Cloud Computing 2026
Lisbon, Portugal

\Orchestrating a brighter world NEC

Agentic Placement of Microservices on the Computing Continuum

Kunal Rao
Researcher
NEC Laboratories America, Inc.
Email: kunal@nec-labs.com

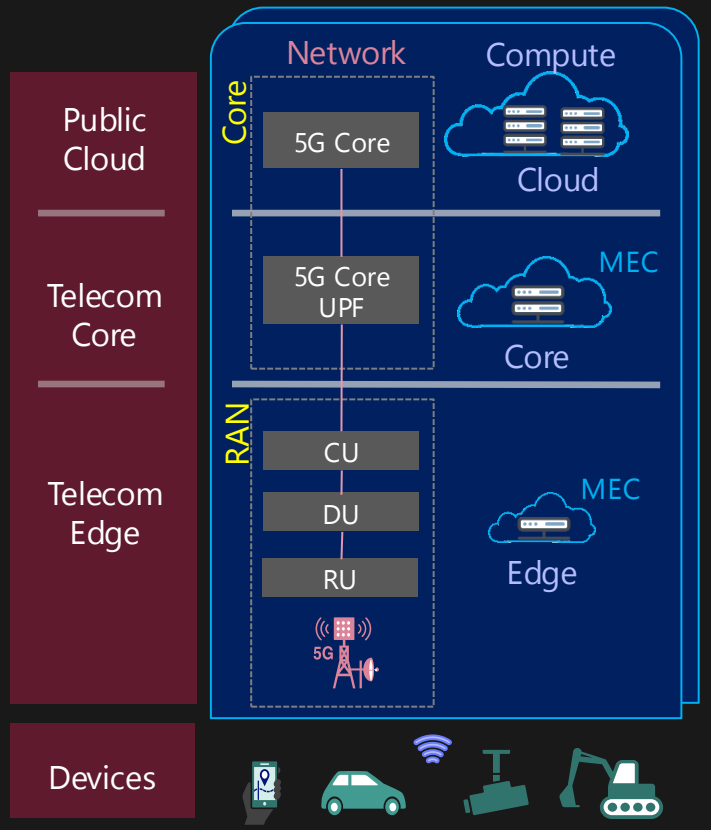
April 22, 2026

Brief bio

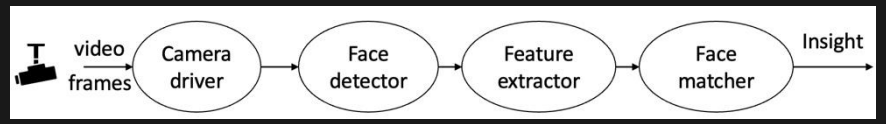
- Work as Researcher at NEC Laboratories America, Inc. in Princeton, NJ
- Current research: Generative AI-based applications and platform, Edge and Cloud computing, leveraging AI/ML models to solve systems problems, gaining utility from AI/ML models for real-world applications
- Past research: High Performance Computing, GPGPU and Xeon Phi computing, Graph analytics, Video Analytics (Computer Vision, Applied Machine Learning)
- 34 granted patents (US and worldwide) and several are published and pending, 35 published papers

Context and Problem (1/2)

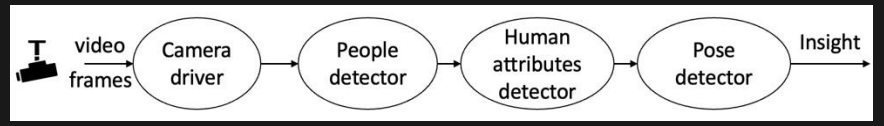
Multi-tiered computing infrastructure



Application pipelines (microservices-based architecture)

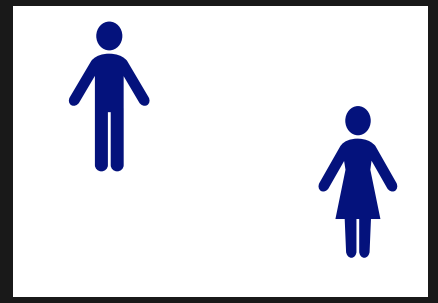


Face recognition application pipeline

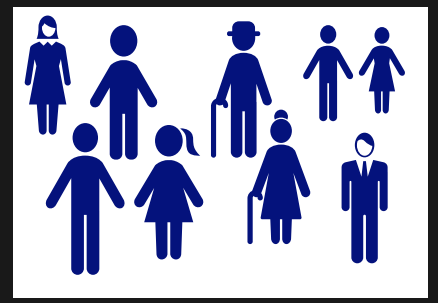


Human attributes detection application pipeline

Video scene (dynamic workload)



Few people

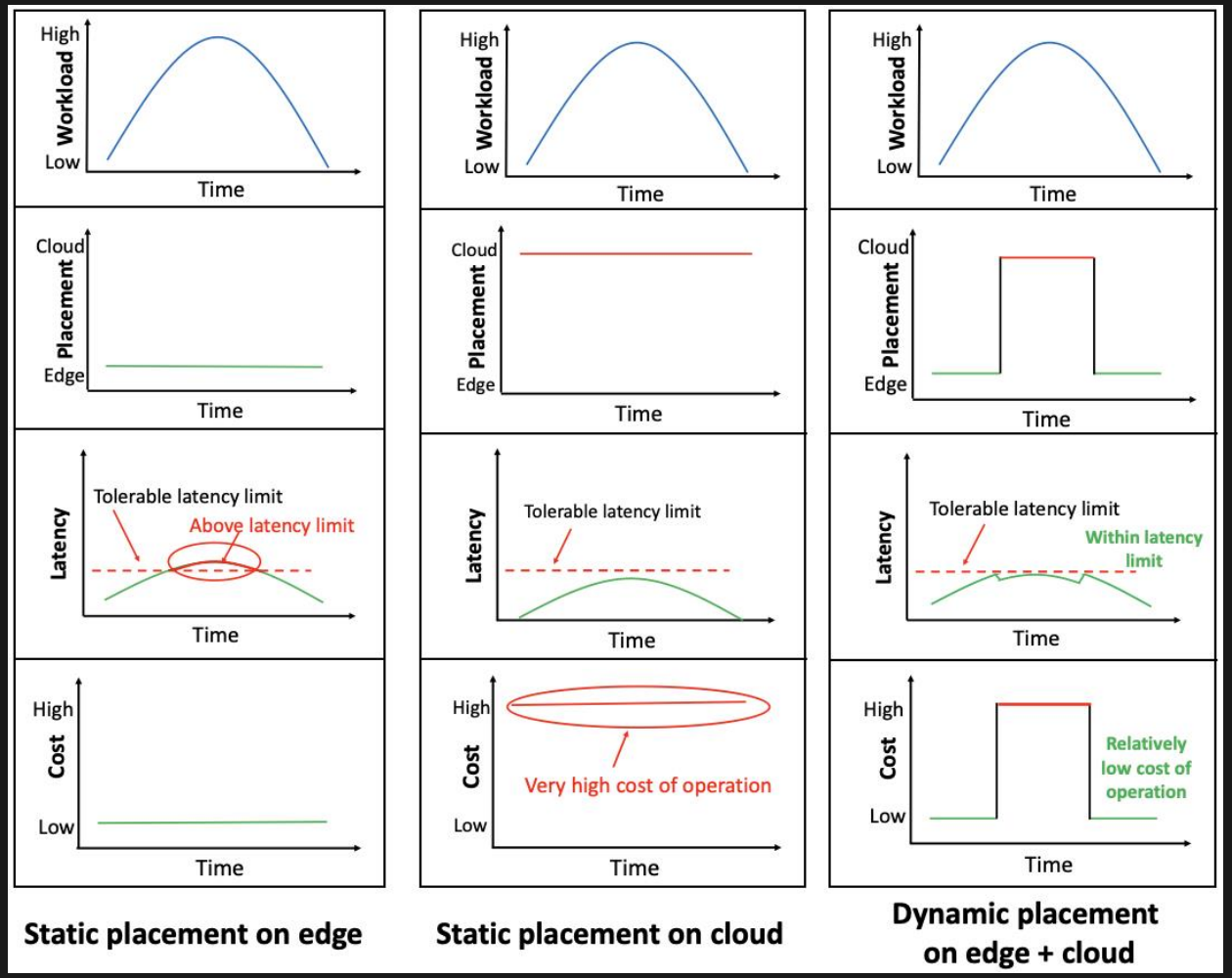


More people

Where to run microservices on this complex multi-tiered infrastructure ?

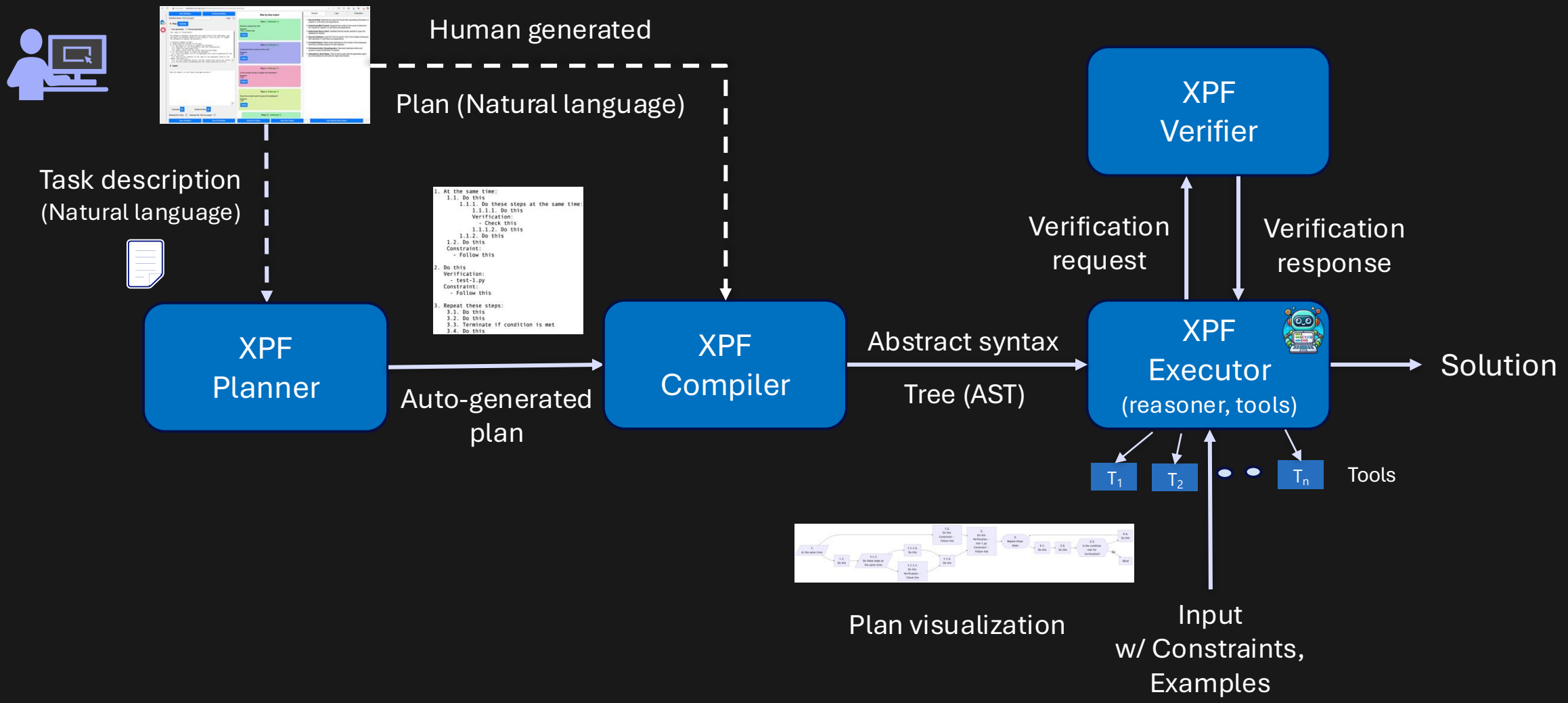
Context and Problem (2/2)

- Static placement on Edge incurs less cost of operation, but may miss latency threshold when workload increases.
- Static placement on Cloud meets latency threshold even for increased workloads, but incurs high cost of operation.
- Dynamic placement on edge+cloud meets latency threshold with relatively low cost of operation.

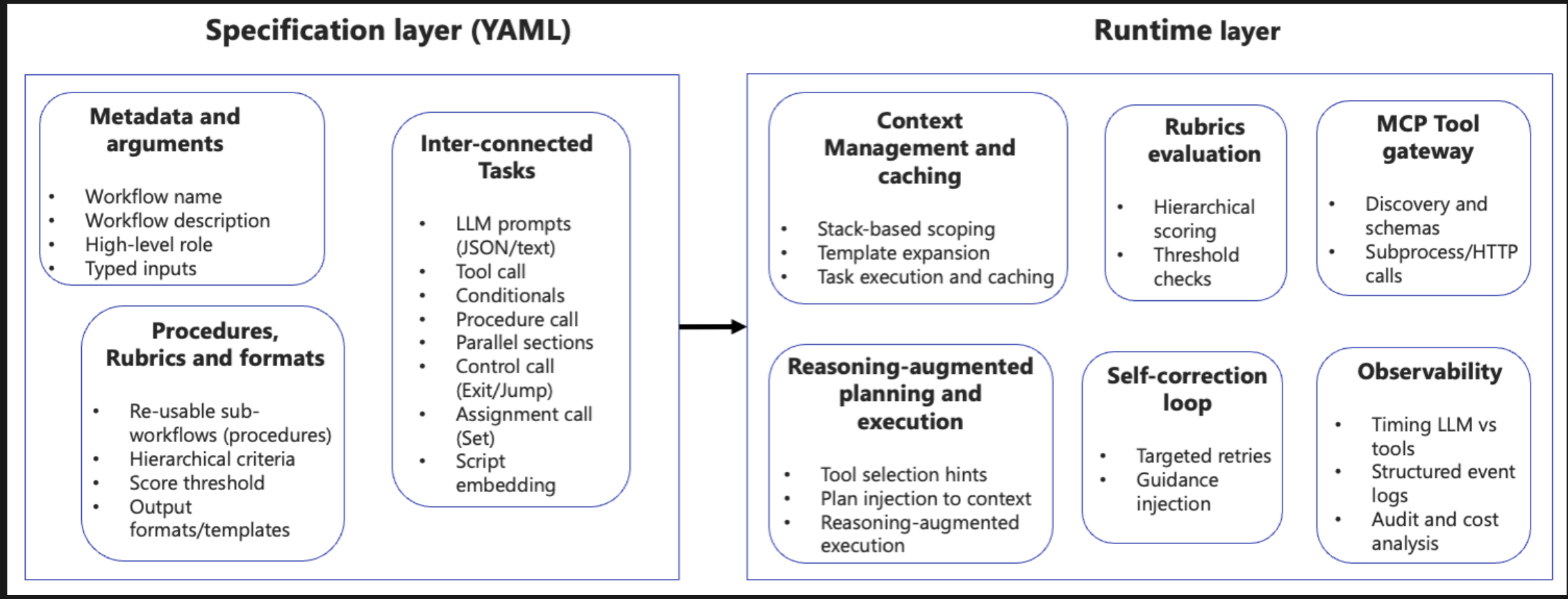


Dynamic placement of microservices is desired as workload changes

XPF: High Level System Architecture



System Architecture



Initial Task Description

Create a placement orchestrator that responds with a JSON object {"placement": "edge" or "cloud"}. Maintain a CSV file latency_report.csv with rows <workload>,<infrastructure>,<latency>.

The orchestrator receives two instruction types:
(1) Recommend placement for workload: <workload> and
(2) Latency on <infrastructure> for workload: <workload> is <latency>.

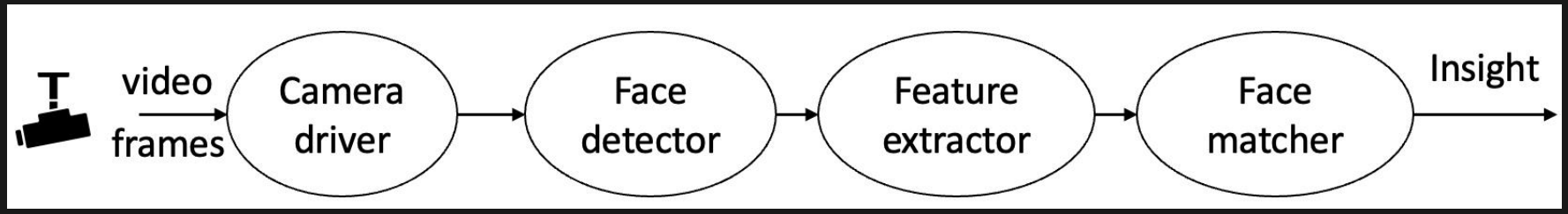
Policy: If a workload has no entries, recommend edge. If exactly one of edge/cloud is recorded, recommend the missing one to collect the other measurement. If both are recorded, recommend cloud only when it is at least 20% faster than edge; otherwise recommend edge.

Generated Placement Orchestrator Agent

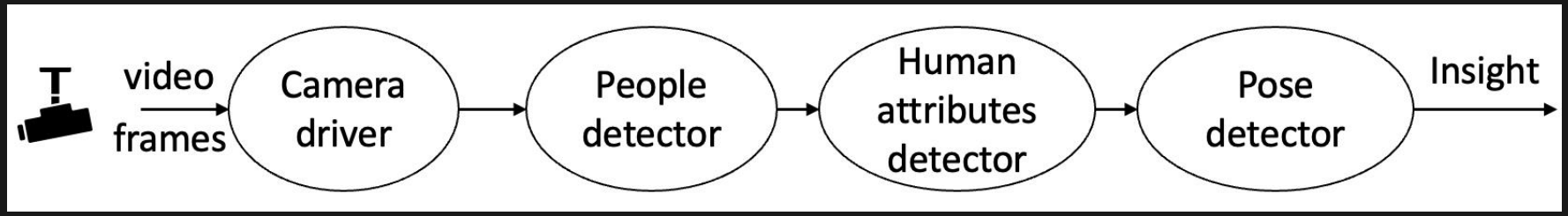
```
title: Placement Orchestrator
description: Orchestrates placement on edge or cloud
arguments:
  instruction:
    type: string
tasks:
  - label: instruction_type
    llm:
      prompt: |
        From the instruction below, return a compact JSON
  object with the following key:
        - type: the type of the instruction, it can be
  either recommend_placement or record_latency.

        The instruction template for type
  "recommend_placement" will be like this:

        "Recommend placement for workload: <workload>"
```

Experimental setup and results

Video Analytics Applications



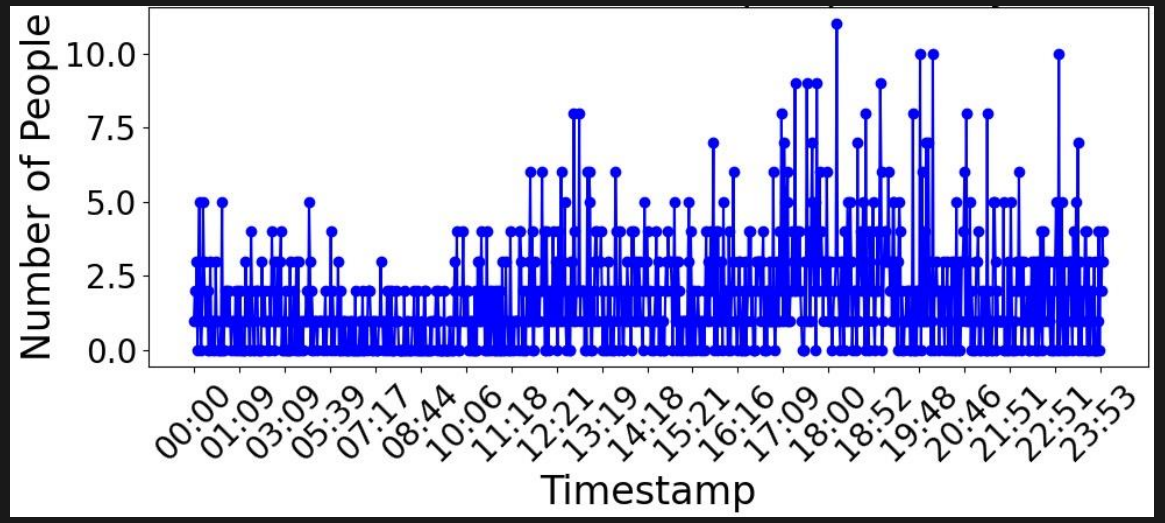
Face recognition application pipeline



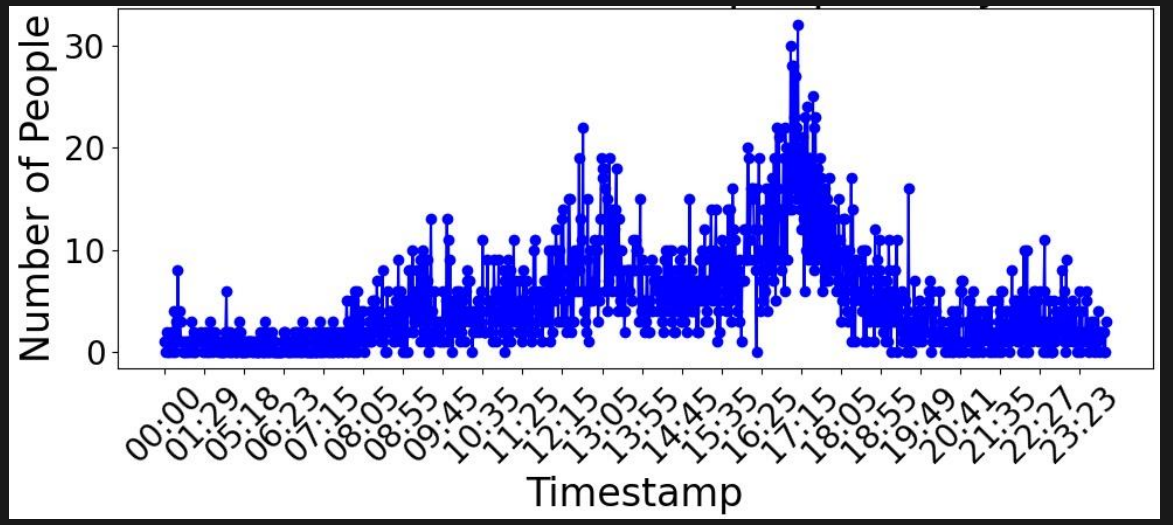
Human attributes detection application pipeline

Workload

Variation in workload (Pedestrian counting system in Melbourne)



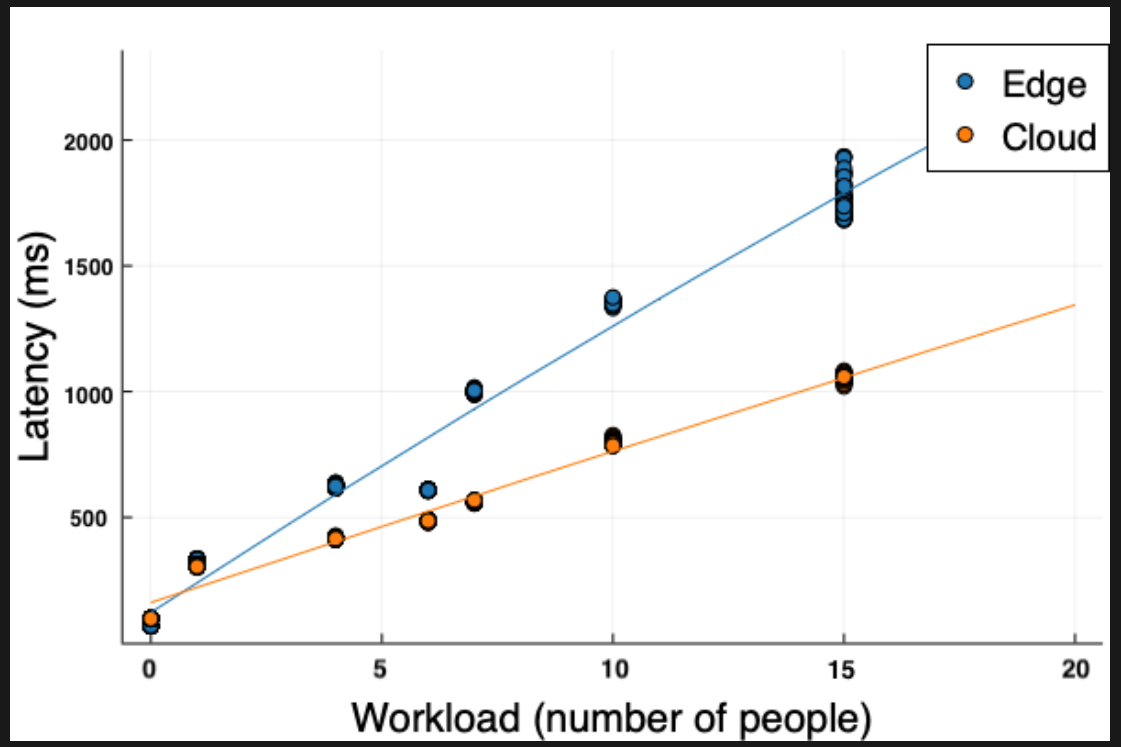
Day 1



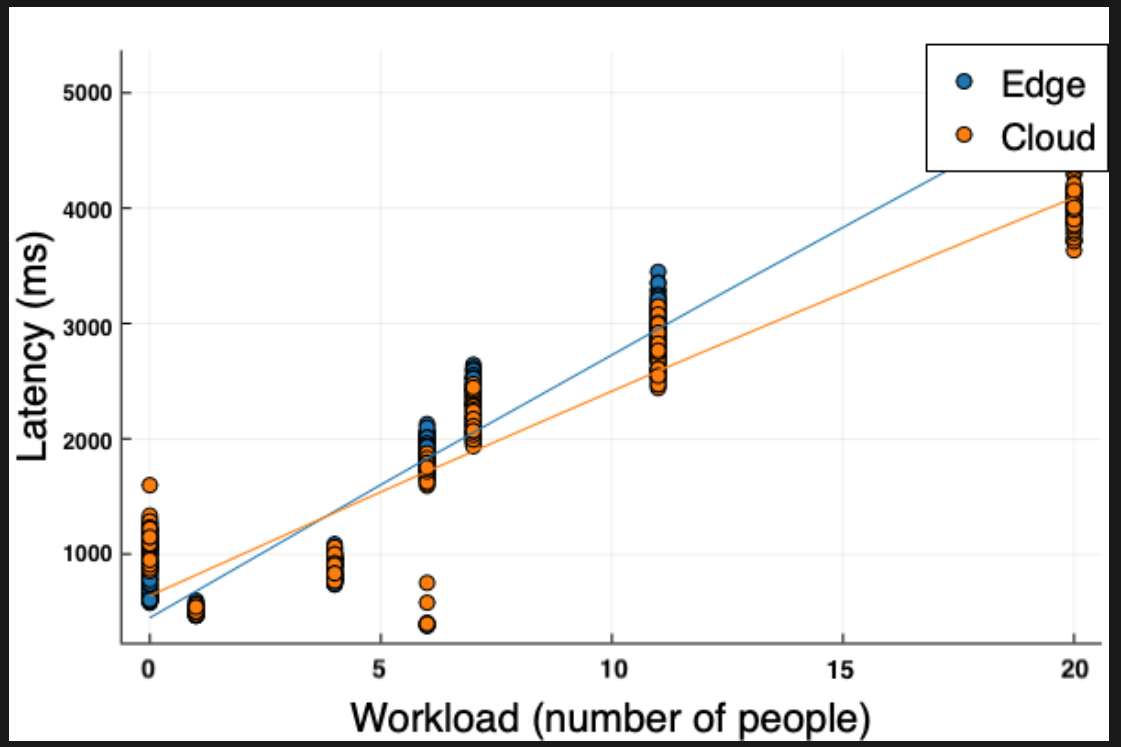
Day 1

Dataset: https://melbournetestbed.opendatasoft.com/explore/dataset/pedestrian-counting-system-past-hour-counts-per-minute/information/?sort=-location_id

Characterization of Application latency



Face recognition



Human attributes detection

Placement Accuracy

	HAD				FR			
TPD	1440				1440			
	GPT-4o		GPT-5		GPT-4o		GPT-5	
	Day1	**Day2**	**Day1**	**Day2**	**Day1**	**Day2**	**Day1**	**Day2**
TIP Edge (E)	42	0	16	8	0	0	0	0
TIP Cloud (C)	35	217	1	0	275	50	10	3
TIP (E+C)	77	217	17	8	275	50	10	3
Error (%)	5.34	15.06	1.18	0.55	19.09	3.47	0.69	0.20
Accuracy (%)	**94.66**	**84.94**	**98.82**	**99.45**	**80.91**	**96.53**	**99.31**	**99.8**

Placement accuracy of the agentic workflow against the oracle baseline (HAD: Human Attributes Detector; FR: Face Recognition; TPD: Total Placement Decisions; TIP: Total Incorrect Placements)

Oracle: Recommends cloud only if the characterized cloud latency is at least 20% lower than the characterized edge latency

Conclusion

- We presented a specification-first system for executing agentic workflows with explicit state, conditional control flow, and verifiable tool calls
- We implemented an LLM-driven agentic placement workflow that
  - interprets natural-language placement and latency-report instructions
  - maintains a persistent CSV latency log
  - recommends edge or cloud placement using a measurement-driven policy
- Across two video-analytics applications and two days of workload variation, the agentic workflow achieves accuracies of 94.66%/84.94% (HAD, Day1/Day2) and 80.91%/96.53% (FR, Day1/Day2) with GPT-4o; with GPT-5, accuracy increases to 98.82%/99.45% (HAD) and 99.31%/99.8% (FR)
- Although we demonstrate with a specific class of applications and using only 2 tiers of computing, future work can investigate the impact with different LLM models on multiple tiers of computing across a diverse set of applications.


Orchestrating a brighter world
NEC
NEC Laboratories America