Designing Reconfigurable Intelligent Systems with Markov Blankets

TU Wien: Boris Sedlak, Victor Casamayor Pujol, Praveen Kumar Donta, Schahram Dustdar

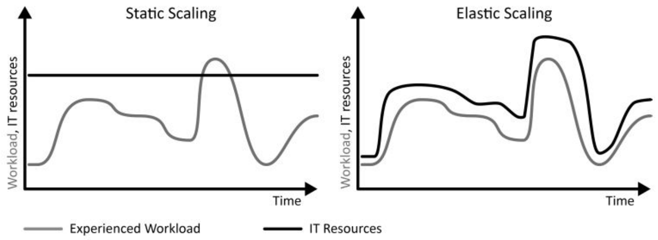
01/12/2023

WWW.TEADAL.EU

About

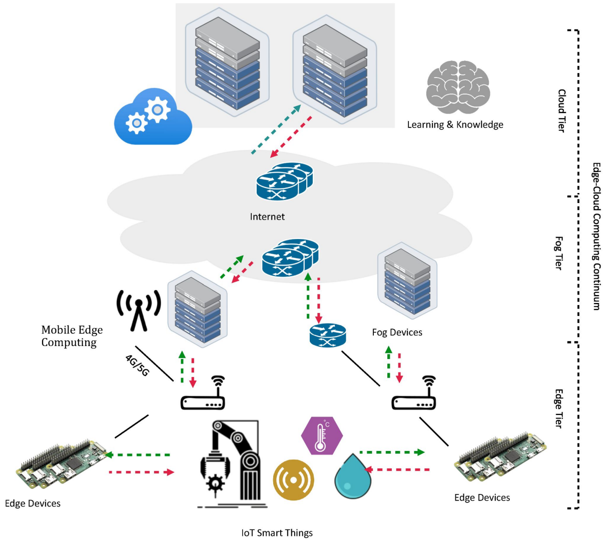
Problem Statement - Environment

https://docs.dynatrace.com/docs/platform-modules/automations
https://www.skylinesacademy.com/blog/2020/3/6/az-900-cloud-concepts-scalability-and-elasticity



- What **elasticity** used to be in the Cloud
- Service Level Objectives (**SLOs**)

- Computing Continuum (**CC**) [1]
- What can elasticity mean for the **Edge**?



[2]

[1] Dustdar, S., Pujol, V.C., Donta, P.K.: On Distributed Computing Continuum Systems. IEEE Transactions on Knowledge and Data Engineering (2023)
[2] Donta, P.K, Pujol, V.C., Murturi, I, Sedak,B, Dustdar, S., .: Exploring the Potential of Distributed Computing Continuum Systems; Computers (2023)

Problem Statement

How can you ensure **QoS**/QoE in such an environment?

Unknown environmental impact → **causality**

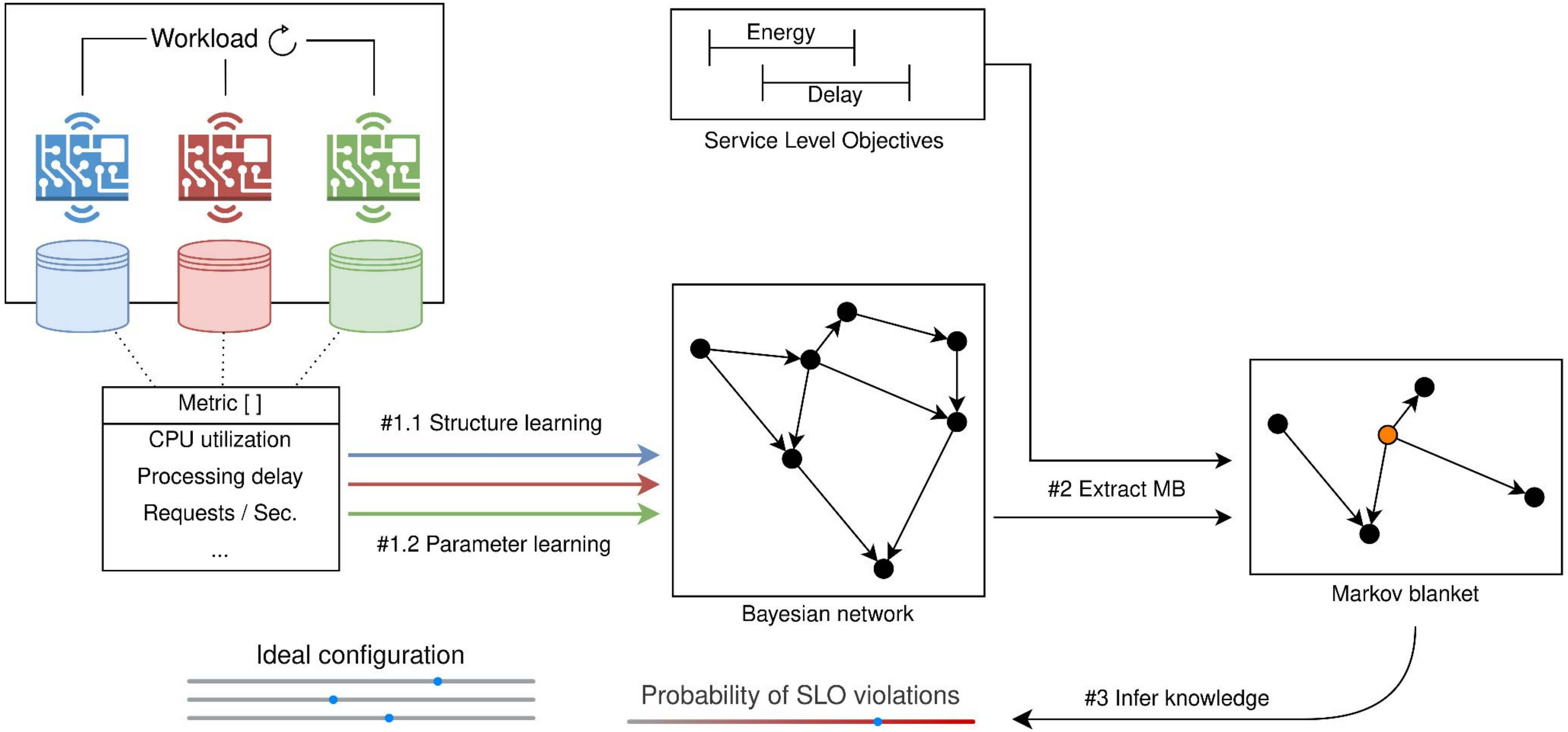
Overall composed complexity → **scope**

Latency/Data towards cloud → **decentralized**

Dynamism of systems → **flexible**

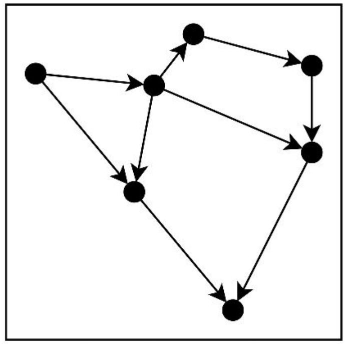
Methodology - Overview

-



Methodology - Details

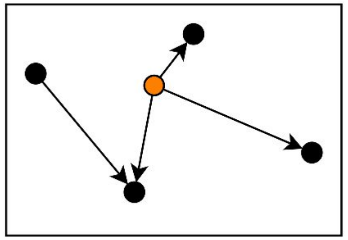
Bayesian Network Learning (BNL)



Bayesian network

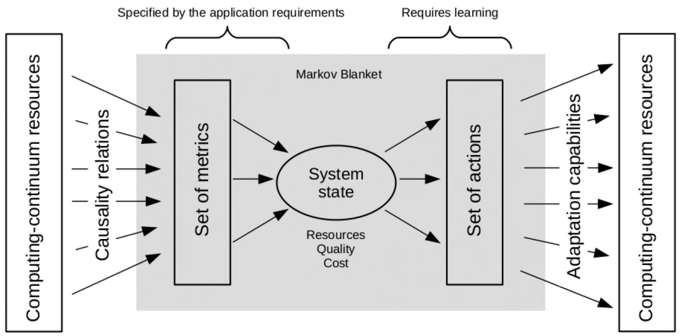
- **Structure Learning**
  Hill-Climb Search (HCS)
  Dir. Acyclic Graph (DAG)
- **Parameter Learning**
  Max. Likelihood Estimation
  Conditional Prob. Table (CPT)

Markov Blanket (MB) Selection



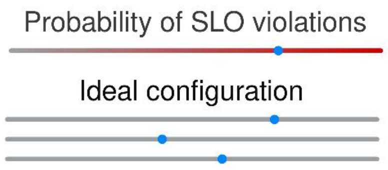
Markov blanket

- Causality filter [3,4]
- Identify relevant variables



[5]

Knowledge Extraction



- P(SLO < x) for all variable combinations
- Find **Bayes-optimal** system configuration

[3] Pearl, J.: Probabilistic reasoning in intelligent systems : networks of plausible inference. San Mateo, Calif. : Morgan Kaufmann (1988)
[4] Friston, K.: Life as we know it. Journal of The Royal Society Interface (Sep 2013)
[5] Casamayor Pujol, V., Raith, P., Dustdar, S.: Towards a new paradigm for managing computing continuum applications. IEEE CogMI (2021)

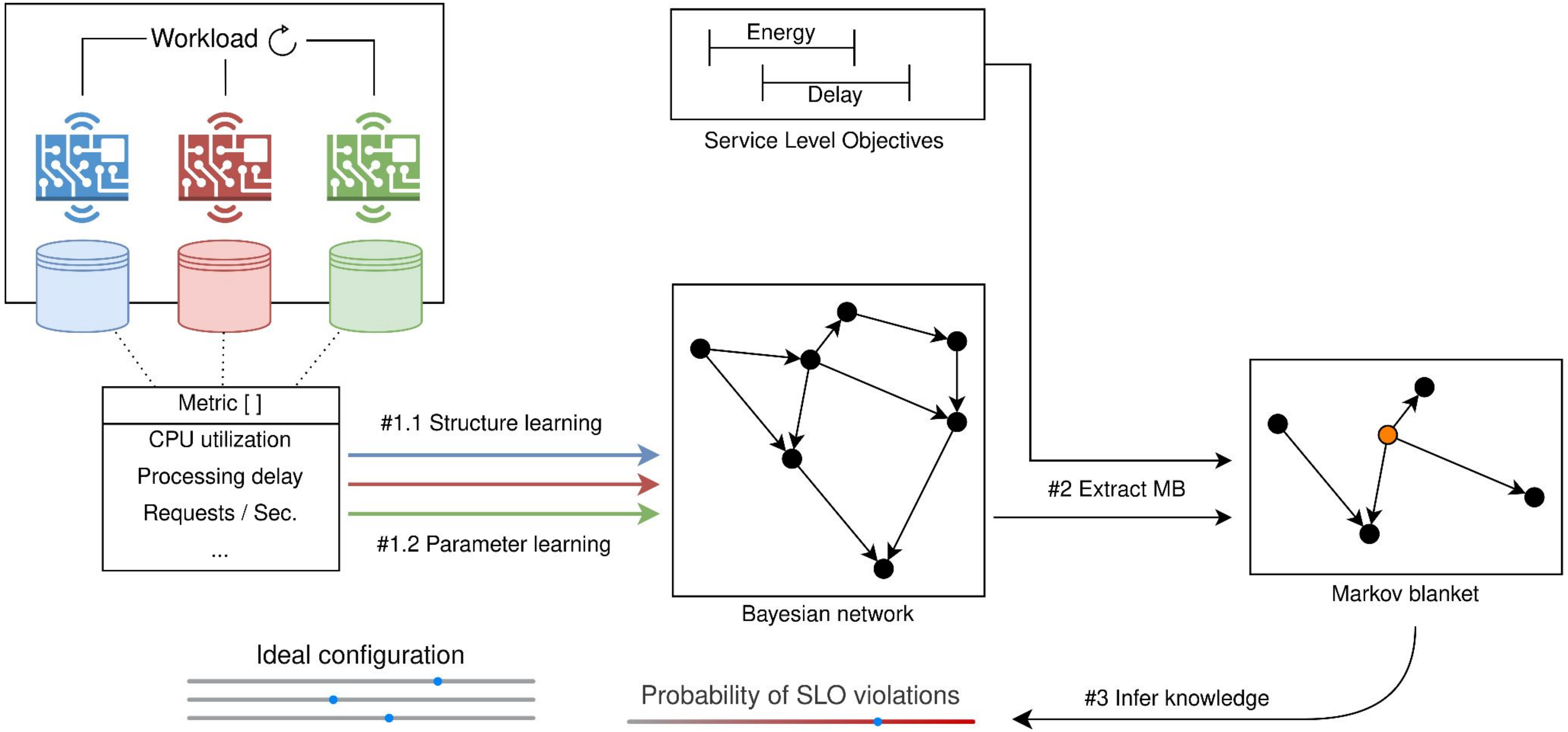
Methodology - Overview

pact → **causality**

mplexity → **scope**

l → **decentralized**

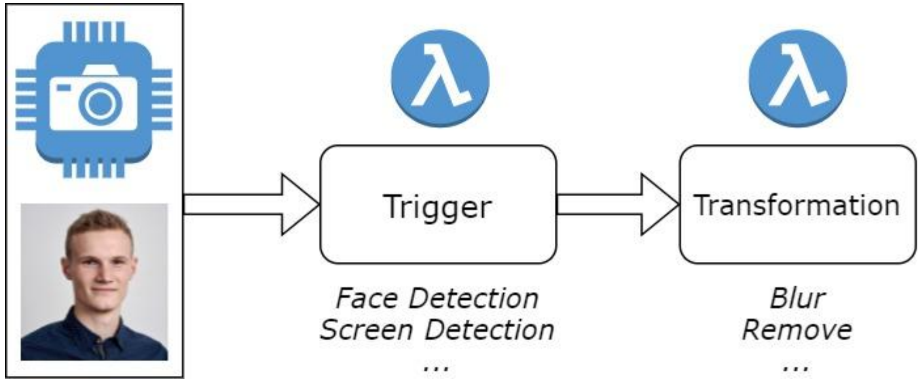
ystems → **flexible**



Case Study - Overview

Distributed video processing on **Edge**

Workload: Privacy-preserving transformation



[6]

11 metrics captured during processing

6 SLOs that must be ensured

[6] Sedlak, B., Murturi, I., Donta, P.K., Dustdar, S.: A Privacy Enforcing Framework for Transforming Data Streams on the Edge. IEEE Transactions on Emerging Topics in Computing (2023)

Case Study - Setup

- Executed on NVIDIA **Jetson Xavier NX**
- **GPU-Acceleration** through NVIDIA CUDA
- Internal **metrics** (10) + consumption

Name	Unit	Description	Param
delay	ms	processing time per frame	No
CPU	%	utilization of the CPU	No
memory	%	utilization of the system memory	No
pixel	num	number of pixel contained in a frame	Yes
fps	num	number of frames received per second	Yes
bitrate	num	number of pixels transferred per second	No
distance	px	relative distance of object between frames	No
transformed	T/F	if the model detected a pattern (i.e., face)	No
GPU	T/F	if the device employs a GPU	No
config	nominal	mode in which the device operates	Yes
consumption	W	energy pulled by the device	No

Table 1: Metrics captured during processing

https://www.nvidia.com/en-sg/autonomous-machines/embedded-systems/jetson-xavier-nx/
https://www.reichelt.com/de/en/wifi-outlet-switch-power-measurement-delock-11827-p262109.html?r=1

Case Study - Setup (2)

network_usage Edge devices have limited network interfaces, and in some cases, limited network bandwidth. Since video streams are transferred over the network, *bitrate* is important to control network congestion.

energy_cons Edge devices are restricted in terms of resources and thus must economize or limit their energy *consumption* while ensuring compliance with the remaining system requirements (i.e., other SLOs).

within_time Video processing introduces a considerable streaming *delay*, which can lead to dropping frames and consequently poorer QoE. Hence, the stream's *fps* can be adjusted to limit/avoid dropping frames.

pixel_distance Measures the quality of the object tracking capacity; we expect the tracked object not to jump, but to have a smooth trajectory. Hence, we define a range for the acceptable *distance*.

transf_success Private or confidential information must not be disclosed; therefore, *transformed* should be maximized to increase the utility of the privacy model transformation.

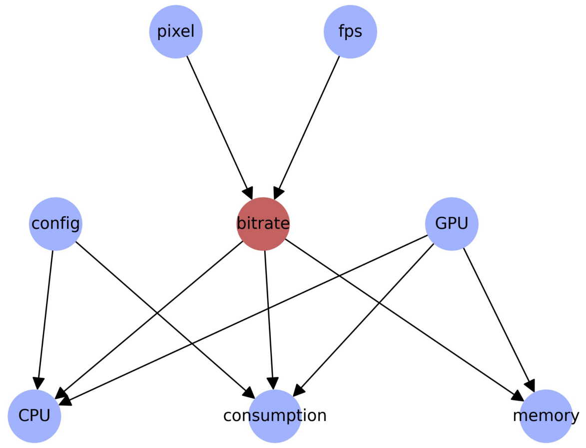
Listing 1: Proposed SLOs for ensuring the service during processing

Case Study - Model Construction

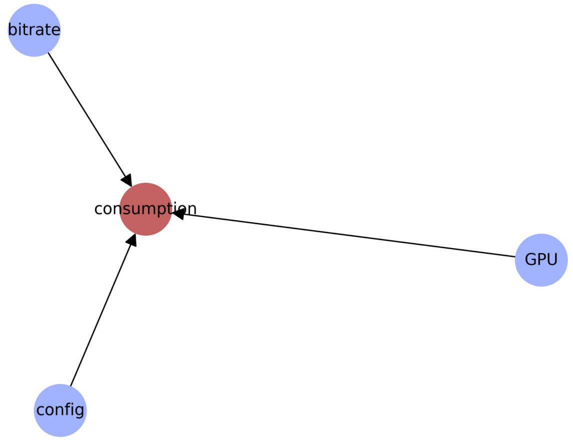
2.5 hours processing → 189,000 metric rows
(Periodically switching *fps* and *pixel*)

Empiric Evaluation of Combinations

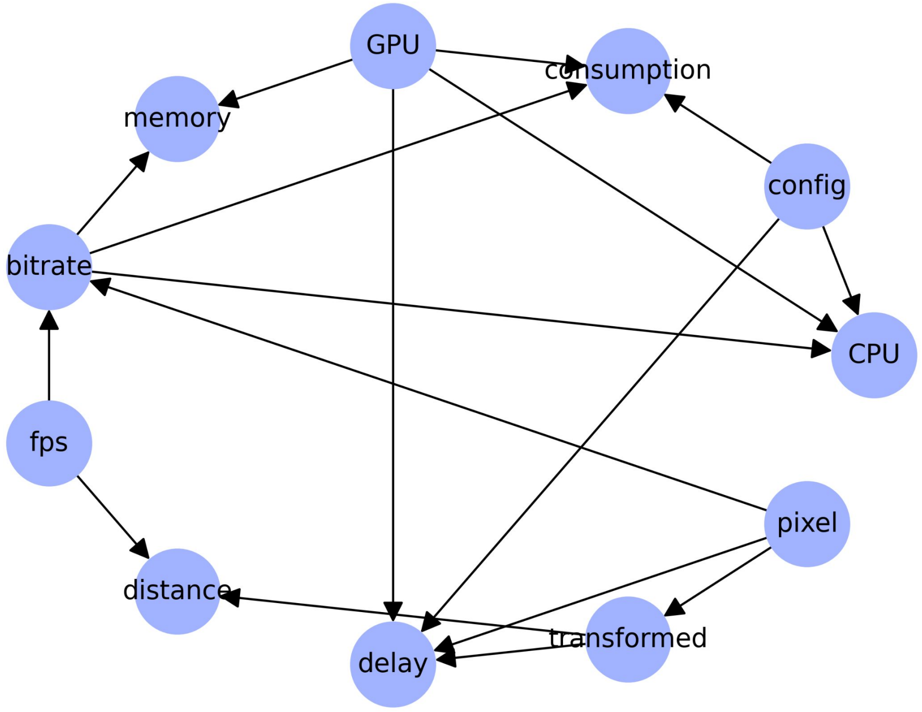
189,000 x 6 config *mode* → 756,000 rows
BNL (= HCS and MLE) → 30 s



MB (**network_usage**)



MB (**energy_cons**)



Full **Bayesian Network**

Case Study - Device Configuration Inference

Parameter Space

- Find device configurations that (maximize) SLO fulfillment
- Configurations **must** include *fps*, *pixel*, *config* mode
  - Respectively 5 (*fps*) * 6 (*pixel*) * 3 (*config*) = **90** configurations
  - Have the format (240p : 20fps : 4C_20W)

Application on Prototype

- Inference for 5 SLOs and 90 configurations; 5 * 90 = 450 queries
- Takes **500** ms on Jetson Xavier NX
- SLO thresholds **parameterizable**; variables **configurable** (e.g., CPU)

Evaluation - Scenarios

Scenario A – Mobile Rendering

- Vehicle capturing street data
- Local transformation & **rendering**
- Live remote inspection

https://www.mosaic51.com/community/best-360-google-street-view-camera/
https://www.amazon.com/Wearable-Recorder-Headband-Streaming-Hands-Off/dp/B0BVPZ95K9

Scenario B – Factory Audit

- Multiple parallel streams provided
- Privacy-preservation (e.g. face, screen)

Scenario	transf_success	distance	network_usage	within_time	energy_cons	GPU
A	$\geq 90\%$	≤ 35	$\leq 8.2\ Mio.\ px/s$	$\geq 95\%$	$min(x)$	No
B	$\geq 98\%$	≤ 60	$\leq 1.6\ Mio.\ px/s$	$\geq 75\%$	$min(x)$	Yes

Table 2: SLO thresholds for the Scenarios (A/B)

Results

From scenarios to device configurations

- **Infer** device configurations (A/B)
- Extend with **Naive**, **Random**

Scenario	Source	Resolution	FPS	Mode	GPU
A	inferred	240p	20	4C_15W	No
	naive	360p	30	6C_20W	
	random #1	120p	16	6C_20W	
	random #2	720p	12	2C_10W	
B	inferred	240p	16	2C_10W	Yes
	naive	180p	26	4C_15W	
	random #1	360p	20	2C_15W	
	random #2	480p	30	6C_20W	

Table 3: Configurations evaluated for comparison

Bayesian Network Learning (BNL)

- Measure **SLO fulfillment** 10m
- *Infer* reported **no** SLO violations

Scenario	Source	transf_success	distance[5]	network_usage	within_time	energy_cons
A	inferred	98%	15 (97%)	2.0 Mio.	100%	6.0W
	naive	100%	10 (100%)	6.9 Mio.	92%	8.0W
	random #1	4%	127 (2%)	0.4 Mio.	100%	7.0W
	random #2	100%	28 (89%)	11 Mio.	100%	6.0W
	average	81%	73 (83%)	6.0 Mio.	81%	7.1W
B	inferred	98%	18(98%)	1.6 Mio.	100%	6.0W
	naive	92%	11(99 %)	1.5 Mio.	100%	6.5W
	random #1	99%	15 (100%)	4.6 Mio.	100%	6.0W
	random #2	100%	10 (100%)	12.3 Mio.	97%	7.5W
	average	81%	73 (86%)	6.0 Mio.	91%	6.7W

Table 4: SLO fulfillment for evaluated scenarios

Summary

- **CC** increases complexity of service provisioning
- Decentralized requirements (i.e., SLOs) assurance
- Causal relations **environment** → **SLO fulfillment**
- **BNL** (1), **MB** Selection (2), and **Inference** (3)

- Prototype with 11 metrics and 5 SLOs
- Inferred device **configurations** conformed with SLOs

TEADAL

TEADAL.EU

@TEADAL_eu

@TEADAL

TEADAL project is funded by the EU's Horizon Europe programme under Grant Agreement number 101070186