Beiträge aus der Informationstechnik

Mobile Nachrichtenübertragung
Nr. 95

Lucas Scheuvens

On Age-of-Information Aware Resource Allocation for Industrial Control-Communication-Codesign

VOGT

Dresden 2022

Bibliografische Information der Deutschen Nationalbibliothek
Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über http://dnb.dnb.de abrufbar.

Bibliographic Information published by the Deutsche Nationalbibliothek
The Deutsche Nationalbibliothek lists this publication in the Deutsche Nationalbibliografie; detailed bibliographic data are available on the Internet at http://dnb.dnb.de.

Zugl.: Dresden, Techn. Univ., Diss., 2022

Die vorliegende Arbeit stimmt mit dem Original der Dissertation „On Age-of-Information Aware Resource Allocation for Industrial Control-Communication-Codesign“ von Lucas Scheuvens überein.

Gesetzt vom Autor

ISBN 978-3-95947-055-1

Jörg Vogt Verlag
Niederwaldstr. 36
01277 Dresden
Germany

Phone: +49-(0)351-31403921
Telefax: +49-(0)351-31403918
e-mail: info@vogtverlag.de
Internet : www.vogtverlag.de

On Age-of-Information Aware Resource Allocation for Industrial Control-Communication-Codesign

Dipl.-Ing.
Lucas Scheuvens

der Fakultät Elektrotechnik und Informationstechnik der
Technischen Universität Dresden

zur Erlangung des akademischen Grades

Doktoringenieur
(Dr.-Ing.)

– genehmigte Dissertation –

Vorsitzender	Prof. Dr. Stefan Mannsfeld
Gutachter	Prof. Dr.-Ing. Dr. h.c. Gerhard Fettweis Prof. Dr. Beatriz Soret
Tag der Einreichung	1. März 2022
Tag der Verteidigung	3. Mai 2022

Lucas Scheuvens
On Age-of-Information Aware Resource Allocation
for Industrial Control-Communication-Codesign
Dissertation

Technische Universität Dresden
Vodafone Chair Mobile Communications Systems
Institute of Communication Technology
Faculty of Electrical and Computer Engineering
01062 Dresden, Germany

Abstract

In industrial manufacturing, *Industry 4.0* refers to the ongoing convergence of the real and virtual worlds, enabled through intelligently interconnecting industrial machines and processes through information and communications technology. Ultra-reliable low-latency communication (URLLC) is widely regarded as *the* enabling technology for *Industry 4.0* due to its ability to fulfill highest quality-of-service (QoS) comparable to those of industrial wireline connections. In contrast to this trend, a range of works in the research domain of *networked control systems* have shown that URLLC's supreme QoS is not necessarily required to achieve high quality-of-control; the co-design of control and communication enables to jointly optimize and balance both quality-of-control parameters and network parameters through blurring the boundary between application and network layer. However, through the tight interlacing, this approach requires a fundamental (joint) redesign of both control systems and communication networks and may therefore not lead to short-term widespread adoption. Therefore, this thesis instead embraces a novel co-design approach which keeps both domains distinct but leverages the combination of control and communications by yet exploiting the age of information (AoI) as a valuable interface metric.

This thesis contributes to quantifying application dependability as a consequence of exceeding a given peak AoI with the particular focus on packet losses. The beneficial influence of negative temporal packet loss correlation on control performance is demonstrated by means of the automated guided vehicle use case. Assuming small-scale fading as the dominant cause of communication failure, a series of communication failures are mapped to an application failure through discrete-time Markov models for single-hop (e.g, only uplink or downlink) and dual-hop (e.g., subsequent uplink and downlink) architectures. This enables the derivation of application-related dependability metrics such as the *mean time to failure* in closed form. For single-hop networks, an AoI-aware resource allocation strategy termed *state-aware resource allocation* (SARA) is proposed that increases the application reliability by orders of magnitude compared to static multi-connectivity while keeping the resource consumption in the range of best-effort single-connectivity. This dependability can also be statistically guaranteed on a system level – where multiple agents compete for a limited number of resources – if the provided amount of resources per agent is increased by approximately 10 %. For the dual-hop scenario, an AoI-aware resource allocation optimization is developed that minimizes a user-

defined penalty function that punishes low application reliability, high AoI, and high average resource consumption. This optimization may be carried out offline and each resulting optimal SARA scheme may be implemented as a look-up table in the lower medium access control layer of future wireless industrial networks.

Kurzfassung

Unter dem Überbegriff *Industrie 4.0* wird in der industriellen Fertigung die zunehmende Digitalisierung und Vernetzung von industriellen Maschinen und Prozessen zusammengefasst. Die drahtlose, hoch-zuverlässige, niedrig-latente Kommunikation (engl. *ultra-reliable low-latency communication*, URLLC) – als Bestandteil von 5G – gewährleistet höchste Dienstgüten, die mit industriellen drahtgebundenen Technologien vergleichbar sind und wird deshalb als Wegbereiter von *Industrie 4.0* gesehen. Entgegen diesem Trend haben eine Reihe von Arbeiten im Forschungsbereich der vernetzten Regelungssysteme (engl. *networked control systems*, NCS) gezeigt, dass die hohen Dienstgüten von URLLC nicht notwendigerweise erforderlich sind, um eine hohe Regelgüte zu erzielen. Das Co-Design von Kommunikation und Regelung ermöglicht eine gemeinsame Optimierung von Regelgüte und Netzwerkparametern durch die Aufweichung der Grenze zwischen Netzwerk- und Applikationsschicht. Durch diese Verschränkung wird jedoch eine fundamentale (gemeinsame) Neuentwicklung von Regelungssystemen und Kommunikationsnetzen nötig, was ein Hindernis für die Verbreitung dieses Ansatzes darstellt. Stattdessen bedient sich diese Dissertation einem Co-Design-Ansatz, der beide Domänen weiterhin eindeutig voneinander abgrenzt, aber das Informationsalter (engl. *age of information*, AoI) als bedeutenden Schnittstellenparameter ausnutzt.

Diese Dissertation trägt dazu bei, die Echtzeitanwendungszuverlässigkeit als Folge der Überschreitung eines vorgegebenen Informationsalterschwellenwerts zu quantifizieren und fokussiert sich dabei auf den Paketverlust als Ursache. Anhand der Beispielanwendung eines fahrerlosen Transportsystems wird gezeigt, dass die zeitlich negative Korrelation von Paketfehlern, die in heutigen Systemen keine Rolle spielt, für Echtzeitanwendungen äußerst vorteilhaft ist. Mit der Annahme von schnellem Schwund als dominanter Fehlerursache auf der Luftschnittstelle werden durch zeitdiskrete Markovmodelle, die für die zwei Netzwerkarchitekturen *Single-Hop* und *Dual-Hop* präsentiert werden, Kommunikationsfehlerfolgen auf einen Applikationsfehler abgebildet. Diese Modellierung ermöglicht die analytische Ableitung von anwendungsbezogenen Zuverlässigkeitsmetriken wie die durschnittliche Dauer bis zu einem Fehler (engl. *mean time to failure*). Für Single-Hop-Netze wird das neuartige Ressourcenallokationsschema *State-Aware Resource Allocation* (SARA) entwickelt, das auf dem Informationsalter beruht und die Anwendungszuverlässigkeit im Vergleich zu statischer Multi-Konnektivität um Größenordnungen erhöht, während der Ressourcenverbrauch im Bereich von konventioneller Einzelkonnektivität bleibt.

Diese Zuverlässigkeit kann auch innerhalb eines Systems von Regelanwendungen, in welchem mehrere Agenten um eine begrenzte Anzahl Ressourcen konkurrieren, statistisch garantiert werden, wenn die Anzahl der verfügbaren Ressourcen pro Agent um ca. 10 % erhöht werden. Für das Dual-Hop Szenario wird darüberhinaus ein Optimierungsverfahren vorgestellt, das eine benutzerdefinierte Kostenfunktion minimiert, die niedrige Anwendungszuverlässigkeit, hohes Informationsalter und hohen durchschnittlichen Ressourcenverbrauch bestraft und so das benutzerdefinierte optimale SARA-Schema ableitet. Diese Optimierung kann offline durchgeführt und als Look-Up-Table in der unteren Medienzugriffsschicht zukünftiger industrieller Drahtlosnetze implementiert werden.

Acknowledgement

This thesis would not have been possible without many people and institutions.

First, I whole-heartedly thank may doctoral advisor Gerhard Fettweis for giving me the opportunity to pursue my PhD at the Vodafone Chair Mobile Communications Systems at the Technische Universität Dresden. You inspired me in your enthusiam, optimism, creativity, and willingness to drive things forward to generate impact. During the time at the chair, you placed trust in me and gave me the chance to prove myself with responsibilities that I had never taken before. Second, I express my gratitude to Beatriz Soret for her unhesitated willingness to act as the second reviewer.

I deeply thank my beloved wife Lea for the ever-lasting support, especially by taking over many family duties in particularly challenging and stressful times. This thesis would have been signficantly more challenging without you.

I consider myself lucky that the German Federal Ministry of Education and Research as well as the European Commission put me in the comfortable position to pursue this work on a full salary. I do not take this for granted and I am highly grateful for it. I thank the Technische Universität Dresden for supplying me with an environment that supports research excellence.

My direct supervisors have changed continuously during my thesis and I owe thanks to all. I especially thank André Noll-Barreto who continously encouraged me and greatly improved my work through his scientific rigour. In chronological order, I additionally thank Meryem, Henrik, David, and Arturo for offering me guidance when I needed it.

My time at the chair has been an exciting adventure. I generated many wonderful memories with the exceptional people I met there. In a non-exhaustive list, I would like to mention Andreas, Behnam, David, Eva, Friedrich, Henrik, Nick, Norman, Philipp, Robert, and Tom. We had a great time with lots of fun and amazing achievements, I am thankful to have been part of the team. I want to express my appreciation for Eva, Carolin, Kathrin, Raffael, Rüdiger, Sylvia, Steffen, and Yaning for keeping the chair running by helping us in all organizational matters.

Contents

1

Introduction

Wireless communication is ubiquitous in today's world. Mobile internet connectivity has increased tremendously in recent years, increasing from connecting 32 % of all people worldwide in 2014 to 49 % in 2019 [GSM20], with a projected penetration of 70 % by 2023 [Cis20]. The societal impact is immense and has manifested clearly during the Covid-19 pandemic [KP20]. Today, most connections still involve at least one human but this will change soon. Machine-to-machine-type communications (M2M) is the by far fastest growing connection type with a compound annual growth rate of 30 %, putting the second-place smartphone sector (7 %) in the shade. It is projected that by 2023 50 % of all wireless connections will not involve humans [Cis20].

Expecting this trend, it comes to no surprise that the integration of M2M communications to wireless networks has been a significant part in the standardization of the fifth generation of cellular networks (5G). Therein, three main pillars are defined [ITU15; Pop+18b].

1. Enhanced mobile broadband
2. Massive machine-type communications (mMTC)/Internet of things (IoT)
3. Ultra-reliable low-latency communications (URLLC)/Mission-critical machine-type communications (μMTC)

While the first is human-centric and mainly evolves around increasing data rates especially for augmented reality and virtual reality applications, the latter two are M2M. The domain of mMTC/IoT is expected to support a vast amount of wireless network devices, e.g., wireless sensors and actuators for smart city, smart logistics, and home applications [Sha+15], usually imposing loose latency requirements and are therefore non-mission-critical. On the other hand, *Tactile Internet* [Fet14; Sim+16; Sch+17] applications such as wireless factory automation, self-driving cars, and real-time remote control, are mission-critical and demand highly reliable connectivity at a minimal latency (URLLC).

In the context of wireless communications, the *reliability* of a connection is defined as the complement of the packet loss ratio (PLR), i.e., $1 - \text{PLR}$, whereby a packet is

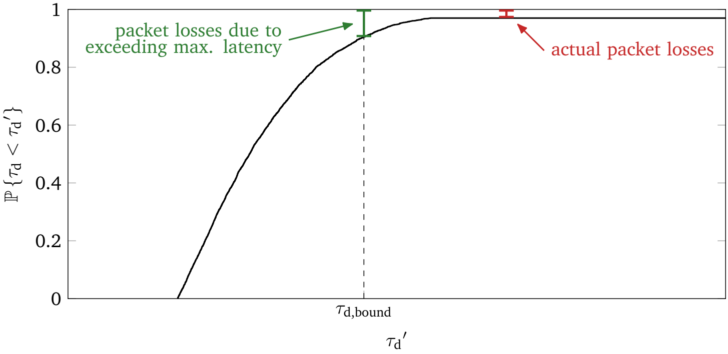

Fig. 1.1.: Latency and packet losses are related quantities (symbolic plot). Packets are also considered *lost* if they arrive after a given threshold $\tau_{d,bound}$ [Pop+18a].

considered lost when it is not correctly received within the time constraint required by the targeted service [Pop+18a; 3GP19b]. This definition relates the PLR to a latency requirement, compare Fig. 1.1.

Reliability may be generally enhanced through *diversity*. If the latency requirement is relaxed, ultra-reliability can be achieved through *time diversity* in the form of (infinitely many) retransmissions of erroneous data. This has been state-of-the-art in wireless standards for decades and is still valid for current wireless local area network (WLAN) and long term evolution (LTE) systems with their respective automatic repeat request (ARQ)/hybrid ARQ (HARQ) mechanisms. However, if the required latency decreases to a minimum, the time budget for a successful transmission is exhausted quickly, narrowing the possibility of retransmissions. For extremely low latencies in the range of <1 ms, only a single (initial) transmission may be viable. In this scenario, a reliability increase may only be achieved through concurrent diversity means, e.g., *frequency diversity* (in the form of multiple simultaneous transmissions in different frequency bands) or *spatial diversity* (in the form of multiple antennas). Concurrent diversity techniques fundamentally differ from time diversity in terms of cost. As the success of a specific wireless transmission cannot be predicted beyond long-term average metrics, wireless resources must be allocated preemptively (frequency diversity) or more antennas need to be installed (spatial diversity), which increases capital and operational expenditure. The authors of [ÖF15] have found that in realistic fading environments, an outage probability <0.001 % in a 1 ms grid may require more than 10 parallel links, which (a) is unrealistic from a hardware perspective and (b) severely challenges the scalability potential of such systems.

While the extreme quality of service (QoS) requirements of URLLC state a packet success probability of $>99.999\,\%$ at a latency bound $\tau_{\mathrm{d,bound}} = 1\,\mathrm{ms}$, first 6G white papers envision “enhanced” [LL19] or “extreme” [NTT21] URLLC, which will signify the resource consumption problem further in future wireless networks.

1.1 The Need for an Industrial Solution

Although URLLC is also relevant for many other application domains, the subdomain of industrial manufacturing is of particular importance in Germany, playing a central strategic role for economic growth and development. The German federal government coined the term *Industry 4.0*, following the nomenclature of software versioning indicating a major upgrade. It describes a shift towards *smart manufacturing* that is broadly characterized by the following points. [Aca13]

1. Individual customer requirements can be met and even one-off items can be manufactured profitably.
2. Agile manufacturing processes pose resilience against unforeseen disruptions and maximize efficiency.
3. Value opportunities through new services are created, e.g., smart algorithms that leverage the resulting (big) data from smart sensors.

Technologically, this shift requires – among other things – an information technology system capable of exchanging data vertically (between manufacturing systems and business processes) and horizontally (within a layer) in the International Society of Automation 95 model (*automation pyramid*), see Fig. 1.2 [Roj17]. Within this model, the width of each layer typically describes the level of heterogeneity and data transmission becomes increasingly mission critical with lower layers.

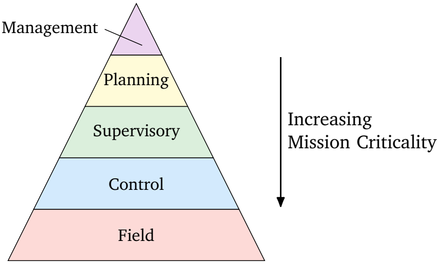

Fig. 1.2.: Automation pyramid according to International Society of Automation 95 model.

5G is often viewed as accelerator (or sometimes even enabler) for *Industry 4.0* as it is the first (wireless) technology to support a range of different QoS classes [Gun+18]. When connecting and operating applications on the factory floor (*field level*), tight requirements are formulated to avoid costly communication-induced production downtime. The motivation is straight-foward: no errors in the communication domain guarantee no communication-induced application failures, as is the case with prevalent production-grade wire-line communication such as Ethercat, SERCOS III, and Profinet IRT. However, the 1:1 replacement of wire-line technologies with wireless counterparts is infeasible due to their fundamentally different operating principles. Multiple key differences stand out:

1. For low-latency applications, the air interface is orders of magnitude less dependable due to path loss, shading, small-scale fading, and interference. Diversity may be used to increase reliability, however, at the cost of increased resource consumption (as discussed above).
2. The air interface is a shared medium while cables are not. An increased number of terminals introduces competition for resources, which not only reduces the individual transmission capacity but also increases the complexity of dynamic scheduling.
3. Commonly used wire-line ring topologies enable to “piggyback” data onto other packets, which reduces communication overhead by sharing header information. With wireless networks, header information must be transmitted separately for every data paket, which further accentuates the capacity issue in wireless networks.

These points indicate that wireless resources are extremely valuable and, therefore, spectral efficiency is crucial to enable real-time applications also in dense scenarios (*scalability*). The conventional URLLC approach does not address this necessity and, hence, might not be ideal to realize the *Industry 4.0* vision.

The need for *dynamic and dependable* industrial applications drives the need for *low-latency and dependable* wireless communication. This thesis contributes to paving the way towards the *Industry 4.0* vision without compromising on spectral efficiency and thereby enabling an increased application density.

1.2 Contributions

This thesis contributes to realizing dynamic, ultra-dependable, and scalable real-time applications over spectrally efficient wireless networks by fundamentally challenging the design target of low-level industrial networks. The age of information (AoI) as a time-based metric is embraced as primary indicator for dependability on an application level, rather than optimizing for long-term average metrics such as the PLR. In detail, the contributions are:

- Chapter 2 provides the foundation of the thesis by presenting the state of the art in the research fields of URLLC, networked control system (NCS), and AoI.
- In Chapter 3, the impact of packet losses on control applications is studied. First, a high-level introduction to linear, time-invariant (LTI) control theory is given. Second, two joint descriptions of control loops with packet losses are discussed to derive actual QoS requirements of closed-loop control applications. The first enables to derive conservative bounds within the LTI domain and therefore provides an easily accessible entry point for such an assessment. The second more advanced approach utilizes markov jump linear system (MJLS) theory to show that limiting the packet loss sequence length has a highly beneficial effect on control utility. The packet loss process is modeled as a discrete-time Markov chain and a coefficient is introduced that describes the temporal correlation between packet losses. The impact of positive and negative correlation on the control loop is studied by means of an example industrial automated guided vehicle (AGV) use case. Lastly, the theoretical results are validated through extensive simulation.
- Chapter 4 presents failure models of real-time applications due to communication errors. First, the (industrial) communication assumptions are presented and causes of failure are discussed. Two network architecture are introduced (single-hop and dual-hop) that build the foundation for all investigations and optimizations of this thesis. Markov chain failure models are developed for both network architectures that link the event of exceeding a certain AoI to an application failure. In the single-hop case, this translates to losing more than a given number of consecutive packets. Multiple key performance indicators (KPIs) are introduced to describe the gains in terms of network performance, resource consumption, and application dependability.
- Chapter 5 exploits the findings of Chapter 3 for an error-prone single-hop wireless network connection. Through the dynamic assignment of resources,

the network can enforce a negative temporal packet loss correlation, limiting the sequence length of consecutive packet losses. A range of different dynamic resource allocation schemes is presented, termed state-aware resource allocation (SARA) and the advantages of SARA with respect to application dependability and network resource consumption are demonstrated. Lastly, the impact of erroneous acknowledgments (ACKs) is investigated, which are otherwise considered ideal.

- In Chapter 6, the framework of Chapter 5 is extended to the multi-agent case, in which a network-wide resource pool must be shared. As system resources might be limited, individual agents may not receive the resources they require and may fail sooner than predicted in the single-agent case. The modeling is performed through a holistic discrete-time Markov chain approach, which includes all individual-agent state transitions. The computational complexity of the approach is investigated and contrasted with required extensive simulation efforts. The single-agent performance metrics of Chapter 4 are lifted to system-level performance metrics, in which the system is considered to fail as soon as a single agent fails. For the case of too few resources in the system, the strategies "random" and "cliff" are investigated, which differ in their prioritization of users. Also, the assignment of spare system resources is investigated. For better tractability, a low-complexity example is presented. The applicability of SARA to a system of agents is discussed thoroughly with a key focus on the required number of system resources that allows for unimpaired KPIs for the individual agent.
- Chapter 7 considers the dual-hop network architecture. An optimization approach is presented that derives an optimal SARA scheme depending on user- and application-specific penalty functions regarding the application dependability, the average resource consumption, and the AoI. Through multiple exemplary sets of penalty functions the dependence of the optimal SARA schemes on the input penalty functions is demonstrated. The interdependence of uplink (UL) and downlink (DL) stemming from a joint AoI optimization is studied as well as the impact of the individual link's packet loss probability.
- Finally, Chapter 8 concludes this thesis and highlights open questions for further research.