Ph.D. Thesis Position 2021

Title: Distributed edge cloud architecture for executing AI based applications.

Host laboratory: LISTIC, Polytech Annecy-Chambéry, 5 Chemin de Bellevue, 74940 Annecy-le-Vieux

Advisor(s):
Francesco Bronzino, Assistant Professor
Université Savoie Mont Blanc, Annecy-le-Vieux `fbronzino@univ-smb.fr`

External Collaborators:
The thesis will be carried on within an ANR collaboration between Université Savoie Mont Blanc, Avignon Université, Inria, and CNAM.

Description. New generations of mobile access networks promise low delay and high-speed throughput data connections paired with in-network processing capabilities. The technology evolution over previous network generations will enable emerging mobile computing services such as the processing of traffic intensive data streams (*e.g.*, IoT data streams), or the control of industrial systems (*e.g.*, autonomous cars). These services are quickly integrating AI-intensive processing tasks in their workloads (*e.g.*, in smart city services or virtual and augmented reality applications). Following the cloud-native microservice architecture, AI applications are conventionally composed by chains of tasks. Each task of the chain executes an operation, often based on an AI model, to process the incoming information and feed it to the next step in the pipeline.

AI pipelines differ from classical microservice chains of tasks due to their use of AI models: as such, they require to rethink how orchestration of edge resources matches app deployment towards system performance and reliability. By nature, AI chains' performance depends on the amount of information contained in each processed flow, which differs and dynamically changes over time. Ultimately, the whole application chain's performance will depend not only on local resources constraints (*i.e.*,, bandwidth and computing power), but also on the current context of the input information (*i.e.*,, the informative content of each processed flow). These new characteristics require the introduction of new levels of dynamicity in resource allocation schemes and orchestration methods for edge computing in order to meet the performance requirements of AI applications.

The proposed Ph.D. thesis, will develop new architectural mechanisms to cope with the aforementioned applications and their dynamicity by re-defining how resource orchestrators distribute their processing pipelines. Taking advantage of ultra-localized computing nodes, orchestrators can gain the ability to split processing pipelines not only hierarchically but horizontally, too. Edge computing frameworks are rapidly evolving to include processing accelerators—*i.e.,* GPUs and TPUs—at the extreme edge of the network. Such solutions aim to offer an extremely scalable approach to support low latency computation services through a pervasive deployment approach. Unfortunately, these architectures are inherently underpowered and relying on them to support AI applications requires to carefully decide how to distribute the processing load. The selected student, will work on defining new algorithmic solutions for the described problem, evaluating them on state of the art hardware platforms.

During the thesis work, the student will first focus on a real world use case to characterize AI applications and evaluate how they adapt to our solutions: real-time video analytics. Video analytics applications process several video cameras feeds in a tagged geographical area to perform sophisticated functionalities such as motion detection or features extraction. Performing these operations conventionally requires executing multiple tasks in sequence (*e.g.,* decoding, background extraction, object detection, plate extraction, vehicle counting). But the sheer size of data generated by camera flows makes it infeasible to centralize the processing. The proposed decentralized edge processing approach becomes then key for their success. To support our hypothesis, we developed a first prototype for experimenting with distributed video analytics on edge

computing nodes. Our initial results [2, 1], suggest that smartly distributing the functions is an efficient way to utilize the resources at the edge.

Candidate Requirements.

- The candidate should have completed a qualifying program by the starting date of the thesis.
- Comfortable speaking English or French (French is not required).
- Good understanding of at least one between computer networks protocols and systems or machine learning methods (preferably both)
- Good proficiency with at least one programming language, preferably Python, Golang, or Rust.

What to submit. An up to date CV, university transcripts, and a letter of motivation clearly stating what the motivations to work on the described subject.

References

[1] F. Faticanti, F. Bronzino, and F. De Pellegrini. The case for admission control of mobile cameras into the live video analytics pipeline. In *Proceedings of the 3rd ACM Workshop on Hot Topics in Video Analytics and Intelligent Edges*, pages 25–30, 2021.

[2] S. P. Rachuri, F. Bronzino, and S. Jain. Decentralized modular architecture for live video analytics at the edge. In *Proceedings of the 3rd ACM Workshop on Hot Topics in Video Analytics and Intelligent Edges*, pages 13–18, 2021.