Network Resource Management for Edge Analytic Services

George Iosifidis

Assistant Professor, TU Delft
Visiting Academic, Amazon

SystemX, Paris, 20 January 2022

A New Era in Wireless Networks

- Mobile network operators face new challenges:
  - Continuously-growing mobile data traffic;
  - New services: mobile augmented reality, video analytics, etc;
  - New clients: autonomous vehicles, drones, robots, Industry 4.0 systems, etc.

- This requires a fundamental shift in the way we design and manage networks.

- Previous network evolution steps:
  1. Increase point-to-point data transfer capacity;
  2. Improve content delivery capacity;
  3. Incorporate NFV solutions and multi-access edge computing;
  4. Support in-network and edge analytics.

A New Era in Wireless Networks

- Mobile network operators face new challenges:
  - Continuously-growing mobile data traffic;
  - New services: mobile augmented reality, video analytics, etc;
  - New clients: autonomous vehicles, drones, robots, Industry 4.0 systems, etc.

- This requires a fundamental shift in the way we design and manage networks.

- Previous network evolution steps:
  1. Increase point-to-point data transfer capacity;
  2. Improve content delivery capacity;
  3. Incorporate NFV solutions and multi-access edge computing;
  4. Support in-network and edge analytics.

A New Era in Wireless Networks

- Mobile network operators face new challenges:
  - Continuously-growing mobile data traffic;
  - New services: mobile augmented reality, video analytics, etc;
  - New clients: autonomous vehicles, drones, robots, Industry 4.0 systems, etc.

- This requires a fundamental shift in the way we design and manage networks.

- Previous network evolution steps:
  1. Increase point-to-point data transfer capacity;
  2. Improve content delivery capacity;
  3. Incorporate NFV solutions and multi-access edge computing;
  4. Support in-network and edge analytics.

A New Era in Wireless Networks

- Mobile network operators face new challenges:
  - Continuously-growing mobile data traffic;
  - New services: mobile augmented reality, video analytics, etc;
  - New clients: autonomous vehicles, drones, robots, Industry 4.0 systems, etc.

- This requires a fundamental shift in the way we design and manage networks.

- Previous network evolution steps:
  1. Increase point-to-point data transfer capacity;
  2. Improve content delivery capacity;
  3. Incorporate NFV solutions and multi-access edge computing;
  4. Support in-network and edge analytics.

A New Era in Wireless Networks

- Mobile network operators face new challenges:
  - Continuously-growing mobile data traffic;
  - New services: mobile augmented reality, video analytics, etc;
  - New clients: autonomous vehicles, drones, robots, Industry 4.0 systems, etc.

- This requires a fundamental shift in the way we design and manage networks.

- Previous network evolution steps:
  1. Increase point-to-point data transfer capacity;
  2. Improve content delivery capacity;
  3. Incorporate NFV solutions and multi-access edge computing;
  4. Support in-network and edge analytics.

A New Era in Wireless Networks

- Mobile network operators face new challenges:
  - Continuously-growing mobile data traffic;
  - New services: mobile augmented reality, video analytics, etc;
  - New clients: autonomous vehicles, drones, robots, Industry 4.0 systems, etc.

- This requires a fundamental shift in the way we design and manage networks.

- Previous network evolution steps:
  1. Increase point-to-point data transfer capacity;
  2. Improve content delivery capacity;
  3. Incorporate NFV solutions and multi-access edge computing;
  4. Support in-network and edge analytics.

Edge Analytics

- Collect the data:
  - From where? How much? How often?
- Transfer the data:
  - To which destinations? Over which routes? How fast?
- Process the data:
  - Where? How much computing? Which AI/ML libraries?

New Decisions: Sampling data sources; compute/memory allocation; ML parameter selection

New Metrics: Accuracy of inferences; number of successful AI tasks; utility of information, etc.

New Trade-offs: Accuracy vs. lifetime vs. volume of tasks vs. resources' consumption

- Energy, in particular, is the common currency all these operations spend!
- **Opportunity:** Softwarization of networks, convergence of comp. & comms.

Edge Analytics

- Collect the data:
  - From where? How much? How often?
- Transfer the data:
  - To which destinations? Over which routes? How fast?
- Process the data:
  - Where? How much computing? Which AI/ML libraries?

New Decisions: Sampling data sources; compute/memory allocation; ML parameter selection

New Metrics: Accuracy of inferences; number of successful AI tasks; utility of information, etc.

New Trade-offs: Accuracy vs. lifetime vs. volume of tasks vs. resources' consumption

- Energy, in particular, is the common currency all these operations spend!
- **Opportunity:** Softwarization of networks, convergence of comp. & comms.

Edge Analytics

- Collect the data:
  - From where? How much? How often?
- Transfer the data:
  - To which destinations? Over which routes? How fast?
- Process the data:
  - Where? How much computing? Which AI/ML libraries?

New Decisions: Sampling data sources; compute/memory allocation; ML parameter selection

New Metrics: Accuracy of inferences; number of successful AI tasks; utility of information, etc.

New Trade-offs: Accuracy vs. lifetime vs. volume of tasks vs. resources' consumption

- Energy, in particular, is the common currency all these operations spend!
- **Opportunity:** Softwarization of networks, convergence of comp. & comms.

Edge Analytics

- Collect the data:
  - From where? How much? How often?
- Transfer the data:
  - To which destinations? Over which routes? How fast?
- Process the data:
  - Where? How much computing? Which AI/ML libraries?

New Decisions: Sampling data sources; compute/memory allocation; ML parameter selection

New Metrics: Accuracy of inferences; number of successful AI tasks; utility of information, etc.

New Trade-offs: Accuracy vs. lifetime vs. volume of tasks vs. resources' consumption

- Energy, in particular, is the common currency all these operations spend!
- **Opportunity:** Softwarization of networks, convergence of comp. & comms.

Edge Analytics

- Collect the data:
  - From where? How much? How often?
- Transfer the data:
  - To which destinations? Over which routes? How fast?
- Process the data:
  - Where? How much computing? Which AI/ML libraries?

New Decisions: Sampling data sources; compute/memory allocation; ML parameter selection

New Metrics: Accuracy of inferences; number of successful AI tasks; utility of information, etc.

New Trade-offs: Accuracy vs. lifetime vs. volume of tasks vs. resources' consumption

- Energy, in particular, is the common currency all these operations spend!
- **Opportunity:** Softwarization of networks, convergence of comp. & comms.

Edge Analytics

- Collect the data:
  - From where? How much? How often?
- Transfer the data:
  - To which destinations? Over which routes? How fast?
- Process the data:
  - Where? How much computing? Which AI/ML libraries?

New Decisions: Sampling data sources; compute/memory allocation; ML parameter selection

New Metrics: Accuracy of inferences; number of successful AI tasks; utility of information, etc.

New Trade-offs: Accuracy vs. lifetime vs. volume of tasks vs. resources' consumption

- Energy, in particular, is the common currency all these operations spend!
- **Opportunity:** Softwarization of networks, convergence of comp. & comms.

Optimizing vBS

- The flexibility and agility of softwarization comes at a cost.
  - A plenitude of configuration options that is difficult to discern and optimize;
  - which may lead to unpredictable resource consumption, e.g., in terms of energy.

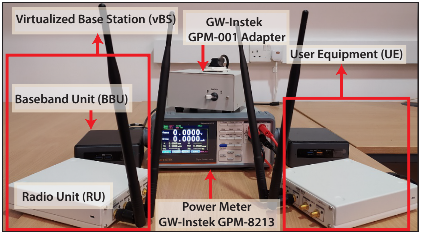
- Two important problems:
  1. How to select transmission power, MCS and airtime for each vBS in order to maximize the served traffic (throughput) and power consumption.
  2. How to select transmission power, MCS and airtime for each vBS in order to maximize the throughput subject to a hard power consumption threshold.

- We need to answer two key questions:
  1. What is the performance and energy consumption profile of vBSs?
  2. How can we optimize their operation using an adaptive and platform-oblivious approach?

J. A. Ayala-Romero, A. Garcia-Saavedra, X. Costa-Pérez, G. Iosifidis: Orchestrating Energy-Efficient vRANs: Bayesian Learning and Experimental Results, IEEE Trans. on Mobile Computing, 2022.

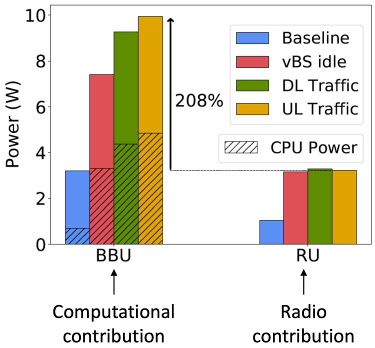
Experimental Evaluation

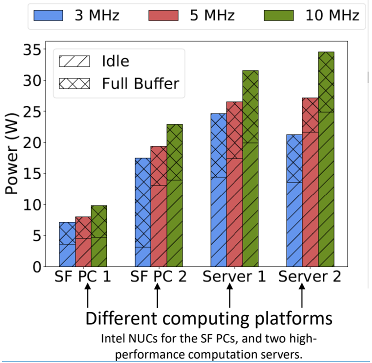
- A testbed with a vBS, user equipement (UE), and a digital power meter.
  - 2 Ettus Research USRP B210 (radio part) and 2 Intel NUCs with CPU i7-8559U (BBU).
  - srsLTE suite to implement the BBU for both the eNB and UE
  - Select the 10 MHz bandwidth.
  - Digital power meter GW-Instek GPM-8213 along with the adapter GPM-001.
  - Integrated O-RAN E2 interface and the ability to change vBS configurations on-the-fly.
  - Generate the traffic load for both DL and UL using mgen.



Experiments (1)

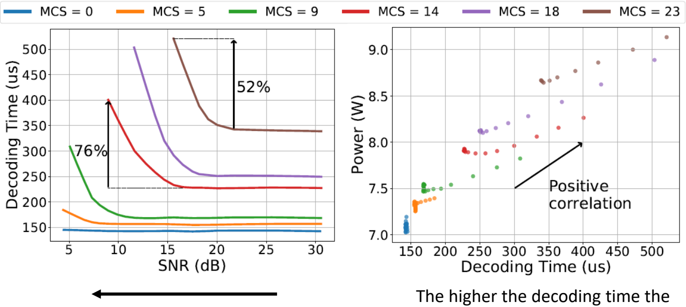
- BBU/CPU cost & Impact of Platform





Experiments (2)

- Impact of SNR and MCS

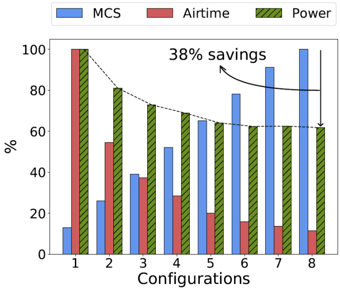


Decrease of channel quality

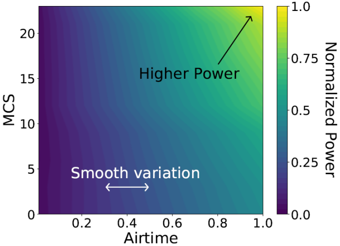
The higher the decoding time the higher the consumed power

Experiments (3)

- Configuration Options and Impact of Scheduler



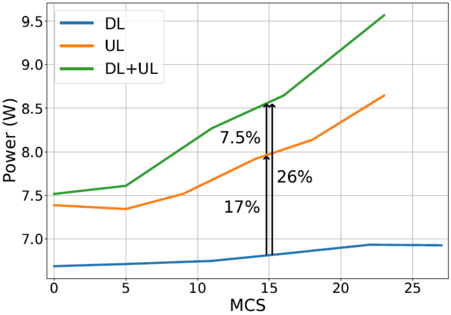
Eight different configurations with the same Throughput in the UL



Joint effect of MCS and airtime on the consumed power

Experiments (4)

- Coupling of UL and DL



Conclusions from Experiments

- Characterizing the vBS power cost is intricate as it depends on traffic, SNR, MCS and airtime.
- There are many DL and UL configurations and some of them present non-linear and non-monotonic relations with power and throughput.
- The power consumption depends on the BBU platform and radio scheduler.
- This hinders the derivation of general consumption models.
- We propose the use of online learning to devise goal-driven configuration policies.

Problem Formulation

- Basic parameters and variables of the model.
  - Context for downlink: $\omega_t^{dl} := [\bar{c}_t^{dl}, \tilde{c}_t^{dl}, d_t^{dl}]$
    - $\bar{c}_t^{dl}$ and $\tilde{c}_t^{dl}$ are the mean and variance of the DL CQI across all users in previous period.
  - Context for uplink: $\omega_t^{ul} := [\bar{c}_t^{ul}, \tilde{c}_t^{ul}, d_t^{ul}]$
  - Actions for downlink: $x_t^{dl} := [p_t^{dl}, m_t^{dl}, a_t^{dl}]$
    - p_t^{dl} is a *transmission power control (TPC) policy* for the max allowed vBS Tx power;
    - m_t^{dl} is the highest MCS eligible (*DL MCS policy*);
    - $a_t^{dl} \in \mathcal{A}^{dl}$ is the maximum vBS transmission airtime (*DL airtime policy*).
  - Actions for uplinkk: $x_t^{ul} := [m_t^{ul}, a_t^{ul}]$
  - Reward:

$$r(\omega_t, x_t) := \log\left(1 + \frac{R^{dl}(\omega_t^{dl}, x_t^{dl})}{d_t^{dl}}\right) + \log\left(1 + \frac{R^{ul}(\omega_t^{ul}, x_t^{ul})}{d_t^{ul}}\right)$$

    where R^{dl}, R^{ul} is the achieved throughput in DL and UL, resp.

Case 1: Balancing Performance & Cost

- The power supply is scarce or the operator needs to reduce OpEx.
- Pareto optimization via scalarization:

$$u(\omega_t, x_t) := r(\omega_t, x_t) - \delta \cdot B\big(P(\omega_t, x_t)\big),$$

- Goal: minimize (contextual) regret:

$$R_T := \sum_{t=1}^{T} \left(\max_{x' \in \mathcal{X}} u(\omega_t, x') - u(\omega_t, x_t) \right),$$

- By finding a sequence of configurations $\{x_t\}_{t=1}^{T}$ such that:

$$\lim_{T \to \infty} R_T / T = 0$$

- Key observation: outcomes of different configurations are correlated.

Case 2: Hard Power Budget

- The vBS operates under a power budget P_{max}, e.g., when PoE operation.
- Find for maximum throughput configuration meeting the budget. Using new regret:

$$R_T^s := \sum_{t=1}^{T} \left(\max_{x' \in \{S_t(\omega_t)\}_t} r(\omega_t, x') - r(\omega_t, x_t) \right)$$

where in this case the decisions are selected from set of *safe configurations*:

$$S_t(\omega_t) = \left\{ x \in \mathcal{X} \;\middle|\; P(\omega_t, x) \leq P_{\max} \right\}.$$

- By finding a sequence of configurations $\{x_t\}_{t=1}^{T}$ such that:

$$\lim_{T \to \infty} R_T^s / T = 0$$

Solution: Bayesian Online Learning

- Use Gaussian Processes (GPs)
  - Context - action pair: $z \in \mathcal{C} = \Omega \times \mathcal{X}$
  - Obtain noisy performance - cost observations $\{u_t\}$ for each $\{z_t\}$.
  - The posterior distribution of the objective function follows a GP with

$$\text{mean} \quad \mu_T(z) = k_T(z)^\top (K_T + \zeta^2 \mathbf{1}_T)^{-1} y_T$$

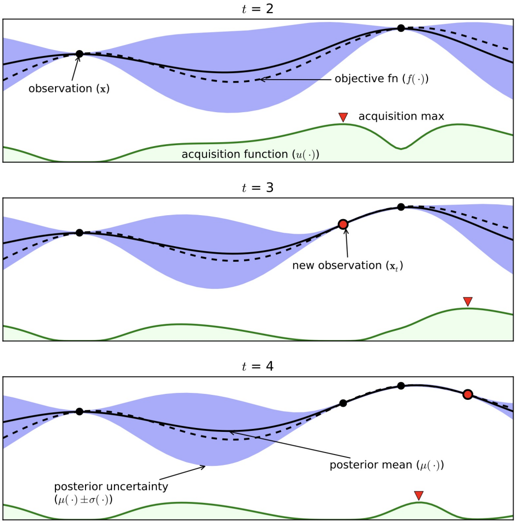
$$\text{variance} \quad k_T(z, z') = k(z, z') - k_T(z)^\top (K_T + \zeta^2 \mathbf{1}_T)^{-1} k_T(z')$$

where $k_T(z) = [k(z_1, z), \ldots, k(z_T, z)]^\top$, $K_T(z)$ is the kernel matrix $[k(z, z')]_{z,z' \in Z_T}$, and $\mathbf{1}_T$ is the T-dimension identity matrix.

- With GPs we can estimate the distribution of unobserved values $z \in \mathcal{Z}$;
  - Thus, to gradually learn the function that we wish to optimize.
- How do we leverage this information? Using the **acquisition function**:

$$x_t = \arg\max_{x \in \mathcal{X}} \mu_{t-1}(\omega_t, x) + \sqrt{\beta} k_{t-1}(\omega_t, x)$$

Solution: Bayesian Online Learning



BP-vRAN Algorithm

Algorithm 1 BP-vRAN: Performance and cost balancing

Inputs: Control Space $\mathcal{X}$, kernel k, β
Initialize: $y_0 = \emptyset$, $Z_0 = \emptyset$
for $t = 1, \ldots, T$ **do**
    Observe the context ω_t
    $x_t = \text{argmax}_{x \in \mathcal{X}} \ \mu_{t-1}(\omega_t, x) + \sqrt{\beta_t}\sigma_{t-1}(\omega_t, x)$
    Measure $R_t^{dl}(\omega_t^{dl}, x_t^{dl})$, $R_t^{ul}(\omega_t^{ul}, x_t^{ul})$ and $P_t(\omega_t, x_t)$ at the end of the decision period t
    Compute $u_t(\omega_t, x_t)$ using (1), (2) and (3)
    Update $Z_t \leftarrow Z_{t-1} \cup [\omega_t, x_t]$
    Update $y_t \leftarrow y_{t-1} \cup u_t(\omega_t, x_t)$
    Perform Bayesian update to obtain μ_t and σ_t
end for

- The algorithm ensures a probabilistic bound for regret:

$$\mathrm{P}\left(R_T \leq \sqrt{C_1 T \beta_T \gamma_T}\right) \geq 1 - \epsilon,$$

Safe BP-vRAN Algorithm

Algorithm 2 SBP-vRAN: Safe online optimization

Inputs: Control Space $\mathcal{X}$, Initial safe set S_0, kernel k, β, $P_{\max}$
Initialize: $y_0^f = \emptyset$, $y_0^c = \emptyset$, $Z_0 = \emptyset$
for $t = 1, \ldots, T$ **do**
  Observe the context ω_t
  $S_t = S_0 \cup \{x \in \mathcal{X} \mid \mu_{t-1}^c(\omega_t, x) + \beta_t \sigma_{t-1}^c(\omega_t, x) \le P_{\max}\}$
  $x_t = \text{argmax}_{x \in S_t} \ \mu_{t-1}(\omega_t, x) + \sqrt{\beta_t}\sigma_{t-1}(\omega_t, x)$
  Measure $R_t^{dl}(\omega_t^{dl}, x_t^{dl})$, $R_t^{ul}(\omega_t^{ul}, x_t^{ul})$ and $P_t(\omega_t, x_t)$ at the end of the decision period t
  Compute $r_t(\omega_t, x_t)$ using (1)
  Update $Z_t \leftarrow Z_{t-1} \cup [\omega_t, x_t]$
  Update $y_t^f \leftarrow y_{t-1}^f \cup r_t(\omega_t, x_t)$
  Update $y_t^c \leftarrow y_{t-1}^c \cup P_t(\omega_t, x_t)$
  Perform Bayesian update to obtain μ_t^f, σ_t^f, μ_t^c and σ_t^c
end for

- We need an additional GP for assessing the safety of the constraint.

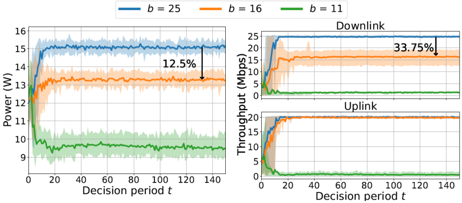
$$S_t = \left\{ x \in \mathcal{X} \;\middle|\; \mu_{t-1}^c(\omega_t, x) + \beta_t \sigma_{t-1}^c(\omega_t, x) \le P_{\max} \right\}.$$

- The configuration is selected using the CGP-UCB policy subject to the safe set:

$$x_t = \operatorname*{argmax}_{x \in S_t} \ \mu_{t-1}^f(\omega_t, x) + \sqrt{\beta_t}\sigma_{t-1}^f(\omega_t, x),$$

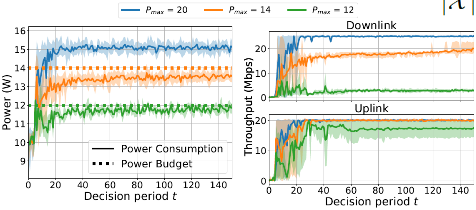
Experimental Evaluation (Convergence)

BP-vRAN

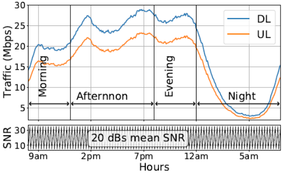


$$|\mathcal{X}| \approx 1.6 \cdot 10^6$$

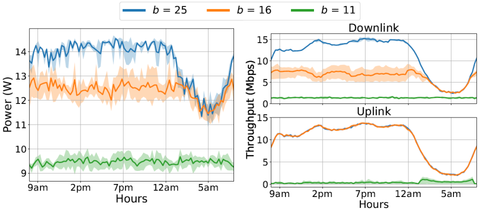
SBP-vRAN



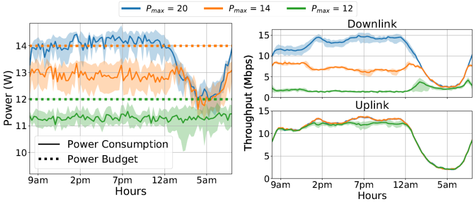
Experimental Evaluation (Convergence)



BP-vRAN



SBP-vRAN

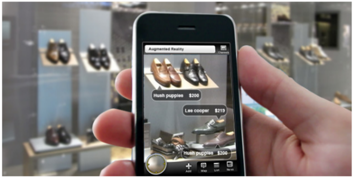


Conclusions

- Presented an in-depth experimental study of the energy behavior of vBSs.
- Found a complex relationship between performance, power cost and vBS config.
- This complexity can only be tamed with data-driven machine-learning solutions.
- We have proposed an online learning framework to achieve two goals:
  - Balance performance and power cost;
  - Maximize performance subject to power constraints vBS, e.g., PoE.
- Theoretical guarantees; high data-efficiency and convergence speed.
- Real-data evaluation verified convergence and efficacy in practice.
- Code and datasets online: `https://jaayala.github.io/`

Edge Analytics

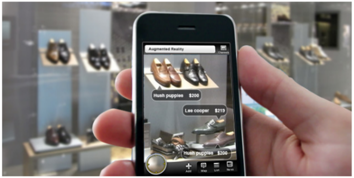
- How to orchestrate the network in order to support real-time video analytics?
  - E.g., capture and process video frames in real time using MEC.



- Challenges:
  - Multiple criteria: fast and accurate inferences; or energy-aware inferences;
  - Multiple decisions: video frame quality; network control; AI pipeline configuration;
  - Need to **jointly** optimize all these decisions;
  - Performance depends on equipment and on the actual processed data.

Edge Analytics

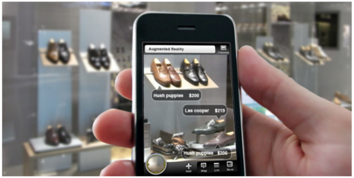
- How to orchestrate the network in order to support real-time video analytics?
  - E.g., capture and process video frames in real time using MEC.



- Challenges:
  - Multiple criteria: fast and accurate inferences; or energy-aware inferences;
  - Multiple decisions: video frame quality; network control; AI pipeline configuration;
  - Need to **jointly** optimize all these decisions;
  - Performance depends on equipment and on the actual processed data.

Edge Analytics

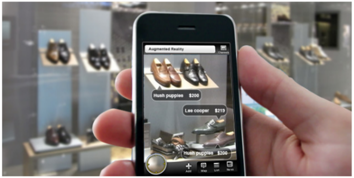
- How to orchestrate the network in order to support real-time video analytics?
  - E.g., capture and process video frames in real time using MEC.



- Challenges:
  - Multiple criteria: fast and accurate inferences; or energy-aware inferences;
  - Multiple decisions: video frame quality; network control; AI pipeline configuration;
  - Need to **jointly** optimize all these decisions;
  - Performance depends on equipment and on the actual processed data.

Edge Analytics

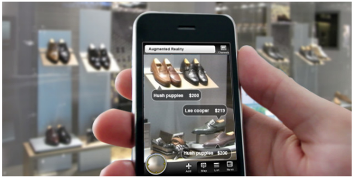
- How to orchestrate the network in order to support real-time video analytics?
  - E.g., capture and process video frames in real time using MEC.



- Challenges:
  - Multiple criteria: fast and accurate inferences; or energy-aware inferences;
  - Multiple decisions: video frame quality; network control; AI pipeline configuration;
  - Need to **jointly** optimize all these decisions;
  - Performance depends on equipment and on the actual processed data.

Edge Analytics

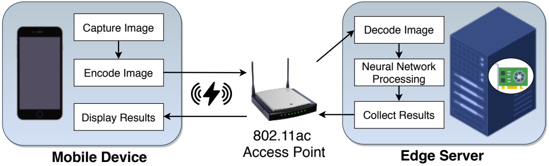
- How to orchestrate the network in order to support real-time video analytics?
  - E.g., capture and process video frames in real time using MEC.



- Challenges:
  - Multiple criteria: fast and accurate inferences; or energy-aware inferences;
  - Multiple decisions: video frame quality; network control; AI pipeline configuration;
  - Need to **jointly** optimize all these decisions;
  - Performance depends on equipment and on the actual processed data.

Edge Analytics over WiFi

- We study a basic scenario – key component of different applications.

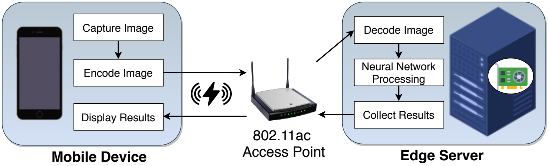


- We have built an exemplary system:
  - An Android app captures and sends images to server for object recognition using YOLO;
  - Bounding boxes of recognized objects returned to the mobile; process repeats;
  - Image encoding rate (at the device) and YOLO NN input-layer size (at the server) affect both the accuracy and latency.

A. Galanopoulos, J. Ayala, D. Leith, G. Iosifidis, Auto-ML for Video Analytics with Edge Computing, IEEE INFOCOM 2021.

Edge Analytics over WiFi

- We study a basic scenario – key component of different applications.

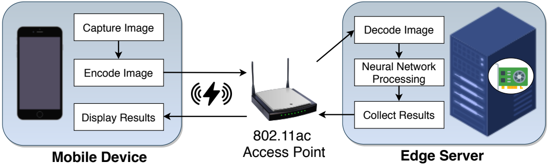


- We have built an exemplary system:
  - An Android app captures and sends images to server for object recognition using YOLO;
  - Bounding boxes of recognized objects returned to the mobile; process repeats;
  - Image encoding rate (at the device) and YOLO NN input-layer size (at the server) affect both the accuracy and latency.

A. Galanopoulos, J. Ayala, D. Leith, G. Iosifidis, Auto-ML for Video Analytics with Edge Computing, IEEE INFOCOM 2021.

Edge Analytics over WiFi

- We study a basic scenario – key component of different applications.

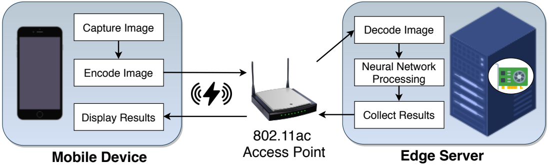


- We have built an exemplary system:
  - An Android app captures and sends images to server for object recognition using YOLO;
  - Bounding boxes of recognized objects returned to the mobile; process repeats;
  - Image encoding rate (at the device) and YOLO NN input-layer size (at the server) affect both the accuracy and latency.

A. Galanopoulos, J. Ayala, D. Leith, G. Iosifidis, Auto-ML for Video Analytics with Edge Computing, IEEE INFOCOM 2021.

Edge Analytics over WiFi

- We study a basic scenario – key component of different applications.

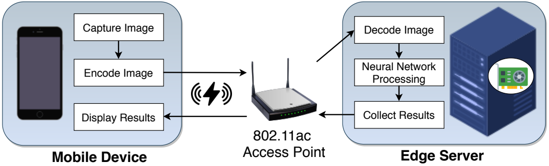


- We have built an exemplary system:
  - An Android app captures and sends images to server for object recognition using YOLO;
  - Bounding boxes of recognized objects returned to the mobile; process repeats;
  - Image encoding rate (at the device) and YOLO NN input-layer size (at the server) affect both the accuracy and latency.

A. Galanopoulos, J. Ayala, D. Leith, G. Iosifidis, Auto-ML for Video Analytics with Edge Computing, IEEE INFOCOM 2021.

Edge Analytics over WiFi

- We study a basic scenario – key component of different applications.

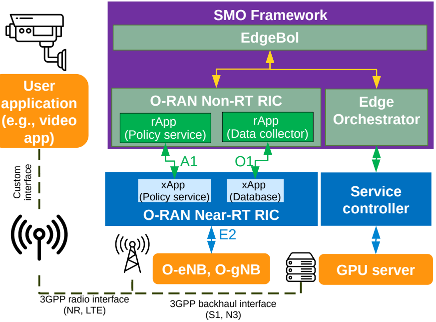


- We have built an exemplary system:
  - An Android app captures and sends images to server for object recognition using YOLO;
  - Bounding boxes of recognized objects returned to the mobile; process repeats;
  - Image encoding rate (at the device) and YOLO NN input-layer size (at the server) affect both the accuracy and latency.

A. Galanopoulos, J. Ayala, D. Leith, G. Iosifidis, Auto-ML for Video Analytics with Edge Computing, IEEE INFOCOM 2021.

Edge Analytics over vBS

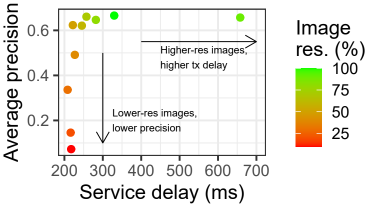
- We also study the case where the network is cellular.
  - How to jointly configure the service (frame features); vBS (MCS, Power); and edge server (GPU power)?



J. Ayala, A. Saavedra, X. Costa-Perez, G Iosifidis, EdgeBOL: Automating Energy-savings for Mobile Edge AI, ACM CoNEXT, 2021.

Edge Analytics over vBS

- Trade offs between service delay and service accuracy.



- Outline of setup:
  - Service delay: image proc. at UE, transmission; GPU processing; return of labels.
  - Mean Average Precision (mAP): typical metric used in object recognition problems.
  - Server power consumption (mainly GPU).
  - BS power consumption *(BBU processing).
  - Control policies:
    - Average image encoding of every image generated; enforced by the service.
    - Radio airtime and MCS.
    - GPU power limit that adapts the GPU speed.

Thank you!