Internship position: energy/performance tradeoffs in ML-enhanced high-speed networks

Description

An internship position is being opened in a joint collaboration between Telecom Paris and Telecom SudParis, both highly ranked engineering schools and founding members of Institut Polytechnique de Paris (IP-Paris) - France. The internship has been funded by the E4C (Energy for climate) interdisciplinary center of the IP-Paris, in the context of the project **ERTON-ML** that started in 2022.

Context

We consider a high-speed network application which is implemented as a piece of code executed on COTS servers [1]. In the context of softwarized networks, we observe that user packets may traverse a main data path (i.e., the usual processing pipeline consisting of several VNFs running on the same COTS server) or an external path which may involve off-chip devices such as GPUs or TPUs. Off-chip devices can be accessed in order to offload some of the functionalities, although this is usually discouraged in high-speed contexts due to the cost of the cross-server interactions. This is especially relevant as machine learning (ML) techniques are starting to be applied within network applications for a plethora of scenarios, ranging from anomaly detection [2], to performance prediction [3].

Specifically, a tradeoff emerges in what regards the placement of the ML processes. On the one hand, if the ML is placed alongside the data path, this may provide new collectible data with a resolution that cannot be reached using legacy equipment with pre built-in monitoring functionalities. On the other hand, this massive data availability comes at a cost: software measurements require CPU cycles that are subtracted to the network functions that compete for the same underlying compute infrastructure: this can highly alter the energy requirements for the packet processing. Furthermore, the deployment of multiple ML and data processing components, may require additional servers to be turned on to host the ML computation, which corresponds to an overall increment of the energy footprint of network applications.

Objectives

In this project, we quantitatively study the energy/performance tradeoff of ML-enhanced high-speed networks, where ML applications are deployed alongside the main data path in standard COTS servers. Computation consolidation techniques have been proposed in order to reduce the resource usage and thus energy footprint of computation. However, when both the ML and the network application compete for the same resources, the execution flow of the application may be severely affected by ML, which may require strong isolation, impeding consolidation. Our aim is to achieve such isolation via offloading ML computation on low-power embedded devices. The objective is to verify that, by doing so, we can decrease the overall energy consumption of computation, while keeping performance.

Tasks and profile of the candidate

In a nutshell, the candidate will quantify the energy/performance tradeoff by means of experiments on a real testbed. The candidate will also implement an optimization process that analyzes placement decisions for newly created VNFs, and decides the offloading of some ML functionalities towards off-chip devices. The optimization process takes into account the application requirements and the ML performance indicators (e.g., accuracy, memory consumption, delay). To reduce the information sent to off-chip ML components, careful design of on-chip lightweight data preprocessing will be considered.

The **profile** of the ideal candidate is as follows:

- Excellent programming skills (preferred languages: C and Python).
- Basic knowledge of frameworks for data analysis and machine learning (e.g., Pytorch).
- Basic understanding of network systems (TCP/IP protocol stack, Software-defined Networking).
- Familiarity with Linux systems.

The project provides a scholarship of **500€ / month** that can be cumulated with other funding sources (e.g., Erasmus mobility or others). The duration of the internship is **8 weeks** (2 months).

How to apply

The candidates must send an email to: `linguaglossa [at] telecom-paris [dot] fr`

which must include an up-to-date CV and the report of the grades during their last year.

References

[1] **Linguaglossa L.**, Lange S., Pontarelli S., Retvari G., Rossi D., Zinner T., Bifulco R., Jarschel M., Bianchi G., "Survey of Performance Acceleration Techniques for Network Function Virtualization," in *Proceedings of IEEE*, 2019.

[2] Putina A., Barth S. et al., "Telemetry-based stream-learning of BGP anomalies," *Big-DaMa*, 2020.

[3] F. Geyer, "Deepcomnet: Performance evaluation of network topologies using graph-based deep learning," *Performance Evaluation*, vol. 130, pp. 1–16, 2019.