SPOTLIGHT ON THE INTERNET OF THINGS
URLLC capabilities to provide deterministic connectivity end to end. This will depend to a large extent on a combination of TSN features and 5G technology for converged networks of Industry 4.0. Future industrial automation will have a variety of new requirements for networks and clouds. At Ericsson, we believe that the best way to address the growing connectivity needs of this industry sector is through a common network solution, as opposed to taking a single-segment silo approach.

## 21 5G-TSN INTEGRATION MEETS NETWORKING REQUIREMENTS FOR INDUSTRIAL AUTOMATION

Time-Sensitive Networking (TSN) is becoming the standard Ethernet-based technology for converged networks of Industry 4.0. Future industrial automation will depend to a large extent on a combination of TSN features and 5G Ultra-Reliable Low-Latency Connectivity (URLLC) capabilities to provide deterministic connectivity end to end.

## 31 END-TO-END SECURITY MANAGEMENT FOR THE IOT

Service providers that want to capitalize on IOT opportunities without taking undue risks need a security solution that provides continuous monitoring of threats, vulnerabilities, risks and compliance, along with automated remediation. We have developed an end-to-end IOT security and identity management architecture that delivers on all counts.

## 39 DISTRIBUTED CLOUD: A KEY ENabler of Automotive and Industry 4.0 USE CASES

Emerging use cases in the automotive industry – as well as in manufacturing industries where the first phases of the fourth industrial revolution are taking place – have created a variety of new requirements for networks and clouds. At Ericsson, we believe that distributed cloud is a key technology to support such use cases.

## 49 BOOSTING SMART MANUFACTURING WITH 5G WIRELESS CONNECTIVITY

5G wireless connectivity has been designed to enable the fully-connected factories of the future. The integration of 5G ultra-reliable low latency communication (URLLC) in the manufacturing process will accelerate the transformation of the manufacturing industry and make smart factories more efficient and productive than ever.

## 59 KEY TECHNOLOGY CHOICES FOR OPTIMAL MASSIVE IOT DEVICES

LTE-M and NB-IoT have enabled the introduction of a new generation of IOT devices that deliver on the promise of scalable, cost-effective massive IOT applications using LPWAN technology. However, a few key technology choices are necessary to create IoT devices that can support the multitude of existing and emerging massive IOT use cases.
SPOTLIGHT ON THE INTERNET OF THINGS

THE INTERNET OF THINGS (IoT) has emerged as a fundamental cornerstone in the digitalization of both industry and society as a whole. It represents a huge opportunity not only in economic terms, but also from a global challenges perspective – making it easier for governments, non-governmental organizations and the private sector to address pressing food, energy, water and climate related issues. With this in mind, we decided to create a special issue of Ericsson Technology Review solely focused on IoT opportunities and challenges.

At its heart, the IoT involves the collection and analysis of insights and the automation of processes involving machines, things, places and people, thus in essence fusing the physical and cyber realms into one system. In so doing, it transforms business models – making it possible to sell services rather than products, for example, or outcomes instead of services – as well as enabling the reengineering of business processes to achieve the same outcome in a more efficient way.

Not a single technology, the IoT is instead composed of a set of key technologies, encompassing devices with sensors and actuators, connectivity, cloud and edge computing, artificial intelligence/machine learning (AI/ML) and security.

5G and the IoT are closely intertwined. One of the biggest innovations within 5G is support for the IoT in all its forms, both by addressing mission criticality as well as making it possible to connect low-cost, long-battery-life sensors.

Supporting the fourth industrial revolution Industry 4.0 – also known as the fourth industrial revolution – is already heavily reliant on IoT technologies. Manufacturing companies have demanding requirements in terms of cost, flexibility, safety and performance, and it is critical that their requirements are addressed in the ongoing development of the IoT. The automotive and transportation industry is another sector that is undergoing fundamental technology changes that require specialized IoT support. Both of these sectors are examined in detail in this issue of the magazine.

Harnessing the full potential Fundamental to any IoT solution is the ability to connect the things of interest. Huge potential is lost when it is not possible to get the relevant things and locations online. When everything is connected, however, a wealth of new data becomes available, raising questions about how it should be handled (and potentially monetized).

The wealth of data that the IoT generates can be used for a wide range of different purposes – everything from controlling robots on a factory floor to tracking and monitoring perishable goods in logistics on a global scale by the creation of Digital Twins. As such, IoT and Cyber-Physical System are converging into one and the same concept. Data must be processed both in the cloud and close to where it is produced and consumed, driven by requirements for reliability, cost and performance. Compute and storage serves as a continuum from the cloud and data center across the network infrastructure to the machines and things. The network itself will become the perfect infrastructure for edge computing for all industries.

Ensuring trust in data integrity and reliability Now that the IoT plays such a key role in the success of so many enterprises, securing data end-to-end has become a top requirement. While reliability and trust are key considerations in all IoT applications, they are of utmost importance in mission-critical applications such as the predictability of data delivery to robots.

I hope that this special IoT issue of Ericsson Technology Review provides you with valuable insights about the IoT-related opportunities available to your organization, along with ideas about how we can overcome the challenges ahead. If you would like to share a link to the whole magazine or to a specific article, you can find both PDF and HTML versions at https://www.ericsson.com/en/ericsson-technology-review.
Driving transformation

IN THE AUTOMOTIVE AND ROAD TRANSPORT ECOSYSTEM WITH 5G

Major mobile network operators around the world have started rolling out 5G cellular networks, with subscriber penetration expected to reach about 20 percent by 2024 [1]. One of the many benefits of these powerful multipurpose networks is their ability to provide reliable, secure and fit-for-purpose cellular connectivity in automotive and transport applications.

Once considered merely “nice to have,” connectivity is rapidly becoming a critical part of road transportation systems. Ericsson predicts that the number of connected cars in operation will rise to more than 500 million in 2025 [9].

Already today, vehicle original equipment manufacturers (OEMs) are increasingly focusing on delivering services in addition to selling vehicles as products. Software is now a critical component of vehicles, and OEMs are investing heavily in automation, architecture simplification and new drivetrain technologies such as electrification.

At the same time, traffic and road authorities are seeking new technology solutions to reduce carbon emissions, traffic congestion and casualties – solutions that are often dependent on vehicle functionality and the ability to provide various types of support for drivers and vehicles. Meeting these diverse needs requires software-defined and network-aware vehicles, combined with advanced network connectivity.

While it is true that many of today’s 2G-4G networks can provide sufficient connectivity for numerous Internet of Things (IoT) applications, the higher data rate, lower latency and improved capacity provided by 5G New Radio (NR) access make 5G systems the ideal choice to maximize the safety, efficiency and sustainability of road transportation.

The purpose of OEM advanced driver assistance systems (ADAS) is to increase road safety by focusing on the driver and driving behavior. They rely primarily on vehicle sensor information and are typically not collaborative across vehicle brands. ADAS services can also benefit from data provided by traffic authorities such as traffic light information. They are expected to evolve to support the driverless vehicles of the future.

Fleet management services are aimed at vehicle fleet owners such as logistics or car-sharing companies. The communication service is primarily used to monitor vehicle locations and the vehicle/driver status. When the fleet consists of driverless vehicles, the fleet management also includes communication support for operations monitoring and remote assistance, which can imply full remote driving.

The primary focus in the logistics and connected goods category is on the tracking of transported objects (commodities, merchandise goods, cargo...
Transforming Transportation with 5G

Examples of connected services trials

In addition to all the connected services already in commercial operation, there are many noteworthy advanced trials on 4G/5G cellular networks, including:

- C-ITS in Europe: https://5gcroco.eu/
- Multi-party information exchange for C-ITS: https://www.nordicway.net/
- Connected traffic light information and driver advice for C-ITS: https://www.talking-traffic.com/en
- ADAS: https://www.ericsson.com/eoneer
- AD-aware traffic control: https://www.driveforward.net/en/events/dom-ad-aware-traffic-control-0
- Tele-operated driving and HD mapping: https://5gcroco.eu/
- Service continuity at border crossings: https://www.ericsson.com/en/blog/2019/5-connected-vehicle-cross-border-service-coverage
- Connected logistics: https://cs.ericsson.net/#/use-cases

Connected road infrastructure services are operated by cities and road authorities to monitor the state of the traffic and control its flow, such as physical traffic guidance systems, parking management and dynamic traffic signs. Each service group contains multiple use cases, and requirements can be diverse within a group. The key connectivity requirements per segment are noted in Figure 1.

5G-enabled network for all services

Connected vehicles and road infrastructure are part of a broader IoT ecosystem that is continuously evolving. To ensure cost efficiency and future-proof support, mobile network operators (MNOs) aim to meet the connectivity demands of multiple industries and verticals, including the automotive and transport industry, using common physical network infrastructure, network features and spectrum resources.

Ericsson divides cellular connectivity for the IoT into four distinct segments: massive IoT, broadband IoT, critical IoT and industrial automation IoT [2].

The three first segments are relevant for automotive and transport services. The colored dots in Figure 1 indicate their relevance for each of the eight service groups, based on key connectivity performance indicators.

Massive IoT

Massive IoT connectivity targets low complexity, narrow bandwidth devices that infrequently send or receive small volumes of data. The devices can be in challenging radio conditions requiring coverage extension capabilities and may rely on battery power supply. Massive IoT is suitable for low data-rate use cases that can be supported with narrow bandwidth modems. These use cases can be found in logistics, telematics, fleet management and connecting parts of road infrastructure, for example:

- Connected road infrastructure services
- Multi-party information exchange for C-ITS
- Connected traffic light information and driver advice for C-ITS
- Service continuity at border crossings

Broadband IoT

Broadband IoT connectivity enables large volumes of data transfer, extreme data rates and low latencies for devices with significantly larger bandwidths than massive IoT devices. Broadband IoT connectivity is also capable of enhancing signal coverage per base station and extending device battery life if requirements on data rate and latency are not stringent. Broadband IoT is vital for the majority of the automotive use cases that require high data rates and low latency, such as infotainment, telematics, fleet management, sensor sharing, basic safety and ADAS.

Critical IoT

Critical IoT connectivity enables ultra-reliable and/or ultra-low latency communication. It aims to deliver messages with strictly bounded low latencies even in highly loaded cellular networks. Critical IoT can enable some very advanced services, such as remote driving of automated commercial vehicles on specific routes.

4G networks already support massive IoT (based on LTE Category M1 and Narrowband IoT access) and broadband IoT (based on LTE access). 5G networks will boost broadband IoT performance and enable critical IoT with the introduction of NR. With the evolution of cellular IoT in the 5G era, cellular networks would enable the full range of existing and emerging automotive applications.

Accelerating the adoption of 5G connectivity

When rolling out 5G networks, MNOs aim to balance investments, new revenues and competitiveness. Decisions about where and when to deploy 5G networks depend not only on commercial factors but also on spectrum availability in different regions. Accelerated adoption of 5G in the ecosystem, including the automotive and transport industry, requires:

- The ability of 5G NR deployments to deliver value from day one
- The ability to efficiently share spectrum resources between 5G NR and 4G LTE
- Operators’ ability to reuse 4G LTE radio base station equipment for 5G NR deployments as much as possible

One of the 5G fundamentals is tight interworking between 4G LTE and 5G NR radio access. This interworking allows 5G-capable devices to simultaneously access 4G LTE and 5G NR carriers. A 5G-capable modem can connect with NR (when in NR coverage) to experience a boost in performance and capacity while maintaining its 4G LTE connection. This approach ensures that 5G NR deployments can deliver value for automotive and transport services from day one.

Both wide-area 5G coverage and automotive sector requirements demand that 5G NR and 4G LTE are able to efficiently share spectrum resources. Lower carrier frequencies where 4G LTE is operational are ideal from a coverage perspective (due to better radio wave propagation characteristics) and very attractive for 5G NR deployments. However, 4G LTE will be required for many years to support legacy devices (such as vehicles with 4G...
modems). To address this, Ericsson has developed fully-dynamic spectrum sharing between NR and LTE on a millisecond level for optimized utilization of spectrum [4].

With respect to operators’ ability to reuse 4G LTE radio base station equipment for 5G NR deployments, the Ericsson Radio System can be fully reused on existing sites following a remote software upgrade, including baseband units, radios and antennas (when NR and LTE share a spectrum band) [6]. This important 5G functionality will facilitate market-driven deployments along most streets and roads. However, in some cases, public incentives can trigger faster road coverage deployment, for example by letting MNOs deploy networks using road authorities’ site assets, or regulating road coverage requirements in spectrum license auctions [5].

The relation between in-vehicle and wide-area connectivity

Figure 2 illustrates how cellular connectivity works for vehicles and roadside equipment. It visualizes vehicles as multipurpose devices in which several connectivity-dependent use cases are executed simultaneously. At the same time, each vehicle also contains an internal network that interconnects in-vehicle sensors, actuators and other devices, including driver and passenger smartphones.

A gateway function (traditionally implemented in the Telematics Control Unit) connects the vehicle-internal network(s) to the external network. Among other things, this gateway function protects the vehicle-internal devices against external misuse. Additional security and traffic separation solutions restrict access to sensitive in-vehicle devices from inside the vehicle as well.

Connectivity to the external network is realized by one or more modems, containing one or more subscriptions (provided by the OEM, for example) has generally been a trade-off between cost constraints and simple service usage. More recently, capacity and redundancy gains have also been taken into consideration.

In some cases, the fleet operator provides connectivity to the transported objects (passengers in this case, as illustrated in Figure 2. Alternatively, the vehicle’s OEM subscription can be used to provide passenger Wi-Fi.

Instead of using the vehicle-mounted connectivity support, infotainment and navigation are often provided by a smartphone with its own subscription that is carried into the vehicle. As future ITS and ADAS services evolve, they too will be available through smartphones, which will increase service penetration to older vehicles.

Achieving global consistency in automotive and transport connectivity

Vehicles all around the world need connectivity to communicate, and, like any other device, a vehicle needs an MNO subscription to access a cellular network. The stark contrast between the global nature of vehicles’ connectivity requirements and the local nature of MNOs presents significant challenges to meet the automotive and transport ecosystem’s connectivity needs, most notably in the areas of subscription provisioning, roaming, local breakout/distributed computing and cost separation/traffic prioritization.

Subscription provisioning

One of the challenges particular to the automotive and transport ecosystem is that the long life cycle of vehicles and their varying roaming needs over time may make it necessary for a vehicle owner and/or OEM to change the subscription multiple times. Since the physical SIM cards that contain the subscription credentials are not easily accessible in vehicles, it is problematic to have to change them.

Embedded SIM (eSIM) technology overcomes this challenge by enabling remote provisioning of MNO subscriptions. An eSIM unit can be soldered into the cellular device which stores the MNO-specific network access credentials (the subscription) as a SIM card profile. The subscriptions can then be changed remotely over-the-air without physically touching the vehicle. To simplify the usage of this technology, the GSMA has developed an eSIM profile specification [6].

Roaming

It is common today for a vehicle to be produced in one country, sold in another, owned in a third, and driven across borders to numerous additional countries or regions, with high requirements on data throughput and latency independent of location. In light of this, roaming is frequently the default operating model for a connected vehicle. Today’s roaming solution, however, is single-human-user-centric – designed to support users traveling outside the coverage of their home mobile networks. It is not designed for connected vehicles on a global scale. As a result, it has a number of limitations in automotive and transport applications.

Terms and abbreviations

Firstly, since roaming fees are only partially regulated, they depend to a large extent on bilateral agreements between two MNOs. As a result, the fees can vary, which can make it difficult to predict the cost for the use of connectivity in certain cases.

Secondly, it has traditionally been the case that only basic connectivity and communication is enabled while roaming, which means that some more advanced service and capacity requirements may not be met when a vehicle connects outside its home network. Roaming agreements between MNOs typically put limitations on how the connectivity can be used, and the visited MNO can disconnect the device if it is not in line with the agreement.

Thirdly, the currently deployed roaming architecture is designed to route traffic to the home network first, which increases latency. This is problematic in automotive use cases that require low-latency or high data throughput. In these cases, fast access to local data centers is required.

Fourthly, the fact that a mobile device loses connectivity for some time (up to about 120 seconds) when being handed over from one MNO to another is a serious issue for many use cases. The reason for the delay is that the mobile device needs to first scan for a suitable network provider and then register itself in the new mobile network. This applies at both international country borders and national coverage borders.

In Ericsson’s view, there are two complementary paths to overcoming roaming challenges in the automotive and transport industry:

1. Enhancing the existing roaming solution through the creation of an alliance of MNOs.
2. Avoiding roaming altogether by using local subscriptions and eSIM technology for provisioning in each local network.

The enhancement of the existing roaming solution would ensure that operators treat roaming users the same way they treat local users — that is, there would be no additional costs and roaming users would have consistent capabilities and support for low-latency and high-volume services. This could be achieved through the creation of an alliance of MNOs that enables the 3GPP roaming architecture “Local breakout in the visited network” [7] which would provide direct, fast access to local data centers.

Alternatively, it is possible to avoid the roaming model altogether by using local subscriptions and eSIM technology for provisioning in each local network. This approach ensures access to all the functionality and capacity provided by the local network, including direct access to local data centers. Some form of coordination of service subscription and cost models between the involved operators would be required to achieve consistency.

Both of these alternatives involve the use of different core networks, which means that there can be variances in service experience and SLA support between operators. This due to the fact that the core network is the entity that controls most of the service-specific parameters and manages the technical SLAs. Full harmonization of services and SLA control requires an alignment of core network functions.

Regardless of which option is chosen, a fast inter-MNO mobility solution is also required to reduce the time for network swap. A combination of network features in a recent trial has been shown to provide fast inter-network service continuity [8].

Local breakout and distributed computing

Several emerging automotive services require vehicles to be connected to the cloud and networks to facilitate the transfer of a large amount of data between vehicles and the cloud. Some of the services may be more time-critical, while other services allow time phasing to a different time slot or another access network. The AECG (Automotive Edge Computing Consortium) addresses the technical realization of such use cases by designing a topology-aware distributed cloud solution on a global scale, to better accommodate the needs of the automotive industry [9, 10].

Cost separation and traffic prioritization

In the automotive and transport ecosystem there is a need to separate the costs for cellular connectivity for different services in the vehicle targeted at different stakeholders — such as the owner of the vehicle or vehicle fleet, the driver/user of the vehicle, the vehicle OEM and traffic/road authorities. For example, one may want entertainment-related costs to be charged to the passengers, while the OEM covers the cost for vehicle-centric sensor data uploads. Support for data traffic prioritization is also essential, particularly at times of high network usage, such as when vehicles are stuck in a traffic jam.

There are two main alternatives for cost separation: multiple subscriptions or multiple connections using a single subscription (also known as dedicated bearers). A vehicle can have multiple subscriptions to connect with one or multiple mobile networks for multiple services. Multiple subscriptions can be active simultaneously when multiple services are needed concurrently. The vehicle can be either natively equipped to support multiple simultaneous active subscriptions through the use of a Dual-Sim Dual Active (DSDA) device, for example, or additional communication devices can be added to the vehicle later (each with its own subscription). These devices could be permanently mounted or they could be temporary devices such as the driver’s smartphone.

A dedicated bearer framework allows separation of traffic flows for differentiated QoS handling and charging using a single subscription and single modem. 3GPP systems support traffic differentiation based on Policy and Charging Control rules. The term ‘policy’ refers to various traffic-handling policies, such as different QoS for different traffic flows.

In 5G networks, the separated data streams are handled as dedicated bearers, which are known as dedicated bearers. The cellular network identifies the traffic flows based on traffic flow templates — typically a 5-tuple in the form of IP addresses, protocol and transport layer ports. The consumed data volumes can be accounted separately for each bearer. Within 5G networks, the separated data streams are handled as different QoS flows.

Figure 3 depicts an end-to-end architecture using dedicated bearers for traffic separation, considering distributed computing with edge clouds.

Figure 3 Usage of dedicated bearers for traffic separation within one vehicle OEM cellular subscription

---

different stakeholders — such as the owner of the vehicle or vehicle fleet, the driver/user of the vehicle, the vehicle OEM and traffic/road authorities. For example, one may want entertainment-related costs to be charged to the passengers, while the OEM covers the cost for vehicle-centric sensor data uploads. Support for data traffic prioritization is also essential, particularly at times of high network usage, such as when vehicles are stuck in a traffic jam.

There are two main alternatives for cost separation: multiple subscriptions or multiple connections using a single subscription (also known as dedicated bearers). A vehicle can have multiple subscriptions to connect with one or multiple mobile networks for multiple services. Multiple subscriptions can be active simultaneously when multiple services are needed concurrently. The vehicle can be either natively equipped to support multiple simultaneous active subscriptions through the use of a Dual-Sim Dual Active (DSDA) device, for example, or additional communication devices can be added to the vehicle later (each with its own subscription). These devices could be permanently mounted or they could be temporary devices such as the driver’s smartphone.

A dedicated bearer framework allows separation of traffic flows for differentiated QoS handling and charging using a single subscription and single modem. 3GPP systems support traffic differentiation based on Policy and Charging Control rules. The term ‘policy’ refers to various traffic-handling policies, such as different QoS for different traffic flows.

In 5G networks, the separated data streams are handled as dedicated bearers, which are known as dedicated bearers. The cellular network identifies the traffic flows based on traffic flow templates — typically a 5-tuple in the form of IP addresses, protocol and transport layer ports. The consumed data volumes can be accounted separately for each bearer. Within 5G networks, the separated data streams are handled as different QoS flows.

Figure 3 depicts an end-to-end architecture using dedicated bearers for traffic separation, considering distributed computing with edge clouds.
The edge cloud servers are shielding the central cloud servers by executing the heavy lifting workloads. The central servers coordinate the heavy workload functions and distribute the load across different edge cloud servers and sites. The central cloud servers steer the vehicle’s connection to an appropriate edge, which supports the service and has sufficient computational capacity. The policy rules for traffic separation can be provided either statically within the policy system of the network or dynamically using the Service Capability Exposure Function (SCEF), which is provided by the mobile network toward the OEM. The SCEF is evolving into the Network Exposure Function in 5G.

In upcoming 5G networks, the network slicing concept [11] may be used for service and cost separation. 5G provides a cost-efficient and feature-rich foundation for a horizontal multiservice network.

Conclusion
The connectivity needs of the automotive and transport ecosystem are diverse and complex, requiring a common network solution rather than a single-segment silo approach. The ongoing rollout of 5G provides a cost-efficient and feature-rich foundation for a horizontal multiservice network. 5G networks (including 2G–4G accesses) offer excellent capabilities that make them the ideal choice to meet the wide variety of needs in the automotive and transport ecosystem. The time-to-market for 5G networks and services is faster than earlier generations, and the connectivity capabilities of 5G networks (including 2G–4G accesses) offer excellent capabilities that make them the ideal choice to meet the wide variety of needs in the automotive and transport ecosystem. The connectivity needs of the automotive and transport ecosystem are diverse and complex, requiring a common network solution rather than a single-segment silo approach. The ongoing rollout of 5G provides a cost-efficient and feature-rich foundation for a horizontal multiservice network.

Further reading
1. Learn more about evolving cellular IoT for industry digitalization at: https://www.ericsson.com/en/networks/insights-and-reports/cellular-iot-evolution-for-industry-digitalization
11. Ericsson, Network Slicing, available at: https://www.ericsson.com/en/digital-services/trending/network-slicing/5t_cat=CkQxCAUgLQg8E71wAzJ4SU-MQ5epjzDOVRH4J3L0CPSrXBi8DRn-SIMyXyD_ tzmLwCFCwQAd_D_BmE

References
3. 5GAA white paper, C-ITS Vehicle to Infrastructure Services: how C-V2X technology completely changes the cost equation for road operators, available at: https://5gaa.org/wp-content/uploads/2019/01/5GAA-BMAC-White-Paper_final2.pdf
11. Ericsson, Network Slicing, available at: https://www.ericsson.com/en/digital-services/trending/network-slicing/5t_cat=CkQxCAUgLQg8E71wAzJ4SU-MQ5epjzDOVRH4J3L0CPSrXBi8DRn-SIMyXyD_tzmLwCFCwQAd_D_BmE
Thorsten Lohmar
joined Ericsson in Germany in 1998 and has worked primarily within Ericsson Research. He specializes in mobile network architectures, focusing on end-to-end procedures and protocols. He is currently working as an expert for media delivery and acts as the Ericsson delegate in different standards groups and industry forums. Recently, he has focused on industry verticals such as automotive and transport. Lohmar holds a Ph.D. in electrical engineering from RWTH Aachen University, Germany.

Ali Zaidi
is a strategic product manager for cellular IoT at Ericsson. He received an M.Sc. and a Ph.D. in telecommunications from KTH Royal Institute of Technology, Stockholm, Sweden, in 2008 and 2013, respectively. Since 2014, he has been working with technology and business development of 4G and 5G radio access at Ericsson. He has co-authored more than 50 peer-reviewed research publications and two books, filed over 20 patents and made several 3GPP and 5G-PPP contributions. He is currently responsible for LTE for machines, NR ultra-reliable low-latency communication, NR Industrial IoT, vehicle-to-everything communication and tactical industrial networks.

Håkan Olsson
has 25 years’ experience of the mobile industry, and its RAN aspects in particular. He joined Ericsson in 1994 and has served the company and the industry in a variety of capacities, mostly dealing with strategic technology development and evolution of 2G to 5G. He is currently head of the System Concept program in Development Unit Networks. He is also codirector of the Integrated Transport Research Lab in Stockholm, founded together with the KTH Royal Institute of Technology and the Swedish vehicle manufacturer Scania. Olsson holds an M.Sc. in physics engineering from Uppsala University, Sweden.

Christer Boberg
serves as a director at Ericsson’s CTO office, responsible for IoT technology strategies aimed at solving networking challenges for the industry on a global scale. He initially joined Ericsson in 1983 and during his career he has focused on software and system design as a developer, architect and technical expert, both within and outside Ericsson. In recent years, Boberg’s work has centered on the IoT and cloud technologies with a special focus on the automotive industry. As part of this work, he founded and drives the Automotive Edge Computing Consortium (AECC) together with industry leading companies.

The authors would like to thank Tomas Nylander, Maciej Muehleisen, Stefano Sorrentino, Michael Meyer, Marie Hogian, Mikael Klem, Anders Fagerholt, Tim Wouda, Fredrik Alriksson, Robert Skog and Henrik Sahlin for their contributions to this article.
Industrial automation is one of the industry verticals that can benefit substantially from 5G, including, for example, increased flexibility, the reduction of cables and support of new use cases [1]. At the same time, factory automation is going through a transformation due to the fourth industrial revolution (also known as Industry 4.0), and this requires converged networks that support various types of traffic in a single network infrastructure.

As it stands, IEEE (Institute of Electrical and Electronics Engineers) 802.1 Time-Sensitive Networking (TSN) is becoming the standard Ethernet-based technology for converged networks of Industry 4.0. It is possible for 5G and TSN to coexist in a factory deployment and address their primary requirements, such as 5G for flexibility and TSN for extremely low latency. Beyond that, 5G and TSN can be integrated to provide solutions to the aforementioned demands of ubiquitous and seamless connectivity with the deterministic QoS required by control applications end to end. Ultimately, integrating these key technologies provides what is needed for smart factories.

5G: adding ultra-reliable low-latency communication
5G has been designed to address enhanced mobile broadband services for consumer devices such as smartphones or tablets, but it has also been tailored for Internet of Things (IoT) communication and connected cyber-physical systems. To this end, two requirement categories have been defined: massive machine-type communication for a large number of connected devices/sensors, and ultra-reliable low-latency communication (URLLC) for connected control systems and critical communication [1][2]. It is the capabilities of URLLC that make 5G a suitable candidate for wireless deterministic and time-sensitive communication. This is essential for industrial automation, as it can enable the creation of real-time interactive systems, and also for the integration with TSN.

Several features have been introduced to 5G in phase 1 (3GPP Release 15) and phase 2 (3GPP Release 16, to be finalized by March 2020) that reduce the one-way latency and enable the transmission of messages over the radio interface with reliability of up to 99.999 percent, achievable in a controlled environment such as a factory.

5G RAN features
5G RAN [3] with its New Radio (NR) interface includes several functionalities to achieve low latency for selected data flows. NR enables shorter slots in a radio subframe, which benefits low-latency applications. NR also introduces mini-slots, where prioritized transmissions can be started without waiting for slot boundaries, further reducing latency. As part of giving priority and faster radio access to URLLC traffic, NR introduces preemption – where URLLC data transmission can preempt ongoing non-URLLC transmissions. Additionally, NR applies very fast processing, enabling retransmissions even within short latency bounds.

Smart factories are being developed as part of the fourth industrial revolution. They require ubiquitous connectivity among and from the devices to the cloud through a fully converged network, supporting various types of traffic in a single network infrastructure, which also includes mobile network segments integrated into the network.
5G defines extra-robust transmission modes for increased reliability for both data and control radio channels. Reliability is further improved by various techniques, such as multi-antenna transmission, the use of multiple carriers and packet duplication over independent radio links.

Time synchronization is embedded into the 5G cellular radio systems as an essential part of their operation, which has already been common practice for earlier cellular network generations. The radio network components themselves are also time synchronized, for instance, through the precision time protocol telecom profile [4]. This is a good basis to provide synchronization for time-critical applications.

Figure 1 illustrates URLLC features. It shows that 5G uses time synchronization for its own operations, as well as the multiple antennas and radio channels that provide reliability. 5G brings in redefined schemes for low latency and resource management, which can be combined to provide ultra-reliability and low latency.

Besides the 5G RAN features, the 5G system (5GS) also provides solutions in the core network (CN) for Ethernet networking and URLLC. The 5G CN supports native Ethernet protocol data unit (PDU) sessions. 5G assists the establishment of redundant user plane paths through the 5GS, including RAN, the CN and the transport network. The 5GS also allows for a redundant user plane separately between the RAN and CN nodes, as well as between the UE and the RAN nodes.

**Time-Sensitive Networking for converged networks**

TSN provides guaranteed data delivery in a guaranteed time window; that is, bounded low latency, low-delay variation and extremely low data loss, as illustrated in Figure 2. TSN supports various kinds of applications having different QoS requirements: from time- and/or mission-critical data traffic, for example, closed-loop control, to best-effort traffic over a single standard Ethernet network infrastructure, in other words, through a converged network. As a result, TSN is an enabler of Industry 4.0 by providing flexible data access and full connectivity for a smart factory.

**Time-Sensitive Networking standards**

TSN is a set of open standards specified by IEEE 802.1 [5]. TSN standards are primarily for IEEE Std 802.3 Ethernet, which means they utilize all the benefits of standard Ethernet, such as flexibility, ubiquity and cost savings.

TSN standards can be seen as a toolbox that includes several valuable tools, which can be categorized into four groups: traffic shaping, resource management, time synchronization and reliability, as shown in Figure 2. Here, we focus only on the TSN tools that are strong candidates for early TSN deployments in industrial automation.

TSN guarantees the worst-case latency for critical data by various queueing and shaping techniques and by reserving resources for critical traffic. The Scheduled Traffic standard (802.1Qv) provides time-based traffic shaping. Ethernet frame preemption (802.1Qv-p), which can suspend the transmission of a non-critical Ethernet frame, is also beneficial to decrease latency and latency variation of critical traffic.

Resource management basics are defined by the TSN configuration models (802.1Qc). Centralized Network Configuration (CNC) can be applied to the network devices (bridges), whereas, Centralized User Configuration (CUC) can be applied to user devices (end stations). The fully centralized configuration model follows a software-defined networking (SDN) approach; in other words, the CNC and CUC provide the control plane instead of distributed protocols. In contrast, distributed control protocols are applied in the fully distributed model, where there is no CNC or CUC.

High availability, as a result of ultra-reliability, is provided by Frame Replication and Elimination for Reliability (FRER) (802.1CB) for data flows through a per-packet-level reliability mechanism. This provides reliability by transmitting multiple copies of the same data packets over disjoint paths in the network. Per-Stream Filtering and Policing (802.1Qp) improves reliability by protecting against bandwidth violation, malfunctioning and malicious behavior.

The TSN tool for time synchronization is the
generalized Precision Time Protocol (gPTP) (802.1AS), which is a profile of the Precision Time Protocol standard (IEEE 1588). The gPTP provides reliable time synchronization, which can be used by other TSN tools, such as Scheduled Traffic (802.1Qbv).

It is important to note that TSN standards are built upon the base IEEE 802.1 bridging standards, some of which have to be supported in TSN deployments as well – including industrial automation.

A special set of TSN standards are the TSN profiles because a profile selects TSN tools and describes their use for a particular use case or vertical.

**Time-Sensitive Networking for industrial automation**


The IEC/IEEE 60802 profile will specify multiple classes of devices. There will be at least two classes of devices for both device types – bridges and end stations. One class is feature rich (currently called Class A), and the other class is constrained (currently called Class B), meaning that it supports a smaller set of features. Bridges and end stations belonging to the same class have the same mandatory and optional TSN capabilities.

The Link Layer Discovery Protocol (LLDP) (802.1AB) is mandatory for all device types and classes for the discovery of the network topology and neighbor information.

Time synchronization is also mandatory for all device types and classes. The current target is to support a minimum of three time domains for Class A and a minimum of two time domains for Class B. Class A devices must support a wide range of TSN functions (such as Scheduled Traffic, Frame Preemption, Per-Stream Filtering and Policing, FRER and TSN configuration), which are optional for Class B devices.

**Integrated 5G and Time-Sensitive Networking**

5G URLLC capabilities provide a good match to TSN features (as illustrated in Figures 1 and 2). The two key technologies can be combined and integrated to provide deterministic connectivity end to end, such as between input/output (I/O) devices and their controller potentially residing in an edge cloud for industrial automation. The integration includes support for both the necessary base-bridging features and the TSN add-ons.

**Terms and abbreviations**

- **5GS** – 5G System
- **5QI** – 5G QoS Indicator
- **AF** – Application Function
- **CNC** – Centralized Network Configuration
- **CUC** – Centralized User Configuration
- **FRER** – Frame Replication and Elimination for Reliability
- **gNB** – Next generation Node B
- **gPTP** – Generalized Precision Time Protocol
- **I/O** – Input/Output
- **IEC** – International Electrotechnical Commission
- **IEEE** – Institute of Electrical and Electronics Engineers
- **IOT** – Internet of Things
- **LLDP** – Link Layer Discovery Protocol
- **NR** – New Radio
- **OFDM** – Orthogonal Frequency Division Multiplexing
- **OPC** – Open Platform Communications
- **PCF** – Policy Control Function
- **PDU** – Protocol Data Unit
- **SDN** – Software-Defined Networking
- **TT** – Time-Sensitive Networking
- **UPF** – User Plane Function
- **URLLC** – Ultra-Reliable Low Latency Communication

**Figure 3** illustrates the 5G-TSN integration, including each TSN component shown in Figure 2. It shows the fully centralized configuration model, which is the only configuration model supported in 5G phase 2 (3GPP Release 16).

The 5GS appears from the rest of the network as a set of TSN bridges – one virtual bridge per User Plane Function (UPF) as shown in the figure. The 5GS includes TSN Translator (TT) functionality for the adaptation of the 5GS to the TSN domain, both for the user plane and the control plane, hiding the 5GS internal procedures from the TSN bridged network.

The 5GS provides TSN bridge ingress and egress port operations through the TT functionality. For instance, the TTs support hold and forward functionality for de-jittering. The figure illustrates functionalities using an example of two user equipments (UEs) with two PDU sessions supporting two correlated TSN streams for redundancy. But a deployment may only include one physical UE with two PDU sessions using dual-connectivity in RAN. The figure illustrates the case when the 5GS connects an end station to a bridged network; however, the 5GS may also interconnect bridges.

The support for base bridging features described here is applicable whether the 5G virtual bridges are Class A or Class B capable. The 5GS has to support the LLDP features needed for the control and management of an industrial network, such as for the discovery of the topology and the features of the 5G virtual bridges. The 5GS also needs to adapt to the loop prevention method applied in the bridged network, which may be fully SDN controlled without any distributed protocol other than LLDP.

**5G supporting Time-Sensitive Networking**

Ultra-reliability can be provided end to end by the application of FRER over both the TSN and 5G domains. This requires disjoint paths between the FRER end points over both domains, as illustrated in Figure 3.
A 5G UE can be configured to establish two PDU sessions that are redundant in the user plane over the 5G network [3]. The SGPP mechanism involves the appropriate selection of CN and RAN nodes (UPFs and 5G base stations (gNBs)), so that the user plane paths of the two PDU sessions are disjoint. The RAN can provide the disjoint user plane paths based on the use of the dual-connectivity feature, where a single UE can send and receive data over the air interface through two RAN nodes.

The additional redundancy – including UE redundancy – is possible for devices that are equipped with multiple UEs. The FRER end points are outside of the SGS, which means that 5G does not need to specify FRER functionality itself. Also, the logical architecture does not limit the implementation options, which include the same physical device implementing end station and UE. Requirements of a TSN stream can be fulfilled only when resource management allocates the network resources for each hop along the whole path. In line with TSN configuration (802.1Qcc), this is achieved through interactions between the SGS and CNC. The green clocks in Figure 3 illustrate the characteristics of the 5G virtual bridge, and for the SGS to establish connections with specific parameters based on the information received from the CNC. Bounded latency requires deterministic delay from a TSN and 5G domain. Note that 5G can provide a direct wireless hop between components that would otherwise be connected via several hops in a traditional industrial wireless network. Ultimately, the most important factor is that 5G can provide deterministic latency, which the CNC can discover together with TSN features supported by the SGS.

The SGS establishes time-controlled packet transmission in line with Scheduled Traffic (802.1Qbb). For the 5G control plane, the TT in the application function (AF) of the SGS receives the transmission time information of the TSN traffic classes from the CNC. In the 5G user plane, the TT at the UE and the TT at the UPF can regulate the time-based packet transmission accordingly. TT internal details are not specified by SGPP and are left for implementation. For example, a play-out (de-jitter) buffer per traffic class is a possible solution. The different TSN traffic classes are mapped to different 5G QoS Indicators (SIs) in the AF and the Policy Control Function (PCF) as part of the QoS alignment between the two domains, and the different SIs are treated according to their QoS requirements.

5G-TSN in Industry 4.0

The logical architecture allows the SGS to connect TSN bridges for their internal operations. Bridges also require time reference if they use a TSN feature that is based on time, such as Scheduled Traffic (802.1Qbb). The green clocks in Figure 3 illustrate a case when both bridges and end stations are time synchronized.

As gPTP is the default time synchronization solution for TSN-based industrial automation, the SGS needs to interwork with the gPTP of the connected TSN network. The SGS may act as a virtual gPTP time-aware system and support the forwarding of gPTP time synchronization information between end stations and bridges through the 5G user plane TTs. These account for the residence time of the SGS in the time synchronization procedure. One special option is the SGS’s clock acts as a grandmaster and provides the time reference not only within the SGS, but also to the rest of the devices in the deployment, including connected TSN bridges and end stations.

Overall, 5G standardization has addressed the key aspects needed for 5G-TSN integration.
5G-TSN INTEGRATION FOR INDUSTRIAL AUTOMATION

The authors would like to thank the following people for their contributions to this article: Shabnam Sultana, Anna Larmo, Kun Wang, Torsten Dudda, Juan-Antonio Ibanez, Marilet De Andrade Jardim, Stefano Ruffini.

János Farkas

◆ is a principal researcher in the area of deterministic networking at Ericsson Research. He is the chair of the IEEE 802.1 Time-Sensitive Networking Task Group, editor and contributor of multiple IEEE 802.1 standards. He is cochair of the IETF Deterministic Networking Working Group and coauthor of multiple drafts. He joined Ericsson Research in 1997. He holds a Ph.D. and M.Sc. in electrical engineering from the Budapest University of Technology and Economics in Hungary.

Balázs Varga

◆ is an expert in multiservice networking at Ericsson Research. He is currently working on 5G-related technologies to integrate mobile, IP, multi-protocol label switching, Ethernet and industrial networks. He is active in related standardizations: 3GPP (RAN2, SA2), MEF Forum (IP Services), IETF (DetNet) and IEEE (TSN). Before joining Ericsson in 2010, he directed and coordinated activities of an R&D group responsible for the enhancement of a broadband service portfolio and related technologies at Telekom. He holds a Ph.D. and M.Sc. in electrical engineering from the Budapest University of Technology and Economics.

György Miklós

◆ is a master researcher at Ericsson Research. Since joining Ericsson in 1998, he has worked on research topics including wireless LAN, ad hoc networking and mobile core network evolution. He has served as an Ericsson delegate in 3GPP for many years for 4G standardization. His current research interests include 5G industrial applications and redundancy support in mobile networks. He holds a Ph.D. and M.Sc. in informatics from the Budapest University of Technology and Economics.

Joachim Sachs

◆ is a principal researcher at Ericsson Corporate Research in Stockholm, Sweden, where he coordinates research activities on 5G for industrial Internet of Things solutions and cross-industry research collaborations. He joined Ericsson in 1997 and has contributed to the standardization of 3G, 4G and 5G networks. He holds an Engineering Doctorate from the Technical University of Berlin, Germany, and was a visiting scholar at Stanford University in the US in 2009.

The authors would like to thank the following people for their contributions to this article: Shabnam Sultana, Anna Larmo, Kun Wang, Torsten Dudda, Juan-Antonio Ibanez, Marilet De Andrade Jardim, Stefano Ruffini.
As the diversity of IoT services and the number of connected devices continue to increase, the threats to IoT systems are changing and growing even faster. To cope with these threats, the ICT industry needs a comprehensive IoT security and identity management solution that is able to manage and orchestrate the IoT components horizontally (from device to service and service user) and vertically (from hardware to application). In addition to this, the ability to address both security and identity from the IoT device all the way across the complete service life cycle will also be essential.

**IoT actors and trust**

IoT systems support new business models that involve new actors in conjunction with traditional telecommunication services. Aside from consumers and mobile network operators, enterprises, verticals, partnerships, infrastructure, and services play increasingly vital roles. All of these actors affect trust.

Figure 1 illustrates the main and supporting IoT actors and their trust relationships. The three main actors in an IoT solution are the IoT service user, the IoT service provider, and the devices that enable the provision of the IoT service. The supporting actors are the IoT platform service provider, whose role is to provide the IoT platform for the IoT service provider, and the connectivity service provider, whose role is to provide connectivity for the IoT devices and service.

The trustworthiness of services and service use depends on how the actors govern identities and data, security and privacy, and the degree to which they comply with the agreed policies and regulations. The combination of the security and identity functions is important for defining the trust level. For example, hardware-based trust does not help if the application does not make use of it. A fully trusted application does not help if the communication cannot be trusted. An E2E approach is therefore essential to ensure trust among all actors across the system.

**E2E IoT security architecture**

The purpose of an E2E IoT security architecture is to ensure the security and privacy of IoT services, protect the IoT system itself and prevent IoT devices from becoming a source of attacks – a Distributed Denial of Service (DDoS) attack, for example – against other systems.

Figure 2 illustrates Ericsson’s view of how security can be managed and deployed in an E2E manner throughout IoT domains to monitor the trustworthiness of services and service use.
and protect system resources and assets. The architecture consists of an E2E security and identity management layer, domain (device, gateway, access, platform and application) specific management layers, and data and identity functions in each domain component. An IoT system spans from the device via different network interfaces to the cloud that hosts the platform and applications that provide services that are consumed by IoT service users. Each element of the chain must be considered when designing an E2E approach to security and identity in the IoT. This approach leverages advanced security analytics and machine learning to provide threat, risk and fraud management at both E2E and domain management layers. To meet industry security and privacy standards, an E2E security management solution must also be in charge of overall security and privacy policies and compliance and be able to coordinate across a multitude of domain management systems through the establishment of cross-domain identities and relevant policies.

Domain management of security and identity functions within domains ensures that security and identities are properly managed, configured and monitored within the domain according to policies, regulations, and agreements. Vulnerability and security baseline management also occurs at the domain management layer based on E2E level policies. According to this approach, the IoT service provider is responsible for managing IoT service security and identities E2E, whereas domain-level management can be delegated to the IoT platform service provider and connectivity service provider. Figure 2 shows how the IoT domains are managed both horizontally and vertically. Horizontal (cross-domain) security is required at two levels: connectivity and application. Depending on connectivity type, security controls such as mutual authentication and encryption of data in transit are provided at the connectivity level. On top of connectivity, security is provided at the application level from device to cloud, based on identification and access management functions and application security policies. Application level security can be independent of or dependent on (federated with) the connectivity level security. Vertical security from hardware to application can be used in every domain to provide hardware-based root of trust, ensuring the integrity of the domain. The domains are built on trusted hardware and software. When required by the industry and the use case, trust is anchored to hardware. The domains include security and privacy functions to handle identity and access management, data protection and right to privacy, network security, logging, key and certificate management, and platform infrastructure security (including virtualization security and hardware-based root of trust). For critical IoT services, the level of security functions must be set high in accordance with the risk management results and service provider security policies. For less critical IoT services, a lower level may be sufficient.

Security policy and compliance management Business-optimal and trust-centric IoT security is dependent on continuous risk management that balances criticality, cost, usability and effectiveness. Risk management, along with security policy and privacy controls, maintain them at a desired level. Security policies. Only in this environment can risk management results and service provider security policies. For less critical IoT services, the level of security functions must be set high in accordance with the risk management results and service provider security policies. For less critical IoT services, a lower level may be sufficient.

Security policy and compliance management Business-optimal and trust-centric IoT security is dependent on continuous risk management that balances criticality, cost, usability and effectiveness. Risk management, along with security policy and privacy controls, maintain them at a desired level. Security policies. Only in this environment can risk management results and service provider security policies. For less critical IoT services, the level of security functions must be set high in accordance with the risk management results and service provider security policies. For less critical IoT services, a lower level may be sufficient.

Domain level security management requires an accurate asset inventory including all the assets that must be protected in the managed domain, such as authorized IoT devices and software. Automation of asset discovery and continuous monitoring is essential to keep the asset inventory updated. The vulnerability information is also correlated with the asset inventory to monitor and remediate the vulnerabilities of protected assets. Rapid detection of attacks and anomalies is critical. Security monitoring and analytics functionalities must have the ability to analyze logs, events and data from IoT domain components combined with external data about threats and vulnerabilities. Machine learning technology makes it possible to learn from and make predictions based on data. Coupling a machine learning analytics engine with central threat intelligence improves the detection of zero day attacks and reduces the response time for known threats. On top of a monitoring and analytics engine, solutions relating to vulnerability, threat, fraud and risk management, along with security policy and orchestration components, are also required to automate security controls and maintain them at desired levels in a changing threat landscape. Combining the information feeds for vulnerability, threat and fraud management results in timely......


**E2E security and identity management**

- **IoT device**
- **IoT gateway**
- **Access and network connectivity**
- **IoT service user**
- **IoT app, platform and cloud**

**Threat intelligence**

**Legend:**

- **Security and identity management**
- **Security and identity functions**
- **Trust anchoring**

**Figure 3** E2E approach to security and identity

and accurate information for evaluating potential risks and helps to direct efforts in protecting the most exposed critical assets. A high degree of automation is necessary to ensure a swift response to any identified threats and anomalies. Since not all security breaches and attacks can be prevented, it is crucial to have an efficient security incident management process that ensures rapid response and recovery. Real-time insights and audit trails from tools such as security monitoring, analytics and log management help to find the root cause of an incident. The same information can be also used as the evidence in digital forensic investigations.

**Identity management**

The main purpose of identity management is to manage the lifecycle of identities and provide identification, authentication and access control services for identities. There are various identities that serve different purposes in the IoT approach, but the main ones are for device and user identification. The others are used for management of devices, functions and services. Identifiers and keys are also used to sign data, including software and firmware. These different device identities are needed to identify the devices for connectivity within the access and network domains, and to identify device applications in the IoT platform and cloud domain.

The level of trust in the device identity depends on the strength of authentication both at the connectivity (for example, 3GPP, Wi-Fi and fixed) and application layers. For device identity to be trusted, strong authentication and follow-up of the device integrity—with the help of hardware-based root of trust in the device, for example—would be needed.

A device will have different identifiers depending on where it is in its life cycle. Lifecycle management of device identities is part of the security management layer. More than one security management domain is involved when provisioning identities. Connectivity and IoT service provider could be different players where each player takes care of its own identity life cycle management.

When a device is manufactured, the vendor will give it an identifier that could have different trust levels. Vendor credentials could be protected in hardware (preferred) or they could be nothing more than a serial number printed on the device. The device has to be authenticated by the IoT system, and newly given identifiers and credentials (bootstrap process) will be used for connectivity and application access. Identifiers and credentials can be changed during the device life cycle depending on different triggers such as expiration of credentials, change of service provider and so on. Connectivity identities are dependent on the connectivity type and have different life cycle management processes. For example, 3GPP access is based on SIM identities (IMSI and ARQ) and SIMs are either physically removable ones or SIMs (i.e. eUICC) that can be remotely provisioned.

The user identities are needed to identify the users of the services within the applications and cloud domain. There may be several different ways to verify (authenticate) the user identities such as single- or multi-factor authentication, federated authentication, or authentication tokens. Each of these provides a certain level of authentication strength.

Due to layered security management architecture and the involvement of several actors (including industries) in the IoT, any identity and access management solution must be able to cooperate with and adapt to external identity and access management systems. On top of identification and authentication, there must also be access control for users so that only the permitted services are authorized.

**Threat intelligence**

Threat intelligence is built and shared in communities. Therefore, a centralized threat intelligence solution must be able to interface with different threat intelligence sources to learn about existing and new threats. Consolidation and correlation of security audit feeds from different domains are necessary to provide a clear view of threat insights across all IoT domains.

Automation and machine learning can be used to great advantage in threat intelligence, to create and share indicators of compromise that are actionable, timely, accurate and relevant to support strategic decision-making and to understand business risks in detail. Targeted threat intelligence feeds are a great way to generate customer-specific threat intelligence.

**Two IoT use cases**

Two concrete examples of how an E2E security management solution can help address IoT challenges are provided below.

**DDoS detection and prevention**

In October 2016, the Mirai botnet exploited a vulnerability in IoT devices to launch a DDoS attack against a critical DNS server that disrupted a number of the internet’s biggest websites, including PayPal, Spotify and Twitter.

Mirai was designed to exploit the security weaknesses of many IoT devices. It continuously scans for IoT devices that are accessible over the internet and are protected by factory default or hardcoded user names and passwords. When it finds them, Mirai infects the devices with malware that forces them to report to a central control server, turning them into bots that can be used in DDoS attacks.

Strong detection and prevention mechanisms are needed against DDoS attacks that attempt...
to saturate the network by exhausting the bandwidth capacity of the attacked site, the server resources or service availability. In our view, an optimal outbound DDoS (botnet) detection and mitigation solution includes remote attestation to verify device trustworthiness and detect malware, monitoring of outbound traffic, anomaly detection, infected entities isolation or blocking and setting of traffic limit policies. Optimal inbound DDoS detection and mitigation includes monitoring of inbound traffic, anomaly detection, setting of traffic limit policies and redirecting malicious traffic to a botnet sinkhole.

The security management layer plays a critical role in detecting and mitigating DDoS attacks. In our framework, DDoS attacks are detected by the security monitoring and analytics functions through the observation of device and network behavior and identification of anomalies. Once an anomaly is detected, immediate mitigation actions can be triggered.

GDPR compliance

There is a legitimate expectation in society that IoT solutions will be designed with privacy in mind. This is becoming especially evident in certain jurisdictions for example, in the European Union with the new General Data Protection Regulation (GDPR) [8].

Data integrity, data confidentiality, accountability and privacy by design are all fundamental to the protection of sensitive personal data. Such data can be protected via appropriate privacy controls. These controls include personal data identification and classification, personal data management and fair data processing practices. When actual personal data might be exposed, additional privacy protective measures will be applied such as data encryption and data anonymization.

Another focus area in the IoT security domain is the privacy breach response. Dedicated privacy logging and audit trail functionality can be used to improve the ability to prevent, detect and respond to privacy breaches in a more prompt and flexible way. Such capabilities will be essential to respond to privacy breaches swiftly (within 72 hours, as prescribed by the GDPR).

Implementing a GDPR compliance tool in the security management layer makes it easier to meet GDPR requirements. To do its job right, it must be able to provide identification and classification of personal data, enforcement of data privacy policies according to the GDPR, demonstration of compliance to the GDPR, and detection, response and recovery from privacy incidents.

Conclusion

The IoT offers a wealth of new opportunities for service providers. Those who want to capitalize on them without taking undue risks need a security solution that provides continuous monitoring of threats, vulnerabilities, risks and compliance, along with automated remediation. Ericsson’s E2E IoT security and identity management architecture is designed with this in mind, managing and orchestrating the IoT domains both horizontally and vertically, and addressing both security and identity from the IoT device throughout the service life cycle.

Keijo Mononen

is general manager of Security Solutions at Ericsson. In this role he is responsible for end-to-end security management solutions including security automation and analytics. Mononen joined Ericsson in 1990 and for the past 15 years he has held leading positions in professional security services and in security technology development. He holds an M.Sc. in computer science and engineering from Chalmers University of Technology in Gothenburg, Sweden.

Timo Suihko

joined Ericsson in 1992 and is currently working as a senior security specialist in the Ericsson Network Security, Security Technologies team, which belongs to Group Function Technology and Emerging Business. He holds an M.Sc. from Helsinki University of Technology.

Patrik Teppo

joined Ericsson in 1995 and is currently working as a security architect with the CTO Office, Architecture and Portfolio team. He is responsible for the privacy part of the Ericsson architecture and leads Ericsson’s IoT security architecture work. He holds a B.Sc. in software engineering from Blekinge Institute of Technology, Sweden.

Terms and abbreviations


References

1. GSMA Remote SIM Provisioning Specifications, available at: https://www.gsma.com/rsp/

Further reading

Distributed cloud

A KEY ENABLER OF AUTOMOTIVE AND INDUSTRY 4.0 USE CASES

Emerging use cases in the automotive industry – as well as in manufacturing industries where the first phases of the fourth industrial revolution are taking place – have created a variety of new requirements for networks and clouds. At Ericsson, we believe that distributed cloud is a key technology to support such use cases.

Both 4G and 5G mobile networks are designed to enable the fourth industrial revolution by providing high bandwidth and low-latency communication on the radio interface for both downlink (DL) and uplink (UL) data. Distributed cloud exploits these features, enabling a distributed execution environment for applications to ensure performance, short latency, high reliability and data locality.

Distributed cloud maintains the flexibility of cloud computing while at the same time hiding the complexity of the infrastructure, with application components placed in an optimal location that utilizes the key characteristics of distributed cloud. The automotive sector and many manufacturing industries already have use cases that make them very likely to be early adopters of distributed cloud technology.

Next-generation automotive services and their requirements

Mobile communication in vehicles is increasing in importance as the automotive industry works to make driving safer, smooth the flow of traffic, consume energy more efficiently and lower emissions. Automated and intelligent driving, the creation and distribution of advanced maps with real-time data, and advanced driving assistance using cloud-based analytics of UL video streams are all examples of emerging services that require vehicles to be connected to the cloud. These services also require networks that can facilitate the transfer of a large amount of data between vehicles and the cloud, often with real-time characteristics within a limited time frame while the vehicle is in active operation.

High data volume

Looking at the automotive industry, we often focus on the real-time use cases for safety, as defined by V2X/C-ITS (Vehicle to Everything/Cooperative Intelligent Transport System), where real-time aspects such as short latency are the most significant requirements. However, the automotive industry’s new mobility services also place high demands on network capacity due to the extreme amount of data that must be transported to and from highly mobile devices, often with near-real-time characteristics. Data needs to be transported within a limited time window (~30 min/day), with a varying geographical concentration of vehicles using a multitude of different network technologies and conditions.

The market forecasts that are generally referred to indicate that the global number of connected vehicles will grow to approximately 700 million by 2025 and that the data volume transmitted between vehicles and the cloud will be around 100 petabytes per month. At Ericsson, however, we anticipate that the automotive services of the near future will be much more demanding. We estimate that the data traffic could reach 10 exabytes or more per month by 2025, which is approximately 10,000 times larger than the present volume. Gartner recently raised the expectations further in its latest report (June 2018), estimating the volume to be as high as one terabyte per month per vehicle [1].

Such massive amounts of data will place new demands on the radio network, as the main part is UL data. New business models will be required, as a result of the high cost of handling massive amounts of data. As explained in the AECC (Automotive Edge Computing Consortium) white paper [2], the current mobile communication network architectures and conventional cloud computing systems are not fully optimized to handle all of this data effectively on a global scale. The white paper suggests many possible optimizations to consider – based on the assumption that much of the data could be analyzed and filtered at an early stage to limit the amount of data transferred.

Definition of key terms

- **Distributed cloud** is a cloud execution environment for applications that is distributed across multiple sites, including the required connectivity between them, which is managed as one solution and perceived as such by the application.
- **Edge computing** refers to the possibility of providing execution resources (compute and storage) with the adequate connectivity (networking) at close proximity to the data sources.
- **The fourth industrial revolution** is considered to be the fourth big step in industry modernization, enabled by cyber-physical systems, digitalization and ubiquitous connectivity provided by 5G and Internet of Things (IoT) technologies. It is also referred to as Industry 4.0.
When dealing with highly mobile devices that connect to a multitude of networks, it must be possible to move execution of the edge application automatically when a more appropriate location for the vehicle is discovered. Some applications require transfer of previously analyzed data and findings to the new location, where a new application component instance will seamlessly take over to serve the moving vehicle.

Distributed computing on a localized network

We have developed the concept of distributed computing on a localized network to solve the problems of data processing and traffic in existing mobile and cloud systems. In this concept, several localized networks accommodate the connectivity of vehicles in their respective areas of coverage. As shown in Figure 1, computation power is added to these localized networks, so that they can process data locally. This reduces the total amount of data exchanged between vehicles and clouds while enabling the connected vehicles to obtain faster responses. The concept is characterized by three key aspects: a localized network, edge computing and data exposure.

A localized network is a local network that covers a limited number of connected vehicles in a certain area. This splits the huge amount of data traffic into reasonable volumes per area of data traffic between vehicles and the clouds.

**INDUSTRY VERTICALS AND COMMUNICATION SERVICE PROVIDERS ARE DEFINING A SET OF NEW USE CASES FOR 5G**

Edge computing refers to the geographical distribution of computation resources within the vicinity of the termination of the localized networks. This reduces the concentration of computation and shortens the processing time needed to conclude a transaction with a connected vehicle. Data exposure secures integration of the data produced locally by utilizing the combination of the localized network and the distributed computation. By narrowing relevant information down to a specific area, data can be rapidly processed to integrate information and notify connected vehicles in real time. The amount of data that needs to be exchanged is kept to a minimum.

Private and local connectivity

As part of the fourth industrial revolution, industry verticals and communication service providers (CSPs) are defining a set of new use cases for 5G [3]. Private deployments and 5G networks provided by CSPs to manufacturing companies, smart cities and other digital industries are on the horizon as well. However, there are two main challenges to mobile network operators’ ability to deliver. The first is the tough latency, reliability and security requirements of these new use cases. The second is figuring out how to shield the industries from the complexity of the infrastructure, to enable ease of use when programming and operating networks.

Secure private networks with centralized operations

Security and data privacy are key requirements for industrial networks. In some cases, regulations or company policies stipulate that the data must not leave the enterprise premises. In other cases, some or all of the data must be available at remote locations for purposes such as production analytics or emergency procedures. A typical industrial environment has multiple applications deployed and operated by different third parties. What this means in practice is that the same on-premises, cloud-edge instance that a factory already uses for business support and IT systems would also need to support the connectivity for its robots to interact with each other. As a result, there is a requirement of multi-tenancy for both the devices and the infrastructure.

Tactile internet and augmented reality

Augmented reality (AR) and machine learning (ML) technologies are widely recognized as the main pillars of the digitalization of industries [4], and research suggests that wide deployment of interactive media applications will happen on 5G networks. Many observers envision the worker of tomorrow as someone who is equipped with eye-tracking smart glasses [5] and tactile gloves rather than screwdriver sets [6]. Human-to-machine applications require low latency while demanding high network bandwidth and heavy compute resources. Running them on the device itself would result in high battery consumption and heat dissipation. At the same time, latency requirements do not allow the running of the complete application in large central databases due to the physical limits of light speed in optical fibers.
Ericsson has developed a distributed cloud solution that provides the required capabilities to support digitalization of industrial operations, with automotive being one of the key use cases. Our solution satisfies the specific security requirements needed to digitalize industrial operations, with automotive being one of the key use cases. Ericsson’s distributed cloud solution provides edge computing, which many applications require. We define edge computing as the ability to provide execution resources (specifically compute and storage) with adequate connectivity at close proximity to the data sources.

As shown in Figure 3, we define the distributed cloud as a cloud execution environment that is geographically distributed across multiple sites, including the required connectivity in between, managed as one entity and perceived as such by applications. The key characteristic of our distributed cloud is abstraction of cloud infrastructure resources, where the complexity of resource allocation is hidden to a user or application. Our distributed cloud solution enables edge computing, which many applications require. In the case of AR/VR and image recognition applications, the placement of applications at various geolocations is determined by the proximity to the data sources.

Our distributed cloud solution provides edge computing and meets end-to-end network requirements as well as offering management, orchestration and exposure of application resources. Edge computing, which many applications require, is being enabled by our distributed cloud architecture. In the automotive use case, the network is designed to support the needs of connected vehicles, with computation resources being hierarchically distributed and layered in a topology-aware fashion to accommodate localized data and to allow large volumes of data to be processed in a timely manner. This infrastructure framework, localized data collected or exchanged across the network is stored in the central cloud and integrated as offering management, orchestration and exposure of application resources.
the geolocation of the moving car, availability of the computation resources and ability to meet regulatory requirements at the edges serving the moving car. Tactile internet and AR applications that are very sensitive to network latency while demanding high bandwidth and high computing power will be deployed at the edges that can fulfill the requirements.

The service orchestration manages the distributed cloud resources as well as the efficient distribution and replication of the applications that utilize the distributed cloud computation and connectivity resources. The service and resource management capabilities are also deployed in a distributed fashion to enable efficient management. For example, the scaling or data functions will be deployed close to the application they supervise.

Service exposure

The applications deployed in the distributed cloud will present their capabilities through the service exposure. With multi-dimensional exposure, each of the layers in the distributed cloud stack will expose its capabilities. The cloud infrastructure layer and the connectivity layer will expose their respective capabilities through the application programming interface(s) (API(s)), which will then be used by applications and consumers of the industries making use of the mobile connectivity. By setting developer needs in focus, the exposed API(s) will be abstracted so that they are easy to use.

Evolution toward the global multi-operator distributed cloud

Global industries such as automotive require solutions that work seamlessly from local to global scale. In light of this, the evolution toward the global multi-operator distributed cloud is no trivial matter. To be part of the globally distributed cloud, the edge clouds that CSPs provide at access and local centers may be dependent on the CSP’s network topology and the requirements of the use cases this applies to central office sites, base stations and new DCs built on industrial sites. This infrastructure should be flexible, so that it is possible to start with a few sites and grow by adding new sites as required.

Management and orchestration

The distributed cloud relies on efficient management and orchestration capabilities that enable automated application deployment in heterogeneous clouds supplied by multiple actors. Figure 3 illustrates how the service and resource orchestration spans across distributed and technologically heterogeneous clouds. It enables service creation and instantiation in cloud environments provided by multiple partners and suppliers. Discovery, onboarding and auto-enrollment of edges are other important capabilities of distributed cloud management.

When deploying an application or a virtual network function (VNF), the placement decisions can be based on multiple criteria, where latency, geolocation, throughput and cost are a few examples. These criteria can be defined either by an application developer and/or a distributed cloud infrastructure provider, serving as input to the placement algorithm. Once a target cloud has been selected, the workload placement continues in any of the subordinated clouds.

In the automotive applications example, the placement decision could be made based on the

The second step will be to gain industry acceptance for the mechanisms, before finally being able to implement the solutions and establish the business models.

One way to evolve the cloud edges that CSPs currently supply is to provide an environment above the current infrastructure that is homogeneous from a consumption perspective but discoverable through APIs and orchestrated in the same way as the CSPs’ infrastructure. This would provide an intermediate step, where CSPs without an edge cloud infrastructure could become a part of the global scale distributed cloud. Following this approach, an industry actor could connect to any CSP access network as opposed to being limited to certain CSPs. While these networks will have the same functional scope, they will not be able to provide full edge characteristics. This will also serve as a catalyst for other CSPs to join the global scale distributed cloud. Otherwise, they will not become preferred suppliers.

Embracing industry initiatives and standardizations

We believe that the evolution toward the global multi-operator distributed cloud is dependent on a few key actions. First, we must take action to address the fact that the current mobile communication networks and core systems of the industries making use of the mobile connectivity are not designed, orchestrated or exposed in a way that can handle the industries’ requirements effectively. We must scrutinize the system architectures and investigate network deployments and preferred profiling to better accommodate the outlined requirements. The architecture evolution will be driven by the relevant standardizations such as 3GPP and ETSI (European Telecommunications Standards Institute) NFV.

Secondly, we believe that it is critical to drive industry alignment by getting reference implementations of edge cloud software. This is why Ericsson has joined the industry collaboration project ONAP (Open Platform for NFV) and OSM (Open Network Automation Platform) [7], which provides the management capabilities of distributed cloud.

Finally, we believe that participating in ecosystems that provide the opportunity for interactions between the industries and vendors is critical to the evolution. This is particularly true for ecosystems that formulate requirements and ways of working, define use cases, agree on a common, easy-to-use reference implementation, and drive alignment in standardization bodies based on those implementations. Examples of such ecosystems are the AECC and SIGAA (the 5G Automotive Association) for automotive and 5G-ACIA (the 5G Alliance for Connected Industries and Automation) [8], Industry 4.0 and the IOT (Industrial Internet of Things) for the fourth industrial revolution.

IT IS CRITICAL TO DRIVE INDUSTRY ALIGNMENT BY GETTING REFERENCE IMPLEMENTATIONS OF EDGE CLOUD SOFTWARE

Conclusion

Distributed cloud is a cornerstone of the intelligent networks that will play a key enabling role in the fourth industrial revolution. A robust distributed cloud solution requires efficient and intelligent management and orchestration capabilities that span heterogeneous clouds supplied by multiple actors. Service exposure will enable monetization and application innovation through integration with the marketplaces and/or integration with the industries’ IT systems.

The evolution toward globally distributed cloud requires action to align the industry both through traditional standardizations as well as active participation in open-source projects aimed at providing reference implementations. Ecosystems such as the AECC play an important role by examining the high-volume data use cases for the automotive industry.

GLOBAL INDUSTRIES SUCH AS AUTOMOTIVE REQUIRE SOLUTIONS THAT WORK SEAMLESSLY FROM LOCAL TO GLOBAL SCALE

Distributed cloud is a cornerstone of the intelligent networks that will play a key enabling role in the fourth industrial revolution. A robust distributed cloud solution requires efficient and intelligent management and orchestration capabilities that span heterogeneous clouds supplied by multiple actors. Service exposure will enable monetization and application innovation through integration with the marketplaces and/or integration with the industries’ IT systems.

The evolution toward globally distributed cloud requires action to align the industry both through traditional standardizations as well as active participation in open-source projects aimed at providing reference implementations. Ecosystems such as the AECC play an important role by examining the high-volume data use cases for the automotive industry.
Malgorzata Svensson  

is an expert in operations support systems. She joined Ericsson in 1996 and has worked in various areas within research and development. For the past 10 years, her work has focused on architecture evolution. Svensson has broad experience in business process, function and information modeling, information and cloud technologies, analytics, DevOps processes and tool chains. She holds an MSc in technology from the Silesian University of Technology in Gliwice, Poland.

Christer Boberg  

serves as a director at Ericsson’s CTO office, responsible for IoT technology strategies aimed at solving networking challenges for the industry on a global scale. He initially joined Ericsson in 1983 and has in his career within and outside Ericsson focused on software and system design as a developer, architect and technical expert. In recent years, Boberg’s work has centered on IoT and cloud technologies with a special focus on the automotive industry. As part of this work, he drives the AECC consortium together with industry leading companies.

Benedek Kovács  

joined Ericsson in 2005 as a software developer and tester, and later worked as a system engineer. He was the innovation manager of the Budapest R&D site 2011-13, where his primary role was to establish an innovative organizational culture and launch internal startups based on worthy ideas. Kovács went on to serve as the characteristics, performance management and reliability specialist in the development of the 4G VoLTE solution. Today he is working on 5G networks and distributed cloud, as well as coordinating global engineering projects. He holds an M.Sc. in information engineering and Ph.D. in mathematics from the Budapest University of Technology and Economics in Hungary.

### References

1. TelecomTV, Gartner says 5G networks have a paramount role in autonomous vehicle connectivity, June 21, 2018, available at: [https://www.telecomtv.com/content/tracker/gartner-says-5g-networks-have-a-paramount-role-in-autonomous-vehicle-connectivity-37356/](https://www.telecomtv.com/content/tracker/gartner-says-5g-networks-have-a-paramount-role-in-autonomous-vehicle-connectivity-37356/)
8. 5G-ACIA, 5G for Connected Industries and Automation, April 11, 2018, available at: [https://www.5g-acia.org/publications/5g-for-connected-industries-and-automation-white-paper/](https://www.5g-acia.org/publications/5g-for-connected-industries-and-automation-white-paper/)

### Further reading

- Ericsson, Going beyond edge computing with distributed cloud, available at: [https://www.ericsson.com/external-services/reading/distributed-cloud](https://www.ericsson.com/external-services/reading/distributed-cloud)

### Terms and abbreviations

- AECC – Automotive Edge Computing Consortium
- API – Application Programming Interface
- APP – Application
- AR – Augmented Reality
- BW – Bandwidth
- CSP – Communication Service Provider
- DB – Database
- DC – Data Center
- ECU – Engine Control Unit
- ETSI – European Telecommunications Standards Institute
- FWA – Fixed Wireless Access
- IoE – Internet of Things
- ML – Machine Learning
- MS – Millisecond
- MS – Mobile Telephone Switching Office
- NFV – Network Functions Virtualization
- UL – Uplink
- VNF – Virtual Network Function
- VR – Virtual Reality
- V2X/C-ITS – Vehicle-to-everything/Cooperative Intelligent Transport System

### The Authors

Nylander.

Thorsten Lohmar

Ola Hubertsson,

Ala Nazari,

to this article:

contribution

the following

the authors

Nylander.

Thorsten Lohmar

Ola Hubertsson,

Ala Nazari,

the authors

Nylander.

Thorsten Lohmar

Ola Hubertsson,

Ala Nazari,

the authors

Nylander.

Thorsten Lohmar

Ola Hubertsson,

Ala Nazari,
Booster smart manufacturing WITH 5G WIRELESS CONNECTIVITY

Industry 4.0 – the fourth industrial revolution – is already transforming the manufacturing industry, with the vision of highly efficient, connected and flexible factories of the future quickly becoming a reality in many sectors. Fully connected factories will rely on cloud technologies, as well as connectivity based on Ethernet Time-Sensitive Networking (TSN) and wireless 5G radio.

The goal of Industry 4.0 is to maximize efficiency by creating full transparency across all processes and assets at all times. Achieving this requires communication between goods, production systems, logistics chains, people and processes throughout a product’s complete life cycle, spanning everything from design, ordering, manufacturing, delivery and field maintenance to recycling and reuse.

The integration of 5G ultra-reliable low-latency communication (URLLC) in the manufacturing process has great potential to accelerate the transformation of the manufacturing industry and make smart factories more efficient and productive.

Today’s state-of-the-art factories are predominantly built on a hierarchical network design that follows the industrial automation pyramid, as shown in Figure 1. The fourth industrial revolution will require a transition from this segmented and hierarchical network design toward a fully connected one. This transition, in combination with the introduction of 5G wireless communication technology, will provide very high flexibility in building and configuring production systems on demand. The ability to extract more information from the manufacturing process and feed it into a digital representation known as the "digital twin" [1] enables more advanced planning processes, including plant simulation and virtual commissioning. Initiatives like the 5G Alliance for Connected Industries and Automation (5G-ACIA) [2] show that industries recognize this need for 5G technology.

The lower section of Figure 1 is often referred to as the operational technology (OT) part of the manufacturing plant, comprising both the field level (industrial devices and controllers) and the manufacturing execution system. The top section is the information technology (IT) part, made up of general enterprise resource planning. For connectivity at field level, a variety of fieldbus and industrial Ethernet technologies are typically used. Ethernet and IP are well established communication protocols at higher levels (IT and the top part of OT).

The OT network domain is currently dominated (>90 percent) by wired technologies [3] and is a heavily fragmented market with technologies such as PROFINET, PROFINET, EtherCAT, Sercos and Modbus. Currently deployed wireless solutions (which are typically wireless LAN based using unlicensed spectrum) constitute only a small fraction of the market.

Figure 1 Hierarchical network design based on the industrial automation pyramid

- Ultra-reliable low-latency communication (URLLC) refers to a 5G service category that provides the ability to successfully deliver a message within a specified latency bound with a specified reliability, such as delivering a message within 1ms with a probability of 99.9999 percent.
- The fourth industrial revolution is considered to be the fourth big step in industry modernization, enabled by cyber-physical systems, digitalization and ubiquitous connectivity provided by 5G and Internet of Things (IoT) technologies. It is also referred to as Industry 4.0.
requirements are non-critical. Wirelessly connecting sensors where communication of the installed base; they mainly play a role for which can only be satisfied with a dedicated local URLLC with ultra-high availability and resilience, requirements that differ significantly from public The manufacturing industry has specific 5G Key manufacturing industry requirements makes it possible to achieve greater floor plan layout improving throughput, capacity and coverage \[4\]. The scalable numerology of NR provides good is compared with the application latency constraint, transmission fails. Thus, the shorter the RAN RTT latency. A packet can be encoded with a very low and be used to achieve a certain level of reliability and latency. A packet can be encoded with a very low and robust code rate, and just be transmitted once, but if the RTT is shorter than the application latency constraint, it can be more efficient to use a higher, less robust initial code rate and perform retransmissions based on feedback in case the initial transmission fails. Thus, the shorter the RAN RTT is compared with the application latency constraint, the higher spectral efficiency (capacity) may be achieved. Licensed spectrum for interference control The availability of spectrum resources is key to meeting requirements on capacity, bitrates and latency. To provide predictable and reliable service levels on the factory shop floor, the spectrum resources need to be managed carefully. The achievable performance depends on several factors: the amount of spectrum available which spectrum is used – low band (below 2GHz), mid-band (2-5GHz) or high band/mmWave (26GHz and above) which licensing regime applies whether the spectrum is FDD or TDD which radio access technology is used the coexistence scenarios that apply for the spectrum. Estimates of spectrum needs are in the range of tens to hundreds of megahertz. Most new mid-band spectrum that is currently being allocated uses TDD, while large parts of the spectrum already allocated to mobile operators are FDD. Latency for an FDD system is inherently lower than that of a corresponding TDD system. Mid-band spectrum is well suited for indoor deployments since its propagation characteristics make it easy to provide good coverage with a limited set of transmission points. Coverage at mmWave is generally spottier, requiring denser radio deployment, but mmWave is still a good complement to mid-band for in-factory deployments since it enables: higher system capacity, as larger bandwidths are available and as advanced antenna systems and beamforming can be implemented in a small form factor suitable for indoor deployment significantly shorter latencies (even though the spectrum is TDD), as a higher numerology with shorter transmission time intervals is used easier management of the coexistence between indoor shop floor networks and outdoor mobile networks, as mmWave radio signals are easier to confine within buildings.

One near-term benefit of leveraging wireless connectivity in factories is the significant reduction in the amount of cables used, which reduces cost, since cables are typically very expensive to install, rearrange or replace. In addition, wireless connectivity enables new use cases that cannot be implemented with wired connectivity, such as moving robots, automated guided vehicles and the tracking of products as they move through the production process. Wireless connectivity also makes it possible to achieve greater floor plan layout flexibility and deploy factory equipment more easily.

The manufacturing industry has specific 5G requirements that differ significantly from public mobile broadband (MBB) services. These include URLLC with ultra-high availability and resilience, which can only be satisfied with a dedicated local network deployment using licensed spectrum. The ability to integrate with the existing industrial Ethernet LAN and existing industrial nodes and functions is another fundamental requirement. Data integrity and privacy are also critical, as well as real-time performance monitoring. In addition, 5G capabilities in terms of positioning, time synchronization between devices, security and network slicing will also be essential for many manufacturing use cases.

Ultra-reliable low-latency communication One of the two service categories of machine-type communication (MTC) in 5G – critical MTC (cMTC) – is designed to meet communication demands with stringent requirements on latency, reliability and availability. Intense standardization and R&D work is ongoing to ensure 5G New Radio (NR) technology is able to fully address the need for URLLC.

With NR we will see large-scale deployments of advanced antenna systems enabling state-of-the-art beamforming and MIMO (multiple-input, multiple-output) techniques, which are powerful tools for improving throughput, capacity and coverage \[4\]. Multi-antenna techniques will also be important for URLLC, as they can be used to improve reliability. The scalable numerology of NR provides good means to achieve low latency, as larger subcarrier spacing (SCS) reduces the transmission time interval.

To further reduce latency and increase reliability, several new MAC (medium access control) and PHY (physical layer) features as well as new multi-connectivity architecture options have been added to the 5G NR specifications in 3GPP release 15, and additional enhancements are being studied in release 16. The goal in release 16 is to enable 0.5-1ms one-way latency with reliability of up to 99.9999 percent. New capabilities include faster scheduling, smaller and more robust transmissions, repetitions, faster retransmissions, preemption and packet duplication \[5\]. All in all, they ensure NR is equipped with a powerful toolbox that can be used to tailor the performance to the demands of each specific device and traffic flow on a factory shop floor.

The achievable round-trip time (RTT) depends both on which features and spectrum are used. For example, the RAN RTT for a mid-band deployment optimized for MBB can be in the order of 3ms (FDD 1.4kHz SCS or TDD 10kHz with DL-DL-DL-UL TDD configuration). The corresponding RTT for a URLLC-optimized millimeter wave (mmWave) deployment (TDD 120kHz SCS, DL-UL TDD configuration) can be below 2ms, thus matching the 3GPP one-way latency goal.

There is a trade-off between latency, reliability and capacity, and different scheduling strategies can be used to achieve a certain level of reliability and latency. A packet can be encoded with a very low and robust code rate, and just be transmitted once, but if the RTT is shorter than the application latency constraint, it can be more efficient to use a higher, less robust initial code rate and perform retransmissions based on feedback in case the initial transmission fails. Thus, the shorter the RAN RTT available to mobile operators are FDD. Latency for an FDD system is inherently lower than that of a corresponding TDD system. Mid-band spectrum is well suited for indoor deployments, since its propagation characteristics make it easy to provide good coverage with a limited set of transmission points. Coverage at mmWave is generally spottier, requiring denser radio deployment, but mmWave is still a good complement to mid-band for in-factory deployments since it enables: higher system capacity, as larger bandwidths are available and as advanced antenna systems and beamforming can be implemented in a small form factor suitable for indoor deployment significantly shorter latencies (even though the spectrum is TDD), as a higher numerology with shorter transmission time intervals is used easier management of the coexistence between indoor shop floor networks and outdoor mobile networks, as mmWave radio signals are easier to confine within buildings.

Terms and abbreviations

- cMTC – Critical Machine-type Communication
- CN – Core Network
- DL – Downlink
- UL – Uplink
- GHz – Gigahertz
- GW – Gateway
- HB – Hertz
- HW – Hirstertz
- IOT – Internet of Things
- kHz – Kilohertz
- LTE – Long Term Evolution
- MBB – Mobile Broadband
- mMTC – Massive Machine-type Communication
- mmWave – Millimeter Wave
- ms – Millisecond
- MTC – Machine-type Communication
- NB-IoT – Narrowband IoT
- NR – New Radio
- OTT – Over-the-Top Technology
- PDCP – Protocol Data Unit
- SCS – Subcarrier Spacing
- SDR – Software Defined Radio
- TDD – Time-Division Duplex
- TSN – Time-Sensitive Networking
- UE – User Equipment
For critical applications, there must be guarantees against uncontrolled interference, which implies that licensed spectrum is necessary. As illustrated in Figure 2, unlicensed technologies such as Wi-Fi and MulteFire cannot guarantee bounded low latency with high reliability as the load increases. This is due to the use of listen-before-talk back-off, which does not perform well during uncontrolled interference. Unlicensed spectrum may nonetheless be relevant for less critical applications.

Licensed spectrum can be provided by operators as part of a local connectivity solution, including network equipment. Operators may also choose to lease parts of their spectrum assets locally to industries without providing the connectivity solution. Another emerging option is for regulators to set aside dedicated spectrum for local licensing to support much more precise positioning. There are several aspects which all contribute to better positioning accuracies below 3m, but NR deployed in a factory environment has the technology potential to support much more precise positioning. These are additional requirements and features of interest.

Integration with industrial Ethernet and TSN
The introduction of 5G on the factory shop floor will happen in steps. When 5G is added to existing production systems, the various parts of the system will be moved to 5G connectivity at different stages, depending on the evolution plan of the production system and where the highest benefits of wireless 5G communication can be obtained. Over time, more parts of the shop floor can be migrated to 5G, in part due to the introduction of new capabilities in future 5G releases. Even in greenfield industrial deployments, not all communication will be based on 5G. The need for wireless connectivity may not be prominent for some subsystems, while others may require performance levels (isochronous sub-millisecond latency; for example) that are not currently addressed by 5G. Consequently, a local industrial 5G deployment will coexist and require integration with wired industrial LANs. To this end, the transport of Ethernet traffic is required, and Ethernet transport has been specified within the release 15 standard of the 5G system.

As part of the ongoing industrial transformation, the wired communication segments of industrial networks are expected to evolve toward a common open standard: Ethernet with TSN support [6]. Therefore, a 5G system needs to be able to integrate with a TSN-based industrial Ethernet, for which 3GPP has defined different study and work items in release 16 of the 5G standards. TSN is an extension of the IEEE 802.1 Ethernet and is standardized within the TSN task group in IEEE 802.1. A profile for TSN in industrial automation is being developed by the IEC/IEEE 60882 project [7]. TSN includes the means to provide deterministic bounded latency without congestion losses for prioritized traffic on an Ethernet network that also transports traffic of lower priority. TSN features include priority queuing with resource allocation mechanisms, time synchronization between network nodes and reliability mechanisms via redundant traffic flows. TSN has the objective to achieve indoor positioning accuracies below 3m, but NR deployed in a factory environment has the technology potential to support much more precise positioning. These are several aspects which all contribute to better positioning accuracies:

- the wide bandwidths of mid- and high-band spectrum enable better measurement accuracy
- beam-based systems enable better ranging and angle-of-arrival/departure estimation
- the higher numerology of NR implies shorter sampling intervals and hence improved positioning resolution
- dense and tailored deployments with small cells and large overlaps improve accuracy and, together with beam-based transmissions, provide more spatial variations that can be exploited for radio frequency fingerprinting.

In 5G release 16, a new requirement is being introduced, whereby the 5G system will be able to synchronize devices to a master clock of one or more time domains [8]. One reason for this is that several

Additional requirements and features of interest
One 5G feature that could have significant importance for manufacturing use cases is positioning. For 3GPP release 16, the objective is to achieve indoor positioning accuracies below 3m, but NR deployed in a factory environment has the technology potential to support much more precise positioning. These are several aspects which all contribute to better positioning accuracy:

- the wide bandwidths of mid- and high-band spectrum enable better measurement accuracy
- beam-based systems enable better ranging and angle-of-arrival/departure estimation
- the higher numerology of NR implies shorter sampling intervals and hence improved positioning resolution
- dense and tailored deployments with small cells and large overlaps improve accuracy and, together with beam-based transmissions, provide more spatial variations that can be exploited for radio frequency fingerprinting.

In 5G release 16, a new requirement is being introduced, whereby the 5G system will be able to synchronize devices to a master clock of one or more time domains [8]. One reason for this is that several
evolving threat landscape, 5G includes new security network. To address new use cases and the users know they are connected to a legitimate communication services, user privacy, every generation to enable confidential combined 5G-TSN network. The time alignment of the 5G system with one or more time domains used in an industrial system. The time alignment of the 5G system with the external industrial LAN is also a basis to enable TSN time-scheduled communication over a wide area networks, enhance service differentiation including isolation of the critical traffic from other service types and enable segmentation into security zones as required for the OT domain. 5G connectivity solution for the factory shop floor A local, on-premises 4G/5G connectivity solution that uses licensed spectrum such as the one shown in Figure 3 is the best way to meet the requirements of the manufacturing industry. This solution can support cMTC, MBB and massive MTC (mMTC) use cases, and it can easily be integrated with mobile operator-provided wide area networks. While cMTC addresses the critical communication needs of the manufacturing industry, mMTC, also included in 5G, is ideal for sensor communication. Narrowband Internet of Things (NB-IoT) and LTE machine-type communication (LTE-M) are examples of mMTC solutions that were developed for 4G and remain well-equipped to support the needs of the manufacturing industry for a long time. MBB and mMTC based on 4G and 5G provide the shop-floor connectivity required by industrial sensors, cameras, smartphones, tablets and wearables to support use cases like data acquisition, predictive maintenance, human-machine interaction and augmented reality. Beyond factories, there are also wide-area use cases like smart logistics that rely on the MBB and mMTC services supplied by mobile operator-provided networks. Network operators are in an excellent position to leverage their spectrum assets, wide area network infrastructure and know-how to address the needs of the manufacturing industry. Alternatively, the solution can be deployed by the industries themselves only by third parties using based or dedicated spectrum. The optimal local connectivity solution requires a well-planned 4G/5G indoor radio system using licensed spectrum to enable ultra-reliable low-latency performance. The virtualization of core network (CN) functions and support of control and user-plane separation enables flexible CN deployments. The CN user plane needs to be deployed in the factory, not only to provide URLLC, but also high availability, local survivability, security and privacy. The requirements on full local control would indicate that CN control functions need to be deployed on premises, but depending on the specifics of the requirements, such as how long survivability duration is required, it may be possible to use more cost-efficient solutions where some of the control functions are provided from a central location, such as a mobile network operator’s CN. An easy-to-use local management system is required to monitor and manage the end-to-end connectivity, including local network infrastructure and connected devices. The local management use cases include both software management and fault, performance and configuration management. The management system also needs to integrate with other elements of the OT systems and the industry IT systems. A low-latency cloud infrastructure is required both for 5G network functions and industrial applications, and all pieces need to be connected using an integrated local transport infrastructure. The resulting solution can provide both IP and Ethernet connectivity to industrial devices and GWs on the shop floor, with performance tailored to each device’s individual needs. The integration between the 5G infrastructure and the industrial Ethernet domain extends beyond simple user plane forwarding of Ethernet frames to include integration with the time synchronization, scheduling and resilience schemes used in the industrial Ethernet domain, using TSN features, for example.

5G includes new security features that benefit industrial deployments

Conclusion
5G is a prime enabling technology to facilitate the industrial transformation to Industry 4.0, providing wireless connectivity in and around the factory based on a global standard with global economy of scale. It can connect a variety of industrial devices with different service needs, including industrial sensors, video cameras or advanced control panels with integrated augmented reality. 5G can also provide deterministic ultra-reliable low-latency communication to bring wireless connectivity to demanding industrial equipment, like industrial controllers and actuators.
A 5G-connected factory is based on a local 5G radio network using licensed spectrum. It can either be provided as a service by a mobile network operator, or it can be operated standalone by a factory owner or system integrator in locally leased or dedicated spectrum. A local core network enables low-latency connectivity, fulfilling strict requirements on availability, local survivability, data security and privacy. The integration of a 5G system with wired industrial LAN equipment—which in future will mainly be based on TSN—is mandatory. Further, 5G enhancements provide additional value to industrial services like precise indoor positioning, and time synchronization for industrial end devices.

References
2. 5G-ACIA, 5G for Connected Industries and Automation, available at: https://www.5g-acia.org/
9. Ericsson white paper, 5G security – enabling a trustworthy 5G system, March 28, 2018, Nakarmi, P.K.; Fogelström, E., available at: https://www.slideshare.net/ericsson/5gsecurityenablingatrustworthy5g-system

Further reading
5G-ACIA white paper, 5G for Connected Industries and Automation, November 22, 2018, available at: https://www.5g-acia.org/index.php?id=5125
The Internet of Things (IoT) represents an ongoing paradigm shift within communications: everything that benefits from a connection can and will be connected.

- Massive IoT refers to applications that are less latency sensitive and have relatively low throughput requirements, but require a huge volume of low-cost, low-energy consumption devices on a network with excellent coverage. The growing popularity of IoT use cases in domains that rely on connectivity spanning large areas, and are able to handle a huge number of connections, is driving the demand for massive IoT technologies.

Through the development of new technologies in the fields of communication, computation, sensors, electronics and batteries, it is now possible to develop battery-powered devices with sensors and actuators and computers that are connected via wide-area communication networks to a cloud-based platform that handles device data and management. These devices can be tailored to fit several specific application areas and deployed in massive numbers, making them fit for use in massive IoT applications. Examples of massive IoT application areas include: wearables (e-health), asset tracking (logistics), smart city (smart homes, environmental monitoring and smart metering (smart buildings)); and smart manufacturing (monitoring, tracking, digital twins). The key device characteristics include:

- low device and deployment cost
- small form factor
- long battery life
- wireless connectivity for challenging locations
- strong application and communication security.

There are two key challenges in the massive IoT device domain: (1) connecting a large volume of devices in a wide area cost-efficiently, and (2) efficiently managing these devices over their complete life cycle. As security and trust are key requirements in most massive IoT applications, the devices must be trusted in terms of both communication and data integrity end-to-end (E2E), from device to application data usage. Many applications also benefit from devices that include local intelligence that can process data before it is further communicated.

To address these challenges, it is necessary to make smart choices in five key technology areas – connectivity, communication protocols, security, identity solutions, and machine intelligence (MI). Connectivity: New massive IoT cellular technologies, such as Narrowband IoT (NB-IoT) and LTE for machine-type communication (LTE-M), are taking off and driving growth in several cellular IoT connections, with a compound annual growth rate of 27 percent expected between 2018 and 2024 [1]. LTE-M and NB-IoT are cellular radio access technologies that provide low-power wide-area (LPWA) IoT connectivity in licensed spectrum, unlike short-range technologies in unlicensed spectrum such as Bluetooth and Zigbee, and LPWA technologies such as Sigfox and LoRaWAN.

The 3GPP release 13 design targets for massive IoT were: long device battery life, low device complexity to ensure low cost, support for massive numbers of devices, and coverage enhancements to be able to reach devices in basements and other challenging locations. Two new cellular technologies...
Massive IoT Devices

Transmission and VoLTE in enhanced coverage, as such as more accurate positioning of UE, multicast things, release 14 features improvements to LTE-M functionalities are introduced in 3GPP LTE-M and more use cases, several enhancements and new coverage, versus +15dB in Cat-M. The UE output enhancement feature, with up to +20dB enhanced of up to 26kbps in the downlink and 63kbps in the uplink. It also includes UE category Cat-NB1 [2].

A Cat-M1 UE supports a reduced bandwidth of 1.6MHz and a data throughput of up to 300kbps in the downlink and 175kbps in the uplink. It also supports mobility and VoLTE services. Therefore, Cat-M1 UEs are suitable for applications such as wearables and asset tracking.

NB-IoT operates in half-duplex mode within the 200kHz bandwidth and supports a data throughput of up to 2.6kbps in the downlink and 6kbps in the uplink. Similar to Cat-M1, NB-IoT offers the coverage enhancement feature, with up to +20dB enhanced coverage, versus +15dB in Cat-M. The UE output power classes are 20dBm and 23dBm, as in Cat-M.

To improve the user experience and to cater to more use cases, several enhancements and new functionalities are introduced in 3GPP LTE-M and NB-IoT releases 14 and 15 [3][4]. Among other things, release 14 features improvements to LTE-M – such as more accurate positioning of UE, multicast transmission and VoLTE in enhanced coverage, as well as higher data rates to serve a wider range of applications, reduce latency and extend battery life. Similarly, release 14 NB-IoT performance is improved with more accurate positioning of UE, multicast transmission, capacity improvement (thanks to the support of paging and random-access procedures on non-anchor carriers), higher peak data rates and a new lower power class (14dBm) that enables reduced power consumption and smaller battery form factors.

In release 15, LTE-M features include support for higher UE velocities, a new lower UE power class, reduced system acquisition time, reduced UE power consumption by early data transmission, wake-up signal for paging monitoring, relaxed monitoring for cell reselection, increased spectral efficiency and improved access control. The main features introduced in release 15 NB-IoT aim to further reduce latency and UE power consumption (early data transmission, wake-up signal and quick Radio Resource Control release, for example). Other features include: UE measurement improvements, support of cell ranges of up to 100km, TDD support, reduced system information acquisition and cell search time, and improved UE differentiation and access control.

The latest compression techniques, such as Static Context Header Compression [5], can compress the IPv6 and other headers into just a few bytes, making it possible for even the most constrained low-power wide-area network (LPWAN) IoT communication to be economically secured with other data models and semantics, as shown in Figure 2.

Communication protocols

While many legacy machine-to-machine (M2M) devices use tailor-made protocol stacks for each specific application, more and more devices today (as well as the vast majority of current ecosystems) use internet protocols as the basis of the IoT protocol stack. That is, they use the Internet Protocol (IP) on top of various link-layer protocols, followed by a selection of standardized transport and relay protocols, ending up at the application layer with data models and semantics, as shown in Figure 2.

Transfer protocols are used over the (secure) transport layer to transfer data objects and provide semantics for operations. Two transfer protocols that reuse the web model are widely used today: Hypertext Transfer Protocol (HTTP) [8] and Constrained Application Protocol (CoAP) [9]. The new version of HTTP, HTTP/2 [10], is also increasingly being adopted. Message Queuing

Figure 2 Structure of an IoT device protocol stack

<table>
<thead>
<tr>
<th>Data link and physical</th>
<th>Transfer</th>
<th>Network</th>
<th>Data and semantics</th>
</tr>
</thead>
<tbody>
<tr>
<td>CoAP / HTTP / HTTP/2 / MQTT / (various)</td>
<td>UDP / TCP / QUIC with transport security</td>
<td>IP</td>
<td>LwM2M &amp; IPSO &amp; SenML / IoT.schema.org / W3C Web of Things / (various)</td>
</tr>
</tbody>
</table>

Terms and abbreviations

IT IS POSSIBLE TO TAKE ADVANTAGE OF PUBLIC-KEY CRYPTOGRAPHIC FUNCTIONS IN SMALL IOT DEVICES

Telephony Transport (MQTT) is a widely-used publish-subscribe protocol for the IoT. In industrial environments, more specialized protocols are often used, and some environments also reuse legacy messaging protocols for IoT. Of all the options, web protocols, and in particular CoAP for the embedded web, have proven to be the best choice, especially for interoperability and scalability. Data models provide common syntax, structure and semantics for the communicating endpoints. A data model can be something very simple – containing a single temperature value, for example – but most real-life systems require the exchange of more information. Traditionally, in many M2M systems this information has been encoded in application-specific ways, but in the IoT, where data is often exchanged with multiple types of loosely coordinated systems, common data models are needed to ensure endpoints understand the meaning of the data. Standardized data models such as Sensor Measurement Lists (SenML) [11] can be used to efficiently interchange batches as well as the time series of sensor and actuator data.

A fully built and operational IoT system also requires life-cycle management capabilities such as automated bootstrapping, configuration and firmware updates. The Open Mobile Alliance SpecWorks Lightweight Machine-to-Machine (LwM2M) device and data management protocol [12] is built on the standard web protocol stack, using IP, UDP/TCP, CoAP, TLS/OSCORE and SenML. Furthermore, IPSO smart objects can be used with LwM2M to enable reusable application semantics. LwM2M and IPSO smart objects provide a full suite to support life-cycle management and applications with interoperability from connectivity to application layer.

Finally, it is possible to bridge the gap between devices from different – and often uncoordinated – ecosystems by using common ways to express device interaction capabilities such as the World Wide Web Consortium’s Web of Things Thing Descriptions (W3C/TD) [13] and common vocabularies for describing things, such as iot.schema.org.

Security

The security of IoT devices is built on functions for secure communication, application security and device security. Together, these functions protect device management, guarantee data ownership and ensure that devices remain trustworthy throughout their entire operational life. Secure communication protocols like TLS, DTLS and OSCORE allow for different algorithms. However, not all supported algorithms are secure – this is the case for TLS v1.2, for example. In addition, IoT devices normally only support a subset of algorithms, which makes it important to select the right ones. Newer protocols like TLS v1.3 are more secure and in many cases also more efficient. IoT devices often only support symmetric key cryptographic algorithms, due to the fact that public-key cryptographic functions are complex and demand large key sizes, which may be problematic for very constrained devices. With proper design (as in IETF Authentication and Authorization for Constrained Environments/OSCORE), however, it is possible to take advantage of public-key cryptographic functions in small IoT devices. The power consumption of complex computations can be reduced by using optimized hardware acceleration of cryptographic functions. It is therefore likely that future small IoT devices will have certain dedicated cryptographic hardware. Persistent cryptographic key material must be stored securely and kept isolated from application software and physical interfaces as much as possible.

IoT devices are increasingly following the smartphone approach of using Trusted Execution Environments (TEE) for this isolation. Recently, ARM’s TrustZone TEE technology was brought to constrained devices. For more powerful devices, there are alternatives such as Intel SGX. Also, dedicated security components like Trusted Platform Modules (TPMs) – or proprietary ASICs (application-specific integrated circuits) – can be used. Such solutions can achieve a high level of security, albeit at higher cost and power consumption levels. In many use cases, integrated TEES will be sufficient and more cost-effective.

To maintain security during their operational life, IoT devices should support secure software/firmware upgrade. Such secure upgrade is often realized by having the software signed prior to release and having a trusted subsystem in the device that performs a verification of the software before it is programmed/loaded into the device. This trusted subsystem is often referred to as the root of trust of a device. New standardization work [14] was recently started for securing updates for software/firmware. Procedures for secure device life-cycle management are not easy and may have to be tailored for a specific use case. The awareness of the importance of device security is growing in the industry, but more efforts are needed to realize well-integrated trustworthy systems that cover the needs of life-cycle management and applications security.

Supporting secure software update is crucial to the creation of trustworthy IoT devices.

Identity solutions

Trustworthiness also depends on secure digital identities. A digital identity can be used for authentication, to maintain data ownership or for software origin verification. For example, a device can prove it is trustworthy – that is, it has been produced by a legitimate manufacturer – through an initial identity. An identity consists of a securely stored secret and an assigned link between this secret and an identifier or name. A well-known way to do this is to use a public key infrastructure (PKI), where the device holds a private key and the identity is a certificate that links this key to an identifier written into the certificate. For IoT devices, traditional PKIs have their problems. Their cryptographic operations can be cumbersome for highly constrained devices, the certificates can be large, and the certificate revocation management is usually so tricky that it is hardly used. Furthermore, traditional PKIs have privacy issues. These issues can be addressed, as they have been in Enhanced Privacy ID, but at significantly higher complexity costs than PKI.

SUPPORTING SECURE SOFTWARE UPDATE IS CRUCIAL TO THE CREATION OF TRUST-WORTHY IOT DEVICES

As an alternative to PKIs, it is possible to use identities based on symmetric key cryptography. This method is already in use for the 2G, 3G and 4G mobile network systems that use SIMs to hold the authentication credentials. SIMs use dedicated hardware chips and are relatively complex, mainly for legacy reasons. More cost-effective solutions are on their way, such as the integrated Universal
Integrated Circuit Card (iUICC), in which the SIM hardware is integrated into the device processors. For 5G mobile network systems, symmetric key-based identities for network access will remain in use, but in 5G it is also possible to use PKI-based identities for network access will remain in use, but in 5G it is also possible to use PKI-based identities for network access will remain in use, but in 5G it is also possible to use PKI-based identities for network access will remain in use, but in 5G it is also possible to use PKI-based identities.

As different device hardware will come with different types of initial identities, Ericsson believes that the federation of identities [13] is important in the bootstrapping of identities that support the device use case. The complexity of identity management can be reduced if identities can be reused. In practice, such reuse may be built on careful derivation techniques, in which a new identity is created and receives trust from an existing one. This is, for example, the case in Generic Bootstrapping Architecture, where a SIM-based key can be used to derive a key for TLS or application security.

A more holistic and distributed approach to handling the trust in device identities can be achieved with blockchains or distributed ledgers.

These options make it possible to link device lifecycle management with that of the device identity in a common framework.

**Machine intelligence**

MI technologies are key to building IoT systems that can improve their own performance of a task as more data becomes available and more knowledge is inferred and retained [16]. In massive IoT, which handles large volumes of data and millions of devices, MI is required to intelligently automate data transmission, routing and data processing. Distributed MI (DMI) concerns the deployment, dynamic composition and lifecycle management of multi-node MI services, which can be chained for provisioning an intelligent system. Orchestrating lightweight DMI components to jointly perform MI tasks that enhance massive IoT operations is a fundamental research topic at Ericsson [17].

One important path in DMI is moving intelligence toward the device end, which will minimize E2E latency, enhance data privacy and lower bandwidth requirements while reducing server-side costs. Such on-device MI (ODMI) efforts go beyond routing IoT data to cloud backends and instead promote horizontal connectivity of devices to edge infrastructure that hosts DMI services. To follow this path, it is essential that the IoT devices are able to perform low power computation close to where the data is generated and the actuation is needed. This requires knowledge of MI-tailored ASICs and of their integration with MI frameworks. In the hardware layer, ODMM has been embodied into graphics processing units, ASICs such as tensor processing units (TPUs), and neuromorphic chips. The main innovation of TPUs relies on efficient complex instruction set implementations for the matrix multiplier unit, which is key for executing modern MI workflows. Neuromorphic chips are low-power hardware where asynchronous brain-inspired manycore meshes are interconnected over sparse and recurrent inter-core communication topologies, thus easing the translation of MI dataflows into instruction flows.

On the software side, many vendors favor the idea of offloading MI computation to hardware accelerators. In this layer, the integration of systems optimization has become widespread, such as compilers and schedulers that can prune and break down MI workflows into distributable task graphs. Scalable massive IoT systems require investment in MI services that can be repurposed to adapt to operational conditions evolving networks, as sensors and actuators are added and removed. Flexibility is then a core design principle in massive IoT systems. Edge and ODMM add such flexibility because they offer more DMI deployment options and control over changing Service Level Agreements.

**ONE IMPORTANT PATH IN DMI IS MOVING INTELLIGENCE TOWARD THE DEVICE END**

Leading the MI and IoT convergence will require intertwining the right competence in unique team setups, bridging system architects, embedded systems designers and distributed system engineers, as well as subject matter experts on MI, security, IoT protocols and systems optimization. At Ericsson, we are taking this multidisciplinary challenge seriously to ensure that we are equipped to apply DMI competently to generate business value in emerging IoT markets.
Conclusion

Rapid technology advances in recent years have been of great benefit to the ongoing realization of massive IoT devices. It is, however, vital for device manufacturers, mobile network operators and other industry players to carefully consider the options and make the right choices when applying new technologies in the device domain. From Ericsson’s perspective, there are five key technology areas that are of particular significance: connectivity, communication protocols, security, identity solutions and machine intelligence (MI).

In terms of connectivity, we are convinced that LTE-M and NB-IoT technologies will further enhance functionality and use-case applicability, improving the possibility to create devices with lower power consumption and a smaller form factor, making the right choices when applying new mass-efficient solutions for LPWAN access will emerge, leveraging the device’s built-in security capabilities. Finally, advances in MI technologies have made it possible to move intelligence toward the device end, which we regard as a great opportunity to minimize E2E latency, enhance data privacy and lower bandwidth requirements, while reducing server-side costs.

With regard to security, we believe that the implementation of cryptographic functions on the device is the optimal approach to achieving strong device security. TEEs will soon be applied to IoT devices to support use cases in which secure storage is crucial and isolation between functionality is required. It is also our view that the use of secure identities will soon become key, as a means to identify the origin of data and to realize secure connectivity. New cost-efficient solutions for LPWAN access will emerge, leveraging the device’s security features.

With regard to security, we believe that the implementation of cryptographic functions on the device is the optimal approach to achieving strong device security. TEEs will soon be applied to IoT devices to support use cases in which secure storage is crucial and isolation between functionality is required. It is also our view that the use of secure identities will soon become key, as a means to identify the origin of data and to realize secure connectivity. New cost-efficient solutions for LPWAN access will emerge, leveraging the device’s built-in security capabilities.

Finally, advances in MI technologies have made it possible to move intelligence toward the device end, which we regard as a great opportunity to minimize E2E latency, enhance data privacy and lower bandwidth requirements, while reducing server-side costs.

Further reading

- Ericsson white paper, March 2018, 5G security – enabling a trustworthy 5G system, available at: https://www.ericsson.com/en/white-papers/5g-security-enabling-a-trustworthy-5g-system
- Ericsson Research blog, March 2017, Smart contracts for identities, available at: https://www.ericsson.com/research-blog/smart-contracts-for-identities/

References

12. OMA SpecWorks, Lightweight M2M (LWM2M), available at: https://www.oma-specworks.org/what-is-oma-specworks/lwm2m/lwm2m
14. IETF, Software Updates for Internet of Things (suit), available at: https://datatracker.ietf.org/wg/suit/about/
the authors

Claes Lundqvist
◆ serves as director of Technology Foresight at Ericsson Group Function Technology. He joined Ericsson in 1996 and has held various positions in R&D and product management, working with technology platforms for mobile devices. His current work focuses on the technology management area, including technologies for mobile devices and the IoT. He holds an M.Sc. in communications engineering from Aalto University in Helsinki, Finland.

Ben Smeets
◆ is a senior expert in trusted computing at Ericsson Research. He holds a Ph.D. in information theory from Lund University, Sweden, where he also serves as a professor. He joined Ericsson Mobile Communications in 1998, and started out working on security solutions for mobile phone platforms. Smeets is currently working on trusted computing technologies in connection with containers and secure enclaves.

Ari Keränen
◆ is an expert in IoT standards and protocols at Ericsson Research in Finland. He joined the company in 2007 and has since worked with various internet technologies ranging from multimedia signaling and peer-to-peer systems to the IoT. He holds an M.Sc. in communications engineering from Aalto University in Helsinki, Finland.

Carlos R. B. Azevedo
◆ joined Ericsson Research’s Brazilian team in 2015. He currently serves as an MI and IoT technologies researcher at Ericsson Research in Stockholm, where he designs the architecture of intelligent, anticipatory and situation-aware systems. He holds a Ph.D. in electrical engineering from the University of Campinas in Brazil.

Peter von Wrycza
◆ joined Ericsson in 2011 and has held different positions in the areas of 3GPP standardization, 5G research and the IoT. He currently serves as head of IoT Technologies Research at Ericsson Research, where he drives the research,

John Fornehed
◆ joined Ericsson in 1991 and currently serves as an IoT export and technical director. He spent many years in Japan, where he was responsible for strategic accounts with mobile operators, among other things. Fornehed’s current work includes serving as an evangelist on IoT device life-cycle management, including secure IDs, for both industry and academia.

Peter von Wrycza
◆ joined Ericsson in 2011 and has held different positions in the areas of 3GPP standardization, 5G research and the IoT. He currently serves as head of IoT Technologies Research at Ericsson Research, where he drives the research.

Ben Smeets
◆ is a senior expert in trusted computing at Ericsson Research. He holds a Ph.D. in information theory from Lund University, Sweden, where he also serves as a professor. He joined Ericsson Mobile Communications in 1998, and started out working on security solutions for mobile phone platforms. Smeets is currently working on trusted computing technologies in connection with containers and secure enclaves.

John Fornehed
◆ joined Ericsson in 1991 and currently serves as an IoT export and technical director. He spent many years in Japan, where he was responsible for strategic accounts with mobile operators, among other things. Fornehed’s current work includes serving as an evangelist on IoT device life-cycle management, including secure IDs, for both industry and academia.

Carlos R. B. Azevedo
◆ joined Ericsson Research’s Brazilian team in 2015. He currently serves as an MI and IoT technologies researcher at Ericsson Research in Stockholm, where he designs the architecture of intelligent, anticipatory and situation-aware systems. He holds a Ph.D. in electrical engineering from the University of Campinas in Brazil.

Peter von Wrycza
◆ joined Ericsson in 2011 and has held different positions in the areas of 3GPP standardization, 5G research and the IoT. He currently serves as head of IoT Technologies Research at Ericsson Research, where he drives the research,