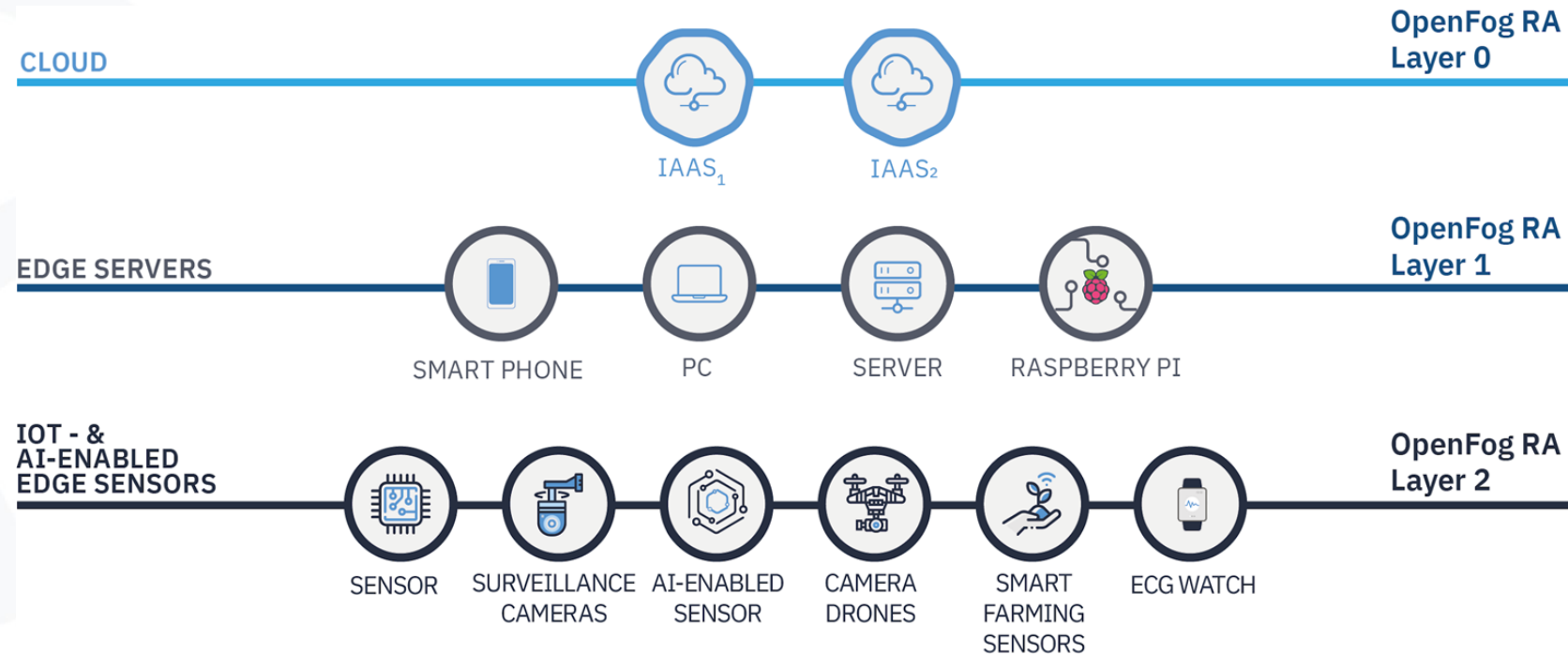Context and setting the scene:
Use case Challenges and needs

Danilo Ardagna,
Politecnico di Milano
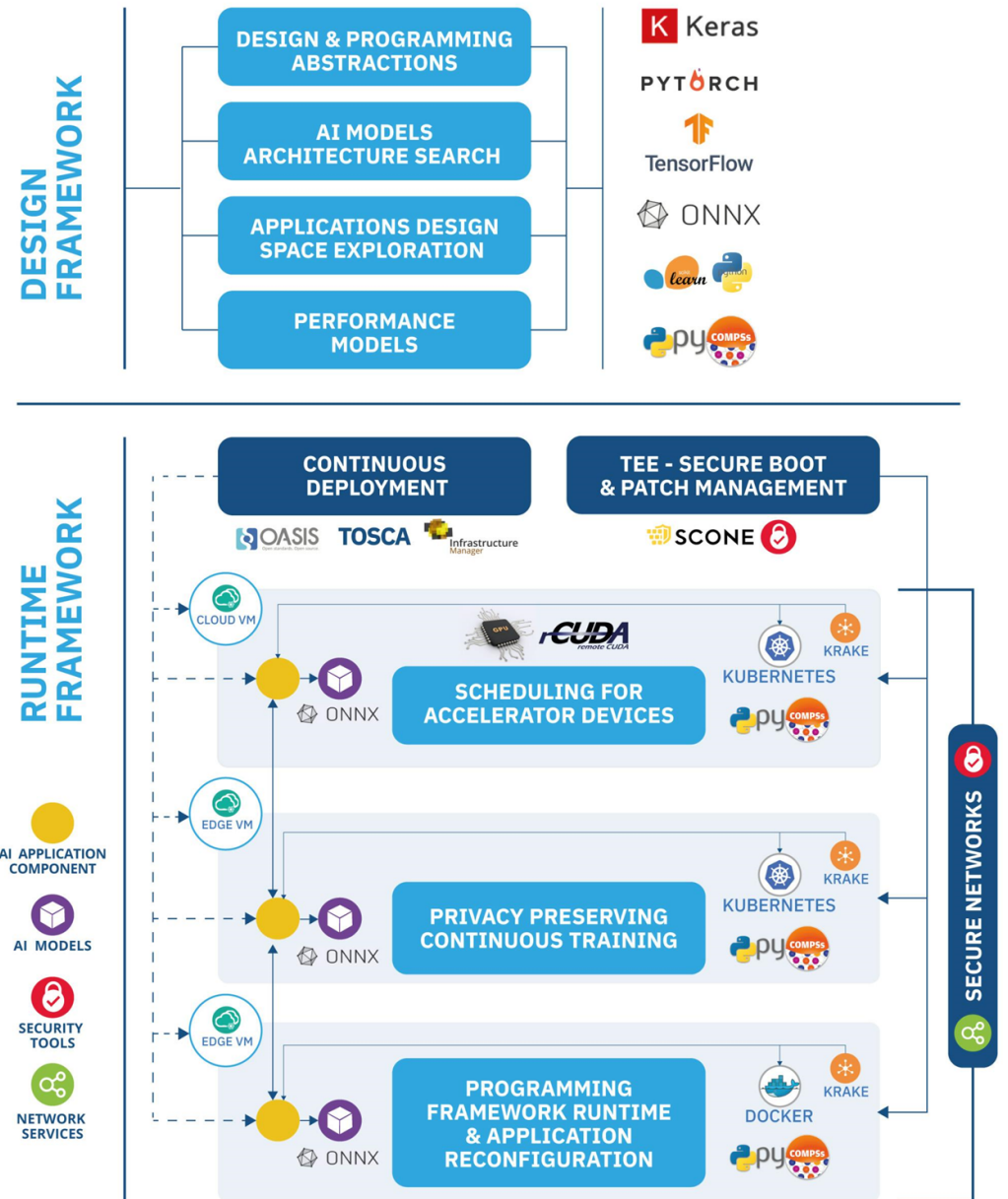
- By 2026, AI worldwide market will approach $900 billion (CAGR 18.6%[1]) while edge computing will reach $324 billion (CAGR 13.6%[2])

- AI needs resources at the edge of the network

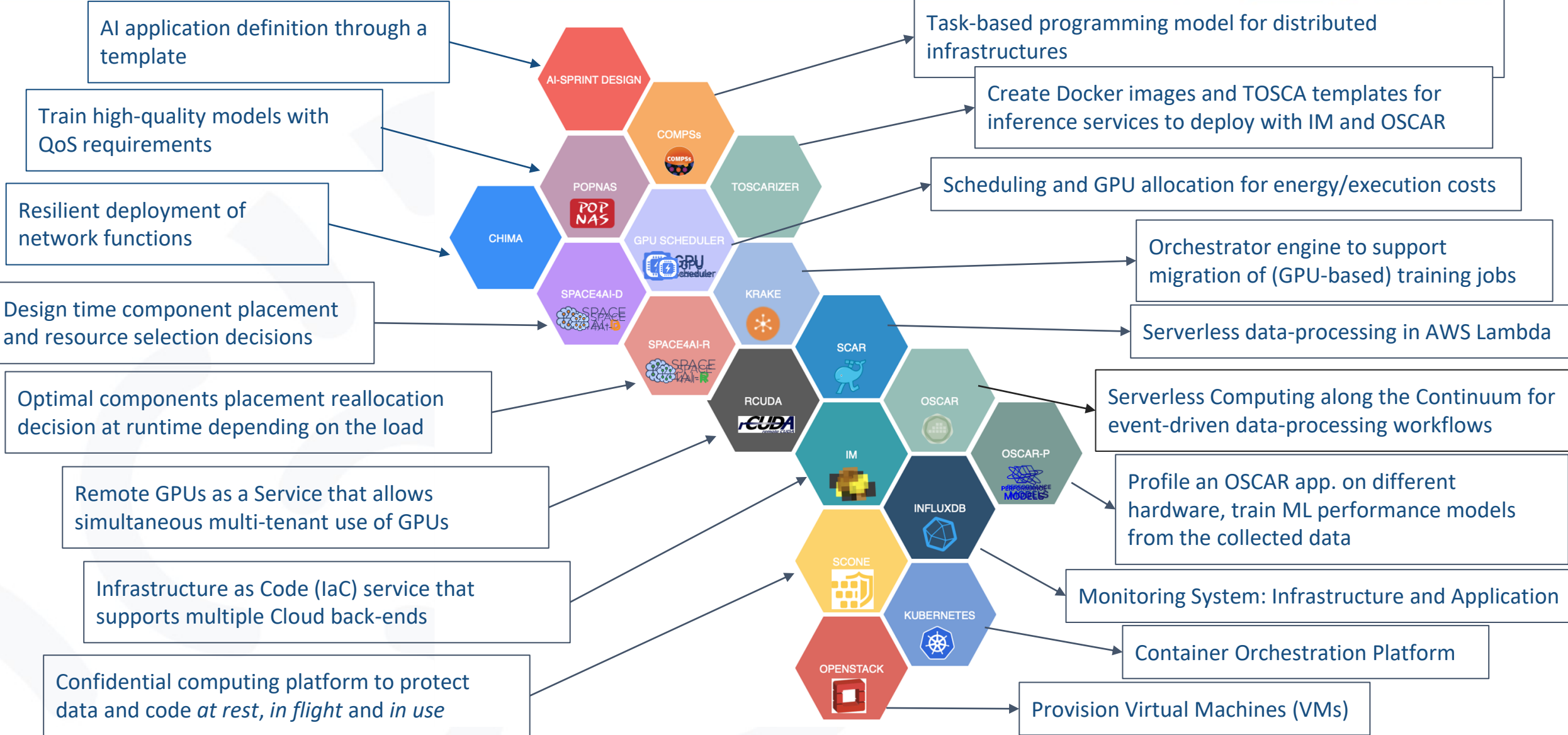- New challenges from the infrastructural perspective

[1]IDC Semiannual Artificial Intelligence Tracker, July 2022

[2]IDC Worldwide Edge Spending Guide, August 2022

CLOUD — IAAS₁ — IAAS₂ — OpenFog RA Layer 0

EDGE SERVERS — SMART PHONE — PC — SERVER — RASPBERRY PI — OpenFog RA Layer 1

IOT - & AI-ENABLED EDGE SENSORS — SENSOR — SURVEILLANCE CAMERAS — AI-ENABLED SENSOR — CAMERA DRONES — SMART FARMING SENSORS — ECG WATCH — OpenFog RA Layer 2

AI-SPRINT objectives

- Simplified programming models

- Automated deployment and dynamic reconfiguration

- Secure execution of AI applications

- Highly specialized building blocks for privacy preservation, distributed training, and architecture enhancement

- Open source

Main Technological Components

AI-SPRINT
AI IN SECURE PRIVACY-PRESERVING COMPUTING CONTINUUM

AI application definition through a template

Train high-quality models with QoS requirements

Resilient deployment of network functions

Design time component placement and resource selection decisions

Optimal components placement reallocation decision at runtime depending on the load

Remote GPUs as a Service that allows simultaneous multi-tenant use of GPUs

Infrastructure as Code (IaC) service that supports multiple Cloud back-ends

Confidential computing platform to protect data and code *at rest*, *in flight* and *in use*

Task-based programming model for distributed infrastructures

Create Docker images and TOSCA templates for inference services to deploy with IM and OSCAR

Scheduling and GPU allocation for energy/execution costs

Orchestrator engine to support migration of (GPU-based) training jobs

Serverless data-processing in AWS Lambda

Serverless Computing along the Continuum for event-driven data-processing workflows

Profile an OSCAR app. on different hardware, train ML performance models from the collected data

Monitoring System: Infrastructure and Application

Container Orchestration Platform

Provision Virtual Machines (VMs)

AI-SPRINT DESIGN
COMPSs
POPNAS
TOSCARIZER
CHIMA
GPU SCHEDULER
SPACE4AI-D
KRAKE
SPACE4AI-R
SCAR
RCUDA
OSCAR
IM
OSCAR-P
SCONE
INFLUXDB
KUBERNETES
OPENSTACK

Personalised Healthcare

Developing an automated system for personalised stroke risk assessment and prevention.
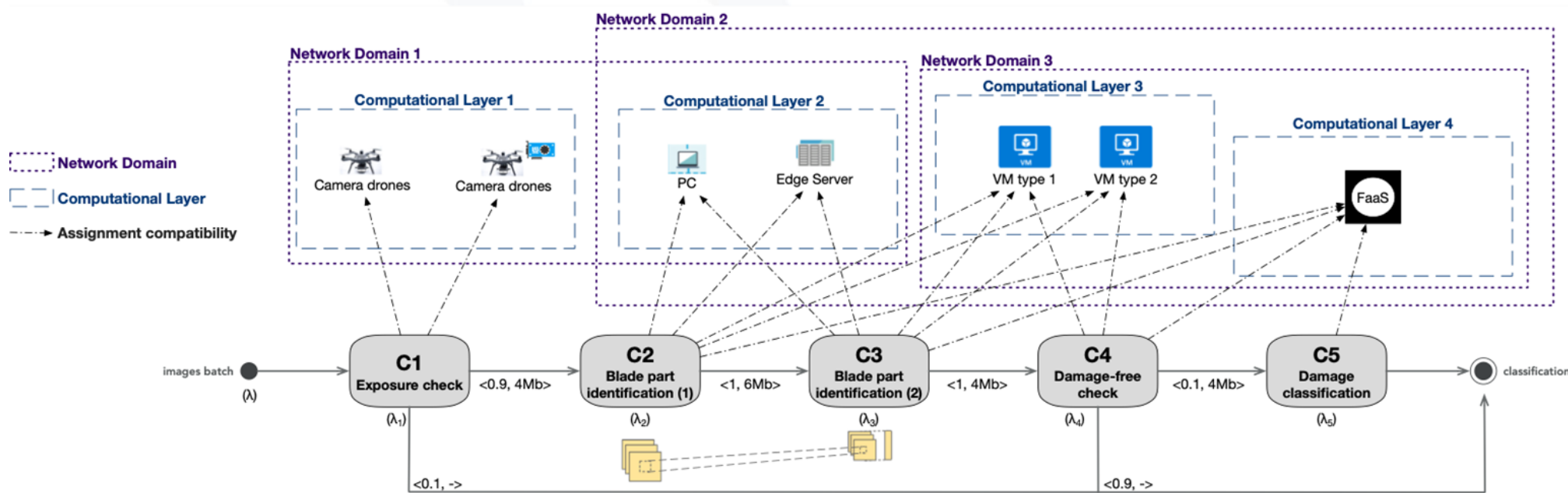

Maintenance & Inspection

Creating an infrastructure that reduces downtime and revenue losses caused by degenerative asset performance.

Farming 4.0

Delivering edge and intelligent sensors to optimise phytosanitary treatments.

https://ai-sprint-project.eu/ai-sprint-alliance