From Edge to Cloud:
Serverless and *Intent-Based* Approaches in the Sustainable Compute Continuum

Adel N. Toosi

***Dis**tributed Systems and **Net**work Applications Laboratory (**DisNet lab.**)*
School of Computing and Information Systems
The University of Melbourne

My home page

Email: adel.toosi@unimelb.edu.au
Homepage: https://adelnadjarantoosi.info/

DisNet Lab

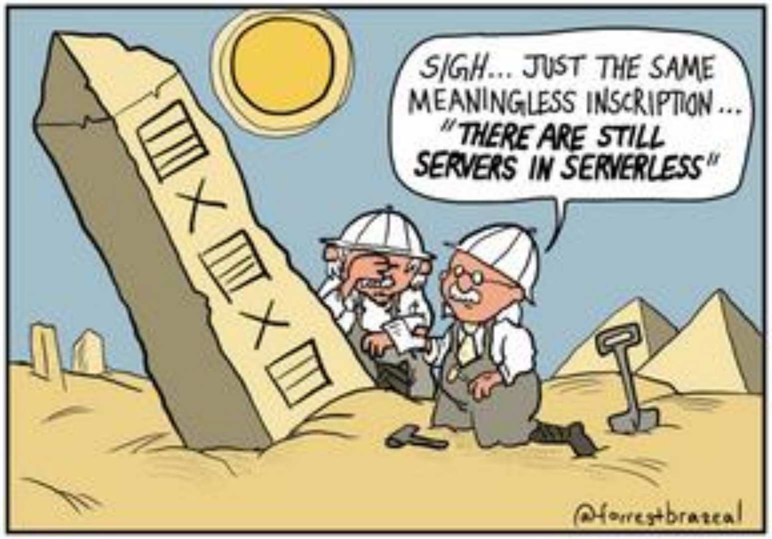
Biography

- **Associate Professor in Computer Systems,** CIS School, University of Melbourne, 2024-
- **Senior Lecturer**, Faculty of IT, Monash University, 2022-2024
- **Lecturer**, Faculty of IT, Monash University, 2018-2022
- **Postdoctoral Research Fellow**, University of Melbourne, 2015-2018
- **PhD**, Computer Science and Software Engineering, 2015
  - Thesis: "*On the Economics of Infrastructure as a Service Cloud Providers: Pricing, Markets, and Profit Maximization*"
- Research Interests
  - Distributed Systems, Cloud/Fog/Edge Computing, Software-Defined Networking (SDN), Serverless Computing, Smart Systems (Smart Agriculture, Smart Transports, etc) Sustainable IT, Energy Efficiency, and Green Computing, Evs.
  - Focused on **Resource Management** and **Scheduling** in Distributed Systems

Outline

- Introduction
- I. Serverless Edge Computing
  - » Con-pi
  - » Performance Evaluation of Serverless Edge
  - » Wattedge
  - » faasHouse
  - » Hedgi
  - » Benchmarking and routing object detection tasks on the edge
- II. Compute Continuum
  - » Intent-based Vehicular Edge Computing
  - » Serverless Vehicular Edge Computing
  - » iContinuum
  - » Empirical Study on Edge-to-Cloud Continuum
  - » IntentContinuum
- Summary

I. Serverless Edge Computing



Edge Computing

- The issues of cloud
  - Cloud data centres reside at a **multi-hop** distance from the sensors and devices
  - Data propagation and transmission can cause significant **delays**
    - » Real-time applications such as autonomous vehicles.
    - » Bandwidth-intensive applications, eg. Video analytics
  - **Privacy** concerns
    - » Secure Healthcare Monitoring
- Edge computing:
  - A **distributed computing paradigm** that brings computation and data storage closer to the sources of data, often on the **edge of the network**.
- Key Points:
  - Reduced Latency
  - Bandwidth Efficiency
  - Enhanced Privacy & Security
  - Real-time Processing

Challenges of Edge Computing

- In remote area applications (smart farming and forestry)
  - **extreme edge:** electricity arrangements to integrate sensory/actuation systems into the edge computing infrastructure are **tedious** and **costly**.
  - Example:
    - » ARC Linkage Project: Precision Pollination and Honeybees Monitoring
- Solution:
  - Battery and energy harvesting (e.g., solar panels)
- Challenge:
  - Edge devices rely on renewable energy sources that are subject to **energy and load variability** which can create an imbalance in their **operational availability**.
- Solution:
  - Resource sharing and task offloading

Con-Pi: Self-Sustained Edge Computing Framework

Task offloading can significantly increase operational availability of devices.

R. Mahmud and A. N. Toosi, "**Con-Pi: A Distributed Container-based Edge and Fog Computing Framework**," in *IEEE Internet of Things Journal*, doi: 10.1109/JIOT.2021.3103053.

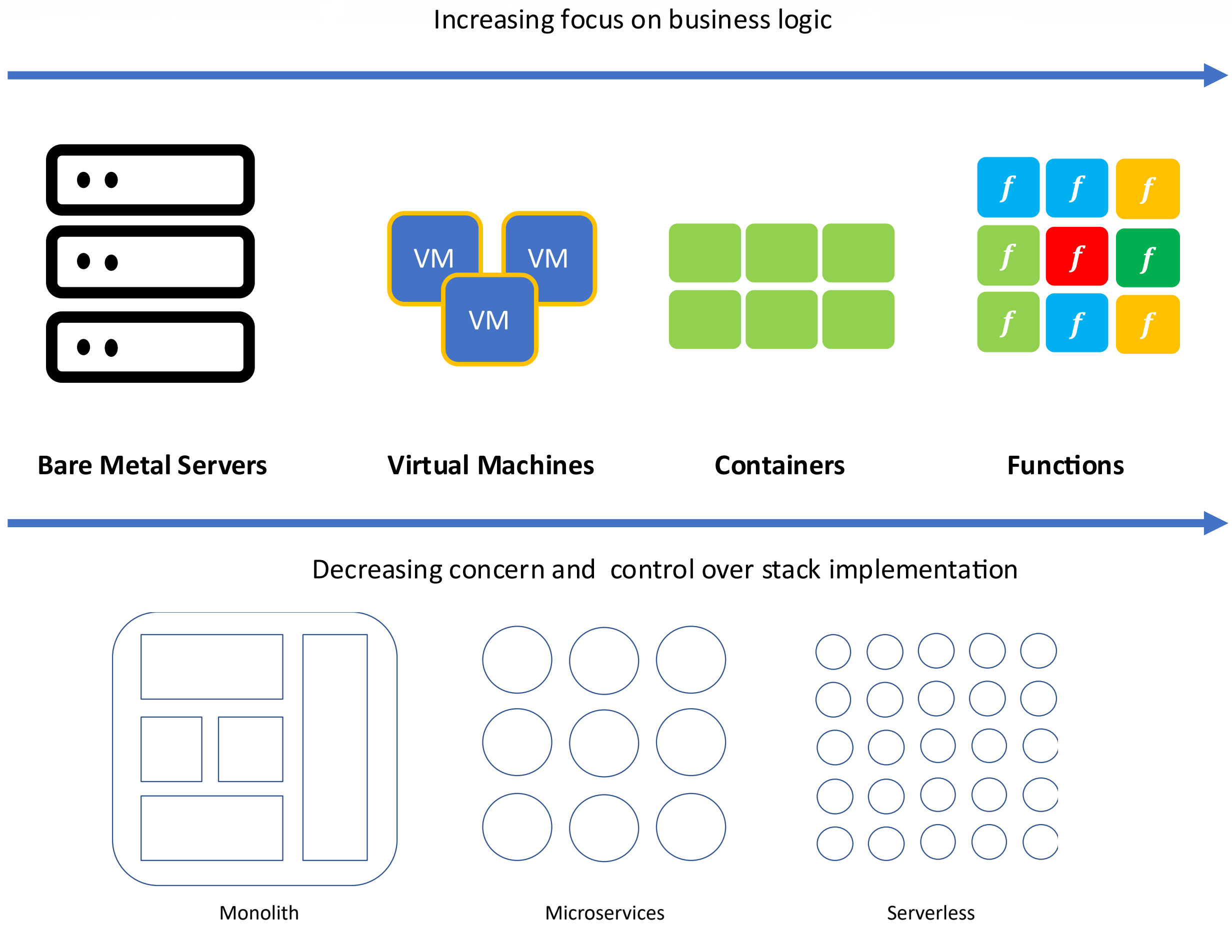
Serverless Edge Computing

- ## Serverless Computing
  - Build applications and services without thinking about the **underlying servers**.
  - Focus on pure application code (**application logic**)
  - It runs in **stateless compute containers** that are **event-triggered**
    - » ephemeral (may only last for one invocation), and fully managed by a third party (cloud provider).
  - One way to think of this is “Functions as a Service” or “FaaS”.

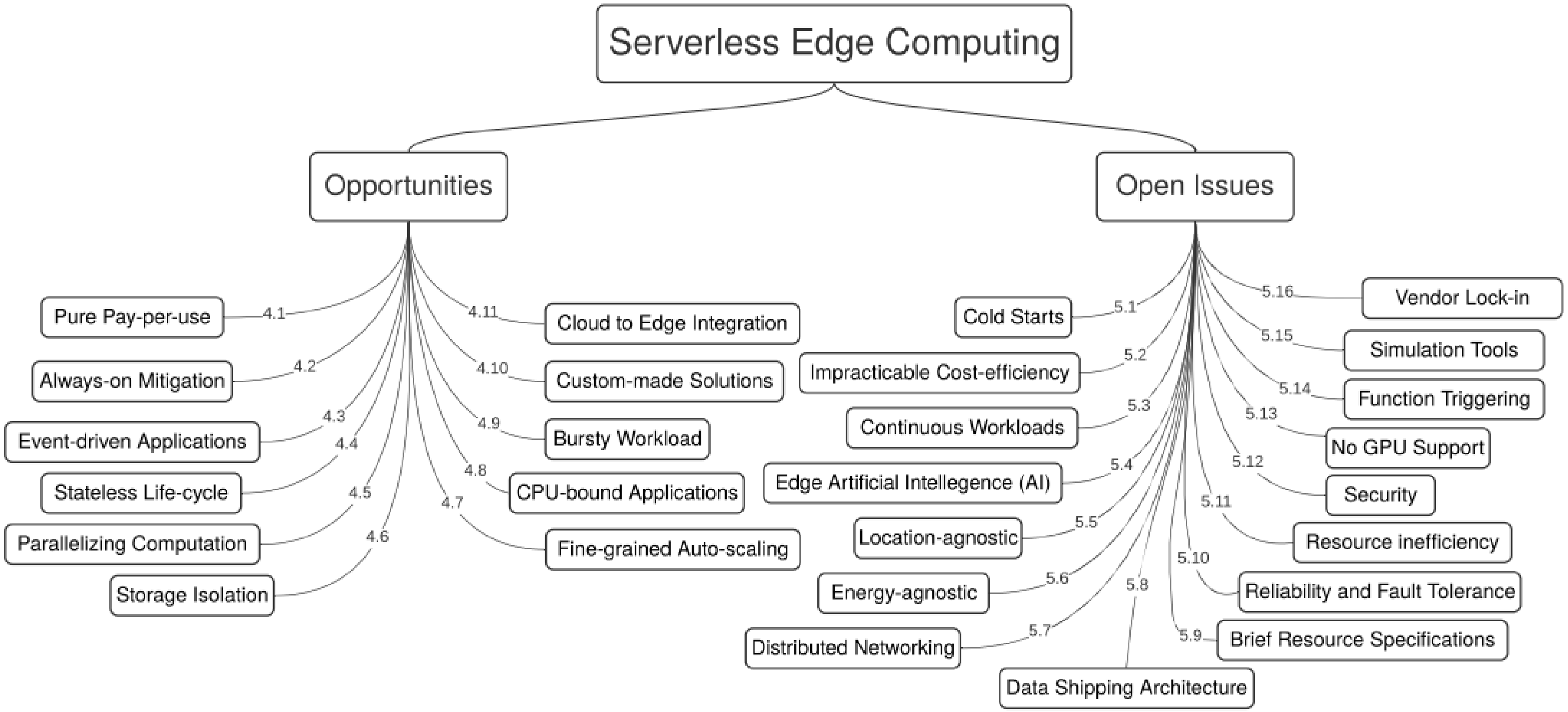
AWS Lambda   Azure Functions   Google Functions

OpenWhisk   OpenFaaS   OracleFN   Kubeless   Fission   Iron Functions   Nuclio

Evolution of Serverless

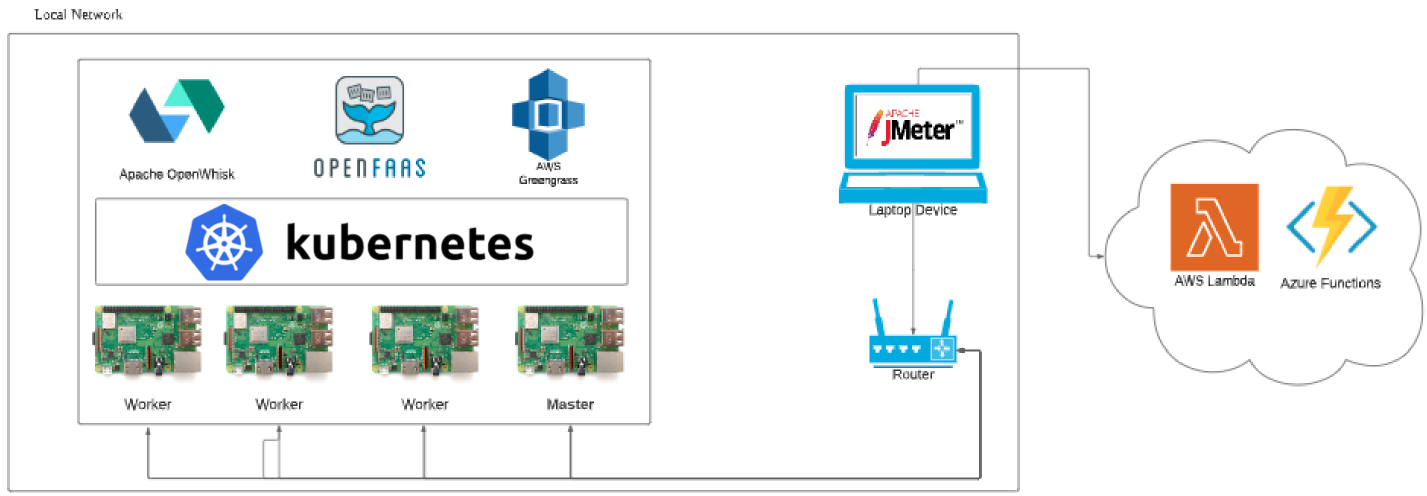


Our vision on Serverless Edge Computing



Mohammad Sadegh Aslanpour, Adel N. Toosi, Claudio Cicconetti, Bahman Javadi, Peter Sbarski, Davide Taibi, Marcos Assuncao, Sukhpal Gill, Raj Gaire, Schahram Dustdar, **Serverless Edge Computing: Vision and Challenges**, *In Australasian Computer Science Week Multiconference (ACSW'21)*, article no 10, Dunedin, New Zealand, 2021, pp. 1-10 doi:10.1145/3437378.3444367, **BEST PAPER AWARD**

Performance Evaluation of Serverless Frameworks on the Edge



OpenFaaS

OpenWhisk

AWS Lambda

AWS GreenGrass

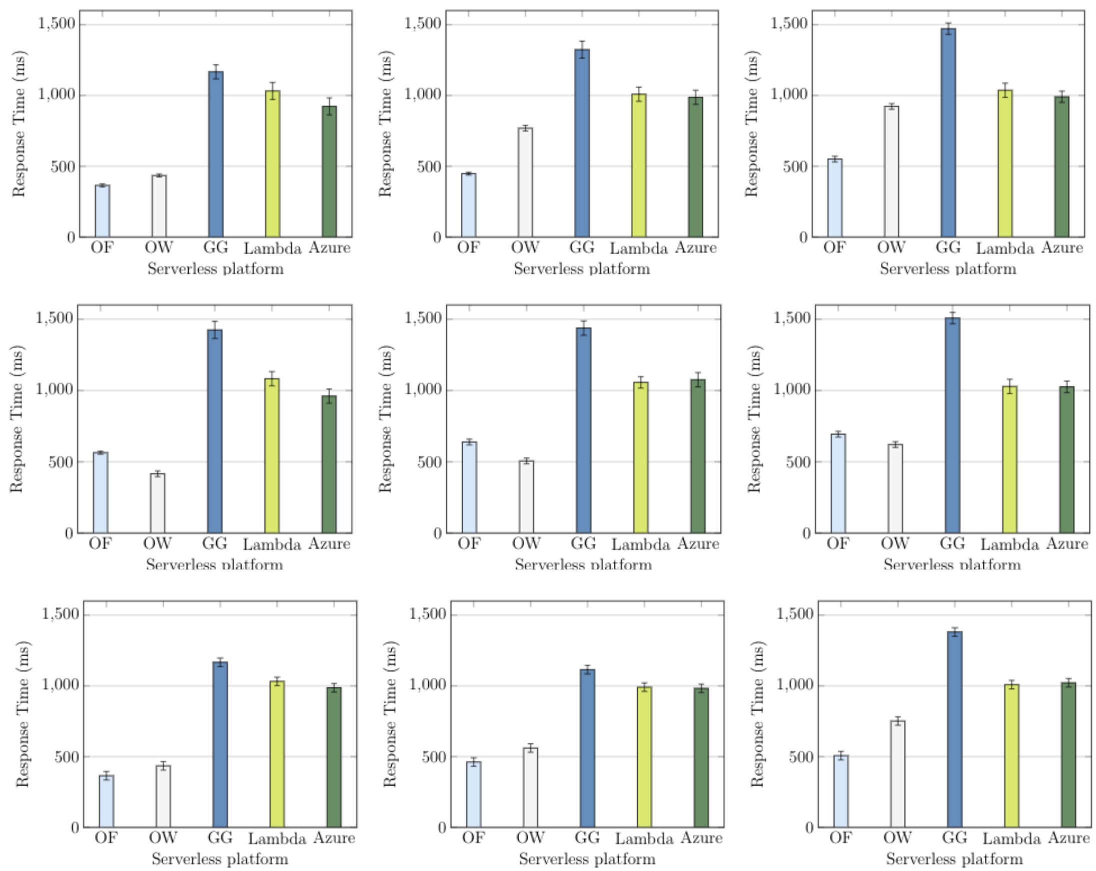
Azure Functions

Performance Evaluation Results

CPU-Intensive

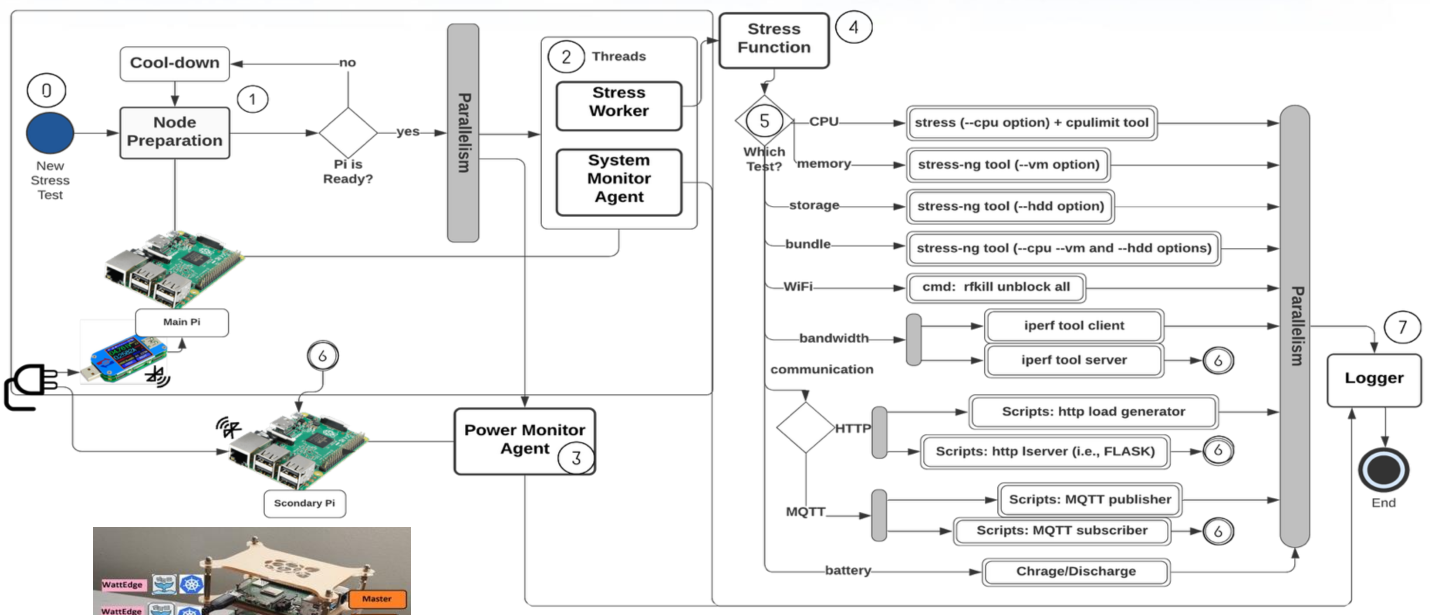
Memory-Intensive

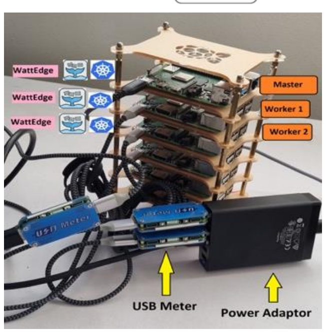
I/O (Disk)-Intensive



(a) 5 concurrent users (b) 10 concurrent users (c) 15 concurrent users

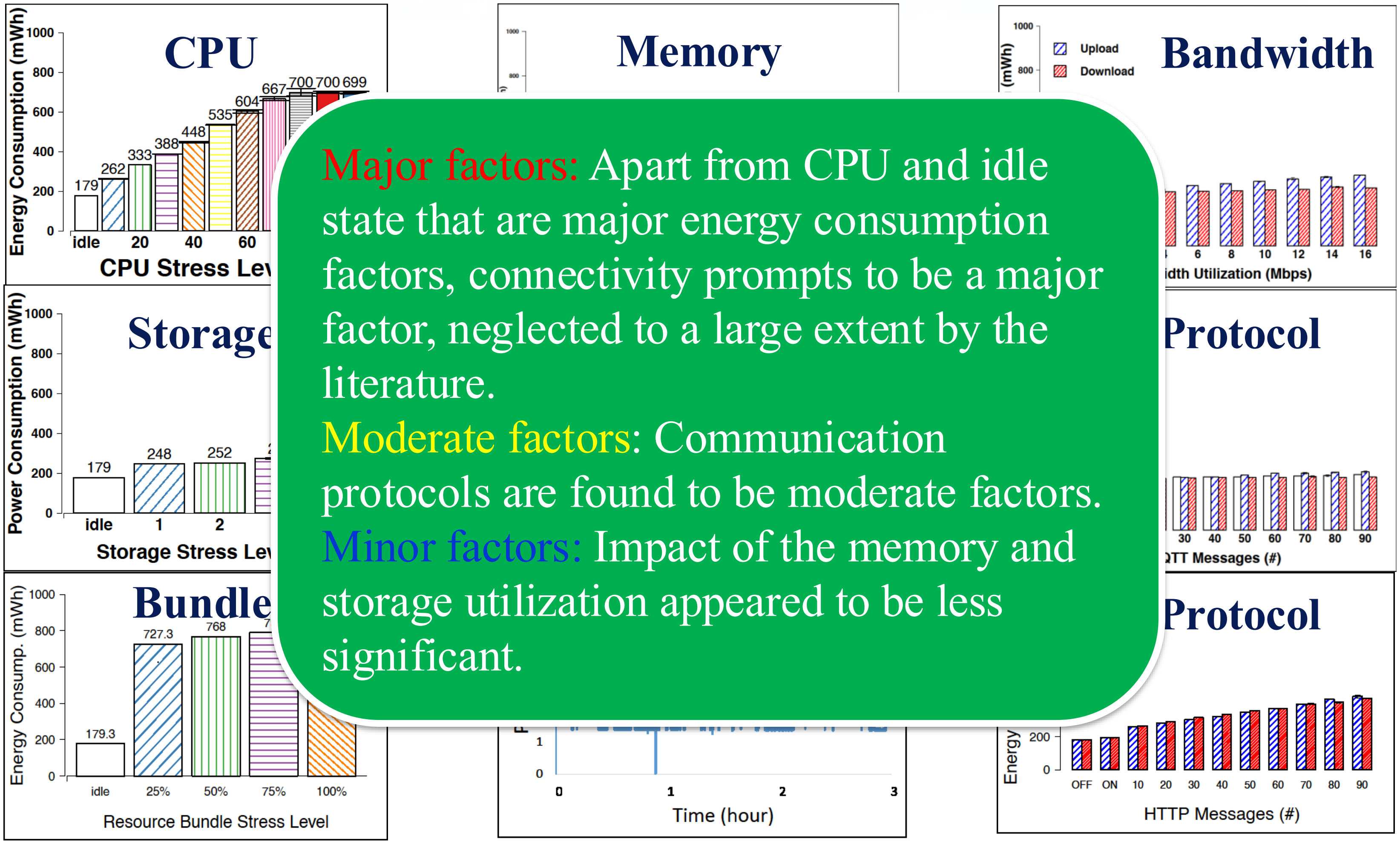
WattEdge





Aslanpour M.S., Toosi A.N., Gaire R., Cheema M.A. (2021) **WattEdge: A Holistic Approach for Empirical Energy Measurements in Edge Computing**. In: Hacid H., Kao O., Mecella M., Moha N., Paik H. (eds) *Service-Oriented Computing. (ICSOC'21)*. Lecture Notes in Computer Science, vol. 13121. Springer. **BEST PAPER CANDIDATE.**

CPU, Memory and Disk



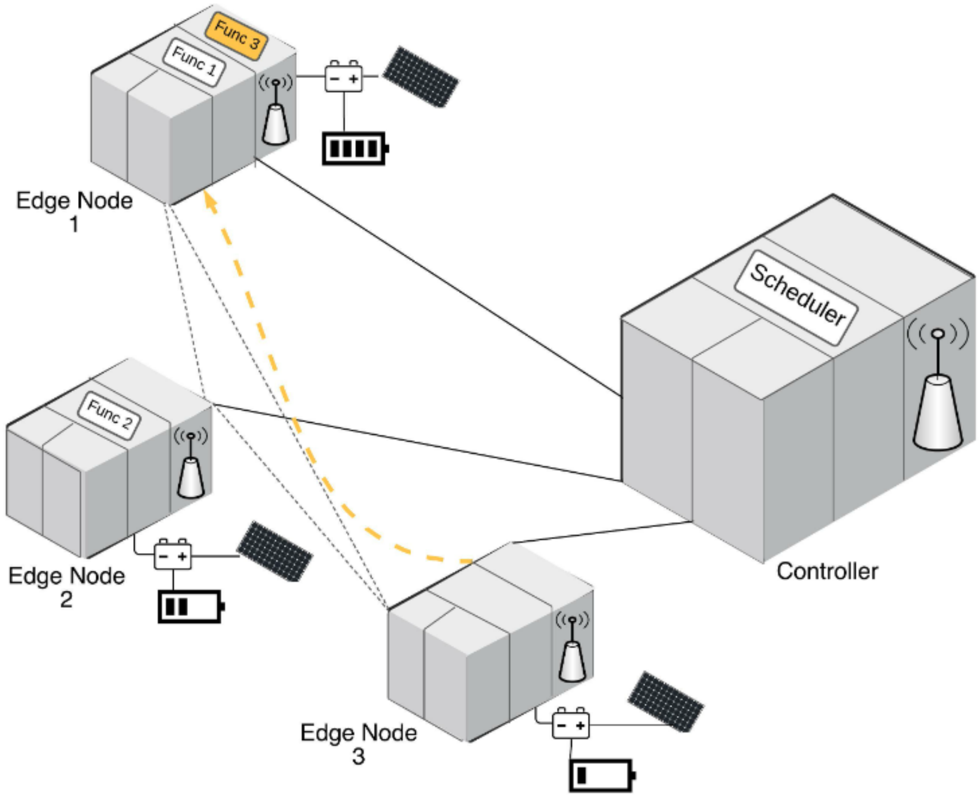
Major factors: Apart from CPU and idle state that are major energy consumption factors, connectivity prompts to be a major factor, neglected to a large extent by the literature.

Moderate factors: Communication protocols are found to be moderate factors.

Minor factors: Impact of the memory and storage utilization appeared to be less significant.

Sustainable Serverless Edge Computing through Energy-aware Resource Scheduling

- Support for various hard/soft requirements
- minimize the number of failures for a node
- minimize wasted energy
- maximize the longest time a node is operational.



Mohammad Sadegh Aslanpour, Adel N. Toosi, Muhammad Aamir Cheema, and Raj Gaire, **Energy-Aware Resource Scheduling for Serverless Edge Computing**, in the proceedings of 22nd IEEE/ACM International Symposium on Cluster, Cloud and Internet Computing (CCGrid'22), pp. 190-199. IEEE, 2022.

Mohammad Sadegh Aslanpour, Adel N. Toosi, Muhammad Aamir Cheema, and Mohan Chhetri, **faasHouse: Sustainable Serverless Edge Computing through Energy-aware Resource Scheduling**, *IEEE Transactions on Services Computing*, 2023, under review.

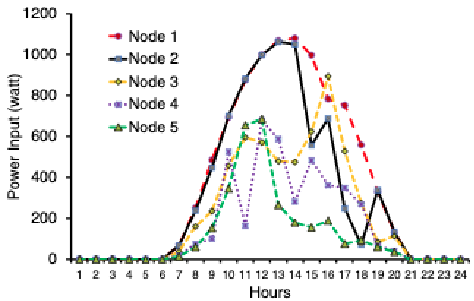
faasHouse

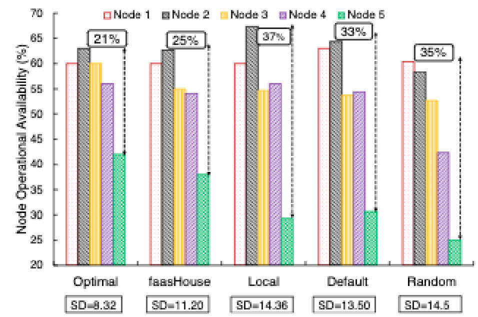
- Scoring
  - Energy, Locality, and Stickiness
- Assignment
  - **House Allocation Problem:** the problem of assigning houses (nodes) to people (functions) considering people's preferences

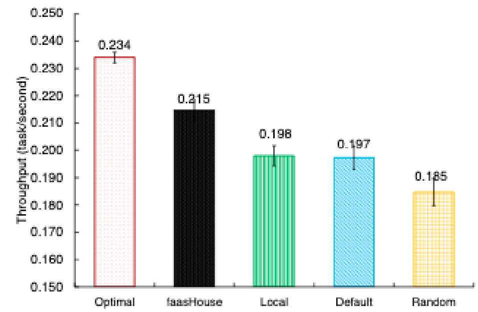
Benchmarks

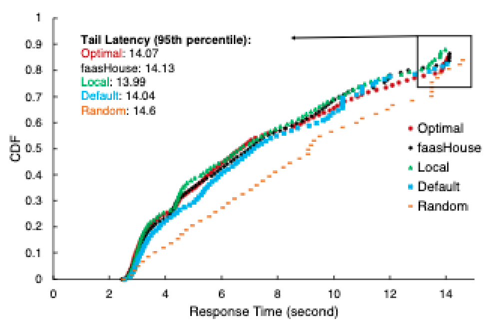
- Evaluated against to the following benchmarks:
- **Optimal**: This is an offline optimal algorithm which requiresthe future knowledge of renewable energy input and incomingworkload for each time slot (constrained optimisation problem)
- **Local**: This baseline algorithm always deploys functions locally. This is worth evaluating to understand the impact of offloading.
- **Default**: This is the default performance-aware scheduler in Kubernetes.
- **Random**: This randomly places functions across the cluster.
- **Zonal:** The proposed approach in CCGrid paper.

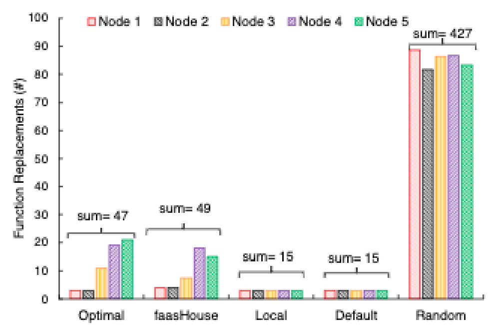
Some Results

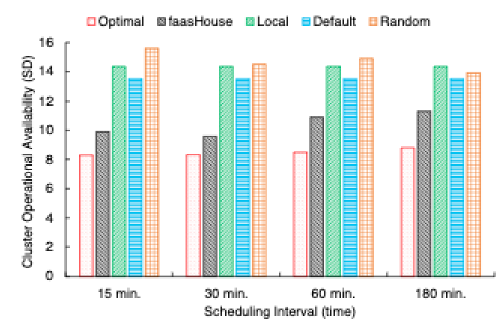






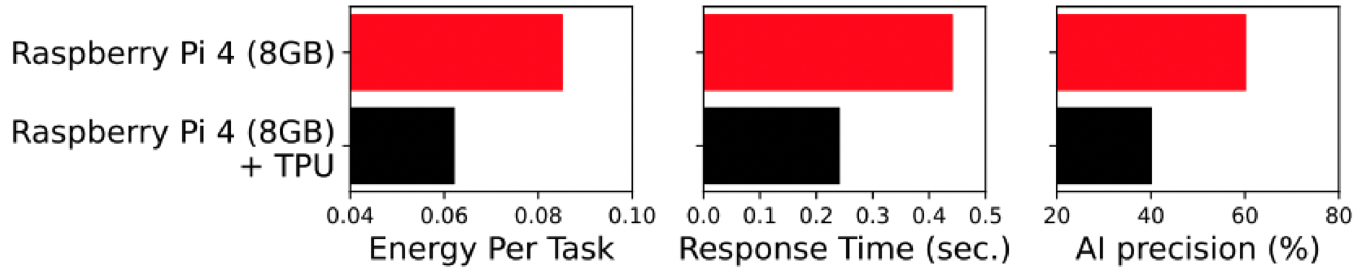






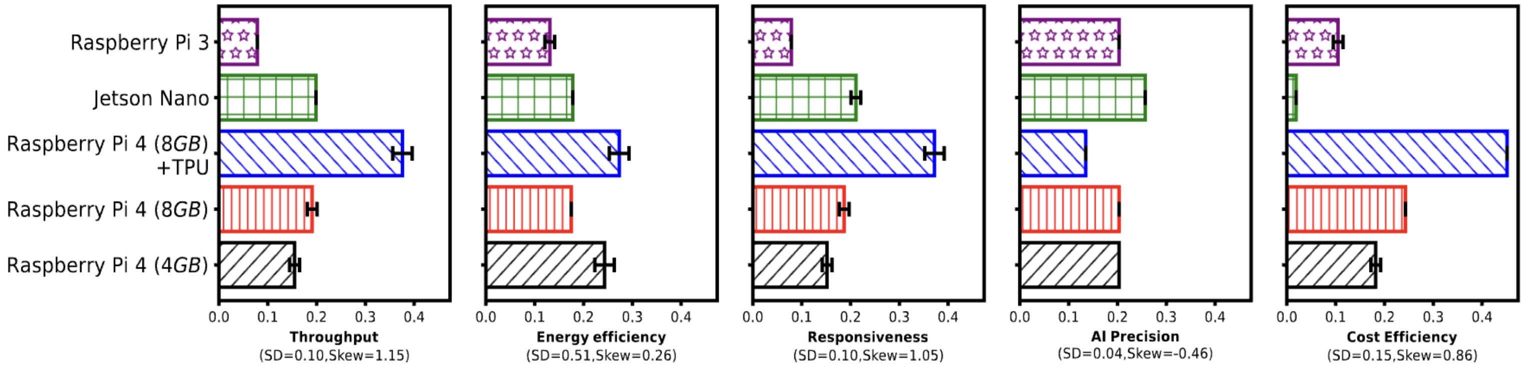
Heterogeneity-aware Resource Scheduler

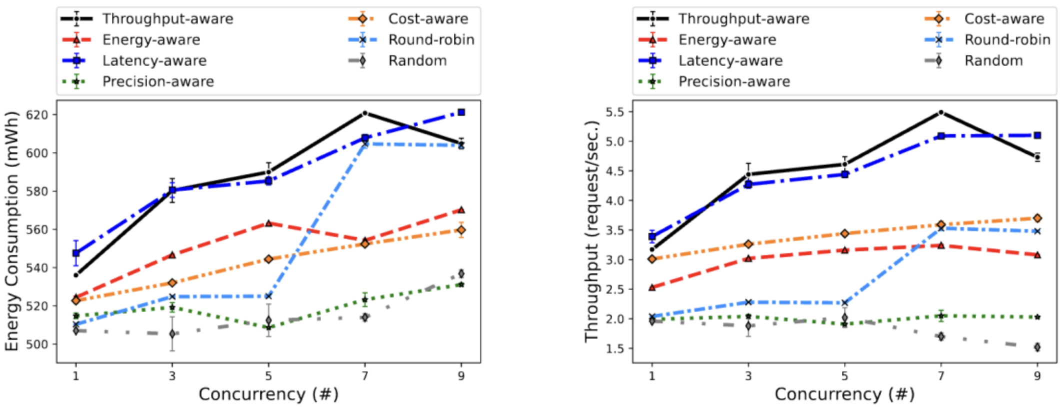
Motivation:



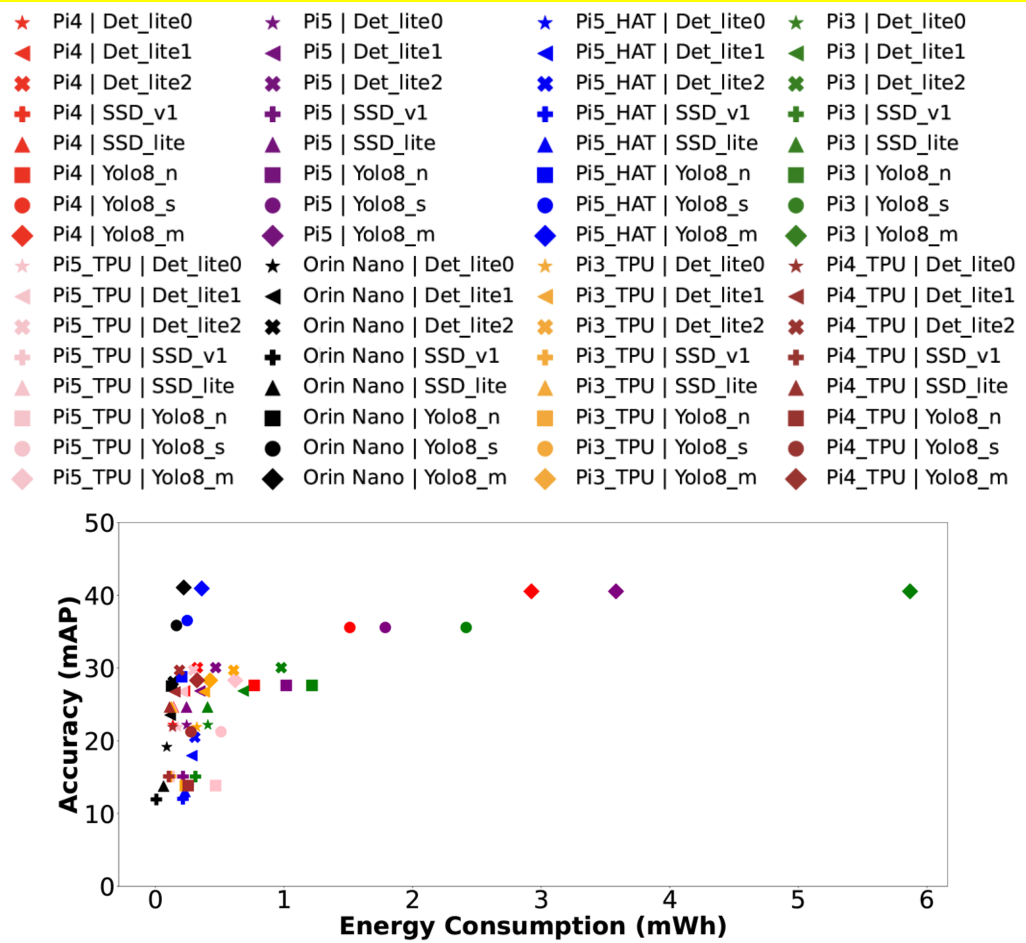
Mohammad Sadegh Aslanpour, Adel N. Toosi, Muhammad Aamir Cheema, Mohan Chhetri, and Mohsen Amini, **Load Balancing for Serverless Edge Computing: A Performance-aware and Empirical Approach**, *Journal of Future Generation Computer Systems*, 2023, CORE A, (R1 Revision).

Performance Characterization





Benchmarking Object Detection Models on Edge Devices



Daghash K. Alqahtani, Muhammad Aamir Cheema, **Adel N. Toosi, Benchmarking Deep Learning Models for Object Detection on Edge Computing Devices**, ICSOC 2024,

Energy-aware Routing for Object Detection Models at the Edge

Why Adaptive Routing?

- Scene complexity varies (crowded vs. empty)
- High complexity → need accurate, energy-hungry models
- Low complexity → lighter models save energy, not as efficient and accurate
- Route requests based on scene difficulty
- Smart trade-off: accuracy where needed, efficiency elsewhere

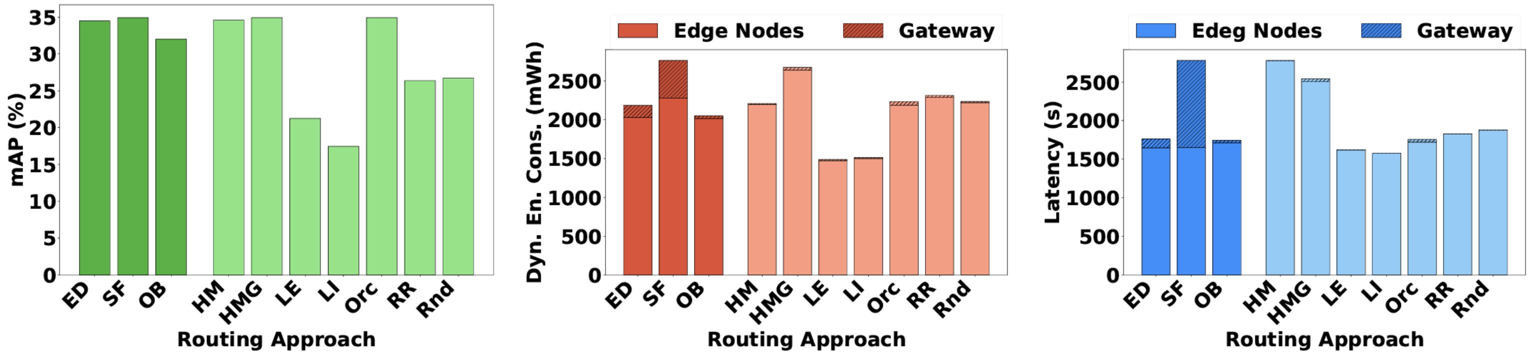
Daghash K. Alqahtani, Maria A. Rodriguez Hamid Rezatofighi, Muhammad Aamir Cheema, **Adel N. Toosi, ECORE: Energy-Conscious Optimized Routing for Deep Learning Models at the Edge**, Submitted to SENSYS, 2026, https://arxiv.org/abs/2507.06011

Proposed Routers

We make routing base on the complexity of the image and number of objects.

- **Edge Detection (ED):** Canny edge detection to count objects; fast but coarse.
- **SSD-Based Front-End (SF):** Lightweight SSD model at gateway; more accurate but costlier.
- **Output-Based (OB):** Reuses previous frame's object count; ideal for video, saves compute.

Some Results



(a) Accuracy (mAP) (b) Energy Consumption (c) Latency

Accuracy(mAP), Latency and Dynamic Energy Consumption for proposed routing approaches against baselines using COCO dataset.

Orc=Oracle, RR=Round Robin, Rnd=Random, LE=Lowest Energy, LI=Lowest Inference, HM=Highest mAP without considering Groups, HMG=Highest mAP Per Group, **ED**=Edge Detection, **SF**=SSD-Based Front, **OB**=Output-Based

δmAP=5.

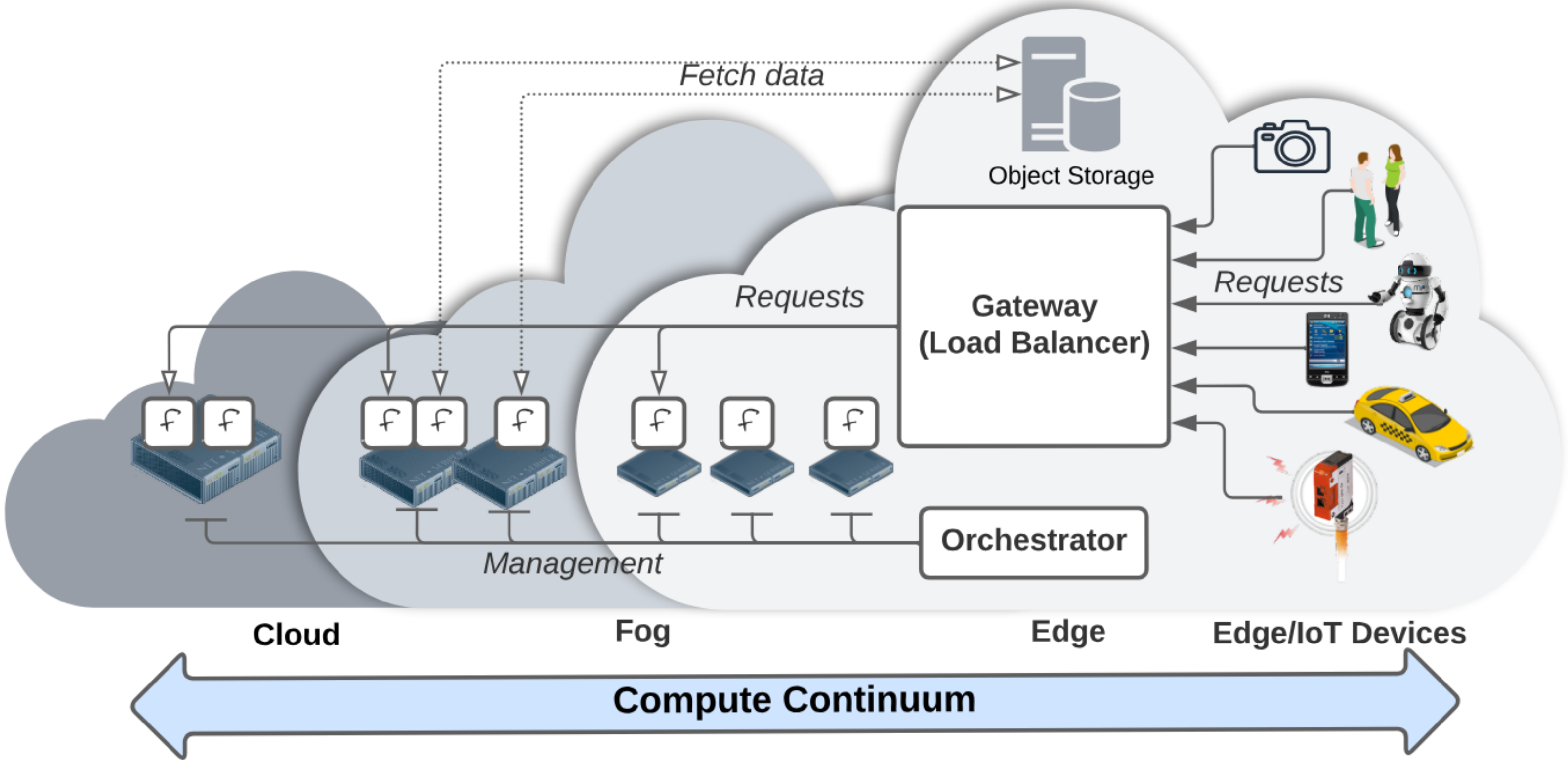
II. Compute Continuum

"Is this computer fast, or what? You just drilled a hole in the Space-Time Continuum! No biggie, though. Just hit 'Control-Escape-Home' and you'll be back to normal!"

Why Compute Continuum?

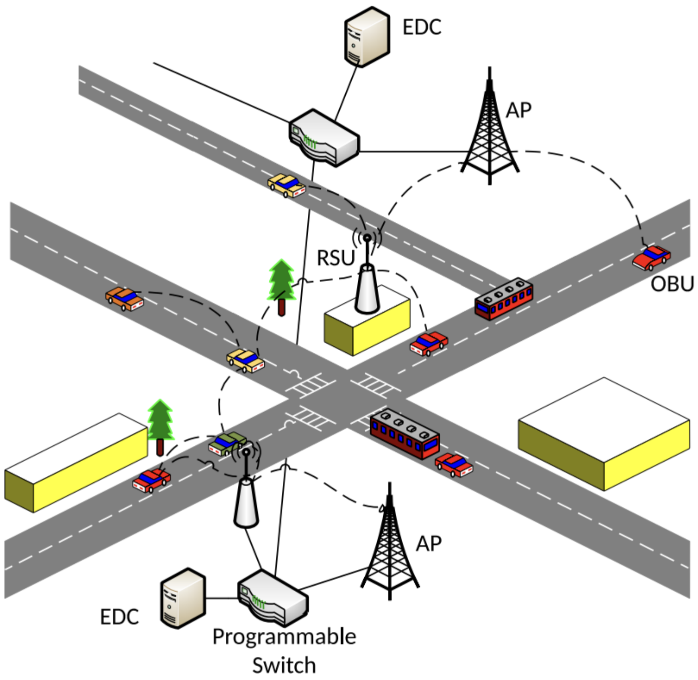
- **Cloud alone may not be suitable for all applications**
  - For real-time processing or strong privacy protections
- **Edge alone is not enough** – it lacks the scalability and power of cloud
- **Bridging Edge, Cloud** for seamless computing

Compute Continuum



Vehicular Edge Computing Overview

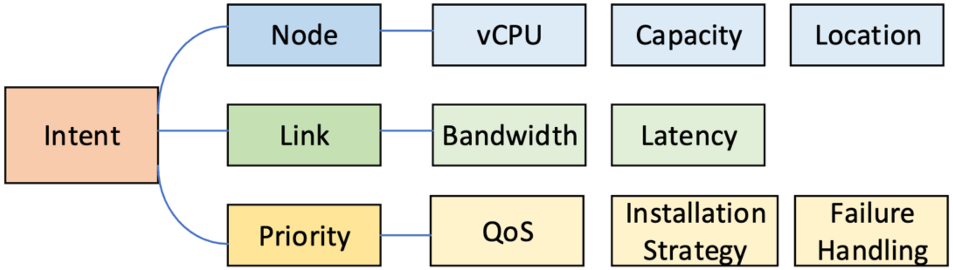
Overview: The automotive industry is one of the fastest-growing industries. In recent years, the increased use of onboard microprocessors such as **On-Board Units (OBUs)** and sensors technology has led to technological advancements that enabled vehicles to provide various safety and driver assistance-related systems.



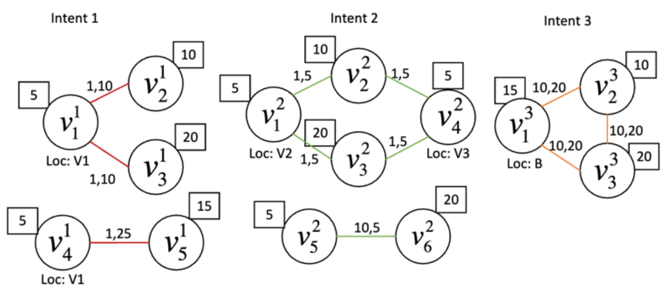
TianZhang He, Adel N. Toosi, Negin Akbari, Muhammed Tawfiqul Islam, and Muhammad Aamir Cheema, **An Intent-based Framework for Vehicular Edge Computing**, *In Proceedings of 2023 IEEE International Conference on Pervasive Computing and Communications (PerCom 2023)*, March. 13 - 17, Atalanta, USA, pp. 121 - 130, doi: 10.1109/PERCOM56429.2023.10099081

Background: Intent-Based Networking (IBN)

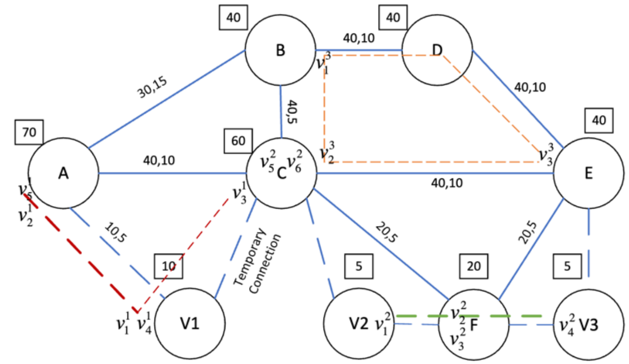
- Intent-Based Networking:
  - Based on **Software-Defined Networking (SDN)**,
  - was introduced to provide the ability to automatically handle and manage the networking requirements of different applications.
- Motivated by the IBN concept, we propose a novel approach to jointly orchestrate networking and computing resources based on user requirements.
- The proposed solution constantly **monitors** user **requirements** and **dynamically re-configures** the system to satisfy the desired states of the application.



Problem Description



An example of intents and compiled requests



An example of intent installation on the substrate network

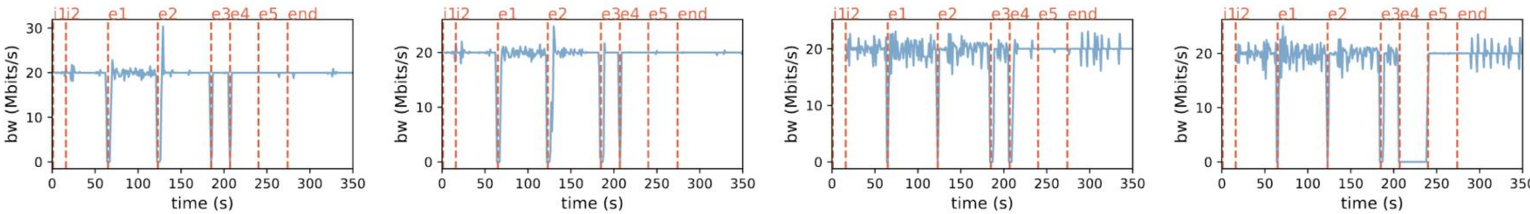
Objectives:

- Maximizes the intent acceptance ratio owing to their priorities
- While efficiently utilizing both computing and networking resources.

Large-scale Simulation configurations

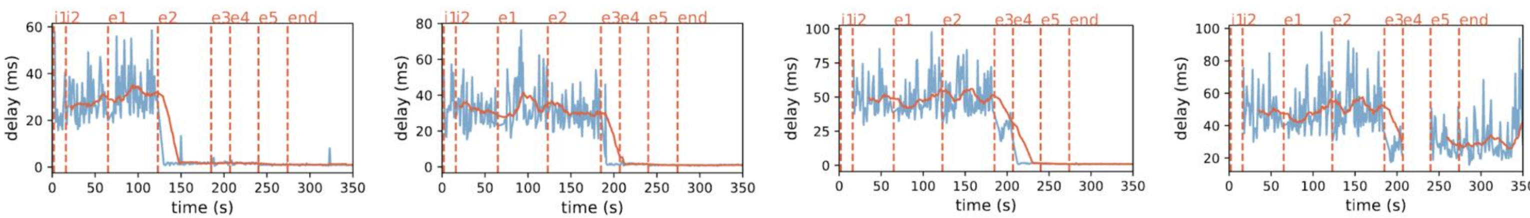
- Real-world taxi GPS dataset in Shanghai (April 1, 2018)
- Locations of base stations from Shanghai Telecom

Emulation/Small-scale Prototype



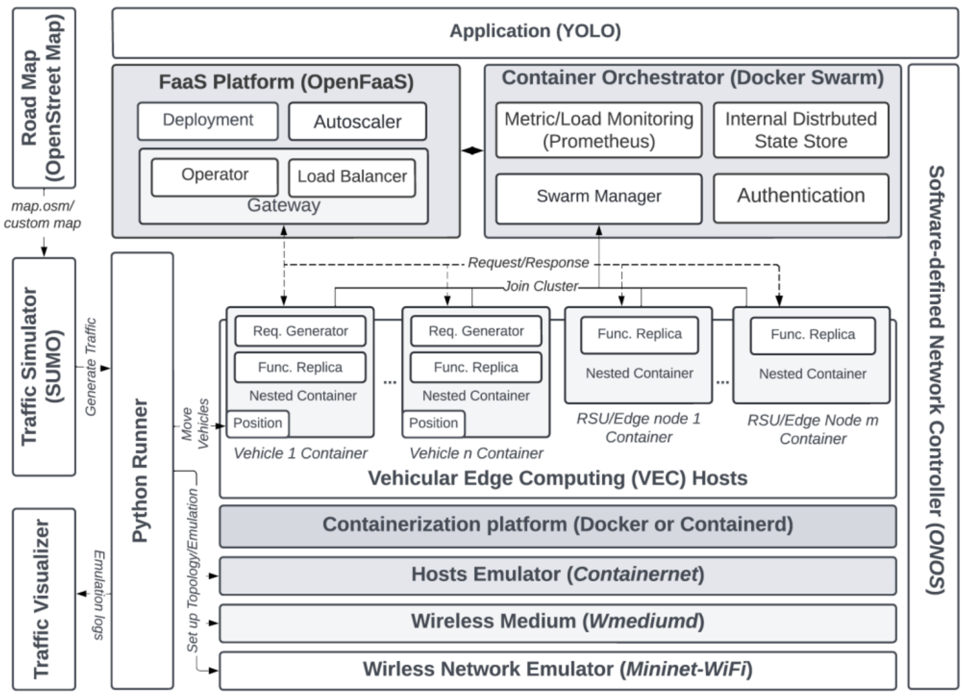
(a) v_1^1 and v_2^1 (b) v_1^1 and v_3^1 (c) v_1^2 and v_2^2 (d) v_1^2 and v_3^2

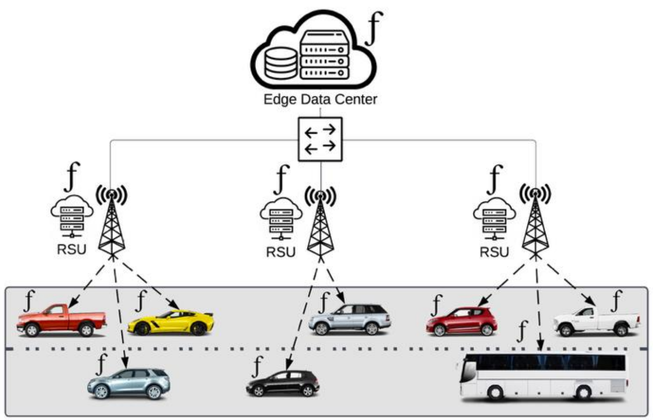
Fig. 11: Emulation performance in bandwidth along the time

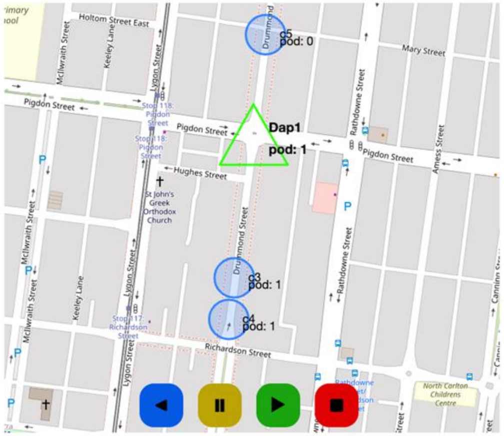


(a) v_1^1 and v_2^1 (b) v_1^1 and v_3^1 (c) v_1^2 and v_2^2 (d) v_1^2 and v_3^2

Serverless Vehicular Edge Computing







Alam F, Toosi AN, Cheema MA, Cicconetti C, Serrano P, Iosup A, Tari Z, Sarvi **A. Serverless Vehicular Edge Computing for the Internet of Vehicles**. *IEEE Internet Computing*. vol. 27, no. 4, pp. 40-51, July-Aug. 2023, doi: 10.1109/MIC.2023.3271641.

Challenge: Testing Applications & Experimentation

- A thorough **testing** of applications leveraging the compute continuum and **experimentation** is challenging before deployment in a production environment
- Solutions:
  - Simulation
    - » Realism limitations
    - » Accuracy concerns
    - » Complexity of network simulation
  - Emulation
    - » More accurate testing
    - » Reduced deployment risks
    - » Improved system reliability

Our proposed method: Emulation

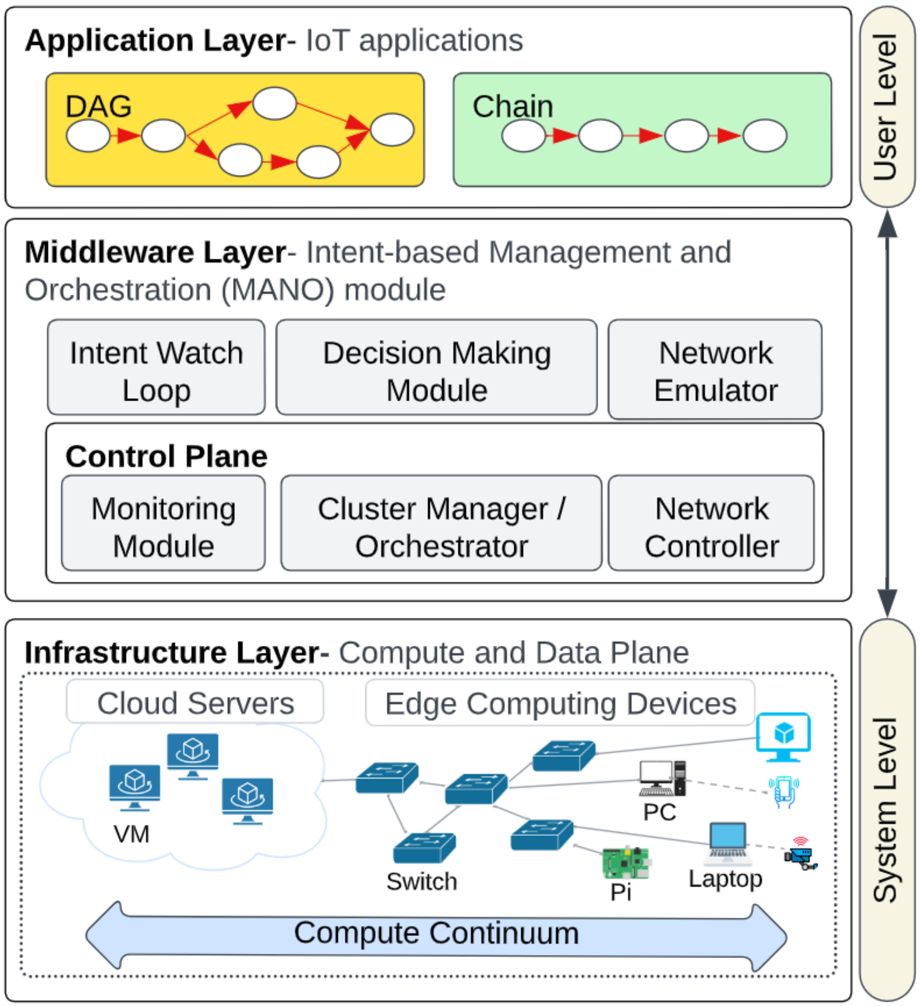
iContinuum

We have fully automated the setup of iContinuum using Ansible, making it incredibly user friendly. This allows iContinuum users to set up a complex edge-to-cloud continuum and application orchestration environment without getting into the complexities of all the proposed tools. All associated codes are available in our GitHub repository

https://github.com/disnetlab/iContinuum

N. Akbari, **A. N. Toosi**, J. Grundy, H. Khalajzadeh, M. S. Aslanpour and S. Ilager, ***iContinuum: An Emulation Toolkit for Intent-Based Computing Across the Edge-to-Cloud Continuum***, *2024 IEEE 17th International Conference on Cloud Computing (CLOUD)*, Shenzhen, China, 2024, pp. 468-474, doi: 10.1109/CLOUD62652.2024.00059.

iContinuum



https://github.com/disnetlab/iContinuum

inMon sFlow-RT

Grafana Prometheus

kubernetes Mininet OpenFlow onos

LOCUST

Another Challenge

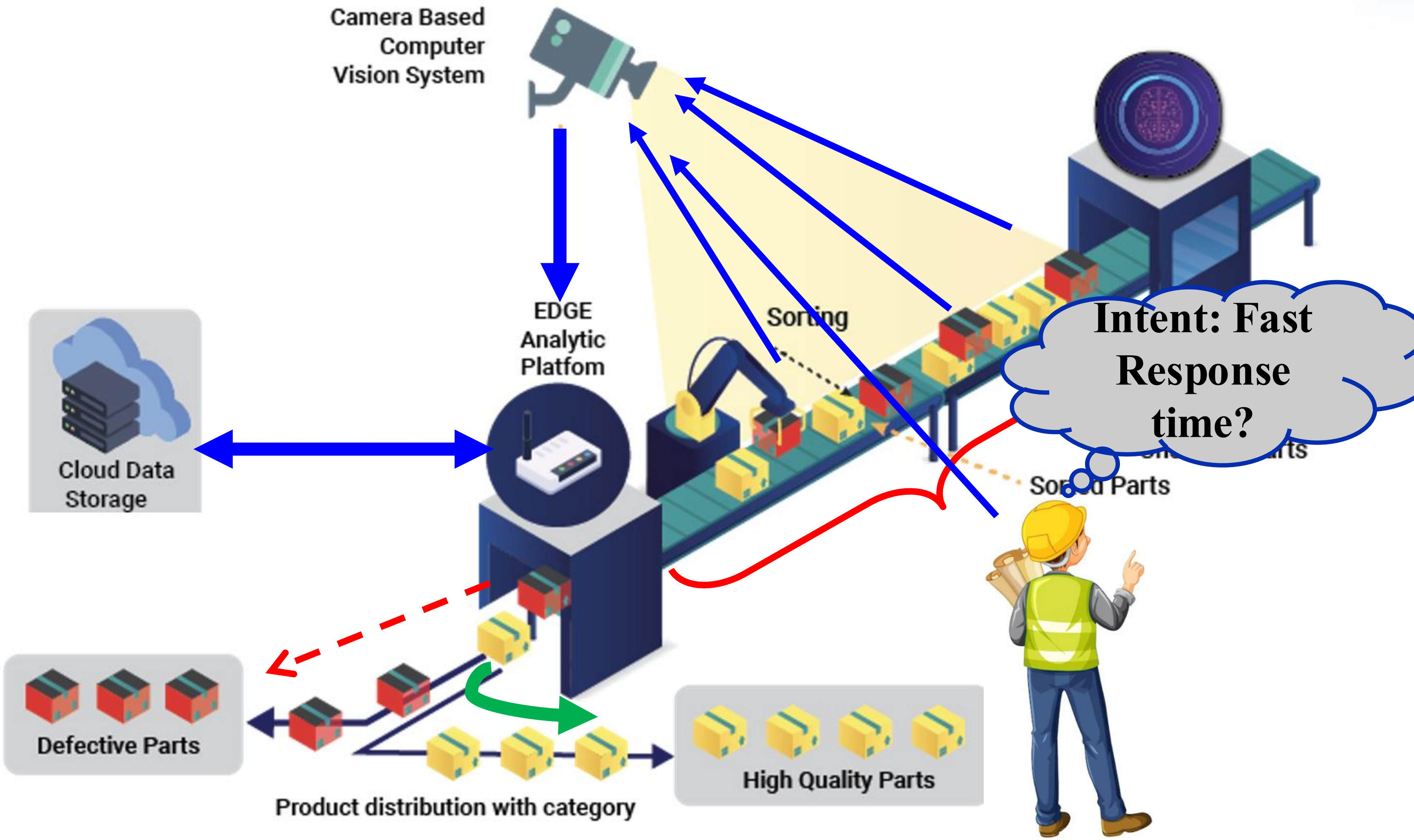
- ***Resource Management as a key challenge***
  - Diverse computing resources (smartphones, IoT sensors, edge servers, cloud data centers)
  - Complex infrastructure management
  - Efficient deployment of distributed applications
- Resource management in the Compute Continuum is challenging (often falling into NP-hard or NP-complete problem classes)
  - heuristic
  - meta-heuristic

Aim (LLM-as-a-Scheduler)

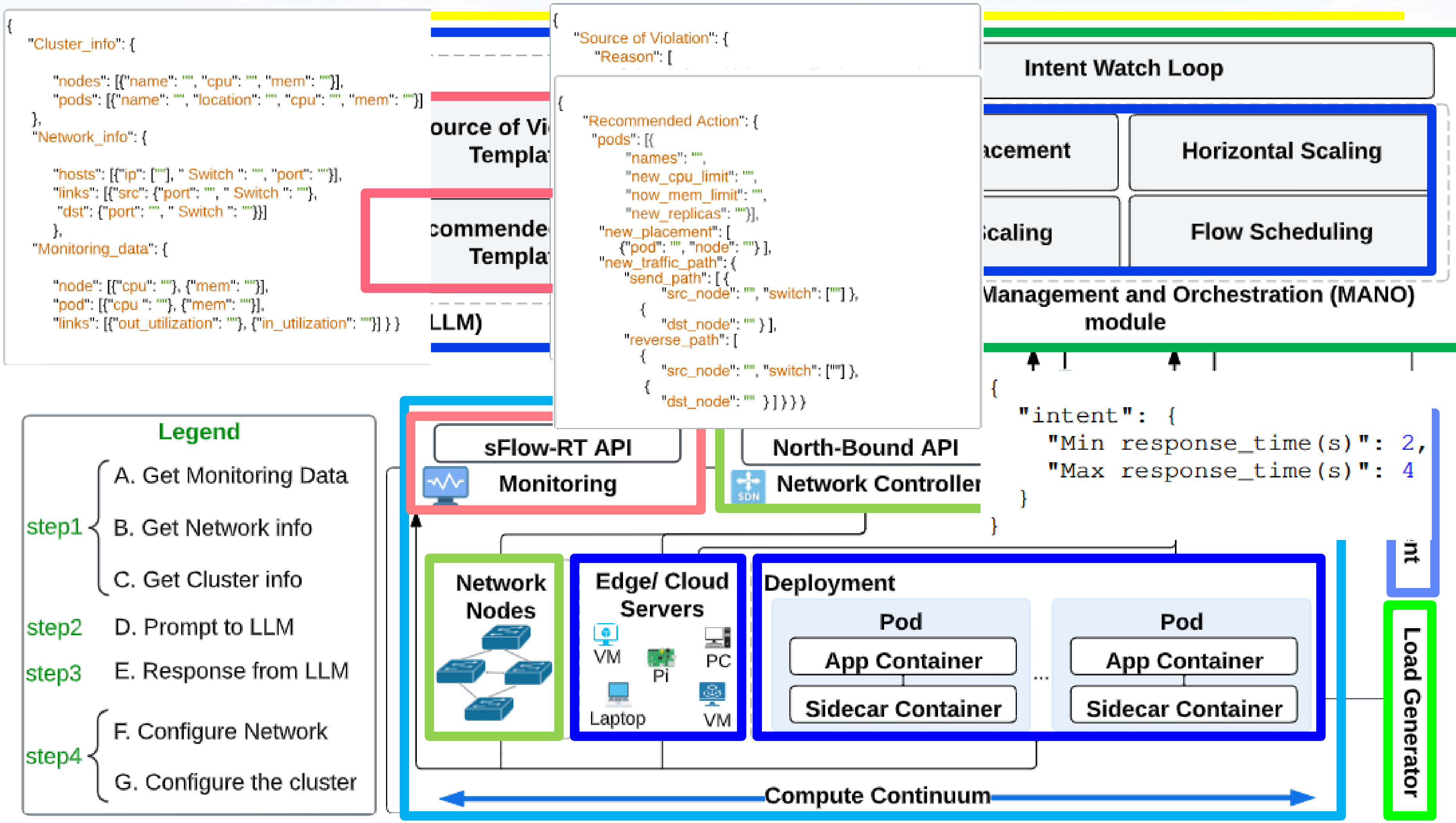
- Using general-purpose LLMs like ChatGPT
- Creating **intent-driven resource management** for Compute Continuum
  - LLMs analyze large data sets to address resource management challenges
  - To reduce manual work and complex rules
  - LLMs enable dynamic, context-aware resource management

N. Akbari, J. Grundy, A. Cheema, **Adel N. Toosi, IntentContinuum: Using LLMs to Support Intent-Based Computing Across the Compute Continuum,** IEEE International Conference on Web Services (ICWS 2025), Helsinki, Finland, 2025,.

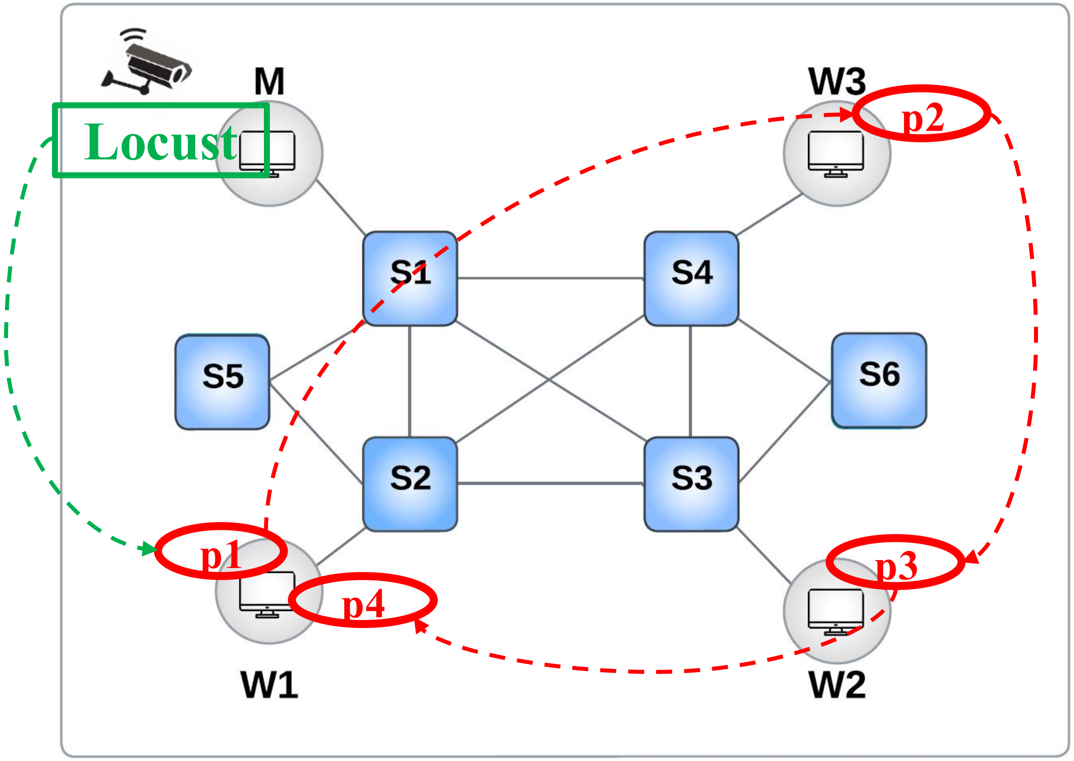
Motivation


Camera Based
Computer
Vision System
EDGE
Analytic
Platfom
Sorting
Cloud Data
Storage
Intent: Fast
Response
time?
Defective Parts
High Quality Parts
Product distribution with category

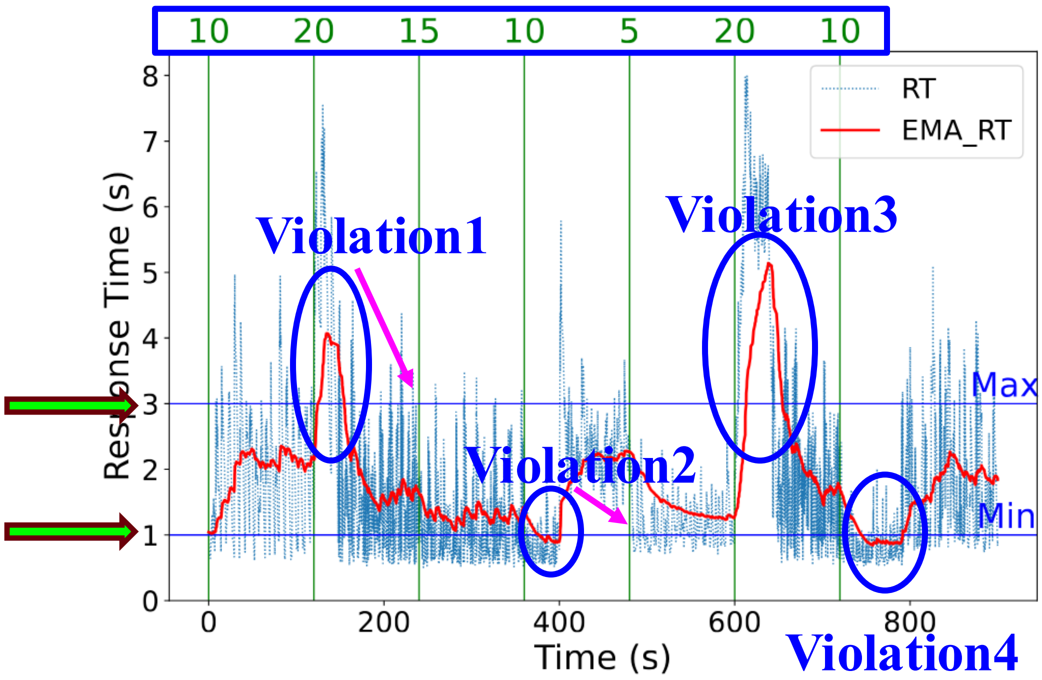
IntentContinuum: Proposed Architecture



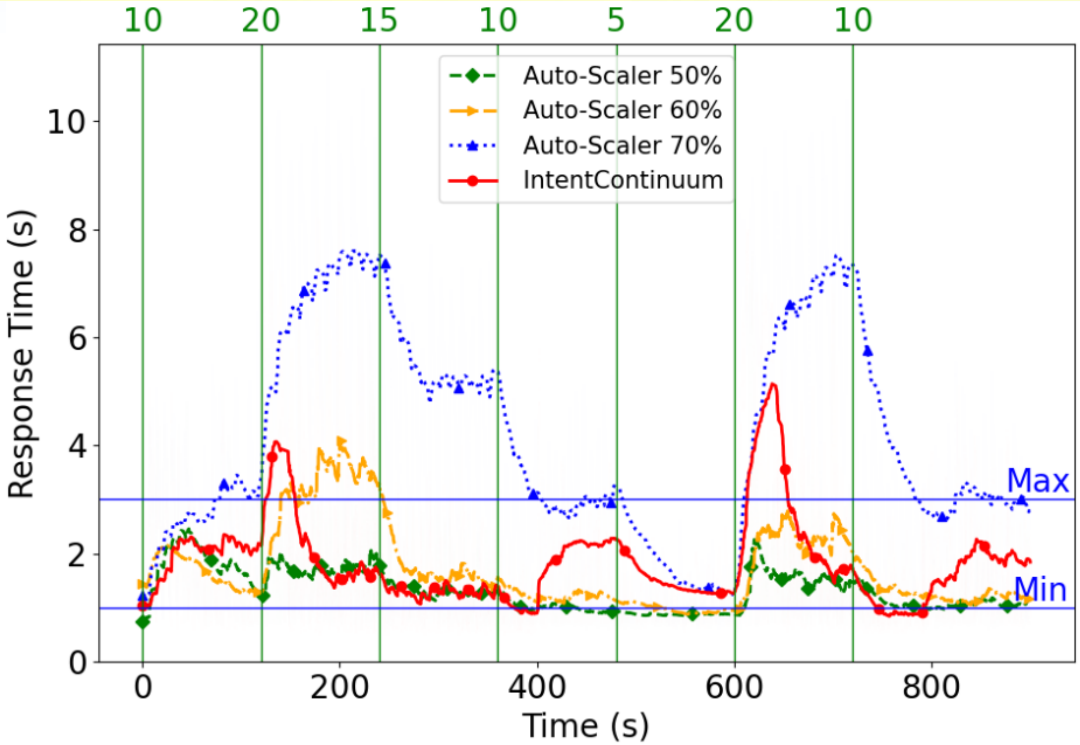
A sample Scenario



Computing Experiment

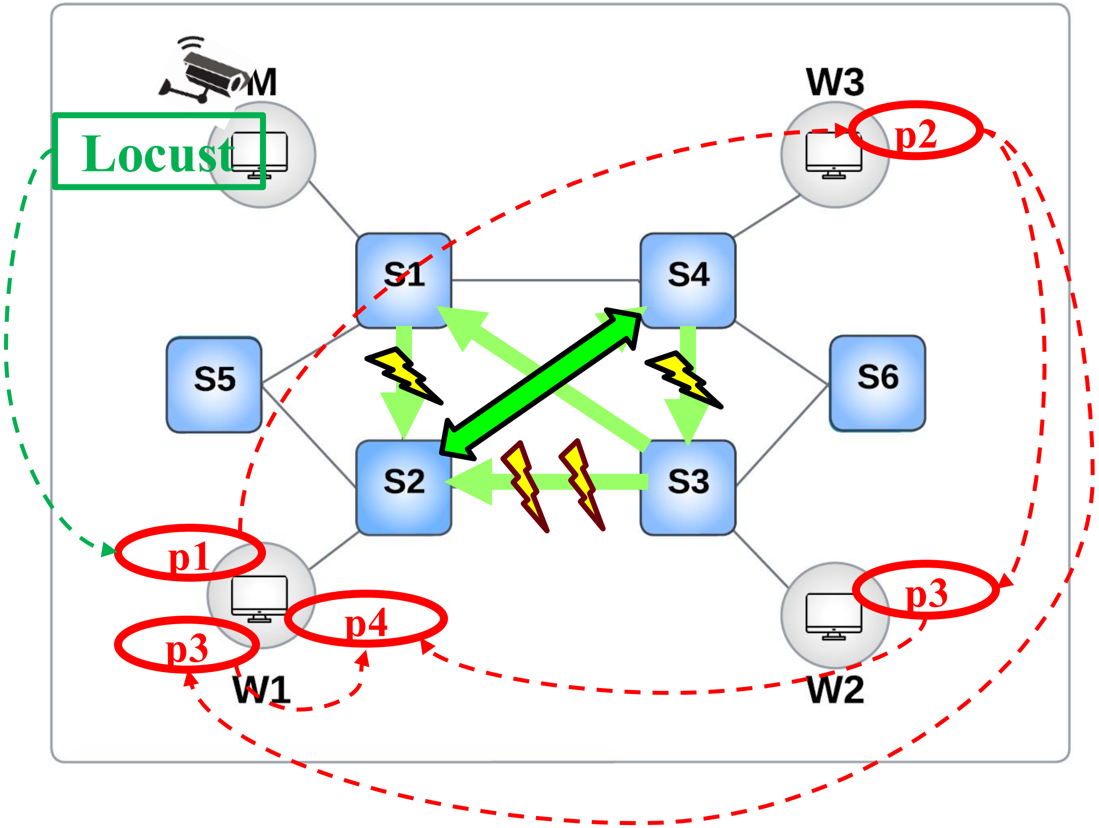


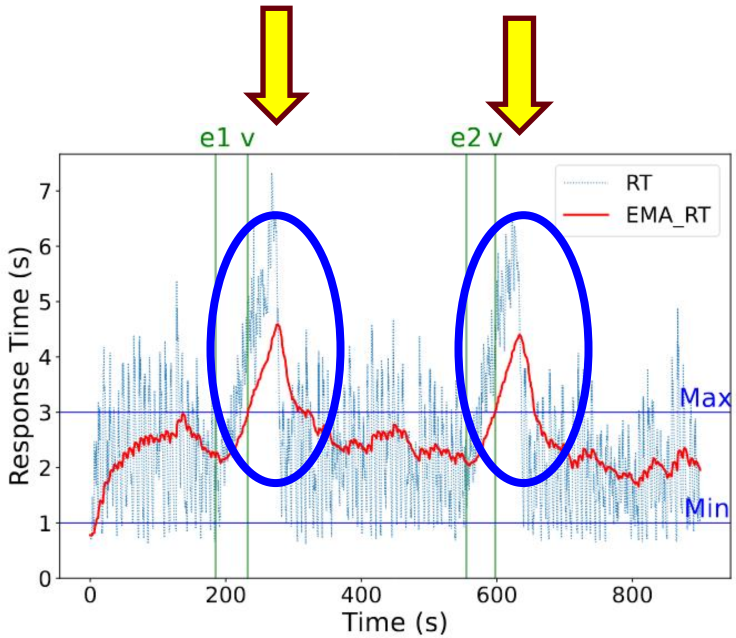
Comparison to Kubernetes HPA



Metrics	***IntentContinuum***	**Autoscaler**		
		70%	**60%**	**50%**
Intent Satisfaction%	**85%**	43%	79.5%	82.5%
Total Amount of Violated Time (s)	**143**	509	184	157

Networking Experiment





Discussions

- ## Strengths:
  - Can dynamically adapt to varying load levels while maintaining the application's response (**Vertical/horizontal scaling**)
  - **best balance** between intent satisfaction and resource utilization compared to various Kubernetes
  - address network issues such as link congestion or link failures by dynamically implementing **flow scheduling** or **pod replacements**
  - Minimizes the intent violations.
- ## Limitations:
  - Dependence on models
  - Limited Transparency and Clarity of LLM Recommendations
  - Scalability (context limits) and Processing Overhead
  - Financial implications

Summary

- **Serverless Edge Computing**: Overview and its significance in modern distributed systems
  - Showcased some ongoing research in Serverless Edge Computing in DisNet lab
- **Compute Continuum**: Concept introduction and its impact on edge-to-cloud integration
  - **Intent-based and Serverless Vehicular Edge Computing**
  - **iContinuum**: a tool for emulating edge-to-cloud continuum
  - **intentContinuum**: LLM as a scheduler for DevOps across the compute continuum

E: adel.toosi@unimelb.edu.au
W: http://adelnadjarantoosi.info
P: (03 90354322 Office)