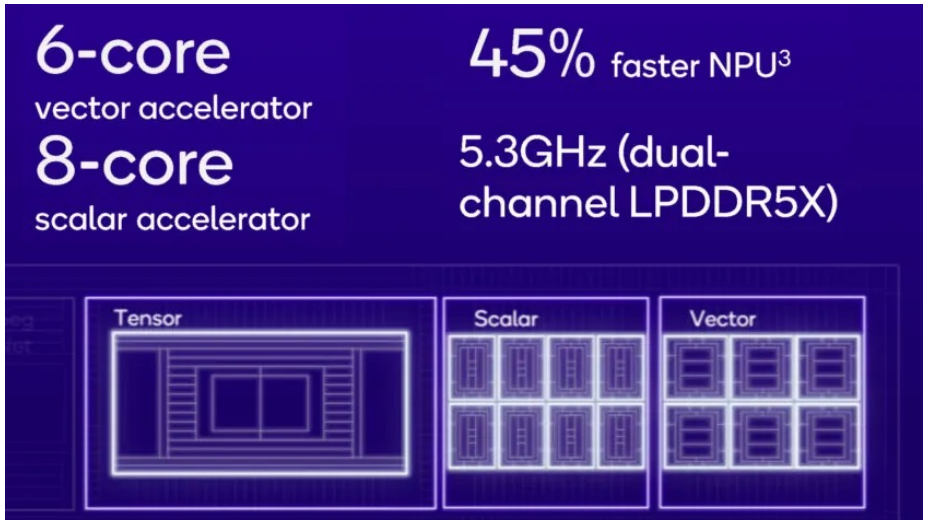


Khadas Edge2

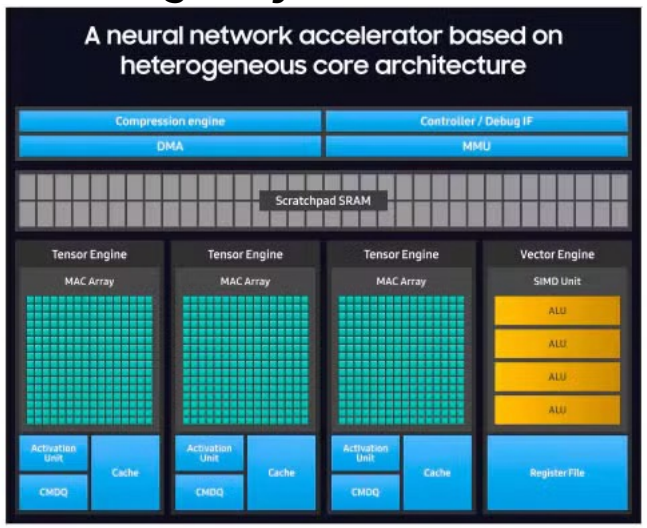
Radxa Rock 5C



Qualcomm's Snapdragon 8 Elite



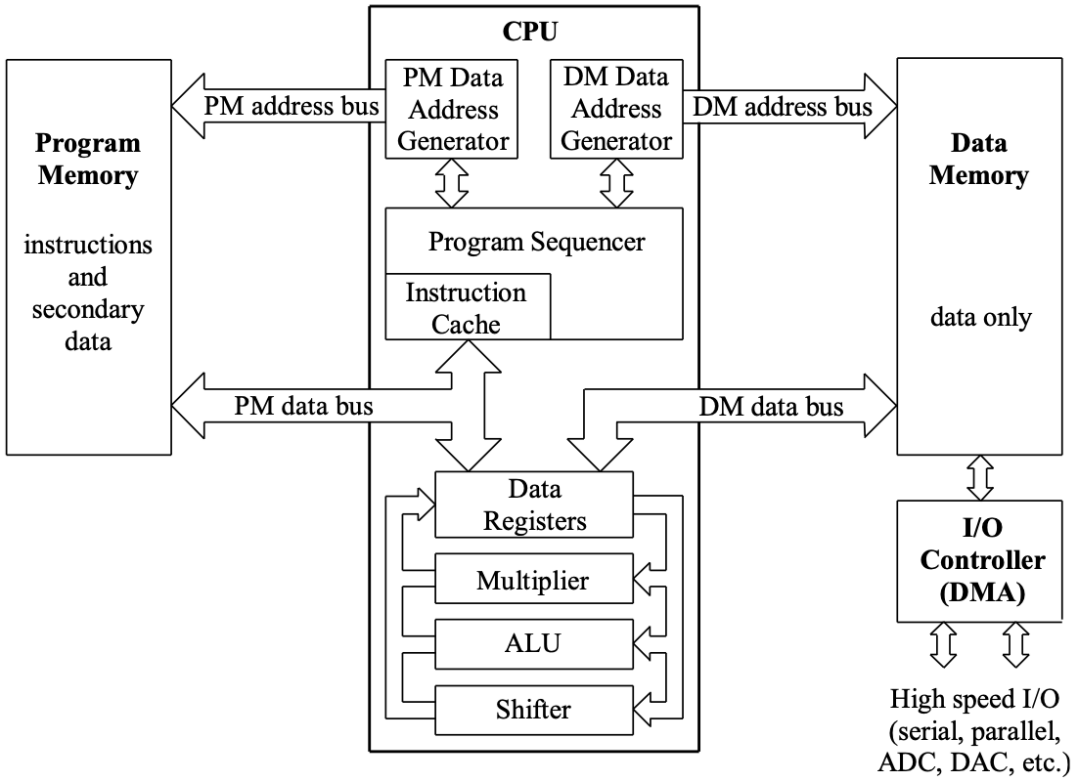
Samsung's Exynos series



… and much more!

Digital Signal Processor (DSP)

Digital Signal Processors (DSP) take **real-world signals like voice, audio, video, temperature, pressure, or position** that have been digitized and then mathematically **manipulate them**. A DSP is designed for performing mathematical functions like "add", "subtract", "multiply" and "divide" very quickly.



https://multimed.org/student/zps/en/02-Architecture.pdf

Data Registers

- Small, fast storage elements within the DSP used to hold operands i.e., data values that the processor operates on.

Multiplier

- Takes two operands from registers, multiplies them, and writes the result back into a register.

ALU (Arithmetic Logic Unit)

- Performs basic arithmetic (addition, subtraction) and logic operations (AND, OR, XOR, NOT). ALUs are designed for high-speed, single-cycle operations.

Shifter

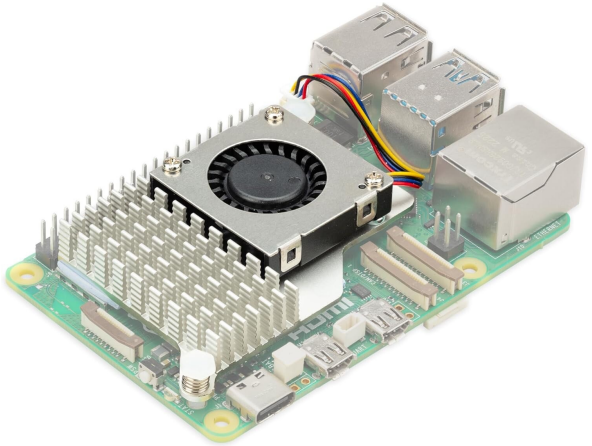
- Allows bit-level manipulation (shifting left or right by multiple bits, rotating, extracting, or aligning data segments) in one cycle.

Digital Signal Processor (DSP) – SOCs Examples

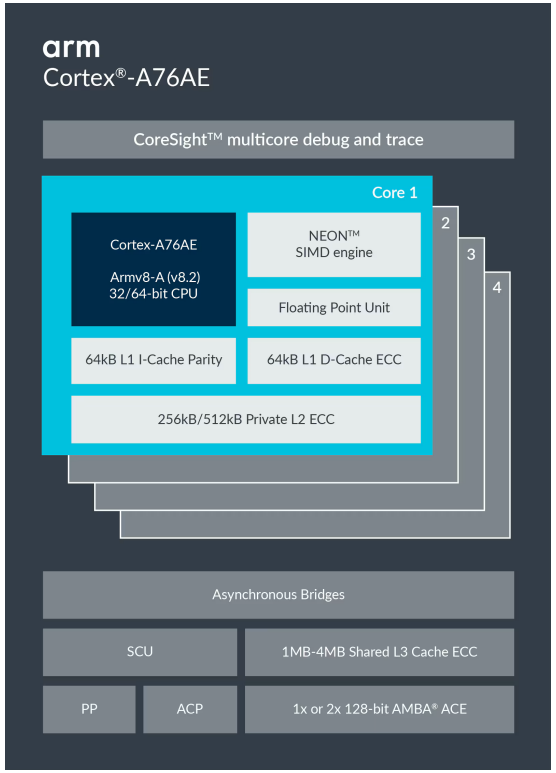
Arm Neon technology (Cortex-A) is an advanced Single Instruction Multiple Data (SIMD) architecture extension used to accelerate signal processing algorithms and functions.

Cortex-M processors with DSP provide a high level of signal processing and integer performance,

Cortex-R processor instruction set includes enhanced DSP instructions to support improved execution performance for arithmetic operations.



RPi5 Arm Cortex-A76 processor



… and DSP processors in mobiles (Texas Instruments, Samsung, Qualcomm, etc.)

FLOPS Demystified

FLOPS = Floating Point Operations Per Second (≠ for FP64,FP32, FP16, BF16),
OPS = Integer Operations Per Second (≠ for INT8, INT4, INT16).
Measure computational performance, especially in CPUs/GPUs for scientific, AI, and HPC tasks.

Key Units

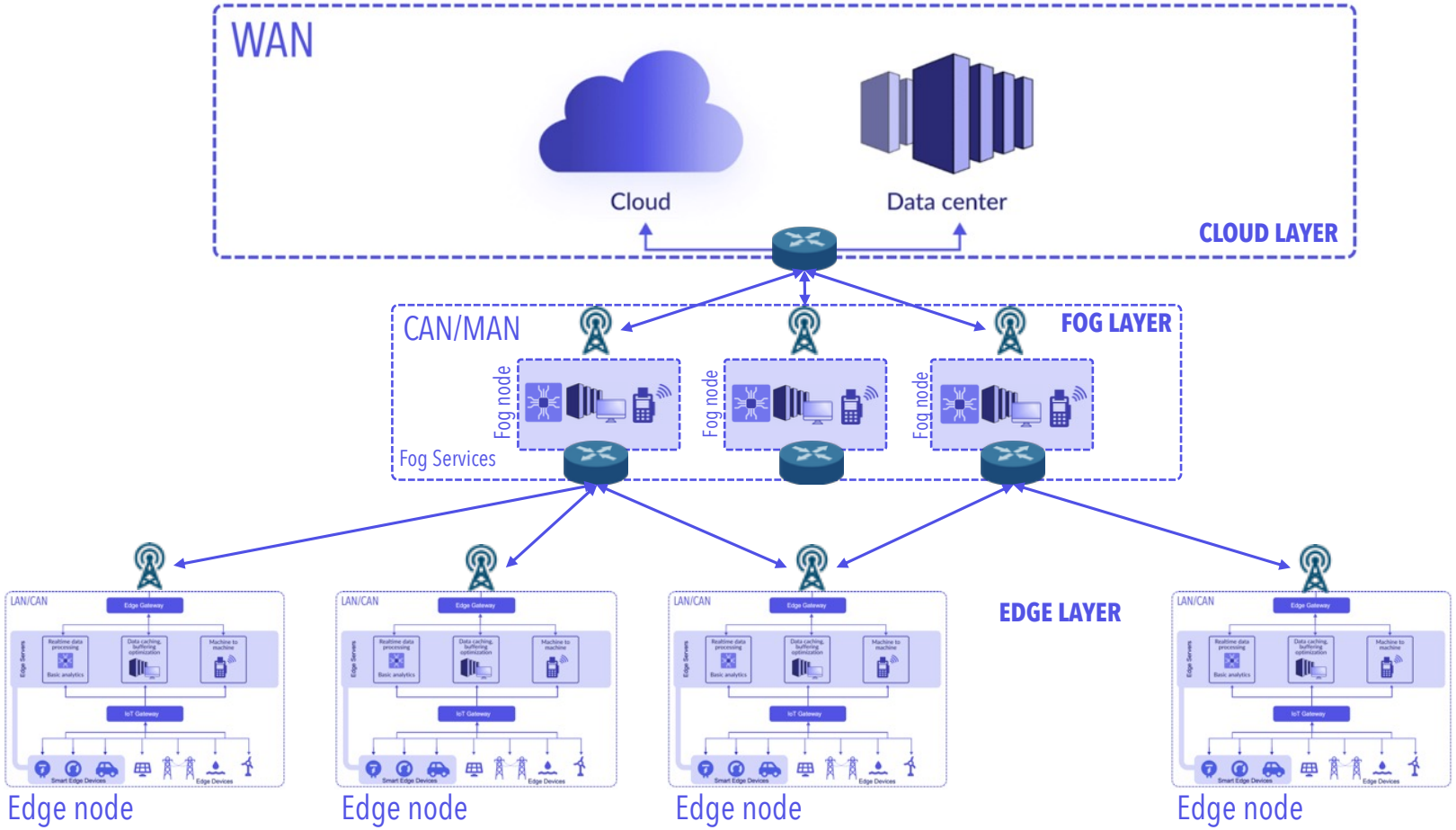
MOPS / MFLOPS = 1 million (10^6) OPS/FLOPS.
GOPS / GFLOPS = 1 billion (10^9)
TOPS / TFLOPS = 1 trillion (10^{12})
POPS / PFLOPS = 1 quadrillion (10^{15})

$$(FL)OPS = cores * \frac{cycles}{seconds} * \frac{(FL)OPs}{cycle}$$

Examples

Hardware Solution	Type	FP64	FP32	FP16	BF16	FP8	INT8	INT4	Memory	Memory Bandwidth	Power (Watts)	Typical Deployment	Use-Case
NVIDIA H100 SXM	GPU	34 TFLOPS 67 TFLOPS (TC)	67 TPLOPS (single precision)	1.98 PFLOPS(TC)		3.9 PFLOPS (TC)	3.9 POPS (TC)	-	80 GB HBM3	3 TB/s	700	Cloud	High-Performance Computing (HPC), AI training, scientific simulations
AMD MI300X	GPU	81.7 TFLOPS (vector) 163.4 TFLOPS (matrix)	163.4 TFLOPS	1.307 PFLOPS		2.615 PFLOPS	2.615 POPS	-	192 GB HBM3	5.3 TB/s	750	Cloud	AI training, HPC, data analytics
Google TPU v4	TPU	-	-	-	275 TFLOPS	-	275 TOPS	-	32 GB HBM	1.2 TB/s		Cloud	Machine learning, AI inference
Intel Habana Gaudi2	AI Accelerator	-	11 TFLOPS		432 TFLOPS	865 TFLOPS	-	-	96 GB HBM2e	2.4 TB/s	600	Cloud	Deep learning training, AI inference
NVIDIA L4	GPU	0.473 TFLOPS	30.3 TFLOPS	121 TFLOPS 242 TFLOPS (sparsity)		485 TFLOPS (sparsity)	485 TOPS	-	24 GB GDDR6	300 GB/s	72	Fog/Edge Servers	AI inference, video analytics
NVIDIA A2	GPU	-	4.5 TFLOPS	18 TFLOPS (TC)		-	36 TOPS (TC)	72 TOPS (TC)	16 GB GDDR6	200 GB/s	60	Fog/Edge Servers	AI inference, edge computing
Qualcomm Cloud AI 100	AI Accelerator	-	-	-	-	-	400 TOPS	-	16GB LPDDR4X		75	Edge Devices	AI inference, cloud AI
RTX A6000 Ada	GPU	1.42 TFLOPS	91.1 TFLOPS	362 TFLOPS	733 TFLOPS	1.46 PFLOPS	1.47 POPS	1.45 POPS	48 GB GDDR6	768 GB/s	300	Fog/Edge Servers	AI inference, professional visualization
AMD Versal AI Edge (VC2802)	AI Accelerator	-	-	-	101 TFLOPS	-	202 TOPS (dense) 405 TOPS (sparsity)	-			75	Edge Devices	AI inference, edge computing
Intel Movidius Myriad X	VPU	-	-	-	-	-	4 TOPS	-	2.5 MB On-Chip		1.5	Edge Devices	AI inference, computer vision
NVIDIA Jetson AGX Orin	AI Accelerator	-	5.3 TFLOPS	-	-	-	275 TOPS	-	64 GB LPDDR5	204.8 GB/s	15-60	Edge Devices	AI inference, robotics, autonomous machines
Hailo-8	AI Accelerator	-	-	-	-	-	26 TOPS	-	8 MB SRAM		2.5	Edge Devices	AI inference, computer vision
Samsung Exynos 2200	NPU	-	-	-	-	-	17.4 TOPS	-				Edge Devices	AI inference, mobile computing
Google Coral TPU	AI Accelerator	-	-	-	-	-	4 TOPS	-	8 MB SRAM		2	Edge Devices	AI inference, edge computing
Khadas Edge 2	AI Accelerator	-	-	-	-	-	6 TOPS	-	8 GB LPDDR4		10	Edge Devices	AI inference, edge computing

Edge AI & AIoT



Large scale AI Models,
Many Watts,
HBMs,
Many Data types.

MODEL LEARNING
INFERENCES

MODELS TRANSFORMATION, SIMPLIFICATIONS

Edge AI, AIoT

Small AI Models,
Few Watts,
LBMs,
Few Data types.

INFERENCES

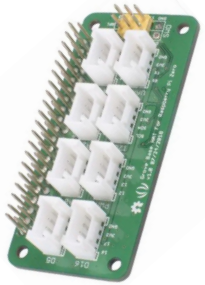
TinyML

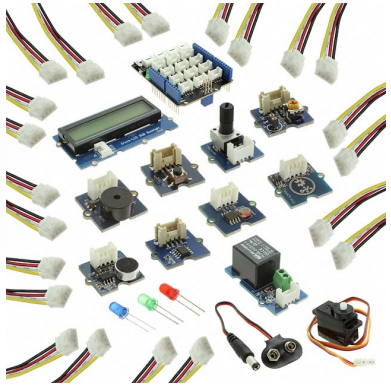
Project Assignment & Development

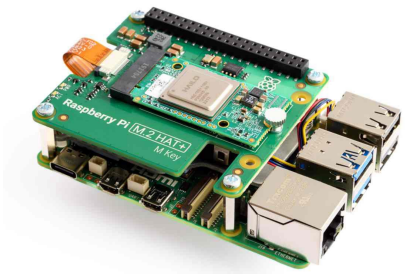
- **Groups of 4 students** (2 IOT-CPS + 2 IA-ID) → 9 groups.
- The project must, where applicable
  - Build on or align with the projects developed during the first bimester (IOT-CPS) → Sensor fusion,
  - Build on or align with the PER subjects,
  - Build on or align with professional projects (FISA) → You bring the necessary equipment,
  - Build on or align with the project of the course "Intelligent Autonomous Systems".
- **Objectives**
  - NN model development,
  - Optimisations,
  - Benchmarks.

Project Assignment & Development

- **Computers**
  - NXP board with NPU (~10 boards),
  - RPi0 (~10 boards),
  - RPi4 (~7 boards),
  - RPi5 + Hailo (5 boards, more to come),
  - Nvidia Jetson Orin (6 units),
  - Google Coral TPU (2 units).
- **Sensors**
  - RGB cameras (NXP + RPi),
  - RGBD cameras (Intel D435i, Luxonis OAK-D),
  - I2C environmental sensors + Rpi hats.







Project Assignment & Development

- **I consider you as engineers and I am your manager,**
- **I will evaluate your work consequently.**

- As such:
  - You are autonomous,
  - You do not ask your manager to fix your issues/problems!
  - Whenever you encounter a problem, you must be able to understand and explain it … and propose alternatives.

Project Assignment & Development

- **Report (15p max)**
  - Context of the project and objectives,
  - Identification of the sensors, their sources of uncertainty, and working ranges,
  - Sensor data pre-processing stage (Data Ingestion Pipeline…),
  - NN architecture and computation graph description,
  - Model optimisation process (quantization, pruning, etc.),
  - Benchmark results (you define relevant criteria!),
  - Github repo and documentation (for results replication).

- **POC presentation**
  - 15mn presentation,
  - 5mn questions.