Lessons on Building Edge AI Solutions towards 6G

Aaron Ding

TU Delft, Netherlands
Outline

• Introduction
• Edge Analytics
• Edge Offloading
• Takeaway
Outline

• Introduction
• Edge Analytics
• Edge Offloading
• Takeaway
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
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)
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

<table>
<thead>
<tr>
<th></th>
<th># Records</th>
<th>Traffic Vol. (TB)</th>
<th># MAC</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>DHCP</td>
<td>CORE</td>
<td>TCP</td>
</tr>
<tr>
<td>Flutes</td>
<td>412.0 M</td>
<td>2.13 B</td>
<td>56.18</td>
</tr>
<tr>
<td>Cellos</td>
<td>101.0 M</td>
<td>4.20 B</td>
<td>73.85</td>
</tr>
<tr>
<td>Total</td>
<td>557.5 M</td>
<td>6.53 B</td>
<td>134.39</td>
</tr>
</tbody>
</table>
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
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 ...* ”

  ...
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
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
FLAMeS

• Feature extraction
  – WLAN logs and NetFlows

Goal 1) Get device type, time and space features
FLAMeS

- Data traffic and mobility interdependency

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

• Towards integrated modeling

Goal III) Should new models be devised? What is the value and utility of an integrated model?
Adjust the Focus

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

Made It!

IEEE INFOCOM 2018
ACM MSWiM 2019
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*. 
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

- Edge Offloading
- Data Acquisition
- IoT Resources

TU Delft
How to Offload to Edge?

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

Lightweight Virtualization
Design and Implementation
Use Cases

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

- **Hardware-oriented**
  - Actuator access
Fine-grained Edge Offloading

Does This Really Work?
Experiments

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

**Test over three types of devices**

<table>
<thead>
<tr>
<th>Device</th>
<th>CPU</th>
<th>RAM</th>
<th>Network</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cubietruck</td>
<td>Allwinner A20 ARM Cortex-A7 dual-core @ 1GHz</td>
<td>1 GB</td>
<td>100Mb Ethernet</td>
</tr>
<tr>
<td>Intel NUC</td>
<td>Intel(R) Core(TM) i5-6260U <a href="mailto:CPU@1.80GHz">CPU@1.80GHz</a></td>
<td>16 GB</td>
<td>1000 Mb Ethernet</td>
</tr>
<tr>
<td>Dell Server</td>
<td>Intel(R) Xeon(R) CPU E5-2640 <a href="mailto:v3@2.60GHz">v3@2.60GHz</a></td>
<td>140 GB</td>
<td>1000 Mb Ethernet</td>
</tr>
</tbody>
</table>
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.
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
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.


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
Outline

• Introduction
• Edge Analytics
• Edge Offloading
• Takeaway
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!
What to Expect Next

EdgeSys 2020
The 3rd International Workshop on Edge Systems, Analytics and Networking
27th April 2020, Heraklion, Crete, Greece

Chairs:
Aaron Ding (TU Delft)
Richard Mortier (Cambridge)