Towards an Internet of Reality

James Gross
School of Electrical Engineering and Computer Science
KTH Royal Institute of Technology, Stockholm, Sweden.
E-mail: jamesgr@kth.se.

Abstract—**Driven by standardization and commercialization, digital infrastructures evolve in waves. Over the last few years, a particular focus has been on realizing ultra-reliable low-latency wireless communications (URLLC), anticipated mostly for rather specific use cases in industrial automation. Even though initial such systems finally exist today - with future network releases advancing URLLC capabilities even more - the broad market impact to date is low. We argue in this paper that an essential missing component for corresponding dependable applications like closed-loop control or human-in-the-loop are nearby compute capabilities provided within the infrastructure, aka edge computing capabilities. Only in conjunction can such future infrastructures support dependable applications to a full extent. Nevertheless, this also leads to unique challenges which will be central to the evolution of networked infrastructures during the current decade. Out of this evolution of networked infrastructures, we finally argue that a new type of networked application class will emerge, resembling the representation of various aspects of reality in the infrastructure at any point in time. We dub this development the Internet of Reality, and discuss further challenges in this context.**

I. Introduction

Contemporary mobile infrastructures are designed according to the end-to-end principle, leading to a distinction between information transmission and processing. While networks realize the transportation of information, the processing of the information is performed at the end points. Over the last decade, these end points for processing have become the plethora of hand-held devices (i.e. smart phones, tablets etc) on the one end and cloud computing centers on the other. This development has been substantially accelerated by the introduction of 4G, allowing the proliferation of mobile devices while providing adequate service levels for the major application classes. These application classes are video streaming, social networking and web surfing [1]. With respect to these classes, the networked infrastructures of today are highly optimized and perform sufficiently well.

The last decade has also witnessed the introduction of machine-to-machine communications. In particular, two different flavors targeting the Internet-of-Things (IoT), as well as the Industrial Internet of Things (IIoT) have been realized. The older of these flavors are networks that target low-power, long-range transmissions for sensor data collection, leading to what is commonly summarized as IoT. Today, two systems have commercial relevance in this domain, namely Long-Range Wide Areas Networks (LoRaWAN) [2] and NarrowBand-IoT (NB-IoT) [3]. Together, these systems represent roughly 90% of the deployed capacity. LoRaWAN is a proprietary technology introduced about a decade ago, and represented today in form of an industrial alliance. In contrast, NB-IoT has its origins in 3GPP, with its currently deployed form being introduced with Release 13. From a technical point-of-view, both these technologies achieve extremely long life times of battery-driven devices by leveraging duty cycling, as well as narrowband signaling in different realizations. While being mature transmission technologies, it is noteworthy that major IoT applications leveraging LoRaWAN or NB-IoT like environmental monitoring, smart buildings, goods tracking as well as smart metering only became commercially significant with the introduction of cloud computing as ubiquitous, scalable data collection peer [4].

In contrast to this flavor, the development around the IIoT is younger and addresses a different technical profile[1]. With the introduction of 5G, so called Ultra-Reliable Low Latency Communication (URLLC) systems [5] have been specified, which serve communication links with extremely high reliability and latency requirements, that are foremost anticipated in industrial automation settings. Target performance indicators have been point-to-point latencies below 1 ms, while guaranteeing simultaneously reliabilities of $1 - 10^{-6}$ packet error rates and below. As of today, early realizations of

[1] In this article, by IIoT we will refer in the following only to latency-critical applications, ie. not to the application cases in industry that are alike to usual IoT cases like low-power sensing and monitoring

such systems are appearing on the market with Release 15 products, while full URLLC capabilities are expected with later releases. Despite this roadmap, the commercial relevance of URLLC is still unclear, with the next years being a proof point for the commercialization.

It is apparent that over the last two decades mobile networks differentiated the air interface, and subsequent communication service models, to account for vastly different application requirements. However, while for traditional, human-based service models the contemporary network architecture is mature, this is less so for machine-type service models. For IoT services, the end-to-end principled approach, with sensor devices on the one end and cloud computing centers on the other, appears more and more as the emerging, dominant architecture. However, for URLLC - broadly speaking - this question is still open. In this paper, we try to address this issue from the perspective of different emerging application classes (Section II). We argue in particular that URLLC systems are a first step into a much more complex, yet powerful, infrastructure evolution with potentially quite surprising consequences and substantial challenges ahead (Section III). Central to this evolution is the notion of infrastructure *responsiveness*, characterizing the ability of an integrated communication and compute path to generate actuation or perceptual feedback from a real-time digital representation (of parts of reality). While evolving towards such infrastructures is a challenging road ahead, a further, more distant challenges arises out of the maintenance of the associated digital representations. With many such representations being maintained at any point in time, a sizable subset of reality might be digitally represented in future infrastructures. This leads to the (speculative) notion of the ”Internet of Reality”.

II. The Emergence of Feedback Systems

Today, the major use case for URLLC systems lies with industrial automation, and therein mostly with factory automation [6]. From the original definition of URLLC, it is clear that such wireless systems have been anticipated as straightforward cable substitutes all over the traditional automation pyramid. However, with the emergence of edge computing, these substitution-driven use cases might become significantly enhanced, pointing towards a more involved direction with respect to the evolution in this domain.

A. *Initial and Evolved Industrial Automation Use Cases*

At its core, many industrial automation uses cases reduce to control-theoretic feedback systems. Such systems are characterized by the continuous (ie. periodic) collection of sensor readings that are provided to a controller. According to the set-point of the control loop and the current reading, the controller based on a control algorithm generates a new actuation command the controlled plant needs to be exposed to. This is then forwarded to the actuator and instantiated. Subsequent sensor readings allow for determining the resulting impact, and adjusting the actuation. With every new sensor reading coming in, the system state represented at the controller is updated, and subsequently a new actuation command is generated. In industrial automation, this principle is deployed in a nested fashion, essentially leading to the automation pyramid. On the lowest level, the so called primary technology likes drives, valves, robots etc. are controlled locally by controllers typically close by. These control loops operate fast but require from somewhere their set-points. This is governed on a second tier, where the set-points are controlled by a further layer of control loops operating at a lower speed, and being governed ultimately by the centralized, and human-operated, control center of the manufacturing site, for instance.

In this context, the introduction of wireless URLLC systems today is mostly motivated for mobility-driven use cases, associated for instance with Automated Guided Vehicles (AGV), robots, or also human operators that require (safety) connections to machines. However, more important use cases relate to the flexible reconfiguration of production sites, where apart from the connections also a major part of the automation pyramid benefits from being realized in a flexible manner by the networked infrastructure. In these use cases, the networked infrastructure holds a digital representation of the production site, ie. a digital twin, and updates this as new sensor data is provided, leading to new actuation commands being generated. Infrastructure resources are allocated on-the-fly to communication and compute-intense parts of the digital representation, but can be reallocated at any time depending on the dynamics of the plant, or potential reconfigurations happening during the operation of the plant.

B. *Human-in-the-Loop and Interactive Applications*

Seemingly unrelated to this evolution of industrial automation, so called interactive applications (or human-in-the-loop systems) follow the feedback principle known from control theory, albeit providing feedback to a human user. The main distinguishing feature from closed-loop control systems is that interactive applications do not enforce direct actuation. Instead, the human user is provided with indications, recommendations or warnings. A well-known and very powerful interactive application type are so called cognitive assistants [7]. A single sensor, or multiple sensors around a human, capture information on the human's action. This sensor data is forwarded to a point of processing, where the data is put into the context of a task model, represented by a state diagram. The goal of the application is to guide a user through a sequence of tasks, without the human having specific experience of the task. Essential is a video feed of the currently executed task, while feedback is provided by visual augmentation or haptic stimulus. Mapping the provided video frames to the task model involves quite heavy computational tasks related to computer vision and feature detection.

It is known that the key metric with respect to the perceived QoE is the system responsiveness [8], defined as the time span between capturing the sensor data of the human's action until a feedback/response is perceived by the user. Several aspects contribute to the responsiveness, constituting also the essential trade-offs in such systems:

1) The policy with which the system is sampling the human's environment, in order to identify for instance the completion of a task [9].
2) The forwarding of the captured video frames over a network as well as of the potential feedback to the human user.
3) The placement of the computational tasks that analyze the incoming state information and potentially generate the feedback to the human user, as well as the contention for and scheduling of the corresponding computational resource.

With respect to the corresponding quality of experience, it is known [8] that thresholds τ_1, τ_2 exist for which a cognitive assistant is either perceived as flawless (beyond which point, ie. $< \tau_1$ there is no need for further optimization of the system responsiveness) or for which a cognitive assistant is perceived as unusable due to a too long responsiveness $> \tau_2$. In between these two thresholds, still interesting and interrelated effects are observed with respect to user quality of experience perception and the corresponding reaction to different degrees of responsiveness [10].

C. Commonalities and Differences

While seemingly unrelated, relevant use cases in industrial automation, as well as human-in-the-loop applications like cognitive assistance have major commonalities that point into the same direction with respect to requirements and subsequent evolution for future networked infrastructures.

- In both cases, sufficient digital representations of a certain, local aspect of the real, physical environment are established and maintained through constant updates. Keeping the digital representation up-to-date is essential for recognizing and reacting to important state changes. In other words, in both cases we are dealing with closed-loop feedback systems. There are relations of these representations to so called digital twins, which have been first established as a digital representation of machinery in the context of predictive maintenance [11].
- The responsiveness of the networked infrastructure is the essential metric of these systems, defined as the interplay between the infrastructure and the application in order to deliver a feedback/actuation given the detection of a state change. Responsiveness relates hence to the time delay between generating/capturing a sensor reading/video frame up to delivering on the "other end" the corresponding actuation/feedback/reaction to the primary device or the human and it's environment. Note that the lag between the actual state change in reality until detecting the state change contributes to the system responsiveness as well.
- All these applications introduce a notion of causal dependence, meaning that the provided actuation/feedback/reaction depends on the accuracy and timeliness of capturing the state in the first place, as well as correctly communicating and processing this state. Failing either in the timely communication/computation, and/or in the accurate communication leads to erroneous actuation or feedback provided.
- The degree of utility and automation that can be achieved by such applications is very high. In the case of industrial automation and digital twins, this is well known and first prototypical implementations have been showing this. However, this is also the case for human-in-the-loop systems, which essentially allow the transfer of knowledge and experience with respect to novel situations, or for training purposes, in an automated fashion.

While many use cases for cognitive assistance exist in the context of entertainment, also with respect to professional training and education, as well as supervision, a plethora of use cases can be thought of with very high utilities to the user.

Despite these commonalities with respect to the principles, there exist also differences between automation control loops, and human-in-the-loop systems. For instance, data rates as well as associated computational loads, tend to be low in the case of legacy automation control loops. In contrast, for human-in-the-loop, at least the uplink data rate requirements are substantial, while the computational load with respect to the video scene analysis is much heavier. Furthermore, while latency and reliability requirements are generally high in both cases, the range of valid latency bounds go much lower in case of automation system due to the potentially much higher dynamics.

III. Evolution towards Suitable Infrastructures and Upcoming Challenges

The commonalities of such feedback systems as well as their high potential utility reveal a set of challenges and requirements, that are to a large extent new for networked infrastructures, and certainly point towards the necessary future evolution of such infrastructures.

The most obvious consequence of the described applications, and their potential ubiquitous instantiation within a networked infrastructure, is the provisioning not only of matching communication links - potentially according to URLLC flavor - but also the provisioning of corresponding computational elements. Publicly accessible compute resources will become commonly provided over the infrastructure, such that processing is not only limited to end devices and cloud centers. The push towards edge computing (alias fog computing) as emerging new ”local” compute paradigm is a manifestation of this evolution [12]. Cloudlets of edge computing infrastructures allow for much better responsiveness of feedback systems, simply through their local proximity (in contrast to cloud compute centers) [8]. However, orchestration and resource management (that depends on the required levels of responsiveness of the application) need to be developed in the future to harness such compute resources. This becomes particularly important for critical feedback systems with safety constraints, where the dependability of the application context mandates resource reservation, isolation and guaranteed responsiveness over the entire feedback loop and underlying communication and computational infrastructure. Not only are quasi-deterministic communication latencies required, as addressed through URLLC systems, but also the execution of (externally developed) application code on cloudlets requires to be upper-bounded in terms of compute delays.

Likewise, future networked infrastructures will have to support the mobility of such applications, which leverage communication and compute resources of the infrastructure for timely feedback generation. The joint support of the handover of communication links as well as compute processes (at run-time of the feedback system) will necessitate a substantial redesign, in particular when still having to provide sufficient levels of responsiveness. Aspects of these challenges have been studied in the past, showing the implications for legacy handover schemes with respect to automation systems [13].

Another crucial aspect is the dynamic variation of the level of responsiveness provided by the infrastructure to feedback systems. Generally speaking, control applications do not require at all times the same level of responsiveness from the infrastructure. It heavily depends instead on the state of the controlled plant and the set-point, or in other words, on the recent evolution of the plant's error and its dynamics. This context dependency has not been shown so far for human-in-the-loop applications, while some of the structural similarities of these applications with closed-loop control imply that such dependencies exist also for them. Recently, nevertheless, we showed [10] certain dependencies between the application pacing of human-in-the-loop systems and the infrastructure responsiveness, which are of temporal variation and have profound implications for example for cross-layer optimization of human-in-the-loop systems. As the responsiveness of the infrastructure deteriorates, the task execution times of humans using cognitive assistance applications deteriorates as well, leading to extended overall application usage durations of up to 50% for executing the *same* task sequence. All these aspects point to a re-consideration of cross-layer optimization between the infrastructure and the applications. The challenge here is obviously that this optimization has to be performed over the entire application path with communication and computational elements, taking the instantaneous requirements of the application into account (for instance, the state of the controlled plant). Such schemes are theoretically difficult to model and analyze [14], while posing additional challenges when trying to implement them.

Finally, tight responsiveness of feedback systems implies a certain relevance only in the direct context of the application. Together with the temporal aspect, semantic relevance of messages (either upstream to the point of computation, or downstream in form of provided feedback) and the specific propagation environment, a potentially much larger optimization space is opened up than what standardized systems with decades-old network stacks allow to leverage. The additional "softwarization" of wireless network stacks can lead to entirely new networked solutions of such feedback systems, that defy to a large extent standardized algorithms and parameter settings. Instead, they might leverage much more statistical characteristics of the application itself and the immediate propagation environment. Cross-layer optimization, paired with end-to-end learning approaches for wireless message exchange could point towards entirely new solutions for feedback systems.

IV. Conclusions and Outlook

With the recent pivot towards URLLC systems, and corresponding applications, a first step has been undertaken towards a scalable and ubiquitous provisioning of infrastructures for closed-loop feedback systems. Such systems are mostly seen today in the specific context of industrial automation, but have a much wider reach also towards human-in-the-loop systems and generally towards any digital twin representation with real-time feedback/actuation provisioning. Central to these emerging application classes is a digital representation of a certain aspect of reality which is instantiated and updated, and from which feedback towards the physical reality is generated. Such applications mandate as key performance metric an appropriate responsiveness of the networked infrastructure, which constitutes of communication and computational delays along the application loop. Responsiveness over a communication and computational infrastructure path is a novel key performance indicator, which nevertheless will have tremendous impact on the design and operation of future networked infrastructures. To date, neither sufficiently situated computational resources, nor corresponding algorithms to guarantee the required responsiveness of the infrastructure exist, posing the most important challenge towards the proliferation of such tightly integrated feedback applications. Beyond the challenge of providing integrated communication and compute paths for closed-loop feedback applications, cross-layer optimization and mobility support will pose further challenges.

Finally, anticipating large-scale deployment of feedback systems in future networked infrastructures, a more distant, but relevant question relates to the usage of the many representations of reality updated by the plethora of applications. We can imagine at any point in time perhaps many millions of closed-loop applications (either in the context of automation or in the context of human-in-the-loop) to be actively running over a networked infrastructure in one decade from now. What value can be extracted from networking these representations, for instance towards an "Internet of Reality"? Which requirements need to be fulfilled to harness these representations, for instance with respect to using unified coding structures to represent aspects of reality? How is this potential value jeopardized by security and especially privacy constraints?

References

[1] P. Cerwall, P. Jonsson, and S. Carson, "Ericsson mobility report," Ericsson AB, Tech. Rep., 2019.

[2] F. Adelantado, X. Vilajosana, P. Tuset-Peiro, B. Martinez, J. Melia-Segui, and T. Watteyne, "Understanding the limits of lorawan," *IEEE Communications Magazine*, vol. 55, no. 9, pp. 34–40, 2017.

[3] S. Grant, "3gpp low power wide area technologies - gsma white paper," GSMA, Tech. Rep., 2016.

[4] L. Columbus, "2017 roundup of internet of things forecasts," 2017.

[5] J. Sachs, G. Wikstrom, T. Dudda, R. Baldemair, and K. Kittichokechai, "5g radio network design for ultra-reliable low-latency communication," *IEEE Network*, vol. 32, no. 2, pp. 24–31, 2018.

[6] M. Lema, A. Laya, T. Mahmoodi, M. Cuevas, J. Sachs, J. Markendahl, and M. Dohler, "Business case and technology analysis for 5g low latency applications," *IEEE Access*, vol. 5, pp. 5917–5935, 2017.

[7] K. H. et al., "Towards wearable cognitive assistance," in *Proc. International Conference on Mobile Systems, Applications and Services (ACM MobiSys)*, 2014.

[8] Z. C. et al., "An empirical study of latency in an emerging class of edge computing applications for wearable cognitive assistance," in *Proc. ACM/IEEE SEC*, 2017.

[9] V. Moothedath, J. Champati, and J. Gross, "Energy-optimal sampling of edge-basedfeedback systems," in *Proc. IEEE International Conference on Communications (ICC)*, 2021.

[10] M. Olguin, R. Klatzky, J. Wang, P. Pillai, M. Satyanarayanan, and J. Gross, "Impact of delayed response on wearable cognitive assistance," *PLOS ONE*, 2021.

[11] E. Negri, "A review of the roles of digital twin in cps-based production systems," *Procedia Manufacturing*, no. 11, pp. 939–948, 2017.

[12] M. Satyanarayanan, "The emergence of edge computing," *IEEE Computer*, vol. 50, no. 1, pp. 30–39, 2017.

[13] D. van Dooren, G. Fodor, J. Gross, and K. Johansson, "Delay analysis of group handover for real-time control over mobile networks," in *Proc. IEEE Global Telecommunications Conference (GlobeCom)*, 2018.

[14] J. Champati, H. Al-Zubaidy, and J. Gross, "Transient analysis of multi-hop wireless networks under static routing," *IEEE/ACM Transactions on Networking*, vol. 28, no. 2, pp. 722–735, 2020.