# Deadline-driven Resource Allocation in Edge Computing with Wireless Communication

October 11, 2022

By 2030, the recent developments of wireless network technologies should have opened a multitude of business opportunities, in particular in the business-to-business market [3]. An analysis of the predictions presented in [3] shows that (i) enhanced video streaming, (ii) real-time automation, and (iii) connected vehicles, are some of the most promising classes of applications.

Figure 1 provides an abstract representation of a typical communication and computation infrastructure that is used in the applications discussed above. It comprises a set of mobile and low-power sensing and actuation devices (henceforth called *end devices*) accessible through a wireless communication network. In this work, we focus on LoRa [1], a low-power wide-area network (LPWAN) technology which was designed for networking low-power Internet-of-Things (IoT) devices that are distributed over geographically wide areas[1]. The wireless gateway is assumed to have edge computation capabilities and is connected with other gateway devices (henceforth called *edge servers*) through a wired backhaul network. Finally, the edge servers also have access to cloud resources through a wired network for offloading heavy computations; different edge servers may have access to different cloud resources.

In this project, we will consider a generic application workload model using parameters such as workload arrival rates, required computation cycles and memory, data sizes to capture communication delays when workload is offloaded to edge or cloud servers, etc. To model tight real-time requirements we will use deadline constraints which denote the time by which workload must be completed so that the results of computation are valid and useful. Additional offloading constraints such as the requirement to offload to only some of the available edge and cloud servers would also be incorporated in this model. We will model the computation servers using available resource capacity (computation cycles as well as memory), and additionally for the edge servers, we will also model the bandwidth capacity of the wired backhaul network to which they are connected. Finally, we will model the LoRa wireless network using channel capacity and performance parameters such as duty cycles, frequency spectrum, back-off interval, number of retries, etc. Note, these model parameters are fairly standard in the literature (e.g., see [6, 4]).

The general objective of this project is to minimize the power consumption of end devices by offloading their computation intensive workload on edge and cloud servers. Formulated as such, this problem has already received a lot of attention in the edge computing literature. *However, the originality of our work lies in the consideration of timing constraints for workload.* This aspect has also been studied in a few research works, but most of them consider simplistic hypothesis when it comes to modelling timing interference among workload. However, such contentions exist when workload share computation and communication resources, and as a result they may incur significant delays in service delivery. Further, these time-sensitive workload could be sharing resources with other workload that do not have deadlines, but nevertheless shortening the response times of such non time-sensitive workload has value to the application. Given the inherent uncertainties in wireless communication, we strive for probabilistic, as opposed to deterministic, guarantees on the deadline constraints in this project.

This problem is ambitious and will require finding a trade-off between optimality and reactivity: on the one hand, resource allocation problems are usually difficult to solve optimally because of their combinatorial

[1] Due to the specific MAC-layer properties of wireless protocols, it is essential to focus on a specific technology when designing resource allocation algorithms for wireless communication.

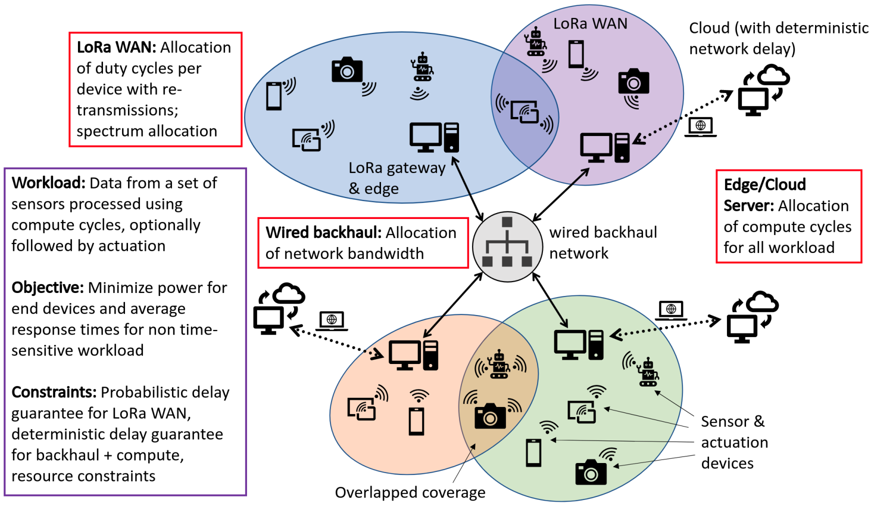
LoRa WAN: Allocation of duty cycles per device with re-transmissions; spectrum allocation
LoRa WAN
Cloud (with deterministic network delay)
LoRa gateway & edge
Workload: Data from a set of sensors processed using compute cycles, optionally followed by actuation
Objective: Minimize power for end devices and average response times for non time-sensitive workload
Constraints: Probabilistic delay guarantee for LoRa WAN, deterministic delay guarantee for backhaul + compute, resource constraints
Wired backhaul: Allocation of network bandwidth
wired backhaul network
Edge/Cloud Server: Allocation of compute cycles for all workload
Sensor & actuation devices
Overlapped coverage

Figure 1: LoRa LPWAN edge-cloud architecture: Associated resource provisioning & scheduling objectives

complexity. On the other hand, service workload characteristics vary over time (notably because of device mobility), which means the allocation framework has to run online to adapt resource allocations to these variations. In the recent past, we have conducted a detailed survey of this research topic [6] and also developed our initial solutions for a specific problem instance [2].

**PhD Objectives:**

1. **Resource Provisioning and Scheduling Problems**: In order to deal with the conflicting objectives of optimality and reactivity, we aim to follow a "divide-to-conquer" approach. In this approach, we assume a decomposition of the infrastructure into segments identifying subsets of the architecture. We then propose to decompose the end-to-end deadlines of time-sensitive workload to derive segment-level deadlines. The definition of the deadline decomposition method, as well as its evaluation, is the focal point of this task. We briefly describe its principles hereafter: for each segment, a *resource provisioning* (mapping workload to resources) and *scheduling* (assigning resources to mapped workload over time) method will be developed in order to export the current state of the segment to a deadline decomposition framework. Using this information, the decomposition framework assigns service deadlines for each segment. A potential methodology for this decomposition framework could be based on the delay composition algebra [5].

2. **Handling Device Mobility**: End devices are often mobile in such applications, which means the routing of data packets between these devices and servers need to consider the current and future location of devices. Additionally, the developed resource provisioning and scheduling algorithms need to dynamically adapt to the changing topology of wireless communication. To address such problems we will consider online algorithms that can react to these changes. Further, decentralized solutions will be sought to facilitate reactivity.

# References

[1] LoRa Alliance. *LoRaWAN™1.1 Specification*. `https://lora-alliance.org/resource_hub/lorawan-specification-v1-1/`. [Online; accessed 10-August-2021]. 2017.

[2] S. Aryaman C. Gao and A. Easwaran. "Deadline-constrained Multi-resource Task Mapping and Allocation for Edge-Cloud Systems". In: *IEEE Global Communications Conference (GLOBECOM)*. 2022. URL: https://arxiv.org/abs/2206.05950.

[3] Ericsson. *5G For Business a 2030 Market Compass*. https://www.ericsson.com/en/5g/5g-for-business/5g-for-business-a-2030-market-compass. [Online; accessed 10-August-2021]. 2017.

[4] A. Gamage et al. "Lmac: Efficient Carrier-Sense Multiple Access for LoRa". In: *International Conference on Mobile Computing and Networking*. 2020, pp. 1–13.

[5] P. Jayachandran and T. Abdelzaher. "Delay Composition Algebra: A Reduction-Based Schedulability Algebra for Distributed Real-Time Systems". In: *IEEE Real-Time Systems Symposium (RTSS)*. 2008, pp. 259–269. DOI: 10.1109/RTSS.2008.38.

[6] S. Ramanathan et al. "A Survey on Time-Sensitive Resource Allocation in the Cloud Continuum". In: *IT - Information Technology* 62.5-6 (2020), pp. 241–255. DOI: doi:10.1515/itit-2020-0013. URL: https://doi.org/10.1515/itit-2020-0013.