Lessons on Building Edge AI Solutions towards 6G

Aaron Ding

TU Delft, Netherlands

TU Delft

Outline

- Introduction
- Edge Analytics
- Edge Offloading
- Takeaway

TU Delft

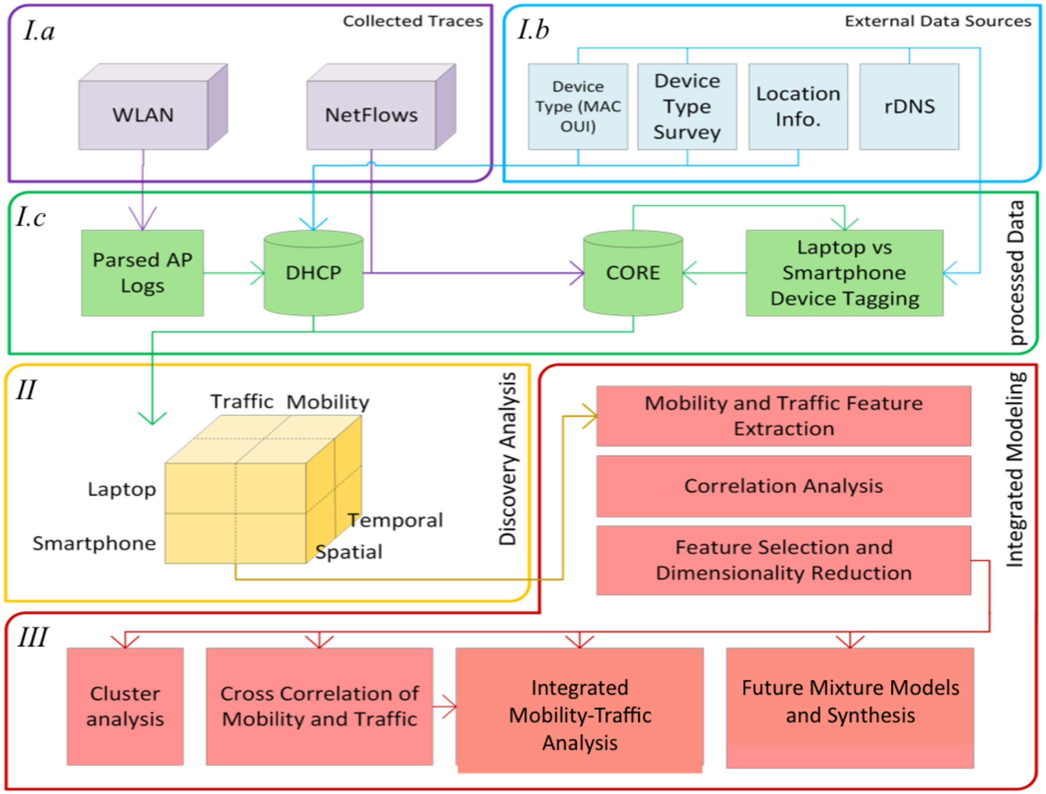
Introduction: Edge AI

Outline

TU Delft

FLAMeS on Wireless Edge Analytics

- Demand for wireless edge analytics
  - Look into the edge
- Mobility and Traffic
  - Interplay
  - Across device types
  - Modeling insights



Output 1: IEEE INFOCOM 2018
Output 2: ACM MSWiM 2019

TU Delft

Flutes vs. Cellos

- Mobile vs Laptop
  - Impact on data traffic and mobility
  - Integrated mobility-traffic models
- Mobility-Traffic Interdependence is not well-studied
  - Usable traces are hard to obtain
  - Privacy concern (GDPR)

VS

Motivations

- Two major factors affecting mobile network performance are **mobility** and **traffic** patterns
  - Mobility and Network usage characterize different aspects of human behavior, e.g., using different devices
  - Simulations, analytical-based performance evaluations, and future predictive caching schemes rely on **models** to approximate factors affecting the network
- Many earlier mobility modeling studies use pre-smartphone WLAN traces (**device types** not considered)
- **Mobility-Traffic Interdependence** is not well-studied

FLAMeS Dataset

- Size of raw dataset
  - 30+ TB, 1760 APs, 138 buildings, over 479 days
  - 76 billion NetFlow records, 555 million AP traces, 316k devices
- Device categorization
  - MAC address survey
  - OUI matching
  - Web domain analysis

	# Records		Traffic Vol. (TB)		# MAC	
	DHCP	CORE	TCP	UDP	WLAN	CORE
Flutes	412.0 M	2.13 B	56.18	4.50	186.0 K	50.3 K
Cellos	101.0 M	4.20 B	73.85	12.90	93.2 K	27.1 K
Total	557.5 M	6.53 B	134.39	17.61	316.0 K	80.0 K

Research Questions

- How different are mobility and traffic characteristics across device types, time and space?
  - Multi-dimensional study
- What are the relationships / correlation?
  - Interdependency
- Should new, integrated mobility-traffic models be devised to capture these differences? What is the value and utility of integrating mobility and traffic?
  - If so, how

Discovery and Insights

- Mobility analysis
  - Session start probability, radius of gyration, visit preference, sessions per building, etc.
- Traffic analysis
  - Flow level, spatial, temporal behavior
- Integrated analysis
  - Feature engineering, modeling insights

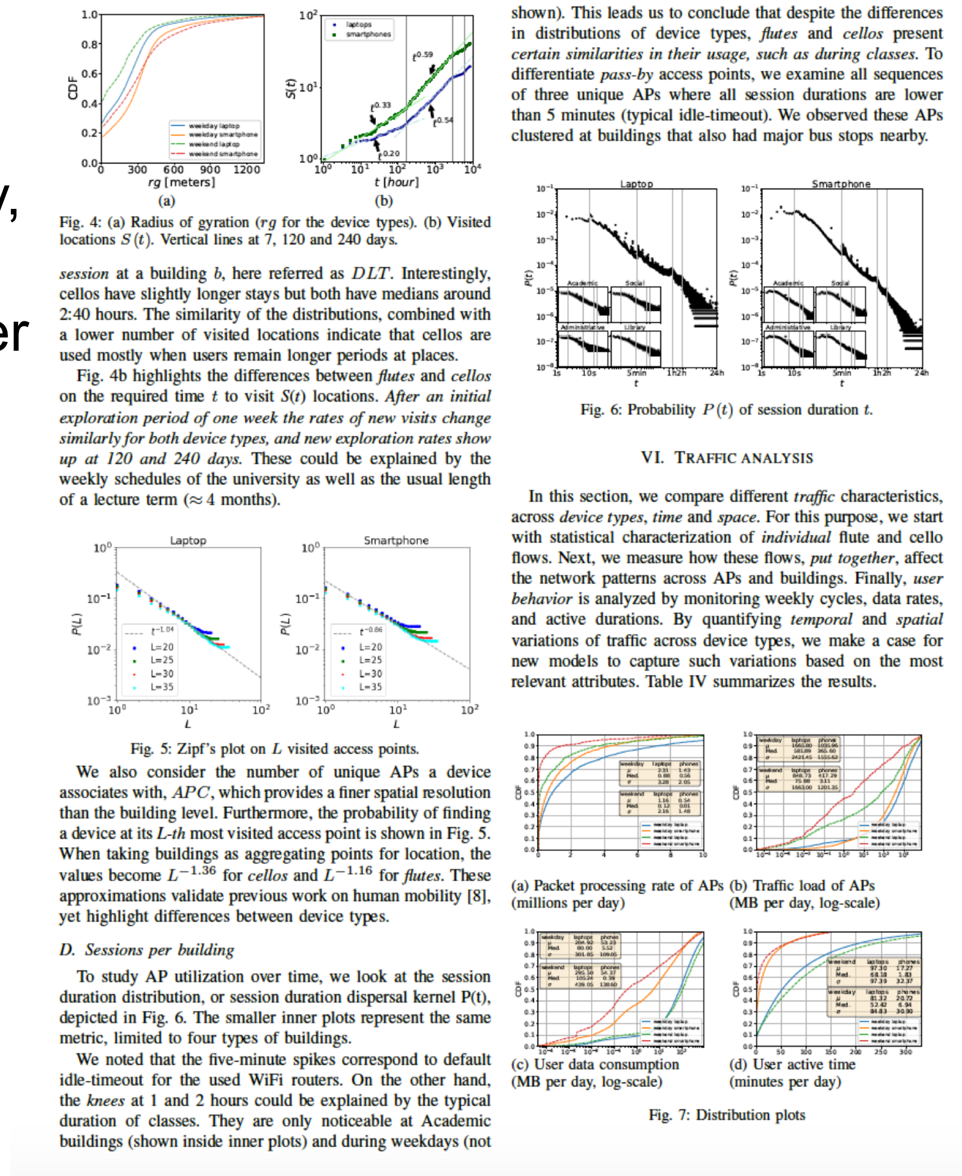


Fig. 4: (a) Radius of gyration (rg for the device types). (b) Visited locations $S(t)$. Vertical lines at 7, 120 and 240 days.

session at a building b, here referred as DLT. Interestingly, cellos have slightly longer stays but both have medians around 2:40 hours. The similarity of the distributions, combined with a lower number of visited locations indicate that cellos are used mostly when users remain longer periods at places.

Fig. 4b highlights the differences between *flutes* and *cellos* on the required time t to visit $S(t)$ locations. *After an initial exploration period of one week the rates of new visits change similarly for both device types, and new exploration rates show up at 120 and 240 days.* These could be explained by the weekly schedules of the university as well as the usual length of a lecture term ($\approx$ 4 months).

Fig. 5: Zipf's plot on L visited access points.

We also consider the number of unique APs a device associates with, APC, which provides a finer spatial resolution than the building level. Furthermore, the probability of finding a device at its *L-th* most visited access point is shown in Fig. 5. When taking buildings as aggregating points for location, the values become $L^{-1.36}$ for *cellos* and $L^{-1.16}$ for *flutes*. These approximations validate previous work on human mobility [8], yet highlight differences between device types.

D. Sessions per building

To study AP utilization over time, we look at the session duration distribution, or session duration dispersal kernel P(t), depicted in Fig. 6. The smaller inner plots represent the same metric, limited to four types of buildings.

We noted that the five-minute spikes correspond to default idle-timeout for the used WiFi routers. On the other hand, the *knees* at 1 and 2 hours could be explained by the typical duration of classes. They are only noticeable at Academic buildings (shown inside inner plots) and during weekdays (not shown). This leads us to conclude that despite the differences in distributions of device types, *flutes* and *cellos* present *certain similarities in their usage, such as during classes.* To differentiate *pass-by* access points, we examine all sequences of three unique APs where all session durations are lower than 5 minutes (typical idle-timeout). We observed these APs clustered at buildings that also had major bus stops nearby.

Fig. 6: Probability $P(t)$ of session duration t.

VI. TRAFFIC ANALYSIS

In this section, we compare different *traffic* characteristics, across *device types*, *time* and *space*. For this purpose, we start with statistical characterization of *individual* flute and cello flows. Next, we measure how these flows, *put together*, affect the network patterns across APs and buildings. Finally, *user behavior* is analyzed by monitoring weekly cycles, data rates, and active durations. By quantifying *temporal* and *spatial* variations of traffic across device types, we make a case for new models to capture such variations based on the most relevant attributes. Table IV summarizes the results.

(a) Packet processing rate of APs (millions per day)

(b) Traffic load of APs (MB per day, log-scale)

(c) User data consumption (MB per day, log-scale)

(d) User active time (minutes per day)

Fig. 7: Distribution plots

TU Delft

Data, Data, Data

- Big shot … grand rejection

Big Data For The Win?

- What were boasted, all **fired back**

> ***" Your data is not new enough "***
>
> ***" Your findings may not reflect the latest situation "***
>
> ***" Your analysis coverage is limited "***
>
> ***" Your insights for modeling are incomplete "***
>
> ***" Your work impact is not … "***
>
> ***…***

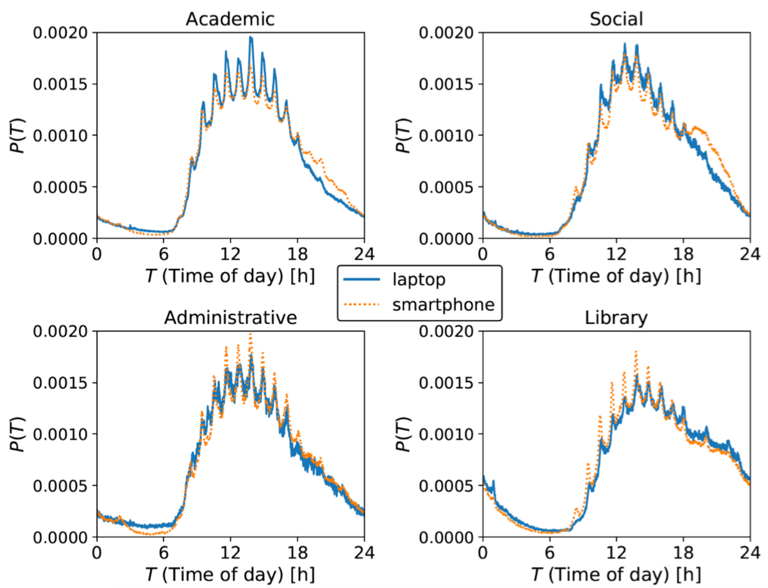
TU Delft

What Went Wrong?

- Reflections
  - Painful but valuable process
  - Comments are actually valid
- Focus adjustment
  - Start over again
  - Rewrite the whole thing

Methodology or Dataset?

- Not just to impress others



	Flutes (F)			Cellos (C)			Ratio (C/F)	
	μ	*mdn*	σ	μ	*mdn*	σ	μ	*mdn*
LJM	435	296	813	178	1	624	0.409	**0.003**
	350	168	683	97	1	312	0.277	0.006
DIA	549	411	874	195	1	642	0.355	**0.002**
	425	179	739	107	1	338	0.252	0.006
TJM	1582	707	2336	378	1	1444	0.239	**0.001**
	1036	279	1793	252	1	1766	0.243	0.004
GYR	396	290	2725	321	191	3265	1.102	1.019
	330	248	1368	178	65.1	1800	1.247	1.4
BLD	5.4	3	5.6	1.8	1	2.1	0.811	0.659
	2.8	2	4.1	1.5	1	1.8	0.539	0.262
APC	11.8	6	13.3	3.7	2	4.8	0.333	0.333
	7.2	4	8.8	3	2	3.8	0.536	0.5
PDT	225	161	219	248	164	254	0.314	0.333
	223	135	272	278	189	292	0.417	0.5
DTL	316	235	302	316	217	305	1	0.92
	326	247	308	316	221	309	0.97	0.89

Start time	Finish time	Duration	Source IP	Destination IP	Protocol	Source port	Destination port	Packet count	Flow size
1334332274.912	1334332276.576	1.664	173.194.37.7	10.15.225.126	TCP	80	60482	157	217708

User IP	User MAC	AP name	AP MAC	Lease begin time	Lease end time
10.130.90.3	00:11:22:33:44:55	b422r143-win-1	00:1d:e5:8f:1b:30	1333238737	1333238741

TUDelft

Back to the Basics

- Wireless edge analytics

Data

- NetFlow
- WLAN Traces
- Device Classes
- DHCP
- Core (merged)

Analysis

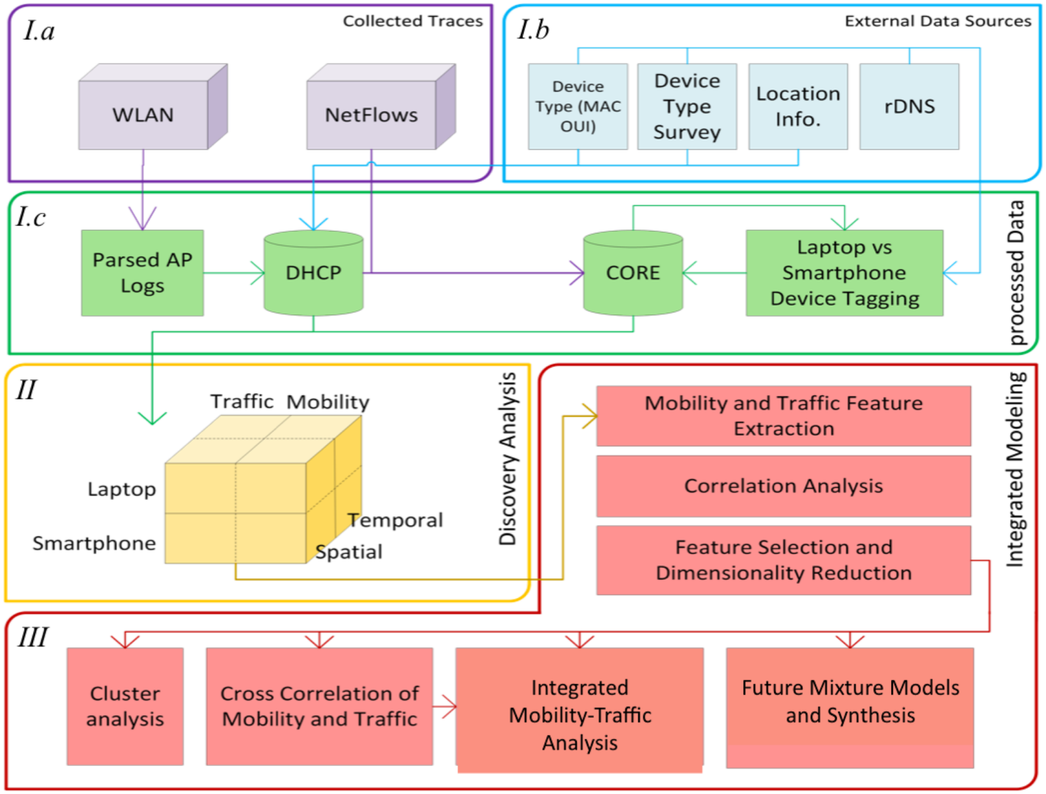
- Cellos vs Flutes
- Spatio-Temporal Characteristics
- Traffic and Mobility Features

Modeling

- Mobility and Traffic feature extraction
- Correlation/Cross-Correlations
- Mixture Models and Synthesis

Framework for Edge Wireless Analytics

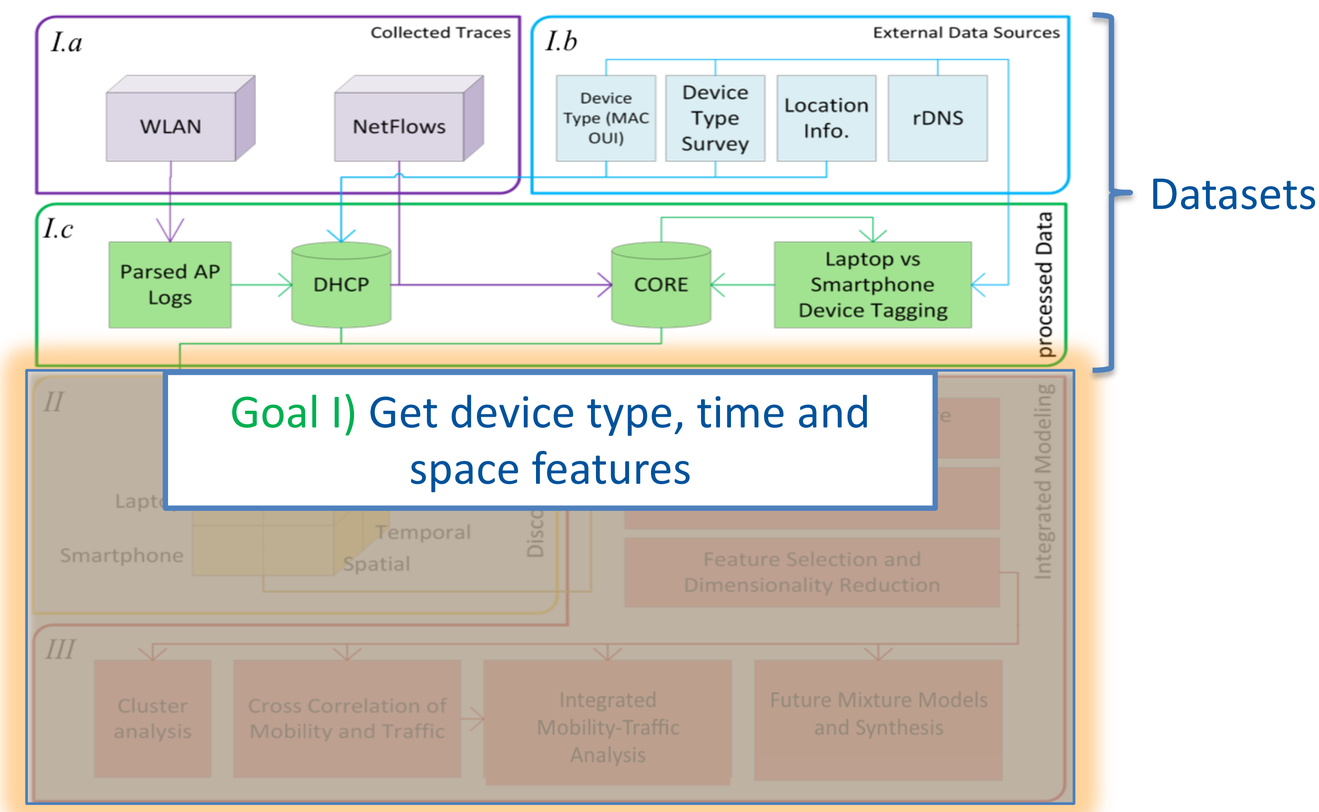
- FLAMeS workflow



TU Delft

FLAMeS

- Feature extraction
  - WLAN logs and NetFlows

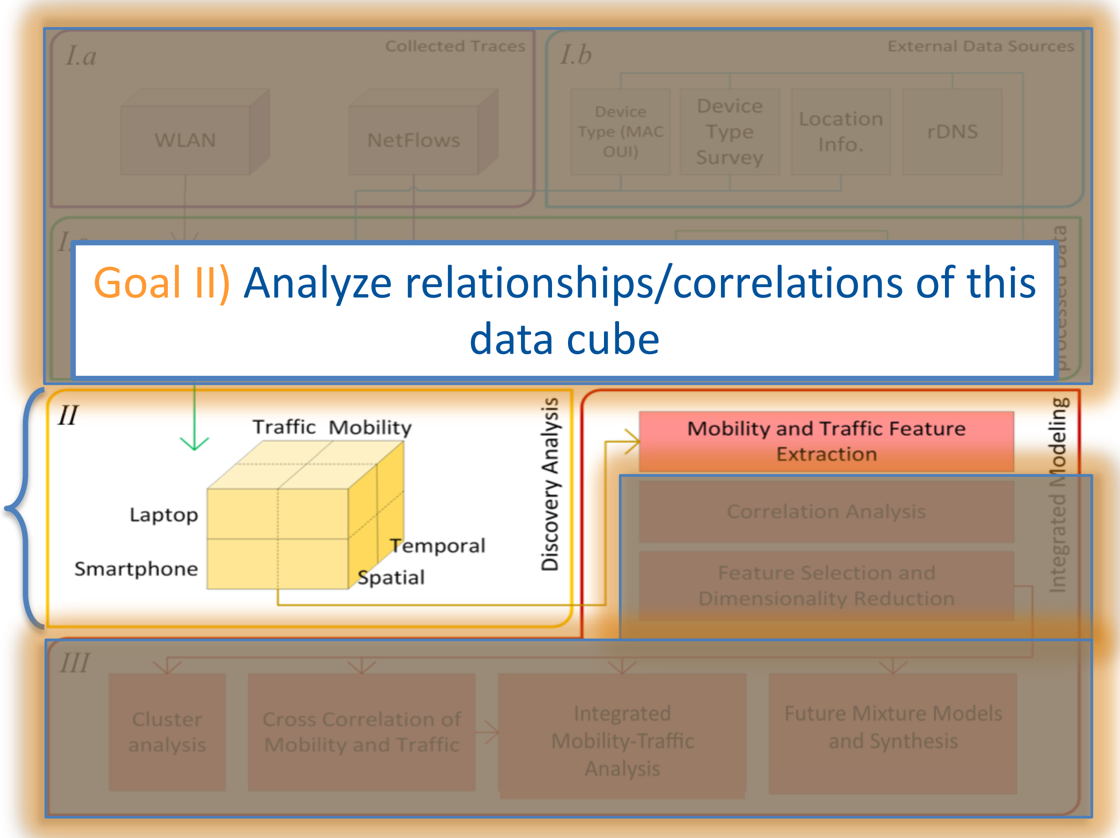


FLAMeS

- Data traffic and mobility interdependency

Data cube, **traffic/mobility** analyzed **temporally**, **spatially**, and **per device type**

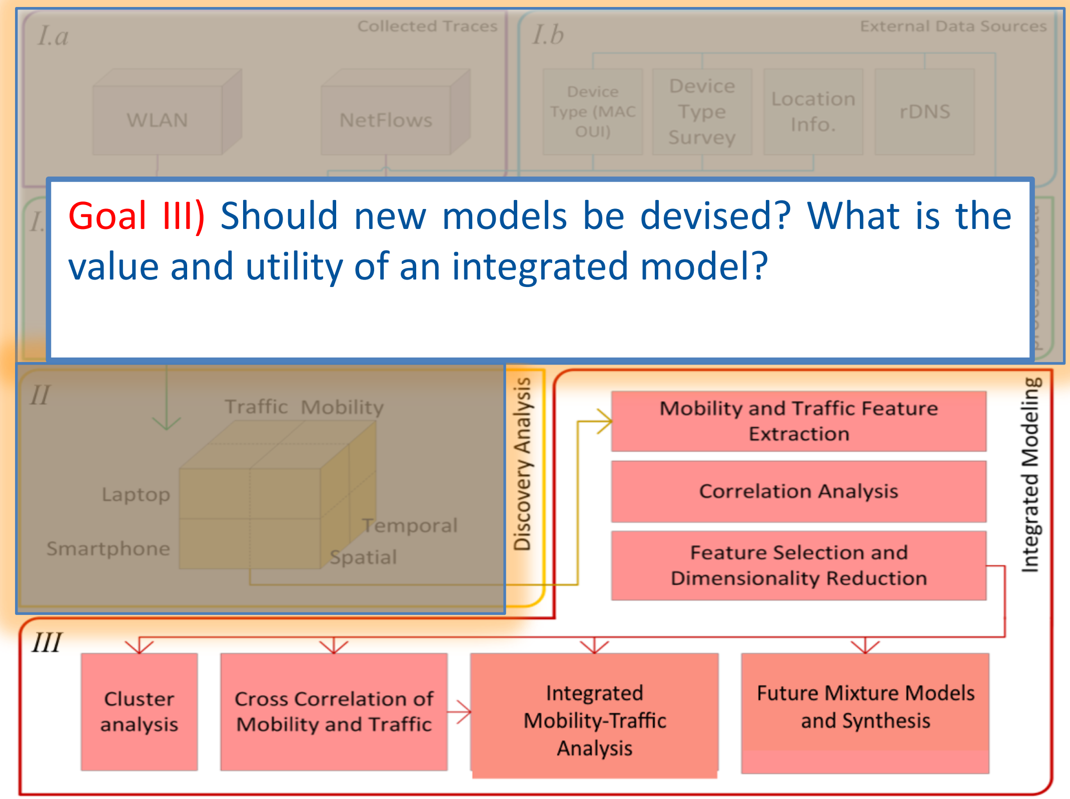
Goal II) Analyze relationships/correlations of this data cube



TU Delft

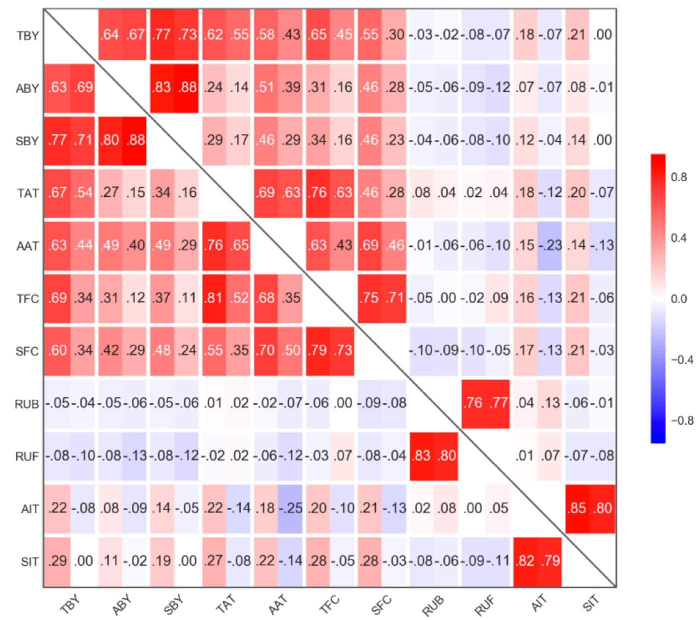
FLAMeS

- Towards integrated modeling



Adjust the Focus

- Methodology and framework
  - Dataset mainly as a tool to verify our assumption and investigations

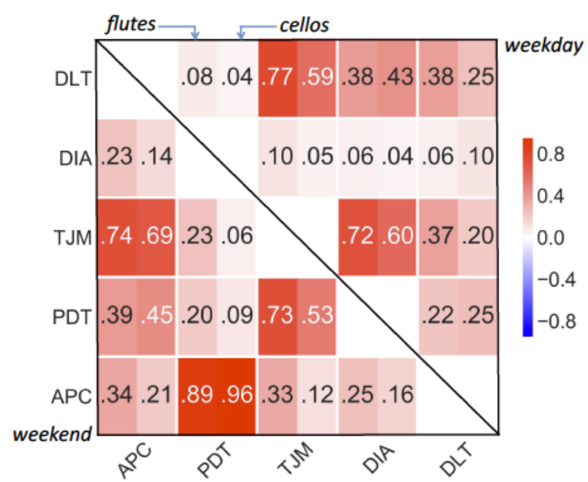


Abbr.	Description
TBY	Total flow bytes
ABY	Avg. flow bytes
SBY	Std. flow bytes
TAT	Total active time
AAT	Avg. active time
TFC	Total flow count
SFC	Std. flow counts
RUB	UDP bytes / total bytes
RUF	UDP flows / total flows
AIT	Avg. IAT
SIT	Std. IAT

Made It!

IEEE INFOCOM 2018

ACM MSWiM 2019



Abbr.	Description
APC	AP Count (unique)
PDT	Preferred building Δt
TJM	Total (sum) jumps
DIA	Diameter of mobility
DLT	Delta time (time of network association)

TU Delft

Remarks

- It is crucial to differentiate **flutes vs. cellos** for both **mobility and traffic** due to their very different nature. Correlations of these features matter, and should be captured in models.

- Traffic generation, **spatial** locations, and **temporal** behavior can be linked per device type and per user "community" (e.g. students of different disciplines at various buildings).

- There is significant potential for an **integrated mobility-traffic model** that captures relationships across **device types**, **time** and **space**.

TUDelft

Lessons

- Risk 1: Boasting dataset value
  - Don't over-estimate, nor over-claim. Otherwise, Over..
  - Correct focus/position is crucial
- Risk 2: Good stuff needs less polishing
  - Will block the work from top venue
  - Balance and structure

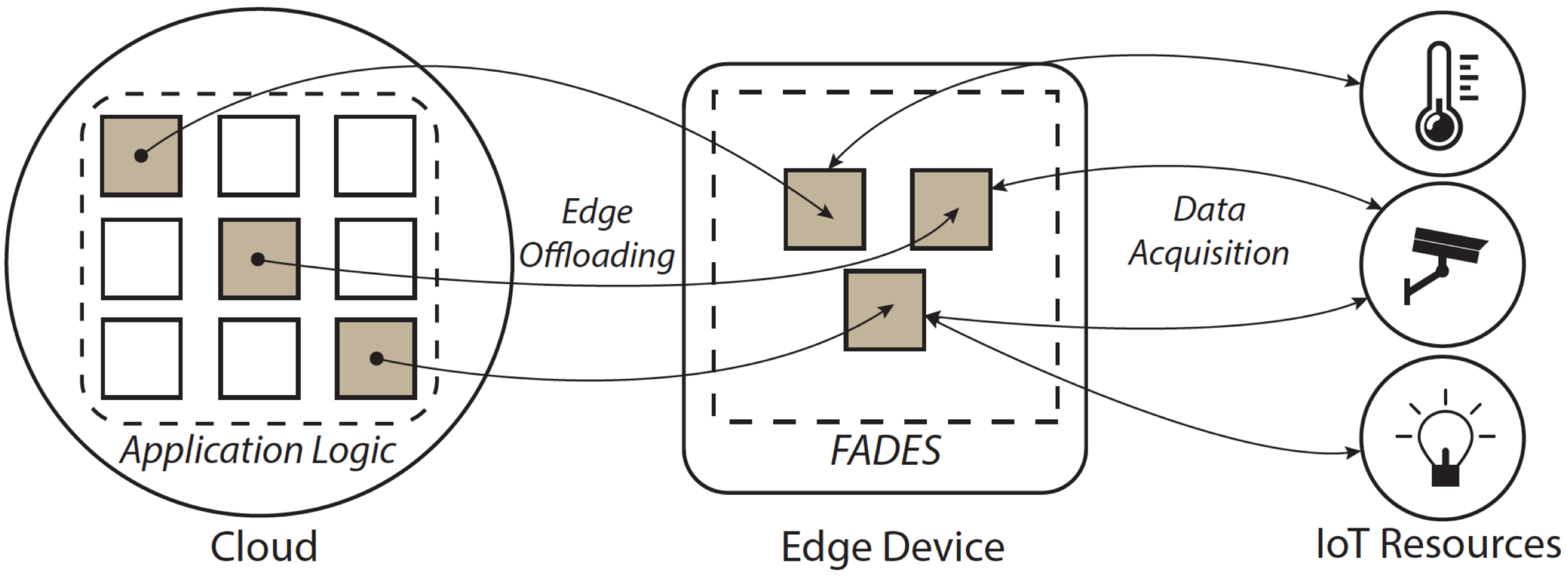
Toolkit and in-depth study are appreciated

Outline

- Introduction
- Edge Analytics
- Edge Offloading
- Takeaway

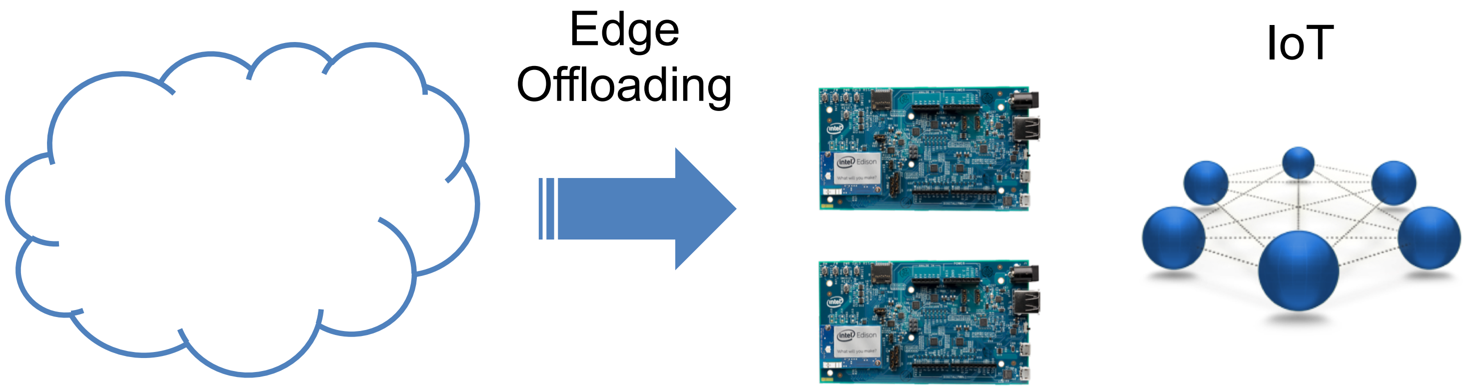
Edge Offloading

- Fine-grained offloading for IoT



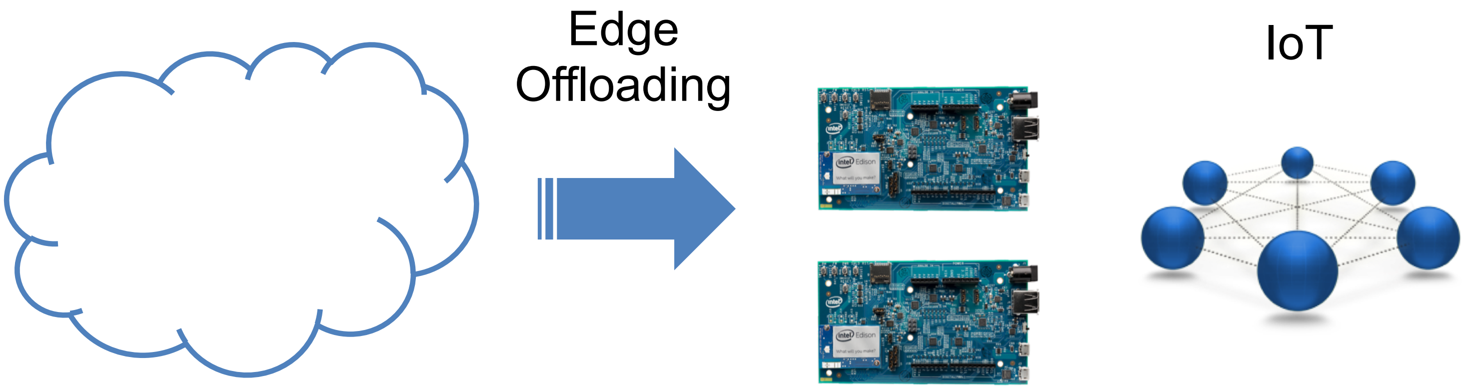
Edge Offloading

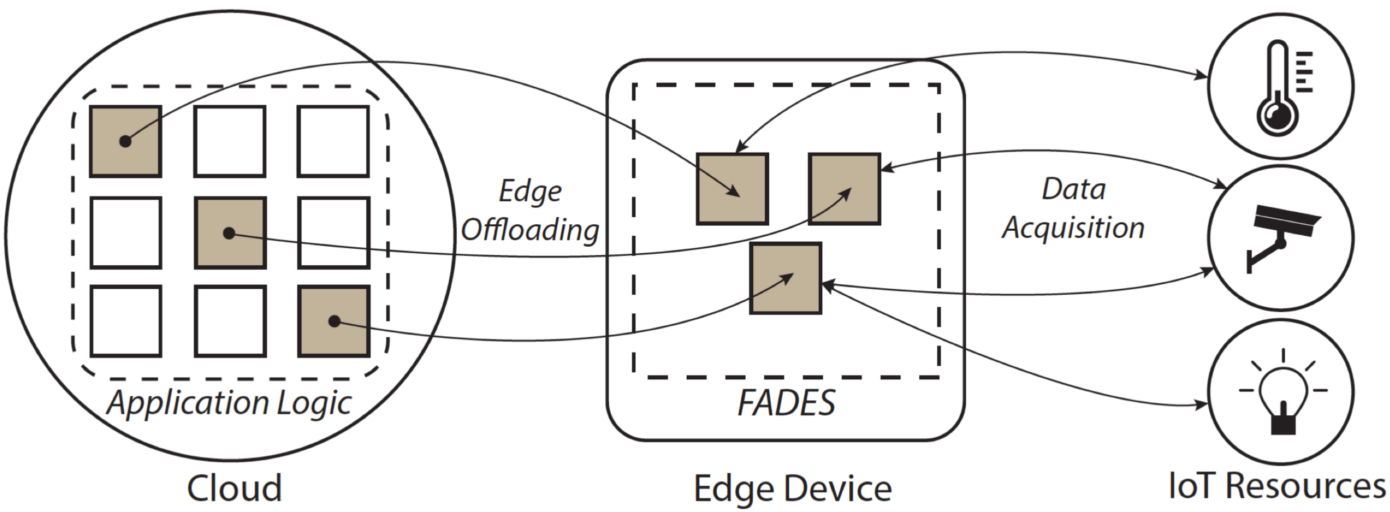
- Reverse direction



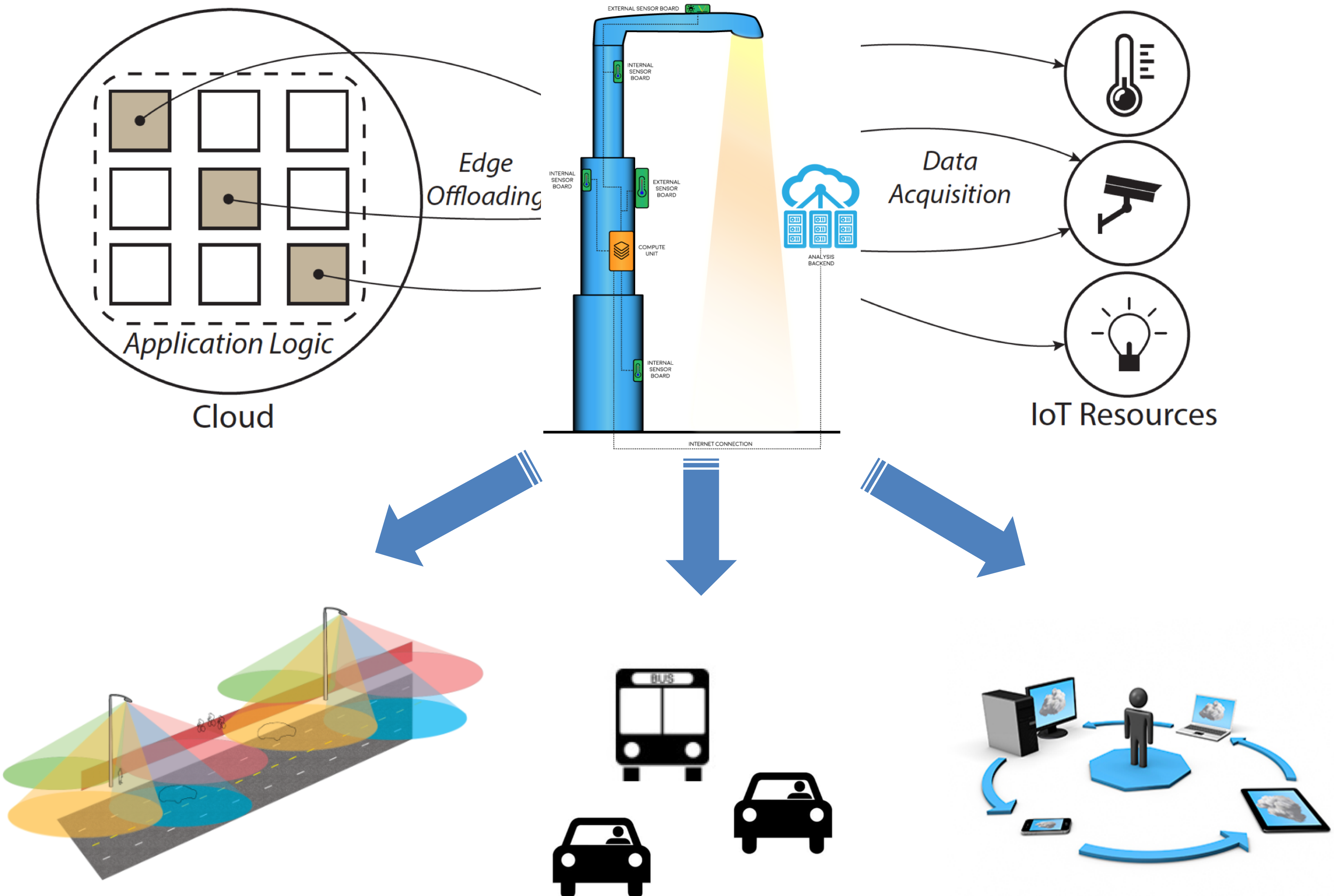
Edge Offloading

- Cloud – Edge – IoT



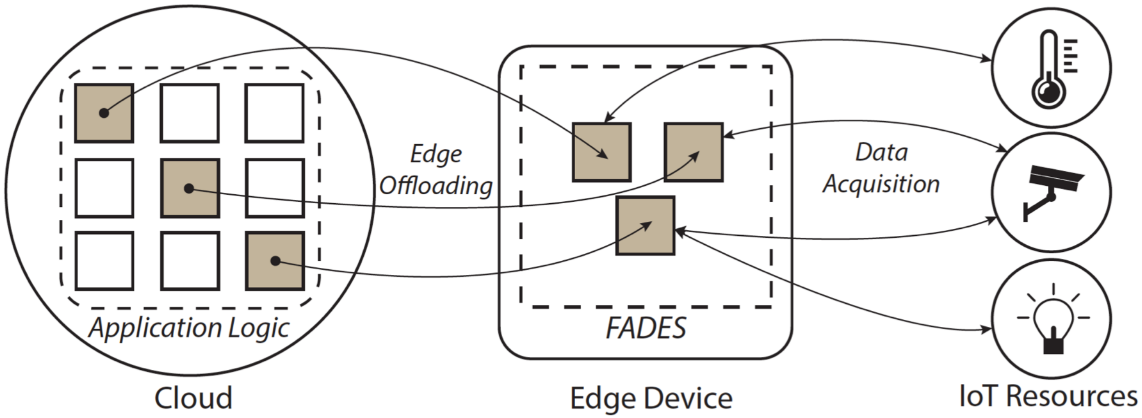


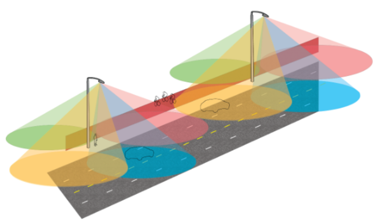
The Real Benefits

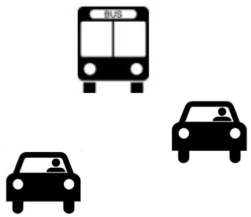


How to Offload to Edge?

- FADES
  - Unikernel
  - MirageOS
  - Single purpose
  - Modular
  - Compact size
  - On demand
  - Isolation

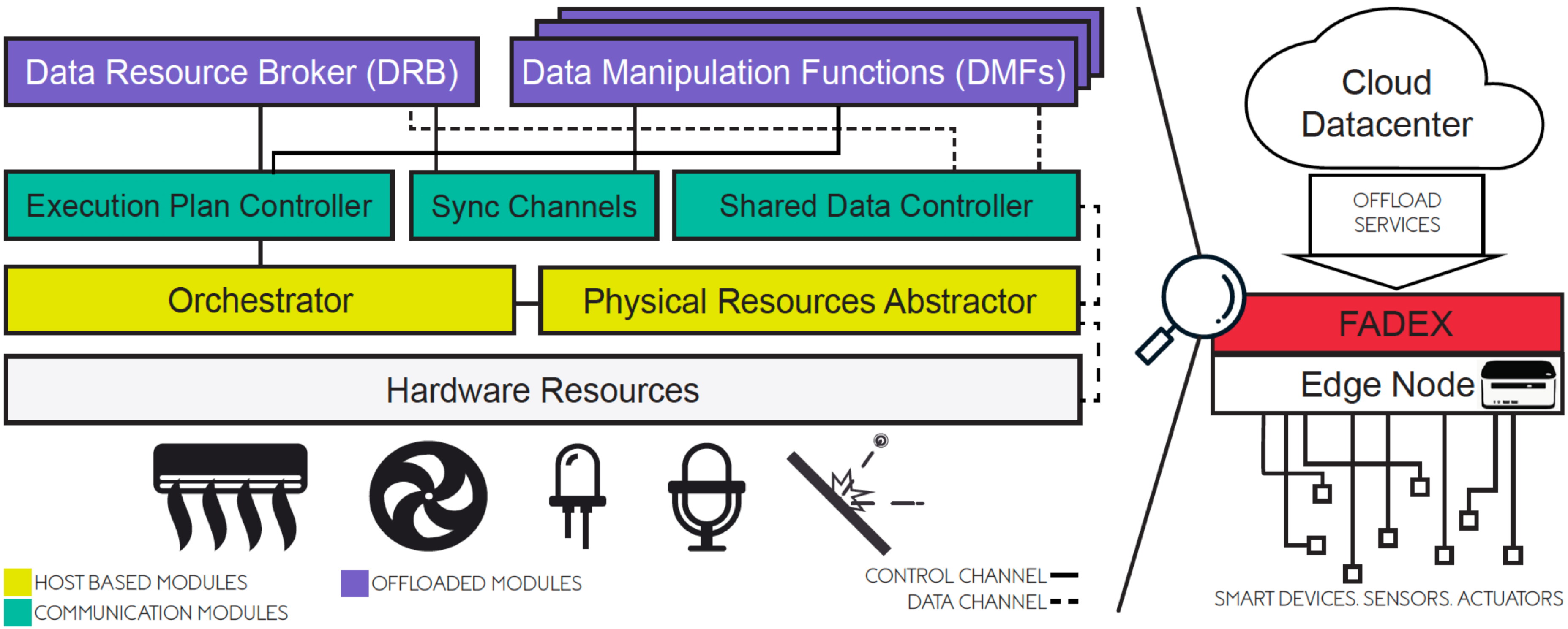






Lightweight Virtualization

Design and Implementation



TU Delft

Use Cases

- Software-oriented
  - IoT sensing data
  - Image
  - Audio
  - Data encryption

- Hardware-oriented
  - Actuator access

Fine-grained Edge Offloading

Does This Really Work?

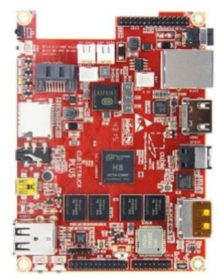
TUDelft

Experiments

- Feasibility
  - System performance and limitation on x86 and ARM
  - Memory utilization, network
  - Does this really work?

Test over three types of devices

Device	CPU	RAM	Network
Cubietruck	Allwinner A20 ARM Cortex-A7 dual-core @ 1GHz	1 GB	100Mb Ethernet
Intel NUC	Intel(R) Core(TM) i5-6260U CPU@1.80GHz	16 GB	1000 Mb Ethernet
Dell Server	Intel(R) Xeon(R) CPU E5-2640 v3@2.60GHz	140 GB	1000 Mb Ethernet

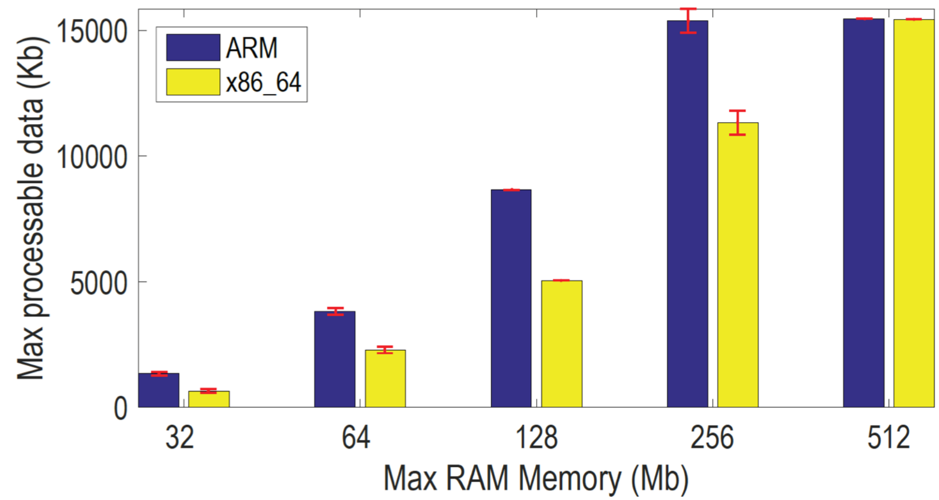


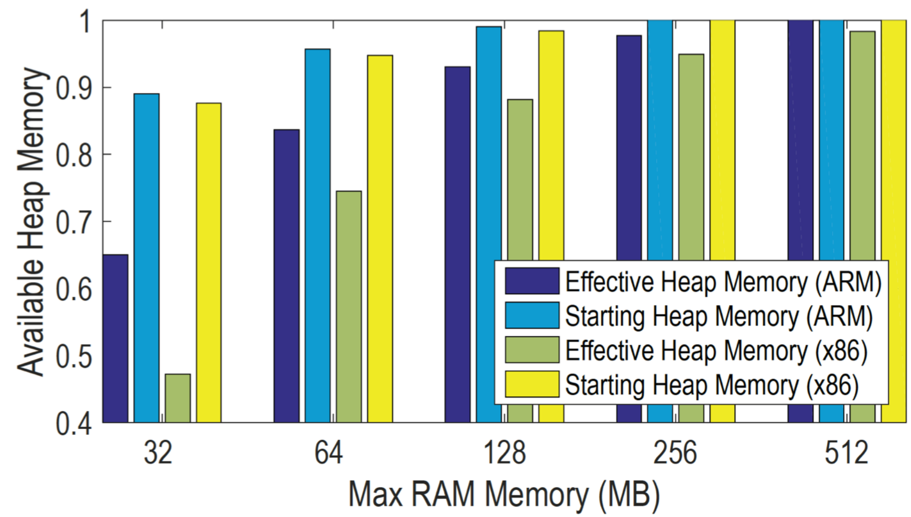
Observations

- On X86 and ARM
  - Micro benchmark
- Immature yet
  - Image size under two arch. affects available runtime memory
  - Low RAM case

Considerable loss of available memory for low RAM Unikernels.

Impact on resource utilization for IoT cases.





TU Delft

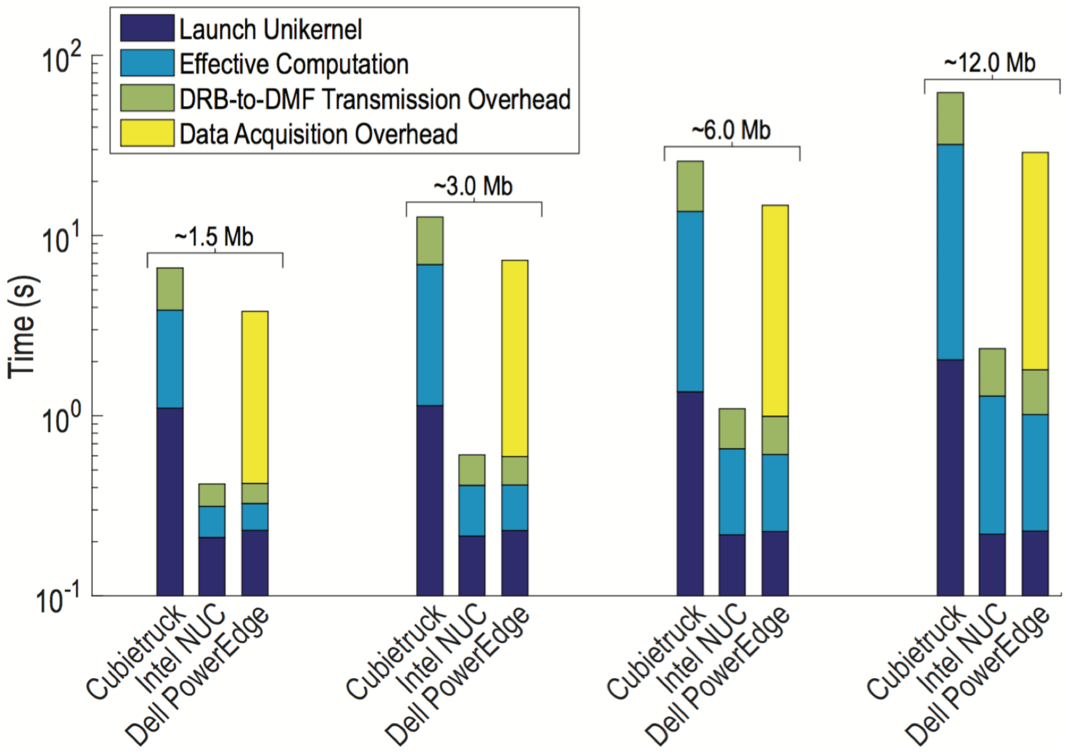
Observations

- Bright side
  - Edge beats the cloud

Cubietruck, Intel NUC have local copy of data (the edge setting)

Dell PowerEdge fetches data from remote location (the cloud setting)

Sufficiently powerful edge device combined with local data makes edge offloading convincing



Observations

Hardware Limitations

- Demanding to find suitable embedded boards that can support Xen and MirageOS.
- Deployment on Cubietruck board was more challenging than on Intel NUC.

Platform Limitations

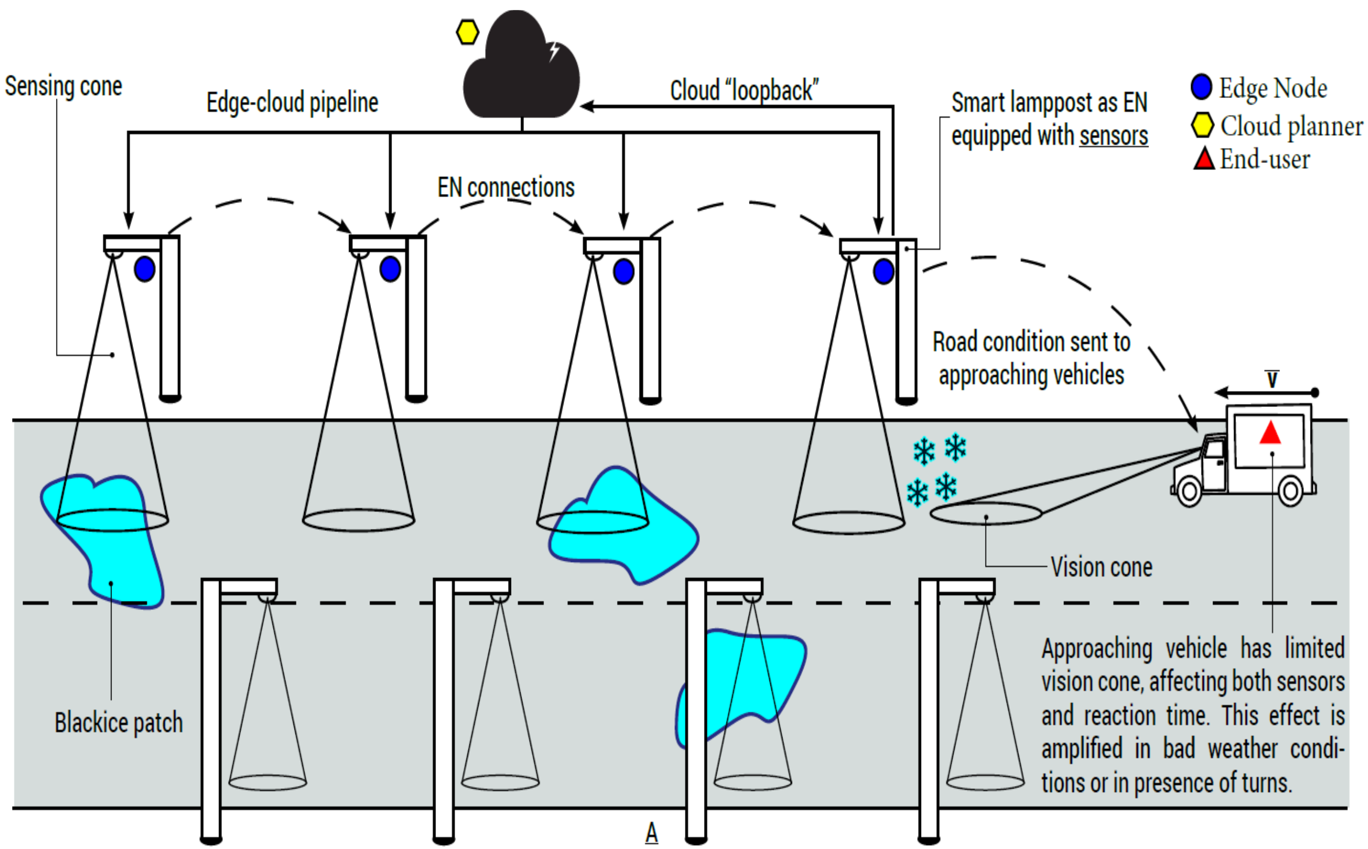
- Issues with the network API when transferring data between two unikernels.
  - Culprit: a bug in the TCP/IP MirageOS stack that doesn't handle properly writing packets larger than the MTU. In consequence, we had to introduce an extra chunking function at the application layer to split, and later reconstruct the data.
- Single CPU considerations

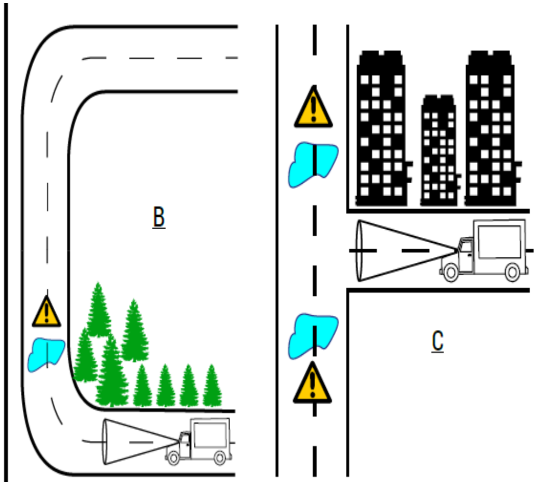
Security Concerns

- Guarantee the authenticity and validity of the offloaded tasks
- Without a signing and validation infrastructure to discriminate legit from tampered unikernels, we might risk executing malicious code and infringe the security requirements
- Side-effects of "decentralizing" control and delegating responsibilities
- Strict control and monitoring are required

Edge Chaining System

Best Paper Award





In this above examples, the presence of a turn combined with reduced field of view due to trees (e.g., mountain road) or buildings negates on-board sensors capability of detecting a black ice patch in time for the driver to react.

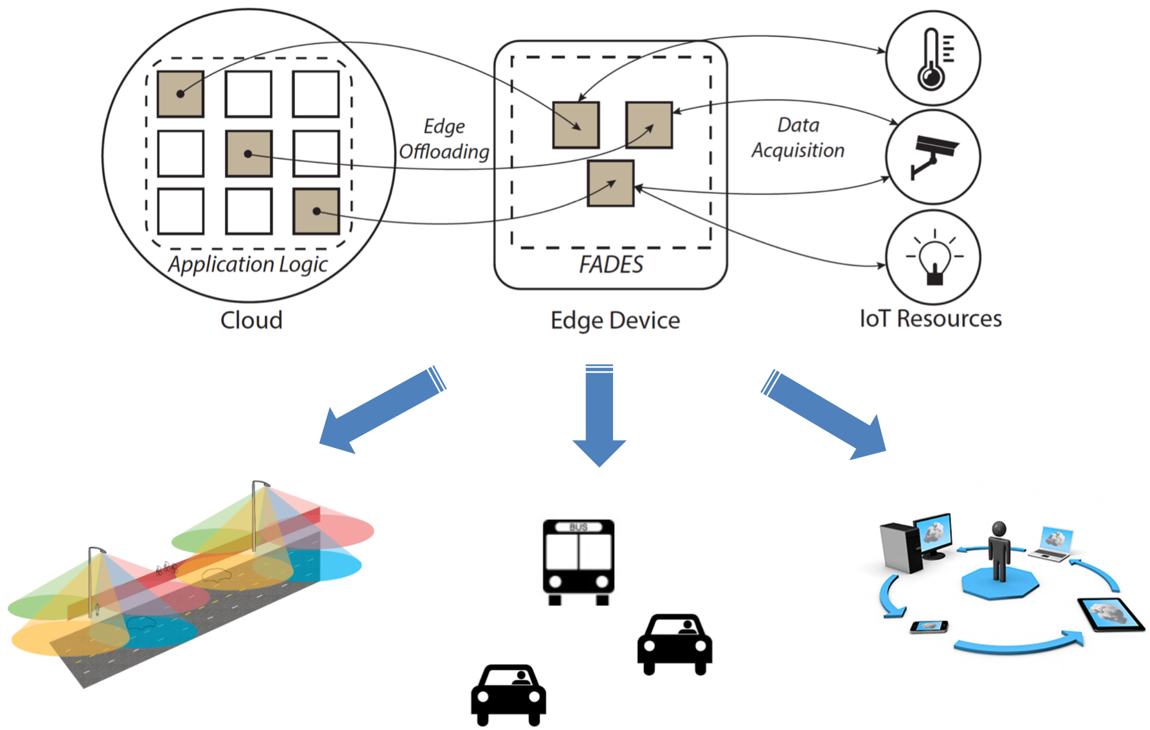
In this conditions, the presence of a fixed infrastructure can be crucial in providing road hazards information.

TUDelft

[1] "Consolidate IoT Edge Computing with Lightweight Virtualization". Volume 32, Issue 1, ***IEEE Network 2018*** **Impact Factor 7.9**

[2] "Edge Chaining Framework for Black Ice Road Fingerprinting". ***ACM EdgeSys 2019*** **Best Paper Award**

[3] "ECCO: Edge-Cloud Chaining for Road Context Assessment". ***ACM/IEEE IoTDI 2020*** **Premier IoT Conference**



Lessons

- Risk 1: Too many options
  - Containers, unikernels
  - System development takes long time
- Risk 2: Worry too much about ‘fancy’ use cases
  - Not the deciding factor
  - Feasible assumption

Advantages of being the First

- Share insights with community
- Even initial work will be appreciated

TU Delft

Outline

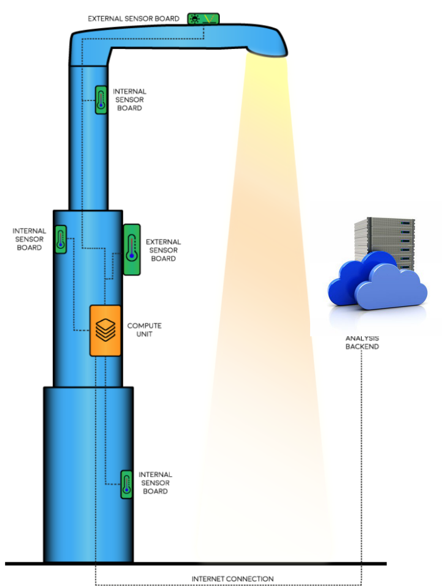
- Introduction
- Edge Analytics
- Edge Offloading
- Takeaway

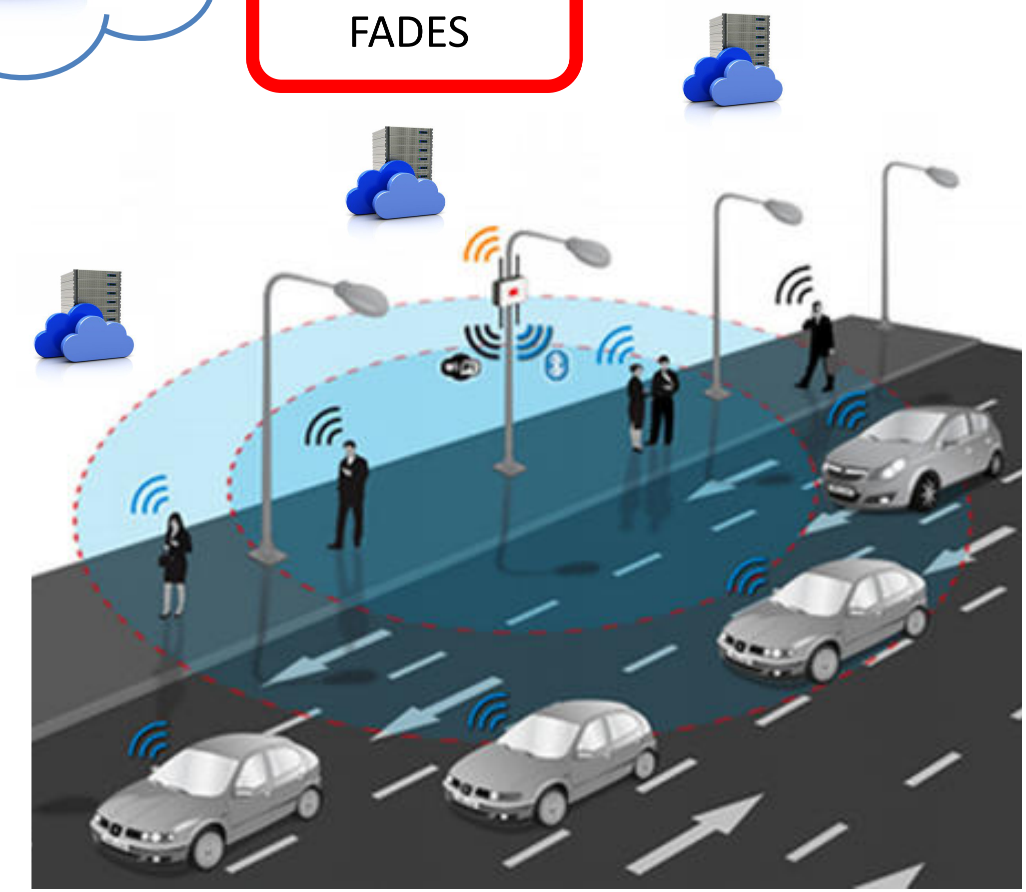
TU Delft

Integrated View

Analytics:
FLAMeS

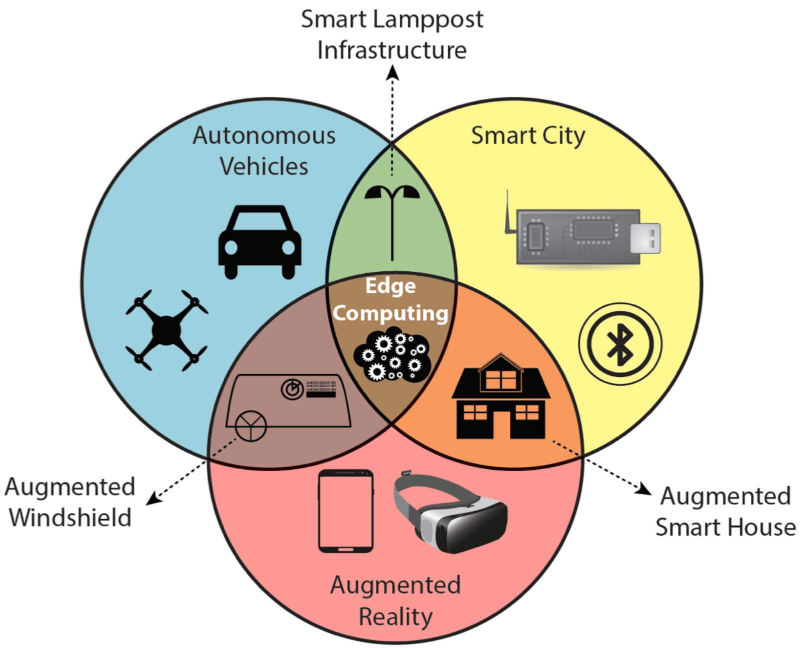
Offloading:
FADES





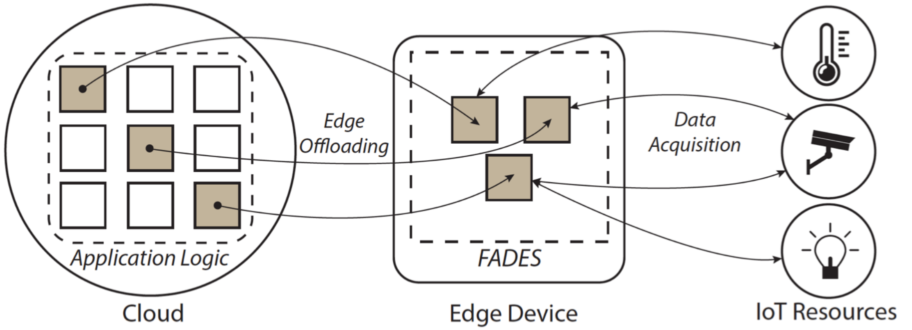
Takeaway

- Dataset
  - Useful but avoid boasting
  - Good work still needs polishing
- Being the first does pay off
  - Analytic and experiment insights



Problems are out there

Research Opportunities !



TU Delft

What to Expect Next