# Poster: Low Latency Networking for Industry 4.0

Xiaolin Jiang, Carlo Fischione
Department of Automatic Control
KTH, Royal Institute of Technology
{xiaolinj, carlofi}@kth.se

Zhibo Pang
Automation Solution Group
ABB Corporate Research
pang.zhibo@se.abb.com

## Abstract

Industry 4.0 poses a stringent delay requirement to the communication network. However, the current protocols cannot satisfy this requirement, as they are mainly designed to achieve high data rates rather than low latency. It is difficult to reduce the delay as the end-to-end delay is accumulated from every layer of the communication network, and the delay introduced by one layer may also be coupled with that from other layers. We analyze the different delay components and investigate the potential techniques that can help reduce one or multiple delay components. To support different latency requirement together with reliability and throughput requirements, the combination and the parameters of potential techniques should be designed in a coordinate way.

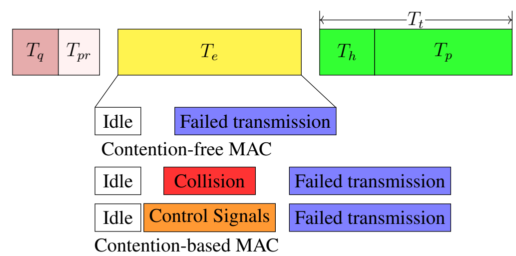
## 1 Introduction

Industry 4.0 is the current trend of the industrial automation. Besides further boosting the production efficiency, Industry 4.0 also envisions to embrace the ability of machines, sensors and people to connect and communicate with each other via the Internet of Things [1]. To enable smooth and efficient operation and fast response to warnings and failures, the communication should be performed with very low latency. Specifically, factory automation requires low latency communications of the order of 0.5-1 ms [2]. Few milliseconds' latencies are also needed for remote control robots, which would thus become a promising alternative to traditional costly autonomous robots. By low latency communications, these remote-control robots would allow, for instance, to have smooth movement in harsh environment (such as construction sites) and to deliver visual and haptic feedback. The challenges for the most critical wireless control applications are comprehensively investigated in [3].

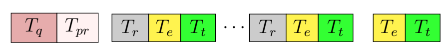
Despite our society being fully connected thanks to wireless networks and Internet, currently there is neither fundamental design method nor technology capable to ensure real-time communications. Such a research and technological gap will hinder the growth and progress of Industry 4.0.

International Conference on Embedded Wireless
Systems and Networks (EWSN) 2017
20–22 February, Uppsala, Sweden

ISBN: 978-0-9949886-1-4




(a) Composition of single hop delay, e.g., from mobile phone to the BS.




(b) Delay components of multiple hops, e.g., from mobile phone to the remote server.

**Figure 1. The delay formation of single hop and multi-hop communication.**

To achieve the goal of low latency communication, we should at first determine where does the delay accumulates and define each involved components of the delay. Define *end-to-end delay* as the time duration between the message generation and when it is correctly decoded by the receiver at the destination. Fig. 1 shows various components of the end-to-end delay. Once a packet is generated, it is put into a queue waiting to be transmitted. Analogously, when a packet arrives at the receiver, it may also be waiting in a queue to be processed. We define the sum of these queuing delay as $T_q$. When there are multiple classes with different priorities, the packets with higher priority usually have lower $T_q$. Similarly, the total processing time during the end-to-end trip is also abstracted together, and is defined as the processing delay $T_{pr}$. For one hop communication, it takes $T_e$ for the packets to establish the channel link, where $T_e$ may include the waiting time for the channel to be idle, the time to send some control signals to reserve the channel, and the failed transmission due to collision or bad channel state. The successful packet transmission time lasts for $T_t$, which includes header transmission time $T_h$ and payload transmission time $T_p$. If the transmission needs multiple hops, then routing delay $T_r$ adds additional delays. We do not take the propagation delay into account, as the propagation delay is merely restricted by

**Table 1. Potential methods to achieve low latency**

| | Methods | Transmission delay $T_t$ | Channel establishment delay $T_e$ | Routing delay $T_r$ |
|---|---|---|---|---|
| Methods having single efffect | Coding & Modulation | √ | | |
| | Sending short packets | √ | | |
| | MAC algorithms | | √ | |
| | Routing algorithms | | | √ |
| Methods having multiple efffects | Waveforms | √ | √ | |
| | HARQ | √ | √ | |
| | Beamforming | √ | √ | |
| | Combined MAC & Routing | | √ | √ |
| | Cloud RAN & Mobile Fronthaul | | √ | √ |
| | Edge Caching & Fog Computing | | √ | √ |
| | D2D | | √ | √ |

the distance and propagation speed.

# 2 Techniques Enabling low Latencies

In this section, we present an overview on the potential techniques to achieve low latency communications. Table 1 shows the most prominent communication techniques that affect different components of delay. We also present brief explanation about how the delay components can be affected and reduced.

## 2.1 Transmission Delay

Transmission delay is the time consumed to convert the message into a serial bit stream to be transmitted over the communication media. For a fixed-length packet, higher transmit rate means lower transmission delay. Modulation and coding is a traditional technique to increase the transmit rate or decrease the bit error rate. Generally speaking, for sufficiently strong wireless links, high modulation order together with light coding schemes can substantially boost the transmit rate. However, when the channel becomes poor, we should reduce the transmit rate (by adopting lower modulation orders or stronger coding schemes) to maintain the target bit error rate. The use of short packets not only decrease the packet size, it also brings difference to the maximum coding rate and the packet error probability [4].

## 2.2 Channel Establishment Delay

The channel establishment delay is defined as the time difference from the instant the node starts trying to send a packet until the beginning of its successful transmission. The channel establishment delay is closely related to the MAC scheduling algorithms, which are divided into two categories: contention-based and contention-free.

Most contention-free protocols impose a constant delay which scales almost linearly with the number of transmitters in the network. TDMA and other contention-free MACs are efficient and can provide better performance than contention-based MAC when the number of devices as well as the traffic pattern are predictable and controllable. However, due to the constant channel establishment delay, contention-free protocols may not be the best options to handle a network with many transmitters each having sporadic short packets. On the other hand, if a lot of empty time slots can be afforded, contention free protocols are also good options.

## 2.3 Routing Delay

When the communication between the transmitter and receiver cannot be completed within one hop, the routing delays due to multiple hops must be added to the end-to-end delay. With fixed topologies, the routing algorithms are divided in to back-pressure-based routing and non-back-pressure-based routing. Back-pressure-based routing algorithms utilize the queue length as the metric for delay, and their optimality is validated when there are stable packets for each destination. However, when the load is light, this queue length metric does not hold, and non-back-pressure-based routing may achieve better delay performance.

## 2.4 Multiple Delay Components

Some of these techniques may only have effect on a single delay component, whereas others may affect multiple delay components. For example, there exists a tradeoff between the transmission delay and channel establishment delay when using HARQ and beamforming. The insight for the techniques affecting both the channel establishment delay and the routing delay is that the resource brought nearer enables the communication to be performed with fewer hops and reduced contention. In all, as different techniques cause various and maybe opposite affect to different delay components, the parameters of selected techniques should be tuned and optimized together to reduce the end-to-end delay.

# 3 Conclusions

Low latency communications are the premise for Industry 4.0 implementation. In this work, we investigated how the delay accumulates from physical layer to transport layer, and we showed how to characterize the end-to-end delay into several components. Then we discussed how different techniques may influence one or multiple delay components. These techniques should be optimized together to reduce the delay while satisfying other requirements such reliability and throughput.

# 4 References


[1] Mario Hermann, Tobias Pentek, and Boris Otto. Design Principles for Industrie 4.0 Scenarios. In *in Proc. IEEE 49th Hawaii Int. Conf. Syst. Sci. (HICSS)*, pages 3928–3937. IEEE, 2016.

[2] Dalimir Orfanus, Reidar Indergaard, Gunnar Prytz, and Tormod Wien. Ethercat-based platform for distributed control in high-performance industrial applications. In *in Proc. IEEE 18th Conf. Emerg. Technol. Factory Autom. (ETFA)*, pages 1–8. IEEE, 2013.

[3] Michele Luvisotto, Zhibo Pang, and Dacfey Dzung. Ultra High Performance Wireless Control for Critical Applications: Challenges and Directions. *IEEE Trans. Ind. Informat.*, 2016.

[4] Giuseppe Durisi, Tobias Koch, and Petar Popovski. Towards massive, ultra-reliable, and low-latency wireless: The art of sending short packets. *arXiv preprint arXiv:1504.06526*, 2015.