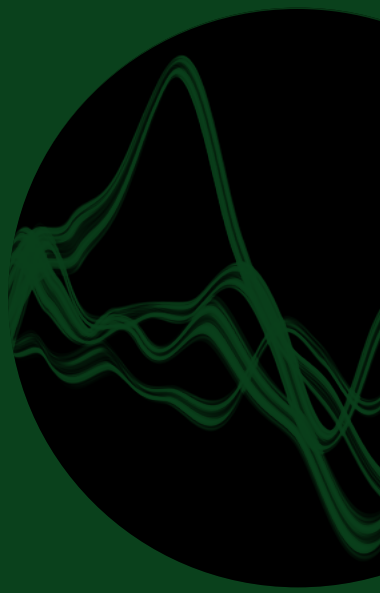
CHARTING THE FUTURE OF INNOVATION | VOLUME 98 I 2019-01

ERICSSON TECHNOLOGY Review



BOOSTING SMART MANUFACTURING WITH 5G

TECHNOLOGY CHOICES FOR MASSIVE IoT DEVICES

DISTRIBUTED CLOUD IN AUTOMOTIVE & INDUSTRY 4.0 USE CASES

ERICSSON

Ericsson Technology Review brings you insights into some of the key emerging innovations that are shaping the future of ICT. Our aim is to encourage an open discussion about the potential, practicalities, and benefits of a wide range of technical developments, and provide insight into what the future has to offer.

ADDRESS
Ericsson
SE-164 83 Stockholm, Sweden
Phone: +46 8 7190000

PUBLISHING
All material and articles are published on the Ericsson Technology Review website: www.ericsson.com/ericsson-technology-review

PUBLISHER
Erik Ekudden

EDITOR
Tanis Bestland (Nordic Morning)
tanis.bestland@nordicmorning.com

EDITORIAL BOARD
Håkan Andersson, Anders Rosengren, Mats Norin, Erik Westerberg, Magnus Buhrgard, Gunnar Thrysin, Håkan Olofsson, Dan Fahrman, Robert Skog, Patrik Roseen, Jonas Högberg, John Fornehed and Sara Kullman

FEATURE ARTICLE
Boosting smart manufacturing with 5G wireless connectivity by Kenneth Wallstedt, Fredrik Alriksson and Göran Eneroth

ART DIRECTOR
Liselotte Eriksson (Nordic Morning)

PRODUCTION LEADER
Susanna O'Grady (Nordic Morning)

LAYOUT
Liselotte Eriksson (Nordic Morning)

ILLUSTRATIONS
Jenny Andersen (Nordic Morning)

CHIEF SUBEDITOR
Ian Nicholson (Nordic Morning)

SUBEDITORS
Paul Eade (Nordic Morning)

ISSN: 0014-0171
Volume: 98, 2019

CAPITALIZING ON THE **POWER OF 5G**

■ **I FIND IT** deeply gratifying to witness the growing enthusiasm among mobile network operators (MNOs) around the globe about the massive growth opportunities that 5G represents for their businesses. In particular, 5G is now widely recognized as a prime enabling technology of the fourth industrial revolution, helping manufacturing companies leverage automation and data exchange technologies that require seamless communication between all the participants and components in industrial processes.

Using 5G effectively in the fully-connected factories of the future is the theme of the feature article in this issue of the magazine. Among other aspects, it explains how 5G can provide deterministic ultra-reliable low-latency communication to bring wireless connectivity to demanding industrial equipment, like industrial controllers and actuators.

Emerging industrial use cases in the automotive and manufacturing sectors, among others, are creating a variety of new requirements for networks and clouds. Our distributed cloud article explains how distributed cloud technology exploits key features in both 4G and 5G networks to enable an execution environment that ensures performance, short latency, high reliability and data locality.

Like 5G, the Internet of Things (IoT) is also playing a pivotal role in Industry 4.0, as well as in transforming business and society in a myriad of other ways. In light of this, MNOs need business support systems (BSS) that can handle IoT use cases, which often involve complex business situations, and optimize outcomes with minimal manual intervention. In this magazine we argue in favor of architectural changes to traditional BSS to fully integrate artificial intelligence.

As IoT use cases continue to grow and spread, it is critically important to take action to ensure that the devices are secure, both in terms of communication and data integrity end-to-end, from device to data usage. It is our opinion that certain key technology choices are necessary to achieve the desired device characteristics and create IoT devices that support the multitude of existing and emerging massive IoT use cases.

While 5G is highly relevant for many industrial (and other) applications that reach far beyond traditional telco, it is also designed to address a myriad of challenges within the traditional telco sphere. One example of this is the way it enables MNOs to overcome the challenge of capacity exhaustion caused by the rapidly increasing data consumption of their subscribers. Rather than densifying 4G networks with new sites, we recommend that operators use 5G technology to add new frequency bands at existing 4G sites.

Of course, one of the most critical aspects of a successful 5G deployment is the operator's ability to support user equipment, radio network, core network and management products that are manufactured by a multitude of device and network equipment vendors. Achieving this can be more difficult than it sounds, however. We propose a smart approach to 5G deployment in this issue that reduces network upgrade cost and time, simplifies interoperability between networks and devices, and enables a faster scaling of the 5G ecosystem.

I hope you will find the articles in this magazine valuable. Please feel free to share them with your colleagues and business associates. You can find all of the articles, along with those published in previous issues, at: www.ericsson.com/ericsson-technology-review

“5G IS NOW WIDELY RECOGNIZED AS A PRIME ENABLING TECHNOLOGY OF THE FOURTH INDUSTRIAL REVOLUTION”

ERIK EKUDDEN
SENIOR VICE PRESIDENT AND GROUP CTO

THE ADVANTAGES OF

Combining 5G NR with LTE

AT EXISTING SITES

5G at mid and high bands is well suited for deployment at existing site grids, especially when combined with low-band LTE. Adding new frequency bands to existing deployments is a future-proof and cost-efficient way to improve performance, meet the growing needs of mobile broadband subscribers and deliver new 5G-based services.

FREDRIC KRONESTEDT,
HENRIK ASPLUND,
ANDERS FURUSKÄR,
DU HO KANG,
MAGNUS LUNDEVALL,
KENNETH WALLSTEDT

The speed expectations and data consumption of mobile broadband (MBB) subscribers continue to grow rapidly. Already today, there are 4G networks in urban areas that are being densified with new sites (macro sites, small cells and indoor solutions, for example) as a result of spectrum exhaustion. Further, in regions such as western Europe and North America, the data demand per smartphone is projected to grow by 30-40 percent yearly [1], resulting in a four- to fivefold increase in five years. Adding new frequency bands at existing sites is a cost-efficient way to meet this demand and improve performance. The ability to achieve indoor coverage is particularly important, because the majority of the traffic is generated indoors [2].

■ Many people in the telecom industry tend to associate the deployment of high-frequency bands with poor coverage, which results in the need for new sites, which leads to high deployment costs. This is, however, not at all the case for 5G New Radio (NR) [3]. 5G NR is designed to make use of frequency bands above 3GHz and offers the possibility to introduce new frequency bands – typically above 3GHz – into existing 4G networks. Taking advantage of this possibility makes it easier to meet the increasing demands from MBB-based services, while simultaneously ensuring that site and backhaul infrastructure investments can be reused. 5G NR is also available for use in new bands below 1GHz and existing 3G/4G bands. Smooth migration from 4G to 5G in existing spectrum in a RAN can be done by means of spectrum sharing, where NR is introduced in parallel with LTE.

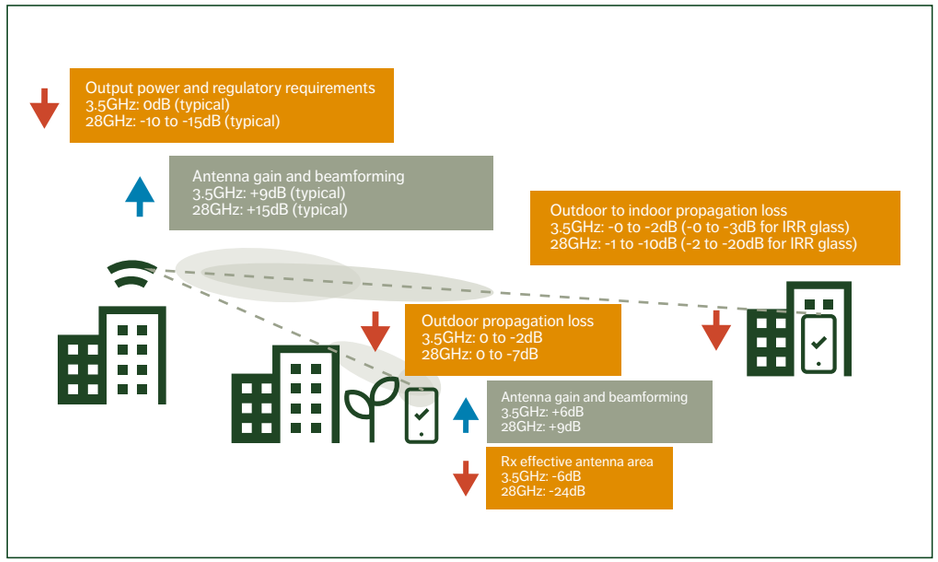


Figure 1 Schematic indication of antenna and propagation factors affecting downlink coverage positively (blue) or negatively (red) compared to coverage at a reference frequency of 1.8GHz. The numbers are indicative and may vary.

The main new NR frequency bands will typically be allocated as TDD in the mid (3-6GHz) and high (24-40GHz) bands. These bands present several interesting challenges and opportunities. By means of measurements and radio network simulations of coverage and capacity, we have demonstrated that it is feasible to deploy both mid and high (also known as millimeter Wave or mmWave) bands on existing sites.

Thanks to beamforming, a fundamental technique in NR, the need for site densification is much smaller than anticipated – particularly when interworking with LTE is applied. Beamforming and massive multiple-input, multiple-output (MIMO) techniques also provide higher capacity from existing 4G sites, which creates room for new 5G-based services and use cases in addition to MBB.

High-frequency challenges and opportunities

The use of mid and high bands for 5G makes it possible to utilize much higher bandwidths. However, the increased carrier frequency can also make it more challenging to provide coverage that is similar to existing low-band deployments. There are three primary reasons for this: (1) physical limits on the power reception capabilities of antennas; (2) radio frequency output power limitations; and (3) increased propagation losses, as shown in *Figure 1.*

> “THANKS TO BEAMFORMING ... THE NEED FOR SITE DENSIFICATION IS MUCH SMALLER THAN ANTICIPATED”

> THE HIGHER PROPAGATION LOSSES CAN BE MITIGATED BY USING HIGH-GAIN ANTENNAS

But the higher frequencies also allow higher antenna gains to be generated without increasing physical antenna size. 5G can utilize these increased antenna gains through beamforming both at the transmitter and at the receiver, which helps mitigate the impact on coverage at higher frequencies.

Additionally, increasing the frequency will allow the antennas to become smaller while maintaining the same antenna gain. It is important to note that any fixed-gain antenna in receiving mode actually captures 20dB less energy for each tenfold increase of the frequency. This is often misunderstood as a propagation loss, when in reality it is a result of a decreasing effective antenna area. If the physical antenna area of the antenna is maintained, its power capture capabilities become independent of frequency, while its antenna gain, for both reception and transmission, grows with the frequency at the same time as the beam width becomes smaller. Thus, at higher frequencies, there is a trade-off between reducing the antenna size and increasing the antenna gain. Coverage and implementation aspects determine the sweet spot.

The achievable output power at higher frequency bands such as mmWave frequencies can also be limited by power amplifier technology and by regulatory requirements [4]. Theoretically, the antenna gain of a fixed-size transmitting antenna would grow by 20dB per decade in frequency (dB/decade), but in practice the increase in EIRP (effective isotropic radiated power) may be smaller due to such constraints.

Electromagnetic wave propagation in cellular networks involves some processes that are strongly frequency-dependent, such as diffraction or transmission through, for example, walls or foliage, but also others such as free space propagation and reflection or scattering that show little to no difference over frequency. Effectively, the outdoor propagation loss is similar or increases slightly with increased frequency, as indicated in Figure 1.

Outdoor-to-indoor propagation losses can be challenging to overcome, especially for buildings equipped with thermally-efficient window glass, which can add up to 20-40dB of additional loss at a given frequency. When increasing the frequency, the outdoor to indoor losses also tend to increase, particularly for deep indoor locations. This increase is small to moderate for regular buildings but can be strong for thermally-efficient buildings, as shown in Figure 1.

The higher propagation losses can be mitigated by using high-gain antennas on both transmitters and receivers. These antennas become directive, forming beams with strong gain in certain directions, and low gain in other directions. The beams need to be set up and maintained to point in the right directions in order to support mobility. In NR, this is supported by beam management. Besides the benefit of amplifying the signal in the desired direction, beamforming also attenuates the signal in other directions, leading to less interference and better channel quality. This can be done to the extent that multiple users, using different beams, can communicate with a base station on the same frequency and time resource. This is known as multi-user MIMO (MU-MIMO), and it enables a significant capacity improvement.

Even with beamforming, using existing site grids, it can be difficult to reach full coverage on higher frequencies. But since a lower frequency band tends to be available, this is not a problem. Users out of coverage on the higher frequencies simply fall back to the lower frequency bands. This can be accomplished by interworking techniques such as dual connectivity or carrier aggregation. The result is a 'forgiving' situation, where a mid- or high-band deployment does not need to be dimensioned for 100 percent coverage. Instead, it simply takes care of the traffic that it covers.

To summarize, the numbers in Figure 1 illustrate that the use of today's technologies, power levels and beamforming gains on the mid band (3-6GHz) provides better downlink (DL) coverage than the 1.8GHz reference. Even so, users in the worst positions require the support of a lower frequency band, especially in the uplink (UL) direction. For high bands (around 30GHz), the situation differs substantially from the reference. Very good outdoor coverage is achieved on existing grids. Outdoor-to-indoor coverage can be achieved by targeted deployments with line-of-sight to the buildings intended to be covered.

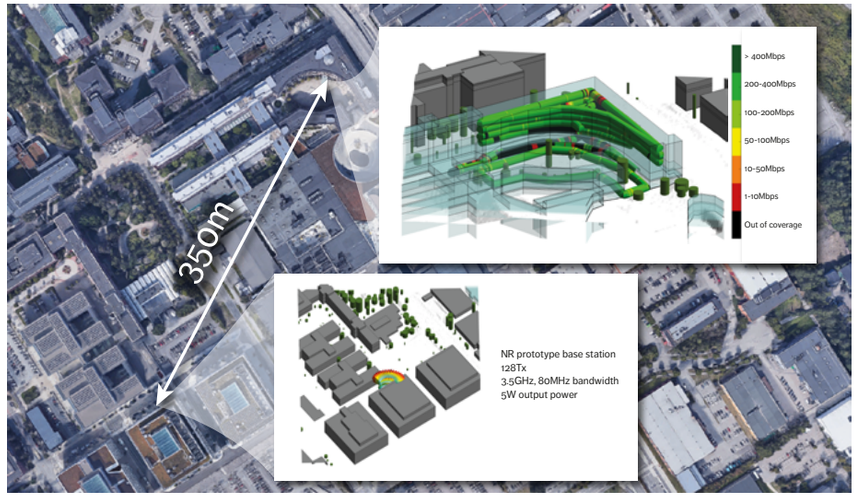


Figure 2 5G outdoor-in throughput measurement results from an NR 3.5GHz radio prototype

Measured beamforming performance and outdoor-in coverage

Early proof points of the 5G concept and its performance can be obtained from measurements in a radio network prototype. Ericsson has developed 5G prototypes for several 5G frequencies, including 3.5GHz and 28GHz. Initial trial deployments are typically set up with a few radio sites and one or a few mobile terminals, allowing for a controlled measurement environment. Test results on beamforming performance are reported in references [5], [6] and [7]. The results demonstrate that high antenna gains can indeed be realized through beamforming, and that the beamforming is able to track fast-moving users with sustained communication quality. Moreover, good indoor coverage can be achieved with 5G at 3.5GHz, proving the feasibility of deploying 5G at existing 4G sites. One example from our measurements is shown in ***Figure 2***, where indoor throughput in a building at the cell edge reaches 200-400Mbps on an 80MHz carrier using conservative rank-2 MIMO transmission.

> THE MID BAND (3-6GHZ) PROVIDES BETTER DOWNLINK COVERAGE THAN THE 1.8GHZ REFERENCE

BENEFITS OF OVERLAYING 5G NR 3.5GHZ AT EXISTING SITES

- Better user data speeds – 95 percent of indoor subscribers have more than 200Mbps with today's typical site grids.
- Higher capacity – adding NR 100MHz TDD (75 percent DL) on top of LTE with 2x50MHz paired spectrum provides an eight times higher DL capacity than using only LTE. Normalized with the 1.5 times higher spectrum usage, NR is thus five times more efficient.

Predicted urban mid-band coverage and capacity

To predict 5G coverage and capacity on a larger scale, we have performed radio network simulations. We chose a part of central London with an inter-site distance of approximately 400m, which is representative of many European urban areas. Similar studies of major cities in other parts of the world, including Asia and the US, indicate that the findings from this study are also applicable in those scenarios. Radio base station characteristics such as beamforming capabilities, power and sensitivity reflect the implementations of the first product generations, and terminals are modeled with expected typical smartphone characteristics for mid and high bands. Four and 32 receive antennas are assumed for terminals in mid and high bands, respectively. For maximal fidelity, a digital 3D map is used together with an accurate 3D site-specific propagation model, explicitly capturing relevant propagation phenomena along the propagation paths [8].

We have modeled LTE systems operating at 800MHz, 1.8GHz and 2.6GHz, as well as an NR system operating at 3.5GHz. This configuration is representative of the non-standalone version of NR that was developed in 3GPP Rel-15. The LTE system uses FDD, 2x10MHz at 800MHz and 2x20MHz at each of 1.8GHz and 2.6GHz adding up to 100MHz paired spectrum, and regular sector antennas. The NR system uses TDD, 100MHz of unpaired spectrum, and a 64T64R antenna array of 8x8 cross-polarized antennas. We applied user-specific digital beamforming, and MU-MIMO with multiplexing of up to four users is supported both in the DL and UL. When LTE and NR systems are evaluated together, carrier aggregation between LTE systems and dual connectivity between LTE and NR carriers are applied for the interworking. Although not considered in this evaluation, there are several interesting possibilities to evolve the LTE systems – with more advanced antennas, for example.

Figure 3 shows DL and UL coverage in terms of achievable data rates in an unloaded network without interference. Eighty percent of users are indoors, and they are shown only from middle floors. When existing LTE rooftop sites are reused with 3.5GHz, both indoor and outdoor users have very good coverage in the DL. The black line in the color bar indicates that 95 percent of the indoor subscribers have coverage for 200Mbps in the DL compared with 50Mbps when aggregating all LTE systems (not shown in the figure). In addition, 95 percent of outdoor users can exceed 500Mbps in the DL with NR 3.5GHz alone. The UL is much more limited with 3.5GHz alone. NR-LTE interworking improves, and many of the blank spots in the 3.5GHz band are covered. The remaining areas with poor coverage are concentrated to inside large buildings with high-loss outer walls. These buildings are suitable candidates for indoor deployments. Comparing the gains from adding 3.5GHz in the DL and UL, it is clear that the gains are larger in the DL. This is due to a DL-heavy TDD asymmetry (75 percent) at 3.5GHz, and the fact that the UL, because of the lower transmit power, is more power-limited and thus gains less from additional bandwidth.

When traffic load increases, more users are active simultaneously, sharing the base station capacity, causing increased interference levels, and leading to a reduction in user throughput compared with the unloaded case. These effects are mitigated by the NR system, using a wider bandwidth, beamforming and MU-MIMO. The ability to serve users in poor coverage areas on a lower band avoids the consumption of extensive resources on the 3.5GHz band, making it more efficient. To quantify the benefit of introducing NR, we measured the maximum traffic load for which (95 percent of) the users still achieve a user throughput exceeding 20Mbps. When adding NR in the DL direction, this maximum traffic load or 'capacity' increases by a factor of eight from 1Gbps/km^2 to 8Gbps/km^2 (corresponding to 135GB/subscriber/month, assuming 10,000 subscribers per km^2 and a busy hour traffic of 8 percent of the daily traffic). In the UL direction, the capacity gain is smaller than the DL due to TDD asymmetry (25 percent for the UL) and a lower transmit power. The capacity gains observed here are typical for a low-rise urban scenario with decent coverage. The gains are scenario-dependent and typically increase with improved coverage and increased vertical spread of users, and decrease with worse coverage and a smaller vertical spread.

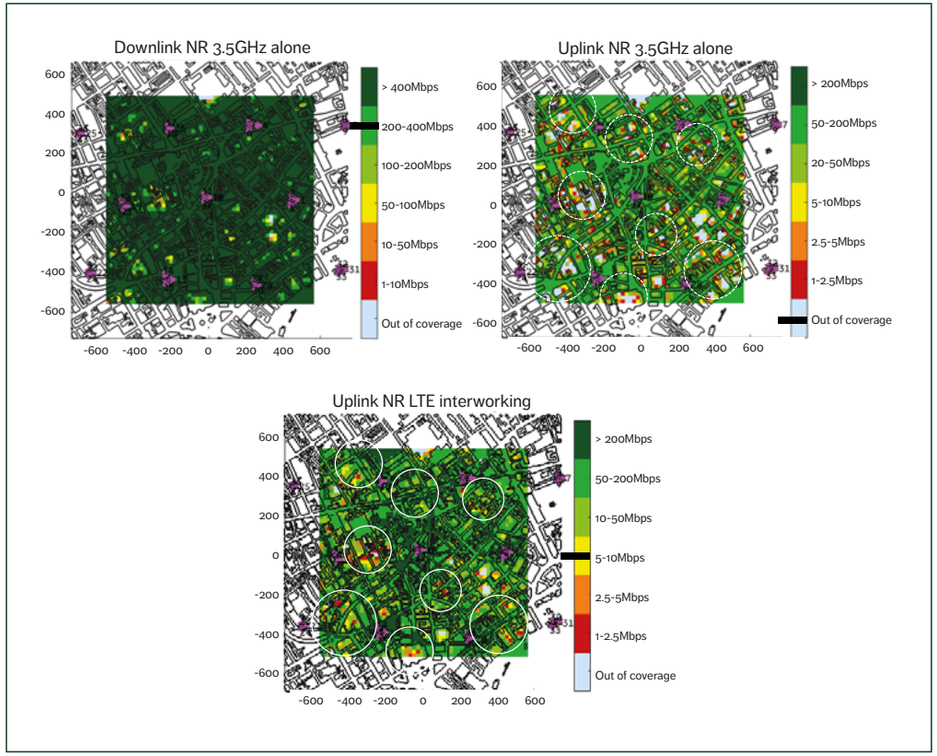


Figure 3 DL and UL coverage maps. The black line in the color legends represents the fifth percentile of an indoor user data rate, and the purple areas indicate antenna positions. The white circles mark indoor areas with limited coverage, improved by interworking.

> APPLYING MMWAVE SPECTRUM AT STREET-LEVEL SITES CAN ALSO BE A GOOD ALTERNATIVE

Predicted urban high-band coverage and capacity

The wide bandwidth available on mmWave spectrum can provide further increased data rates and additional capacity on top of the combined 3.5GHz mid-band NR and LTE system. Higher frequencies allow a higher gain of antenna array at the same physical area – both in a base station and at a user terminal side – so as to increase the maximum antenna gain.

Simulation studies in the central London scenario show that an NR 200MHz TDD system at 26GHz with the 256T256R antenna array of 16x16 cross-polarized antennas can provide very good DL coverage to outdoor users – for example, 50-60 percent approaching 1Gbps. With larger spectrum allocations such as 400MHz, it is possible to reach multi-Gbps speeds. When there is line-of-sight from the base station to a building and the building is a low-loss type, there is also a good chance that indoor users will be well covered.

Our results show that deploying the 3.5GHz and 26GHz band on existing macro sites can provide a capacity improvement of approximately 10 times compared with the LTE systems in low and mid bands. This additional gain is because 26GHz offloads the lower frequency bands by letting good-coverage users utilize an additional 200MHz, which thereby improves overall performance.

Applying mmWave spectrum at street-level sites can also be a good alternative. By placing antennas on lampposts, outer walls and the like, it is possible to avoid typical diffraction losses from rooftops and achieve shorter distances to users on outdoor hotspots or in targeted buildings. Our simulation studies in the London scenario indicate that the street-level radio deployment of an NR system with 64T64R antenna array provides good coverage both in nearby outdoor areas and for indoor users in low-loss buildings with line-of-sight to the base station.

Suburban and rural deployment considerations

Despite the typically larger cells in suburban and rural scenarios, it is possible to achieve similar results to those that we have seen in urban scenarios due to differences in the radio propagation. While the urban environment is characterized by relatively low antennas, frequent large obstacles and large, highly attenuating buildings, the suburban and rural environments have taller antennas, fewer obstacles and smaller sized buildings with wall types that are easier to penetrate. This compensates for the differences in cell range, and as a result it is typical to achieve very good performance in suburban and rural scenarios as well.

Indoor deployments

In-building deployments play a central role in providing good indoor performance in many parts of the world today. Large buildings with high building entry losses are an example of a coverage-driven in-building deployment, whereas a crowded public venue like a train station or a stadium would be a good example of a capacity-driven one. Passive distributed antenna systems (DASs) are currently the most common solution used for indoor deployments.

The hardware components of a passive DAS often have an operating frequency range that is limited to bands below 3GHz, which means that adding the new NR mid or high bands requires a new 5G indoor solution. A radio dot [9] solution at 3.5GHz provides good coverage and much higher speeds than current LTE bands at the same radio node density, as well as consuming less power than a DAS. For extreme demands in terms of user speeds or capacity, an indoor solution based on mmWave small cells might be the best choice. In this case, it is important to deploy a mid-band coverage complement.

Conclusion

The key benefits of deploying 5G New Radio with mid bands (3-6GHz) at existing 4G sites are that doing so results in a significant performance boost and allows for maximal reuse of site infrastructure investments. By adding NR with 100MHz unpaired spectrum, it is possible to achieve eight times higher downlink capacity relative to LTE (2x50MHz paired spectrum) along with improved downlink data rates – both outdoors and indoors – by means of massive MIMO techniques such as beamforming and multi-user MIMO. Uplink coverage deep indoors is maintained through interworking with LTE and/or NR on low bands using dual connectivity or carrier aggregation (new, refarmed or by using LTE/NR spectrum sharing). As a result of these possibilities in 5G NR, growing data demands can be met with limited site densification.

> IT IS POSSIBLE TO ACHIEVE EIGHT TIMES HIGHER DOWNLINK CAPACITY RELATIVE TO LTE

Further speed and capacity increases can be attained by deploying 5G NR at high bands (26-40GHz), also known as mmWaves. The high bands are particularly effective outdoors and inside buildings with line-of-sight from the deployed radio node and with low wall loss properties. Buildings that have or need dedicated indoor solutions due to high penetration loss and interior losses can be successfully upgraded with upcoming NR bands for higher speeds and capacity at similar radio node density to those used for LTE in-building deployments today.

Terms and abbreviations

DAS – Distributed Antenna System | **DL** – Downlink | **IRR** – Infrared Reflective | **MBB** – Mobile Broadband | **MIMO** – Multiple-input, Multiple-output | **mmWave** – Millimeter Wave | **MU-MIMO** – Multi-User Multiple-input, Multiple-output | **NR** – New Radio | **RAN** – Radio Access Network | **Rx** – Radio Receiver | **Tx** – Radio Transmitter | **UL** – Uplink

References

1. **Ericsson Mobility Report, June 2018, available at:** *https://www.ericsson.com/en/mobility-report/reports/june-2018*
2. **Ericsson ConsumerLab report, Liberation from Location, October 2014, available at:** *https://www.ericsson.com/res/docs/2014/consumerlab/liberation-from-location-ericsson-consumerlab.pdf*
3. **5G NR: The Next Generation Wireless Access Technology, 1st Edition, August 2018, Dahlman, E; Parkvall, S; Sköld, J, available at:** *https://www.elsevier.com/books/5g-nr-the-next-generation-wireless-access-technology/dahlman/978-0-12-814323-0*
4. **GSMA, 5G, the Internet of Things (IoT) and Wearable Devices: What do the new uses of wireless technologies mean for radio frequency exposure?, September 2017, available at:** *https://www.gsma.com/publicpolicy/wp-content/uploads/2017/10/5g_iot_web_FINAL.pdf*
5. **IEEE, Beamforming Gain Measured on a 5G Test-Bed, June 2017, Furuskog, J; Halvarsson, B; Harada, A; Itoh, S; Kishiyama, Y; Kurita, D; Murai, H; Simonsson, A; Tateishi, K; Thurfjell, M; Wallin, S, available at:** *https://ieeexplore.ieee.org/document/8108648/*
6. **IEEE, High-Speed Beam Tracking Demonstrated Using a 28 GHz 5G Trial System, September 2017, Chana, R; Choi, C; Halvarsson, B; Jo, S; Larsson, K; Manssour, J; Na, M; Singh, D, available at:** *http://ieeexplore.ieee.org/document/8288043/*
7. **IEEE, 5G NR Testbed 3.5 GHz Coverage Results, June 2018, Asplund, H; Chana, R; Elgcrona, A; Halvarsson, B; Machado, P; Simonsson, A, available at:** *https://ieeexplore.ieee.org/document/8417704/*
8. **Proceedings of the 12th European Conference on Antennas and Propagation (EuCAP 2018), A set of propagation models for site-specific predictions, April 2018, Asplund, H; Johansson, M; Lundevall, M; Jaldén, N,**
9. **Ericsson Radio Dot System, available at:** *https://www.ericsson.com/ourportfolio/radio-system/radio-dot-system*

Further reading

- **5G deployment considerations, available at:** *https://www.ericsson.com/en/networks/trending/insights-and-reports/5g-deployment-considerations*
- **Massive MIMO increasing capacity and spectral efficiency, available at:** *https://www.ericsson.com/en/networks/trending/hot-topics/5g-radio-access/massive-mimo*
- **Going massive with MIMO, available at:** *https://www.ericsson.com/en/news/2018/1/massive-mimo-highlights*
- **Superior indoor coverage with 5G Radio Dot, available at:** *https://www.ericsson.com/en/networks/offerings/5g/5g-supreme-indoor-coverage*

THE AUTHORS

Fredric Kronestedt

◆ joined Ericsson in 1993 to work on RAN research. Since then he has taken on many different roles, including system design and system management. He currently serves as Expert, Radio Network Deployment Strategies, at Development Unit Networks, where he focuses on radio network deployment and evolution aspects for 4G and 5G. Kronestedt holds an M.Sc. in electrical engineering from KTH Royal Institute of Technology, Stockholm, Sweden.

Henrik Asplund

◆ received his M.Sc. in engineering physics from Uppsala University, Sweden, in 1996, and joined Ericsson the same year. His current position is Master Researcher, Antennas and Propagation, at Ericsson Research, with responsibility for propagation measurements and modeling within the company and in cooperation with external organizations such as 3GPP and ITU-R. He has been involved in propagation research supporting predevelopment and standardization of all major wireless technologies from 2G to 5G.

Kenneth Wallstedt

◆ is Director, Technology Strategy, in Ericsson's CTO office, where he focuses on the company's radio and spectrum management strategy. He joined Ericsson in 1990 and since then he has held various leading positions in Ericsson's research, development and market units in Canada, Sweden and the US. He holds an M.Sc. in electrical engineering from KTH Royal Institute of Technology in Stockholm, Sweden.

Du Ho Kang

◆ joined Ericsson Research in 2014 and currently serves as a Senior Researcher. He holds a Ph.D. in radio communication systems

from KTH Royal Institute of Technology, Sweden, and an M.Sc. in electrical and electronics engineering from Seoul National University, South Korea. His expertise is concept developments of 4G/5G radio networks and performance evaluation toward diverse international standardization and spectrum regulation bodies including 3GPP RAN, CBRS alliance, Multifire alliance (MFA), ETSI BRAN and ITU-R. Kang's particular interest at present is developing solution concepts for internetworking and massive MIMO for 5G base station products.

Magnus Lundevall

◆ is Expert, Radio Network Performance, in Ericsson's R&D organization, where he currently focuses on 5G radio network deployment and evolution strategies. He joined Ericsson in 1998 and has 20 years of experience in radio network modeling, simulation and performance analysis. He holds an M.Sc. in electrical engineering from KTH Royal Institute of Technology in Stockholm, Sweden.

Anders Furuskär

◆ joined Ericsson Research in 1997 and is currently a senior expert focusing on radio resource management and performance evaluation of wireless networks. He holds an M.Sc. in electrical engineering and a Ph.D. in radio communications systems, both from KTH Royal Institute of Technology in Stockholm, Sweden.

Simplifying the 5G ecosystem

BY REDUCING ARCHITECTURE OPTIONS

Previous mobile generations have taught us that industry efforts to reduce fragmentation yield massive benefits. In the case of 5G, an industry effort to focus deployment on a limited set of key connectivity options will be critical to bringing it to market in a timely and cost-efficient way.

TORBJÖRN CAGENIUS,
ANDERS RYDE,
JARI VIKBERG,
PER WILLARS

The multiple connectivity options in the 3GPP architecture for 5G have created several possible deployment alternatives. Initial deployments focus on options 3 (non-standalone New Radio) and 2 (standalone New Radio). However, the deployment of several additional options would create a level of complexity that impacts the whole 5G ecosystem – across operator network operations, equipment vendors and user equipment (UE) chipset vendors as well as spectrum assets. To avoid ecosystem fragmentation, we believe that the best approach is to limit the number of options that are deployed.

There is much more to introducing 5G than simply deploying New Radio (NR) technology. For a successful 5G launch, the operator needs to secure a network that includes end-to-end (E2E) capabilities aligned across devices, RAN, core and management systems. 5G is also a technology transformation for operators striving for more flexibility and speed in network deployment – and with an expectation of being able to address new business opportunities with use cases beyond mobile broadband (MBB). One of the key strategic topics that operators need to decide on is which connectivity options to support in the network to address the targeted use cases.

5G connectivity options

In Release 15, the 3GPP [1] has defined multiple architectural options for a UE to connect to the network, using LTE/eLTE and/or NR access to connect to Evolved Packet Core (EPC) or 5G Core (5GC) networks. A new use of dual connectivity has also been applied to use LTE/eLTE and NR as the master or secondary radio access technology (RAT) in different combinations. This has resulted in six connectivity options for a UE, as shown in ***Figure 1***. Note that while the option terminology is not explicitly used in the 3GPP standards specifications, it originates from the 5G study phase of 3GPP Release 15 and is widely used in the industry.

Connectivity option	Core network	Master RAT	Secondary RAT	3GPP term	3GPP release
Option 1	EPC	LTE	-	LTE	Rel. 8
Option 3	EPC	LTE	NR	EN-DC	Rel. 15, Dec 2017
Option 2	5GC	NR	-	NR	Rel. 15, June 2018
Option 4	5GC	NR	eLTE	NE-DC	Rel. 15, March 2019
Option 5	5GC	eLTE	-	eLTE	Rel. 15, June 2018
Option 7	5GC	eLTE	NR	NGEN-DC	Rel. 15, March 2019

Figure 1 UE connectivity options

The six connectivity options shown in Figure 1 define how any single UE is connected to the network at a given time. In most cases, a network will support a set of such options simultaneously. One base station may have different UEs connected via different connectivity options, as well as moving a UE connection between the options depending on factors such as radio conditions. Legacy LTE/EPC (option 1) is the baseline, and the industry has an aligned view that the initial 5G deployments are based on options 3 and 2. The next step, therefore, is to establish industry alignment on the potential use of options 4, 5 and 7.

The need for industry alignment

Mobile network operators that deploy 5G must be able to support UE, radio network, core network and management products that are manufactured by a multitude of device and network equipment vendors. With multiple connectivity options, and even more possible combinations of options, there is a high risk that different operators will deploy different options, in a different order. If that happens, chipset, device and network equipment vendors are likely to get contradictory requirements from different operators or markets. This would cause significant product and integration complexity, as well as creating interoperability issues that prolong the time it takes to establish a complete ecosystem that supports the deployed options.

The complexity caused by a multitude of deployed connectivity options would also have an impact on the E2E testing of services in the operator network,

including both existing services like voice as well as new ones. Further, the higher the number of options deployed, the more complex and time consuming it will be for the operator community to establish 5G roaming in the industry.

Network deployments based on options 3 and 2

Option 3 is the best short-term alternative for 5G deployment, as it relies on existing LTE/EPC (option 1). Option 3 will provide good performance in several aspects, allowing optimized transmission on NR when NR coverage is good, extending NR downlink (DL) usage on a higher band by combining with a lower-band LTE for uplink (UL) data, and, if needed, aggregating throughput over both NR and LTE spectrum. It also provides reliable and smooth mobility based on anchoring in LTE/EPC, even if the NR coverage is spotty. The use of dual connectivity has, however, introduced some challenges on the UE side with dual transmitters, which, in some cases, will limit performance and coverage.

One of the main drivers for going beyond option 3 is to provide 5GC-enabled capabilities like enhanced network slicing, edge computing support and operational benefits, even though EPC can also support these services to some extent (slicing based on DECOR, for example). Another main driver for going beyond option 3 is to be able to deploy standalone NR and get the radio performance benefits of an NR-only based radio interface. Option 2 (standalone NR) is the first 5GC-based option available in UEs and networks.

Even if general NR coverage is limited, option 2 can initially be deployed for specific use cases in local areas, where devices stay within good NR coverage on a mid or high band. Examples include industrial deployments with ultra-reliable low latency communication requirements, and fixed wireless access (FWA), even if the latter is also well served via option 3.

Key enablers

- **LTE-NR spectrum sharing**
  3GPP specifications allow efficient sharing of operator spectrum, so that one carrier appears as an NR carrier to NR UEs, and an LTE carrier to LTE UEs. Resources are pooled and distributed dynamically between the two RATs, according to instant needs. There is no impact on legacy LTE UEs, and the impact on LTE capacity is very small. Compared with classic refarming, this provides a smooth migration of spectrum from LTE to NR as NR-capable UE penetration increases, enabling NR to be rolled out on new and legacy bands.
- **Spectrum regulation**
  Spectrum is becoming technology neutral in most of the world except for a few markets and frequency bands where the spectrum license is currently tied to a specific RAT, prohibiting NR to operate in existing frequency bands. It is important that regulators acknowledge the need for NR deployment in all bands. This is a key enabler for migration to wide area coverage of services like MBB/voice and cMTC over 5G, depending on the possibility to deploy NR in lower frequency bands.
- **Dual-mode core network**
  4G devices will be the major device type and traffic consumer for a long time [2]. In addition, operators are introducing new 5G devices depending on both EPC (option 3) and 5GC (option 2). A "dual-mode" core network with both EPC and 5GC functionality will support the evolving device fleet in the network and enable a smooth network transformation. To ensure service coverage during the migration period, the dual-mode core network will provide tight interworking between EPC and 5GC for seamless 4G-5G mobility.

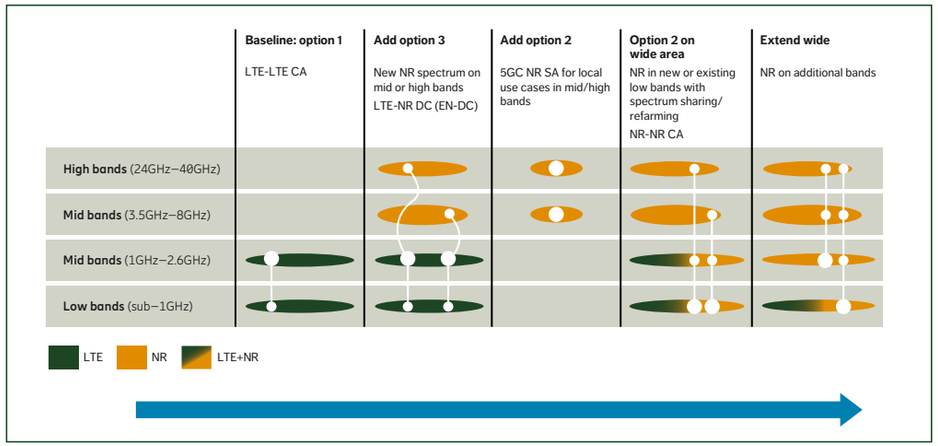


Figure 2 Spectrum migration steps for the 5G network

Figure 2 illustrates the evolution of spectrum usage in a network, starting with LTE deployed on sub-1GHz and 1–3GHz bands. First, NR is deployed on 3.5GHz and/or mmW and with LTE bands using option 3. The next step is to deploy option 2 for specific use cases in local areas – such as for FWA and industrial deployments.

Expanding standalone NR coverage and capacity

When deploying option 2 for wide-area use cases like MBB, it is important to ensure continuous NR coverage within the targeted area (initially urban for example). Spotty NR coverage would result in frequent mobility events between NR and LTE for wide-area use cases, even though intersystem mobility between option 2 and LTE/EPC will be well supported. For these use cases, option 2 requires a sufficiently low NR band in relation to the site grid. In many cases, the site grid for a 3.5GHz deployment will give good DL coverage both outdoors and indoors, but not enough UL coverage. NR on 3.5GHz should therefore typically be combined with NR on low band to provide continuous coverage in both the UL and DL [3]. The low NR band can be new, refarmed or an existing LTE band that is shared between NR and LTE. With refarming or sharing, a key enabler is that the spectrum license allows NR deployment (see fact box on page 4, spectrum regulation).

To support option 2 for MBB in an area, it is also advisable to deploy NR in one or more legacy LTE bands using LTE-NR spectrum sharing (see fact box on page 4, LTE-NR spectrum sharing). Together with NR on low and mid/high bands, this maximizes the throughput via NR carrier aggregation (CA). This is essential to provide good MBB performance, especially in areas without DL coverage from new NR bands. While NR deployment is limited, mobility to option 2 should only be triggered when the UE

> WITH REFARMING OR SHARING, A KEY ENABLER IS THAT THE SPECTRUM LICENSE ALLOWS NR DEPLOYMENT

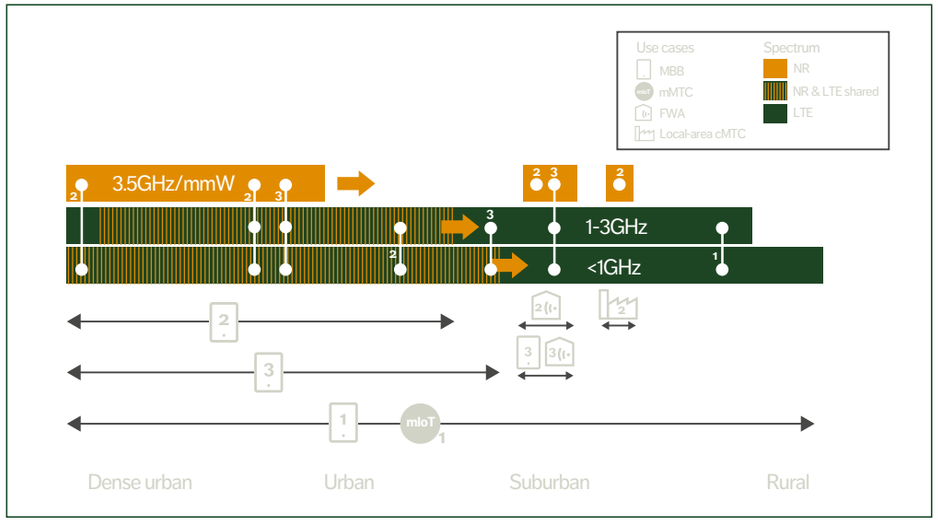


Figure 3 Network deployment during migration, including supported use cases

has enough coverage of sufficient NR spectrum, which can be handled with thresholds and offsets. The possibility of aggregating bands using CA, with a single UL transmitter in the UE, is an important benefit of option 2, compared with the dual connectivity used in options 3, 4 and 7. The third step of Figure 2 shows the use of NR on multiple legacy LTE bands using LTE-NR spectrum sharing.

Options 1 and 3 provide good support for smartphones and MBB. Moving MBB traffic to option 2 requires support for voice telephony. This means that NR must be able to support voice natively, as well as supporting seamless mobility via handover to LTE/EPC when leaving the option 2 coverage area. As an intermediate step before NR supports (and is dimensioned for) voice, the voice service can rely on EPS fallback to LTE/EPC. Tight 5GC-EPC interworking is needed for both voice solutions, and this will also provide good intersystem mobility for other services (see fact box on page 4, dual-mode core network).

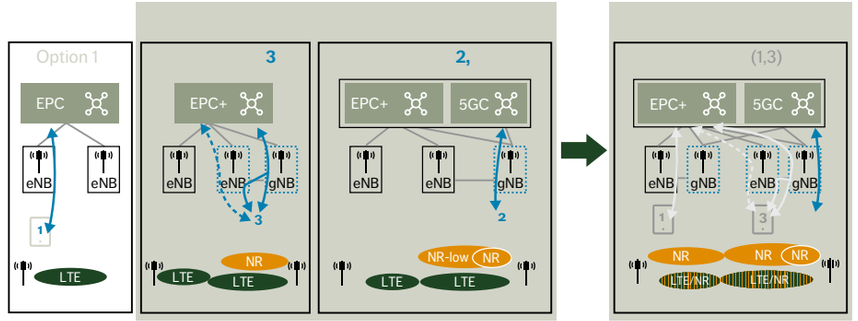
When a UE leaves an area where the NR coverage is not good enough for option 2, the network can trigger intersystem mobility to EPC, either to option 3 or 1. The support of option 2 can thus be extended gradually in ever-larger areas in an operator's network, starting with dense urban areas. By deploying option 2 in high-traffic areas first, a significant amount of traffic can be migrated from EPC to 5GC, even if the geographic coverage is initially more limited.

Many LTE sites will be modernized with more advanced radios for improved performance (such as 4T4R) or by adding modern baseband hardware, and will then typically be prepared to support NR on the LTE bands. The deployment of option 2 in a RAN capable of option 3 is then done with a software upgrade. The same gNB will serve some UEs in the same NR cell with option 3 and others with option 2.

Figure 3 illustrates network deployment during the migration from LTE to NR. In selected urban areas, NR (in orange) is deployed on 3.5GHz or mmW to add capacity to the network. NR is also deployed, on a sub-1GHz band (to complement UL coverage), and in legacy LTE bands with LTE-NR spectrum sharing.

The horizontal black lines in Figure 3 represent the coverage of options 1, 2 and 3. Option 1 is used in a large part of the network to support MBB and act as the main solution for Massive Machine Type Communication (mMTC) – specifically, Narrowband Internet of Things (NB-IoT) and LTE-MTC standard (LTE-M). Option 3 can be used anywhere there is NR coverage. Early option 2 deployments in local areas include FWA and industrial deployments. Option 2 for general MBB is supported where there is low-band NR and sufficient NR bandwidth (mid/high band and/or on 1-3GHz).

The orange arrows in Figure 3 indicate that areas of good NR coverage are expanded geographically, covering more urban areas, and in time also extending into suburban areas and beyond.



Standalone NR and 5GC with appropriate features and coverage for addressed use cases.

Full 5GC and NR coverage NR+5GC mainstream LTE/EPC for legacy devices

Figure 4 Migration steps toward target architecture

With the support of options 1, 2 and 3, key use cases such as mMTC, MBB and industrial critical-MTC (cMTC) will be supported in the near- and mid-term with good performance.

Target architecture

The industry has specified a new radio access technology – NR – and a new core network – 5GC – as the foundation for the evolution of 3GPP networks, which, in our view, makes option 2 the long-term target architecture for the industry. In the long-term target network, option 2 is deployed with wide coverage, used broadly in most devices, and should be the basis for future investments and feature growth.

Figure 4 illustrates the migration steps to the 5G target architecture for the mobile industry, recognizing option 2 as the long-term target. The first step is to add option 3, followed by option 2 in selected areas. By gradually expanding the areas where option 2 is deployed, the operator and the

> THE INDUSTRY HAS DECIDED TO BASE THE INITIAL DEPLOYMENT OF 5G ON OPTIONS 3 AND 2

industry will always invest in steps leading to the long-term target architecture. Eventually, the option 2 coverage will be sufficient to also support wide-area cMTC use cases that will benefit from both NR and 5GC.

At some point in the future there will also be mMTC solutions based on NR/5GC. However, many mMTC services are already adequately served by the existing mMTC solutions NB-IoT and LTE-M. The mMTC services in the low-end Low Power Wide Area (LPWA) segment are just one example. To avoid fragmentation, the best alternative for these use cases is continued use of NB-IoT and LTE-M for a long time.

The timing to reach this long-term target may vary between markets. It should be noted that even when the target is reached, networks will need to continue to support a set of legacy devices (LTE/EPC-based), in particular in the area of mMTC. When the UE penetration for NR support is high enough, selected bands can be fully refarmed to NR-only, as shown in the last step of Figure 2.

Analysis of options 5, 7 and 4

The industry has decided to base the initial deployment of 5G on options 3 and 2. While options 5, 7 and 4 may initially seem beneficial for specific operators' deployment cases, it is important to recognize that none of them are direct steps leading toward the long-term target architecture. Further, the use of options 5, 7 and 4 would add unnecessary complexity in the target architecture, in the interaction with other network functions, and in the evolution of new features, which needs to take the combination of all existing options into account.

After a thorough analysis of options 5, 7 and 4 that encompassed the main drivers, potential benefits and drawbacks, we have come to the conclusion that all three can and should be avoided. We have also identified preferred alternative solutions for each option.

Option 5

The main driver for deploying option 5 is to allow devices that move outside the area covered by option 2 to remain connected to 5GC, which would also increase the 5GC coverage to eLTE areas.

A key question to consider is: which use cases require nationwide 5GC coverage? Traditional MBB/voice obviously requires wide-area support, but this is well supported with intersystem mobility during the build-out of NR coverage, as it was in previous generation shifts. 5GC provides a range of new values but the need for other wide-area 5GC-based services in the near term is undefined. In the longer term, we expect wide-area option 2 to enable the new use cases that emerge.

Option 5 could be used to increase wide-area 5GC coverage, but reaching full wide-area 5GC coverage would take time and investment, as it would require new UEs, new RAN functionality and retesting the system. Option 5 would have a major impact on the UEs in terms of supporting the 5GC non-access stratum and the new parts of the eLTE radio interface, as legacy LTE devices are not supported. In addition, substantial interoperability retesting between networks and UEs would be required to ensure the operation of legacy features and services, including VoLTE. Further, option 5 requires substantial upgrades of the eNB software and, in many cases, the eNB baseband hardware as well.

In summary, option 5 is unlikely to provide a faster route to 5GC wide-area coverage than the wide deployment of option 2. Wide deployment of option 2 is the better alternative, particularly since option 5 would mean investing in technology that does not capitalize on the benefits of the latest radio technology (NR).

Option 7

Option 7 builds on option 5 and cannot exist without it. If option 5 were to be used, it is very likely that option 7 would also be supported in areas with NR. The driver for option 7 is the same as for option 3; that is, to use dual connectivity to aggregate NR and LTE bands to enhance capacity, but in this case for a UE connected via eLTE to 5GC. According to the same logic explained in the Option 5 section above, we recommend using option 2 instead.

> THE MOBILE INDUSTRY HAS AN OPPORTUNITY TO SIMPLIFY THE 5G ECOSYSTEM

Option 4

Option 4 is an addition to option 2, using dual connectivity to add eLTE to an NR anchor. It is primarily relevant when serving MBB traffic via 5GC. The driver for option 4 is to maximize throughput when the amount of NR spectrum is limited. An example of this type of situation would be if NR is deployed on 700MHz, 3.5GHz and mmW, but the UE is outside coverage of the two higher bands.

In terms of drawbacks, option 4 would require new software support in eNB, gNB and UE, with related interoperability testing. Further, the future evolution of features would need to consider option 4, and its use would require continued investments in eLTE deployments for a long time.

Option 4 is not necessary, and performs worse than option 2 with NR-NR CA and enough NR spectrum, in areas serving MBB via 5GC. Using option 2 instead of option 4 also focuses investments on the rollout of the long-term target architecture.

Conclusion

Our analysis shows that the mobile industry has an opportunity to simplify the 5G ecosystem by focusing network deployments on connectivity options 3 and 2, which are capable of delivering all the 5G benefits without adding unnecessary complexity and cost (as in options 5, 7 and 4). The flexible design of radio and core networks supports a smooth migration with LTE-NR spectrum sharing and dual-mode core technologies. The regulation of frequency bands should allow NR deployment in existing LTE bands that are in sync with the required spectrum migration.

Operators have the opportunity to avoid connectivity options 5, 7 and 4 by implementing a proactive spectrum migration strategy that considers NR for new low bands, and by refarming or introducing LTE-NR spectrum sharing in existing low/mid bands. This approach will reduce network upgrade cost and time, simplify interoperability between networks and devices, and enable a faster scaling of the 5G ecosystem.

A 5G deployment approach based exclusively on options 3 and 2 ensures that investment is focused on the long-term target architecture, leveraging full 5GS capabilities. Early key use cases for wide-area, like MBB including voice services, are fully supported during the migration period, along with services to existing devices.

Terms and abbreviations

4T4R – 4-Branch Transmit/Receive Antenna and Radio Arrangement | **5GC** – 5G Core | **5GS** – 5G System | **CA** – Carrier Aggregation | **cMTC** – Critical Machine Type Communication | **CN** – Core Network | **DC** – Dual Connectivity | **DECOR** – Dedicated Core Network | **DL** – Downlink | **E2E** – End-to-end | **eLTE** – Evolved LTE | **eNB** – Evolved Node B | **EN-DC** – E-UTRA – NR Dual Connectivity | **EPC** – Evolved Packet Core | **FWA** – Fixed Wireless Access | **gNB** – Next Generation Node B | **IoT** – Internet of Things | **LPWA** – Low Power Wide Area | **LTE-M** – LTE-MTC Standard | **MBB** – Mobile Broadband | **mMTC** – Massive Machine Type Communication | **mmW** – Millimeter Wave | **NB-IoT** – Narrowband Internet of Things | **NR** – New Radio | **RAT** – Radio Access Technology | **UE** – User Equipment | **UL** – Uplink

References

1. **3GPP Release 15 specifications, e.g. TS 23.501, TS 38.401, available at:** *http://www.3gpp.org/release-15*
2. **Ericsson Mobility Report, available at:** *https://www.ericsson.com/en/mobility-report*
3. **Ericsson Technology Review, November 2018, The advantages of combining 5G NR with LTE, available at:** *https://www.ericsson.com/en/ericsson-technology-review/archive/2018/the-advantages-of-combining-5g-nr-with-lte*

Further reading

- **Ericsson, 5G deployment options, 2018, available at:** *https://www.ericsson.com/assets/local/narratives/networks/documents/5g-deployment-considerations.pdf*
- **Ericsson, Core evolution from EPC to 5G Core, download available from:** *https://pages.digitalservices.ericsson.com/core-evolution-to-5g*

THE AUTHORS

Torbjörn Cagenius

◆ is a senior expert in network architecture at Business Area Digital Services. He joined Ericsson in 1990 and has worked in a variety of technology areas such as fiber-to-the-home, main-remote RBS, fixed-mobile convergence, IPTV, network architecture evolution, software-defined networking and Network Functions Virtualization. In his current role, he focuses on 5G and associated network architecture evolution. He holds an M.Sc. in electrical engineering from KTH Royal Institute of Technology in Stockholm, Sweden.

Anders Ryde

◆ is a senior expert in network and service architecture at Business Area Digital Services, based in Sweden. He joined Ericsson in 1982 and has worked in a variety of technology areas in network and service architecture development for multimedia-enabled telecommunication, targeting both enterprise and residential users. This includes the evolution of mobile telephony to IMS and VoLTE. In his current role, he focuses on bringing voice and other communication services into 5G, general 5G evolution and associated network architecture evolution. He holds an M.Sc. in electrical engineering from KTH Royal Institute of Technology in Stockholm, Sweden.

Jari Vikberg

◆ is a senior expert in network architecture and the chief network architect at CTO office. He joined Ericsson in 1993 and has both wide and deep technology competence covering network architectures for all generations of RANs and CNs. He is also skilled in the application layer and other domains, and the impact and relation these have to mobile networks. He holds an M.Sc. in computer science from the University of Helsinki, Finland.

Per Willars

◆ is an expert in network architecture and radio network functionality at Business Area Networks. He joined Ericsson in 1991 and has worked intensively with RAN issues ever since. This includes leading the definition of 3G RAN, before and within the 3GPP, and more lately indoor solutions. He has also worked with service layer research and explored new business models. In his current role, he analyzes the requirements for 5G RAN (architecture and functionality) with the aim of simplifying 5G. He holds an M.Sc. in electrical engineering from KTH Royal Institute of Technology.

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.

CHRISTER BOBERG, MALGORZATA SVENSSON, BENEDEK KOVÁCS

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 applications.
- **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.

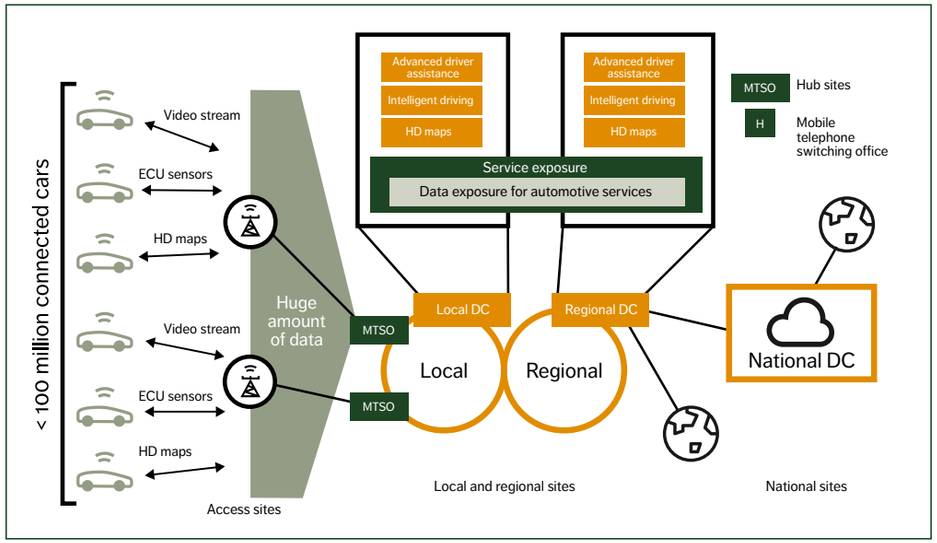


Figure 1 High-volume data automotive services and their characteristics

Topology-aware cloud computing and storage is an example of one such solution that provides what we call a global automotive distributed edge cloud. The limitation on the amount of data that can be effectively transported over the cellular network must not be allowed to affect the service experience negatively, as that would hinder the evolution of new automotive services. It is therefore necessary to increase capacity, availability and coverage as well as finding appropriate mechanisms to limit the amount of data transferred. Orchestrating applications and their different components running in a multitude of different clouds from different vendors is one of the challenges. Vehicles connecting to networks without an existing application edge infrastructure is another.

The placement of application components at edges depends on the behavior of the application and the available infrastructure resources. 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.

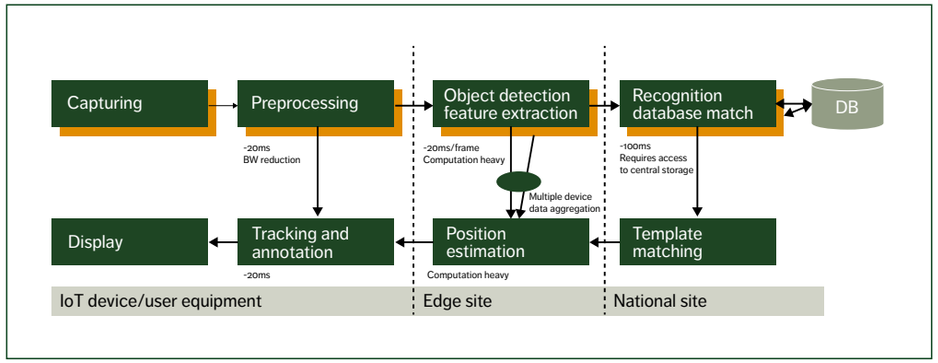


Figure 2 An AR application and its modules optimized for edge computing

A simple AR application and its main components are shown in *Figure 2*. The components of the application could be executed either on the device itself, the edge server or in the central cloud. Deploying application components at the network edge may make it possible to offload the device while maintaining short latency. Edge compute is also optimizing the flow when coordination is required – for example, when using multiple real-time camera feeds to determine the 3D position of objects, also as shown in Figure 2. Furthermore, advanced cloud software as a service – ML, analytics and DBs as a service, for example – may also be provided on the edge site.

Our distributed cloud solution

Ericsson has developed a distributed cloud solution that provides the required capabilities to support the use cases of the fourth industrial revolution, including private and localized networks. 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 and meets end-to-end network requirements as well as offering management, orchestration and exposure for the network and cloud resources together.

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 is based on software-defined networking, Network Functions Virtualization (NFV) and 3GPP edge computing technologies to enable multi-access and multi-cloud capabilities and unlock networks to provide an open platform for application innovations. In the management dimension, distributed cloud offers automated deployment in heterogeneous clouds. This could be provided by multiple CSPs, where workload placement is policy driven and based on various externalized criteria.

To enable monetization and application innovation, distributed cloud capabilities are exposed on marketplaces provided by Ericsson, third parties and CSPs. The distributed cloud capabilities can be offered according to various business and operational models. One example of a possible scenario is for a CSP to offer connectivity and a cloud execution environment to enterprises as a service. In this case, a CSP manages the computation and connectivity resources, but these are located at the enterprise premises. The application characteristics determine the placement of applications at various geolocations. In the case of AR/VR and image recognition applications used by technicians to fix a broken power station, for example, it would be most effective to place them close to the broken power station.

Edge computing

Our distributed cloud solution enables 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.

> OUR DISTRIBUTED CLOUD SOLUTION ENABLES EDGE COMPUTING, WHICH MANY APPLICATIONS REQUIRE

In the automotive use case, the network is designed to split data traffic into several locations that cover reasonable numbers of connected vehicles. The computation resources are 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. In this infrastructure framework, localized data collected via local and wide area networks is stored in the central cloud and integrated

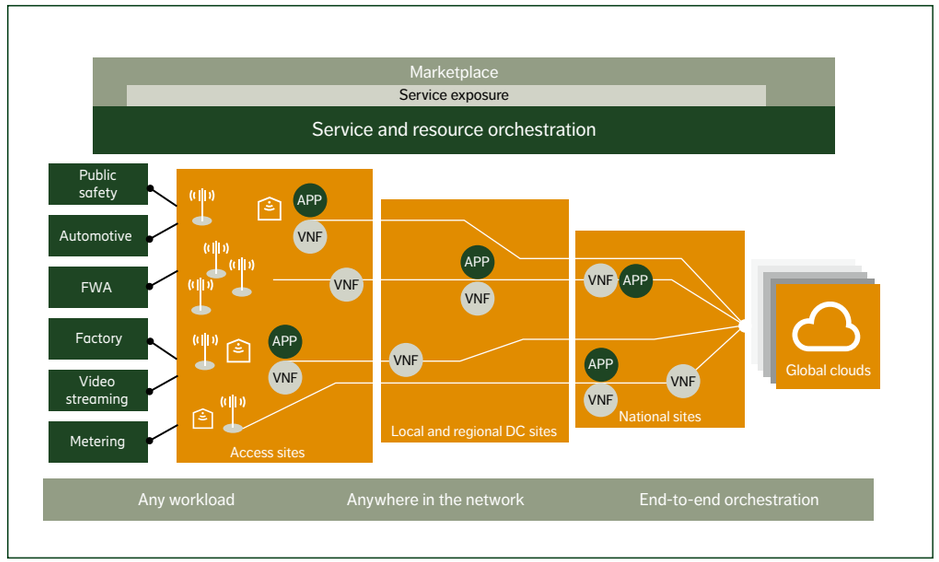


Figure 3 Distributed cloud architecture

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

on the edge computing architecture to provide real-time information necessary for services of connected vehicles.

The exact locations of the micro and small data 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 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 application developers 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 sites must support a stringent set of functions and APIs. This implies that CSPs must join forces to create a federated model. Doing so will require significant effort, with the first step being to reach an agreement on the standard mechanisms to use.

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 network architectures and conventional cloud computing systems 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 OPNFV (Open Platform for NFV) and ONAP (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 5GAA (the 5G Automotive Association) for automotive and 5G-ACIA (the 5G Alliance for Connected Industries and Automation) [8], Industry 4.0 and the IIoT (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.

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-31356/*
2. **AECC White Paper, available at:** *https://www.ericsson.com/res/docs/2014/consumerlab/liberation-from-location-ericsson-consumerlab.pdf*
3. **Government Technology, Making 5G a Reality Means Building Partnerships — Not Just Networks, June 5, 2018, Descant, S, available at:** *http://www.govtech.com/network/Making-5G-a-Reality-Means-Building-Partnerships--Not-Just-Networks.html*
4. **Wired, Eye tracking is coming to VR sooner than you think. What now?, March 23, 2018, Rubin, P, available at:** *https://www.wired.com/story/eye-tracking-vr/*
5. **Think Act, Digital factories – The renaissance of the U.S. automotive industry, Berger, R, available at:** *https://www.rolandberger.com/en/Publications/pub_digital_factories.html*
6. **Ericsson, Technology Trends 2018, Five technology trends augmenting the connected society, 2018, Ekudden, E, available at:** *https://www.ericsson.com/en/ericsson-technology-review/archive/2018/technology-trends-2018*
7. **Ericsson Technology Review, Open, intelligent and model-driven: evolving OSS, February 7, 2018, Agarwal, M; Svensson, M; Terrill, S; Wallin, J, available at:** *https://www.ericsson.com/en/ericsson-technology-review/archive/2018/open-intelligent-and-model-driven-evolving-oss*
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/*

Further reading

- **Ericsson Consumer & IndustryLab, 5G business value: A case study on real-time control in manufacturing, April 2018, available at:** *https://www.ericsson.com/assets/local/reports/5g_for_industries_report_blisk_27062018.pdf*
- **Ericsson, Turn on 5G: Ericsson completes 5G Platform for operators, February 8, 2018, available at:** *https://www.ericsson.com/en/press-releases/2018/2/turn-on-5g-ericsson-completes-5g-platform-for-operators*
- **Ericsson, Going beyond edge computing with distributed cloud, available at:** *https://www.ericsson.com/digital-services/trending/distributed-cloud*
- **Ericsson/KTH Royal Institute of Technology, Resource monitoring in a Network Embedded Cloud: An extension to OSPF-TE, available at:** *https://www.ericsson.com/assets/local/publications/conference-papers/03-cloud-ospf-camera-ready.pdf*
- **M2 Optics Inc., Calculating Optical Fiber Latency, January 9, 2012, Miller, K, available at:** *http://www.m2optics.com/blog/bid/70587/Calculating-Optical-Fiber-Latency*
- **Application function placement optimization in a mobile distributed cloud environment, Anna Peale, Péter Kiss, Charles Ferrari, Benedek Kovács, László Szilágyi, Melinda Tóth, in Studia Informatica - Issue no. 2/2018, pp37-52, available at:** *http://www.studia.ubbcluj.ro/arhiva/abstract_en.php?editie=INFORMATICA&nr=2&an=2018&id_art=15974*

The authors would like to thank the following people for their contributions to this article: Carlos Bravo, Ala Nazari, Stefan Runeson, Ola Hubertsson, Thorsten Lohmar and Tomas Nylander.

THE AUTHORS

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 M.Sc. 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.

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 | **IoT** – Internet of Things | **ML** – Machine Learning | **MS** – Millisecond | **MTSO** – 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

FEATURE ARTICLE – 5G AND SMART MANUFACTURING ✱

BOOSTING 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.

JOACHIM SACHS,
KENNETH WALLSTEDT,
FREDRIK ALRIKSSON,
GÖRAN ENEROTH

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.

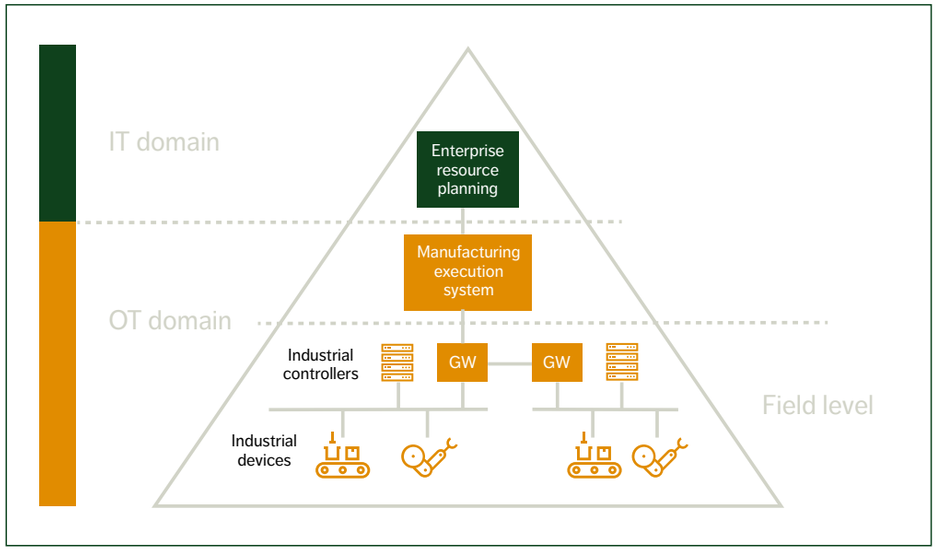


Figure 1 Hierarchical network design based on the industrial automation pyramid

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 PROFIBUS, PROFINET, EtherCAT, Sercos and Modbus. Currently deployed wireless solutions (which are typically wireless LAN based using unlicensed spectrum) constitute only a small fraction

Definition of key terms

- **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.

of the installed base; they mainly play a role for wirelessly connecting sensors where communication requirements are non-critical.

Today, the field level consists of connectivity islands that are separated by gateways (GWs), which helps to provide the required performance within each connectivity island. The GWs are also needed for protocol translation between the different industrial networking technologies. However, this segmented design puts limitations on the digitalization of factories, as information within one part of the factory cannot be easily extracted and used elsewhere.

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.

Key manufacturing industry requirements

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 5ms (FDD 15kHz SCS or TDD 30kHz 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 is compared with the application latency constraint, the higher spectral efficiency (capacity) may be achieved.

> MMWAVE IS A GOOD COMPLEMENT TO MID-BAND FOR IN-FACTORY DEPLOYMENTS

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.

Terms and abbreviations

cMTC – Critical Machine-type Communication | **CN** – Core Network | **DL** – Downlink | **GHz** – Gigahertz | **GW** – Gateway | **IoT** – Internet of Things | **kHz** – Kilohertz | **LTE-M** – LTE Machine-type Communication | **MBB** – Mobile Broadband | **mMTC** – Massive Machine-type Communication | **mmWave** – Millimeter Wave | **ms** – Millisecond | **MTC** – Machine-type Communication | **NB-IoT** – Narrowband IoT | **NR** – New Radio | **OT** – Operational Technology | **RTT** – Round-trip Time | **SCS** – Subcarrier Spacing | **TSN** – Time-sensitive Networking | **UE** - User Equipment | **UL** – Uplink | **URLLC** – Ultra-reliable Low-latency Communication

FEATURE ARTICLE – 5G AND SMART MANUFACTURING ✱

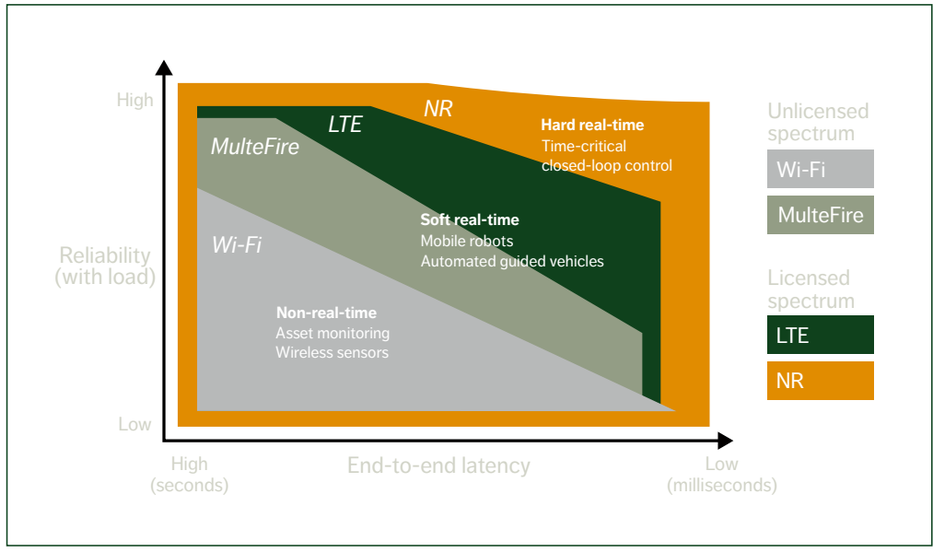


Figure 2 Latency and reliability aspects of spectrum and technology choice

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 industries, as is under consideration in some European countries such as Germany and Sweden on 3.7-3.8GHz.

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 802.3 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 60802 joint 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.

5G enhancements include support of redundant transmission paths, which can be combined with the TSN feature 'Frame replication and elimination for reliability' (FRER) that is standardized in IEEE 802.1CB. One of the resource allocation features of TSN for bounding the latency for periodic control traffic is 'Time-aware scheduling' (standardized in IEEE 802.1Qbv), for which transmission queues are time-gated in every switch on the data path to create a protected connection. This requires all Ethernet switches to be time-synchronized according to IEEE P802.1AS-Rev. Features that are being developed in 5G standardization to support time-aware transmission across a mixed TSN-5G network are to time-align the 5G system with the TSN network and provide 5G transmission with deterministic latency.

Keeping things local

On top of URLLC performance and integration with industrial Ethernet networks, many manufacturers also require full control (that is, independent of external parties) of their critical OT domain connectivity in order to fulfill system availability targets. Full control can be expressed as requirements on keeping things local:

- local data – the ability to keep production-related data locally within the factory premises for security and trust reasons
- local management – the ability to monitor and manage the connectivity solution locally
- local survivability – the ability to guarantee the availability of the connectivity solution independently of external factors (for example, shop-floor connectivity must continue uninterrupted even when connectivity to the manufacturing plant is down).

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. There 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

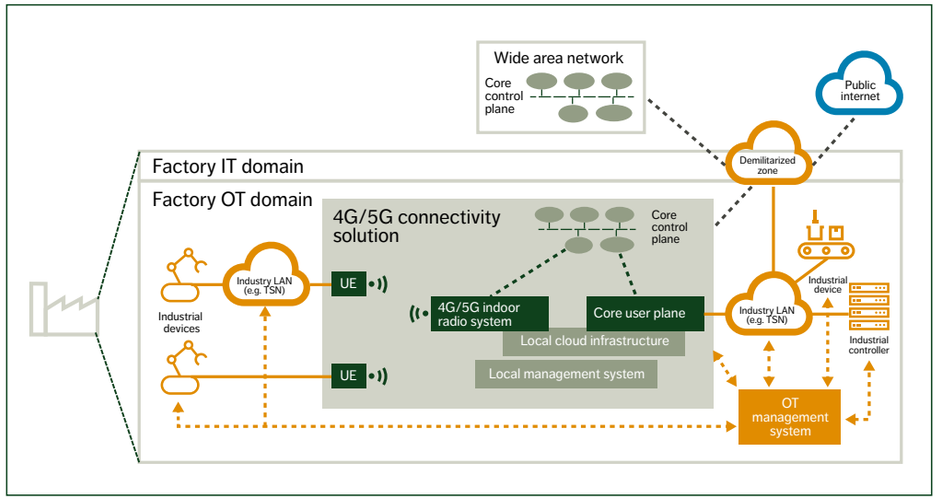


Figure 3 5G manufacturing solution architecture

industrial applications require time-synchronized actions of multiple machines. This can be a collaborative common task performed by multiple industry robots, where the control of the different robots needs to be coordinated in time. NR in release 16 will supply the capability for a base station to provide precise timing references to devices down to microsecond precision. It will also make it possible to relate this time reference to the reference clocks of 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 combined 5G-TSN network.

Security in cellular networks has matured with every generation to enable confidential communication services, user privacy, authentication of users for network access and accountability, and authentication of the network so users know they are connected to a legitimate network. To address new use cases and the evolving threat landscape, 5G includes new security features that benefit industrial deployments [9]. Examples include improved confidentiality of user-plane data achieved by both the encryption and integrity protection of data to prevent eavesdropping and modification as it passes through the 5G system. With 5G, industrial networks gain additional options for device authentication supporting both SIM-based and certificate-based authentication. Lastly, 5G standards prevent IMSI (International Mobile Subscriber Identity) catching attacks, as the user's or device's long-term identifier is never transmitted over the radio interface in clear text [10].

5G's network slicing capabilities enable the provision of a dedicated slice both locally and in 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 will 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 or by third parties using leased 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.

FEATURE ARTICLE – 5G AND SMART MANUFACTURING ✱

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

1. **Ericsson Technology Review, Industrial automation enabled by robotics, machine intelligence and 5G, February 15, 2018, Sabella, R.; Thuelig, A.; Carrozza, M.C.; Ippolito, M., available at:** *https://www.ericsson.com/en/ericsson-technology-review/archive/2018/industrial-automation-enabled-by-robotics-machine-intelligence-and-5g*
2. **5G-ACIA, 5G for Connected Industries and Automation, available at:** *https://www.5g-acia.org/*
3. **HMS Industrial Networks, 2018, available at:** *https://www.hms-networks.com/press/2018/02/27/industrial-ethernet-is-now-bigger-than-fieldbuses*
4. **Ericsson white paper, Advanced antenna systems for 5G networks, Von Butovitsch, P.; Astely, D.; Furuskär, A.; Göransson, B.; Hogan, B; Karlsson, J; Larsson, E, available at:** *https://www.ericsson.com/en/white-papers/advanced-antenna-systems-for-5g-networks*
5. **Proceedings of the IEEE, Adaptive 5G Low-Latency Communication for Tactile Internet Services, September 5, 2018, Sachs, J.; Andersson, L. A. A.; Araújo, J.; Curescu, C.; Lundsjö, J.; Rune, G.; Steinbach, E.; Wikström, G., available at:** *https://ieeexplore.ieee.org/document/8454733*
6. **Technical report, OPC UA TSN: A new Solution for Industrial Communication, Bruckner, et al., available at:** *https://www.automationworld.com/sites/default/files/opc_ua_tsn_whitepaper_1.pdf*
7. **IEC/IEEE 60802 joint project webpage, IEC/IEEE 60802 TSN Profile for Industrial Automation, available at:** *https://1.ieee802.org/tsn/iec-ieee-60802-tsn-profile-for-industrial-automation/*
8. **3GPP, "Service requirements for cyber-physical control applications in vertical domains," technical specification TS 22.104, January 2019, available at:** *http://www.3gpp.org/ftp//Specs/archive/22_series/22.104/22104-g00.zip*
9. **Ericsson white paper, 5G security – enabling a trustworthy 5G system, March 28, 2018, Norrman, K.; Nakarmi, P. K.; Fogelström, E, available at:** *https://www.ericsson.com/en/white-papers/5g-security---enabling-a-trustworthy-5g-system*
10. **Ericsson Blog, 3GPP release 15 – an end to the battle against false base stations, January 18, 2019, Nakarmi, P. K.; Ben Henda, N.; Tsiatsis, V., available at:** *https://www.ericsson.com/en/blog/2019/1/3gpp-release15*

The authors would like to thank the following people for their contributions to this article: Jonathan Olsson, Jari Vikberg, Juan-Antonio Ibanez, Kurt Essigmann, Lisa Boström and Filip Mestanov.

THE AUTHORS

Joachim Sachs

◆ is principal researcher at Ericsson corporate research in Stockholm, where he coordinates research activities on 5G for industrial IoT 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 a Dr-Ing. from Technical University Berlin, Germany, and was a visiting scholar at Stanford University in the US in 2009.

Kenneth Wallstedt

◆ is director of technology strategy in Ericsson's CTO office, where he focuses on the company's radio and spectrum management strategy. He joined Ericsson in 1990 and since then he has held various leading positions in Ericsson's research, development and market units in Canada, Sweden and the US. He holds an M.Sc. in electrical engineering from KTH Royal Institute of Technology in Stockholm, Sweden.

Fredrik Alriksson

◆ is a research engineer at Development Unit Networks, where he coordinates strategic technology and concept development within IoT & New Industries. He joined Ericsson in 1999 and has worked in R&D with architecture evolution covering a broad set of technology areas including RAN, Core, IMS and VoLTE. He holds an M.Sc. in electrical engineering from KTH Royal Institute of Technology in Stockholm, Sweden.

Göran Eneroth

◆ is a product development leader at Development Unit Networks, where he leads strategic technology and concept development within IoT & New Industries. He joined Ericsson in 1983 and has held a variety of leading positions in Ericsson's R&D units, as well as in standardization and industry collaborations. He holds an M.Sc. in electrical engineering from KTH Royal Institute of Technology in Stockholm, Sweden.

Further reading

- **Ericsson Smart Wireless Manufacturing, available at:** *https://www.ericsson.com/en/internet-of-things/solutions/smart-wireless-manufacturing*
- **Ericsson Consumer & IndustryLab Insight Report, 5G business value: A case study on real-time control in manufacturing, April 2018, available at:** *https://www.ericsson.com/assets/local/reports/5g_for_industries_report_blisk_27062018.pdf*
- **Ericsson Mobility Report, Realizing smart manufacturing through IoT, June 2018, available at:** *https://www.ericsson.com/en/mobility-report/reports/june-2018/realizing-smart-manufacturing*
- **Ericsson Blog, 5G meets Time Sensitive Networking, December 2018, available at:** *https://www.ericsson.com/en/blog/2018/12/5G-meets-Time-Sensitive-Networking*
- **5G-ACIA white paper, 5G for Connected Industries and Automation, November 22, 2018, available at:** *https://www.5g-acia.org/index.php?id=5125*

KEY TECHNOLOGY CHOICES FOR

optimal massive IoT devices

The latest cellular communication technologies LTE-M and NB-IoT enable 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.

CLAES LUNDQVIST,
ARI KERÄNEN,
BEN SMEETS,
JOHN FORNEHED,
CARLOS R. B. AZEVEDO,
PETER VON WRYCZA

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 home, environmental monitoring and smart metering (smart building); 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.

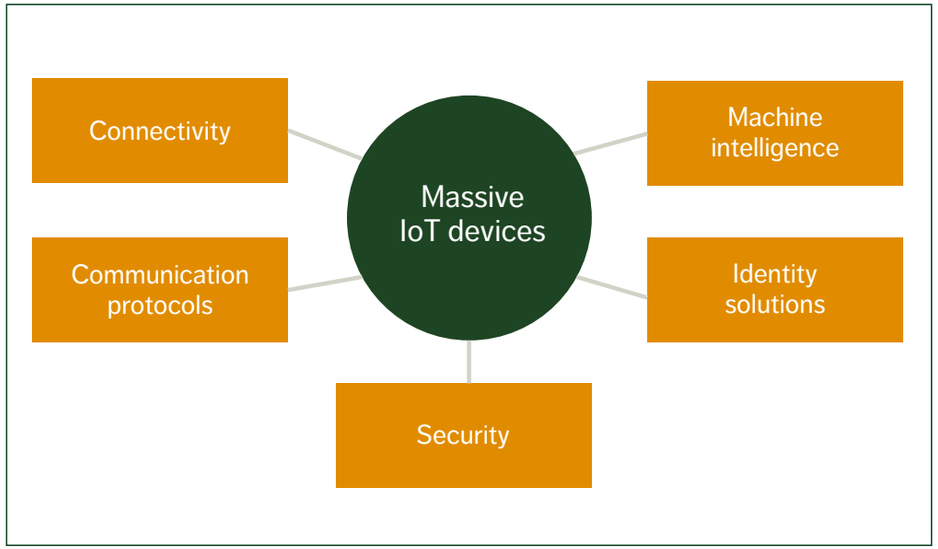


Figure 1 Key technologies for massive IoT devices

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) – as shown in *Figure 1*. Carefully considered choices in these five areas make it possible to achieve the desired key device characteristics and create IoT devices that support the multitude of existing and emerging massive IoT use cases.

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

were introduced in 3GPP release 13: LTE-MTC (LTE-M), which includes a new user equipment (UE) category called Cat-M1, and NB-IoT, which includes UE category Cat-NB1 [2].

A Cat-M1 UE supports a reduced bandwidth of 1.4MHz and a data throughput of up to 300kbps in the downlink and 375kbps 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 26kbps in the downlink and 63kbps 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, a 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.

Terms and abbreviations

ASIC – Application-Specific Integrated Circuit | **CoAP** – Constrained Application Protocol | **DMI** – Distributed Machine Intelligence | **E2E** – End-to-end | **EAP** – Extensible Authentication Protocol | **HTTP** – Hypertext Transfer Protocol | **IETF** – Internet Engineering Task Force | **IoT** – Internet of Things | **IPSO** – Internet Protocol for Smart Objects | **iUICC** – Integrated Universal Integrated Circuit Card | **LoRaWAN** – Long Range Wide-Area Network | **LPWA** – Low-Power Wide-Area | **LPWAN** – Low-Power Wide-Area Network | **LTE-M** – LTE for Machines | **LwM2M** – Lightweight M2M | **M2M** – Machine-to-Machine | **MI** – Machine Intelligence | **MNO** – Mobile Network Operator | **MQTT** – Message Queuing Telemetry Transport | **MTC** – Machine Type Communication | **NB-IoT** – Narrowband Internet of Things | **ODMI** – On-Device Machine Intelligence | **OSCORE** – Object Security for Constrained RESTful Environments | **PKI** – Public Key Infrastructure | **QUIC** – Quick UDP Internet Connections | **SenML** – Sensor Measurement Lists | **SGX** – Software Guard Extensions | **TCP** – Transmission Control Protocol | **TEE** – Trusted Execution Environment | **TLS** – Transport Layer Security | **TPU** – Tensor Processing Unit | **UDP** – User Datagram Protocol | **UE** – User Equipment | **WoT-TD** – Web of Things Thing Descriptions

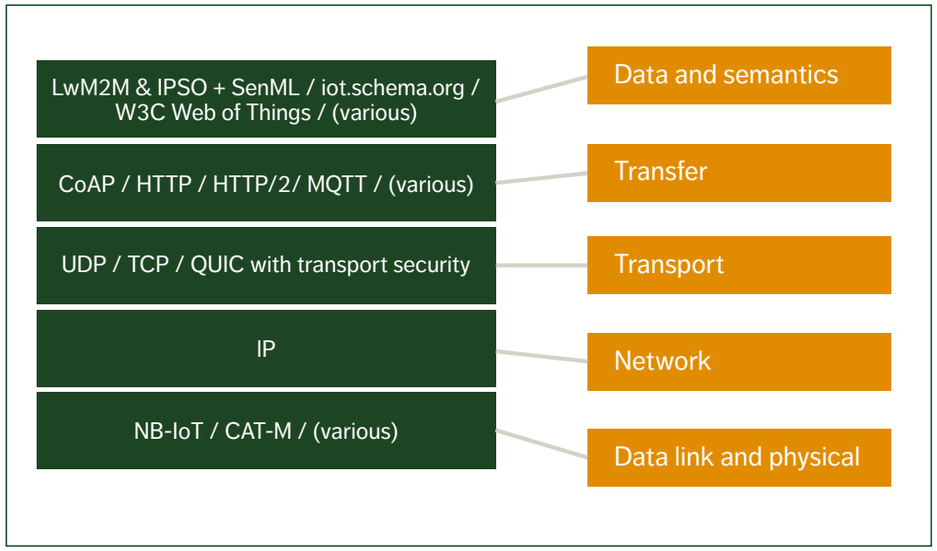


Figure 2 Structure of an IoT device protocol stack

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 data link protocols, followed by a selection of standardized transport and transfer protocols, ending up at the application layer with data models and semantics, as shown in *Figure 2*.

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 systems to use IPv6. On top of IPv6, User Datagram Protocol (UDP) or Transmission Control Protocol (TCP) is usually used at the transport layer. More recently, the QUIC protocol [6], combining features from UDP and TCP, is attracting interest for IoT scenarios as well.

IoT E2E communication is usually secured with Transport Layer Security (TLS). Recently, the Internet Engineering Task Force (IETF) finished the standardization of TLS v1.3. This latest version enables faster connection setup, more resiliency to address changes and stronger security. When E2E security through middleboxes, such as proxies, is needed, IoT communication can be secured with Object Security for Constrained RESTful Environments (OSCORE) [7].

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

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

Telemetry 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. Out 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 (WoT-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 (TEEs) 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 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

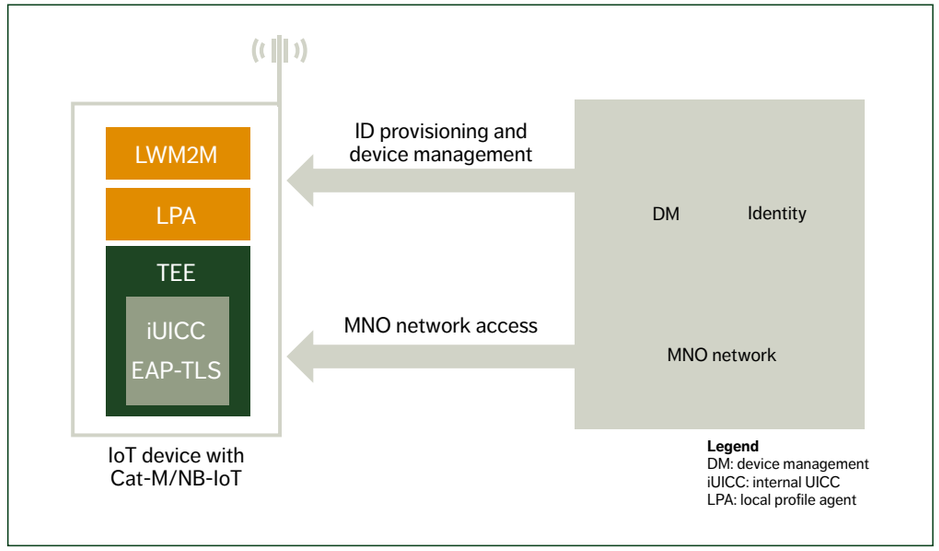


Figure 3 EAP-TLS ID management and use for network access

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 via Extensible Authentication Protocol (EAP)-TLS. *Figure 3* illustrates EAP-TLS ID management and use for network access.

Beyond mobile networks, other network technologies also require identities, and applications may need identities too. Therefore, depending on the device use case, a single device may need several identities. This can be problematic for constrained devices, and it makes identity management difficult. As different device hardware will come with different types of initial identities, Ericsson believes that a federation of identities [15] 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 life-cycle 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 life-cycle 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, ODMI 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 in evolving networks, as sensors and actuators are added and removed. Flexibility is then a core design principle in massive IoT systems. Edge and ODMI 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, at a lower cost. It is also our opinion that the best way to ensure the interoperability of IoT devices from communication to application layer is through the use of protocol stacks based on standardized internet protocols and data models with efficient capabilities for data transfer and device management.

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 web page, Internet of Things, available at:** *https://www.ericsson.com/en/internet-of-things*
- **Ericsson white paper, January 2016, Cellular networks for Massive IoT – enabling low power wide area applications, available at:** *https://www.ericsson.com/en/white-papers/cellular-networks-for-massive-iot--enabling-low-power-wide-area-applications*
- **Ericsson white paper, June 2017, IoT security – protecting the networked society, available at:** *https://www.ericsson.com/en/white-papers/iot-security-protecting-the-networked-society*
- **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/en/blog/2017/10/smart-contracts-for-identities*
- **Ericsson Technology Review, November 2017, End-to-end security management for the IoT, available at:** *https://www.ericsson.com/en/ericsson-technology-review/archive/2017/end-to-end-security-management-for-the-iot*

References

1. **Ericsson Mobility Report, November 2018, available at:** *https://www.ericsson.com/en/mobility-report/reports/november-2018*
2. **Academic Press, Cellular Internet of Things: Technologies, Standards and Performance, 1st edition, 2017, O. Liberg, M. Sundberg, E. Wang, J. Bergman, J. Sachs**
3. **IEEE Network, volume 31, issue 6, Overview of 3GPP Release 14 Enhanced NB-IoT, November/December 2017, A. Höglund et al.**
4. **IEEE Communications Standards Magazine, volume 2, issue 2, Overview of 3GPP Release 14 Further Enhanced MTC, June 2018, A. Höglund et al.**
5. **IETF, June 2018, LPWAN Static Context Header Compression (SCHC) and fragmentation for IPv6 and UDP, available at:** *https://tools.ietf.org/html/draft-ietf-lpwan-ipv6-static-context-hc-16*
6. **IETF, October 2018, QUIC: A UDP-Based Multiplexed and Secure Transport, available at:** *https://tools.ietf.org/html/draft-ietf-quic-transport-15*
7. **IETF, August 2018, Object Security for Constrained RESTful Environments (OSCORE), available at:** *https://tools.ietf.org/html/draft-ietf-core-object-security-15*
8. **IETF, June 2014, Hypertext Transfer Protocol (HTTP/1.1): Message Syntax and Routing, available at:** *https://tools.ietf.org/html/rfc7230*
9. **IETF, June 2014, The Constrained Application Protocol (CoAP), available at:** *https://tools.ietf.org/html/rfc7252*
10. **IETF, May 2015, Hypertext Transfer Protocol Version 2 (HTTP/2), available at:** *https://tools.ietf.org/html/rfc7540*
11. **IETF, August 2018, Sensor Measurement Lists (SenML), available at:** *https://tools.ietf.org/html/rfc8428*
12. **OMA SpecWorks, Lightweight M2M (LWM2M), available at:** *https://www.omaspecworks.org/what-is-oma-specworks/iot/lightweight-m2m-lwm2m/*
13. **W3C, October 21, 2018, Web of Things (WoT) Thing Description, available at:** *https://www.w3.org/TR/wot-thing-description/*
14. **IETF, Software Updates for Internet of Things (suit), available at:** *https://datatracker.ietf.org/wg/suit/about/*
15. **Intel, October 15, 2018, Intel and Arm Share IoT Vision to Securely Connect Any Device to Any Cloud, Lorie Wigle, available at:** *https://newsroom.intel.com/editorials/intel-arm-share-iot-vision-securely-connect-any-device-any-cloud/*
16. **Ericsson Technology Review, April 2017, Tackling IoT complexity with machine intelligence, available at:** *https://www.ericsson.com/en/ericsson-technology-review/archive/2017/tackling-iot-complexity-with-machine-intelligence*
17. **Ericsson white paper, May 2018, Artificial intelligence and machine learning in next-generation systems, available at:** *https://www.ericsson.com/en/white-papers/machine-intelligence*

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 electrical engineering from KTH Royal Institute of Technology in Stockholm, Sweden.

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.

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 expert 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, development and standardization activities for the IoT. Von Wrycza holds a Ph.D. in telecommunications from KTH Royal Institute of Technology in Stockholm.

BSS and artificial intelligence

–TIME TO GO NATIVE

The growing need to support disruptive services emerging from the Internet of Things (IoT) and 5G requires a fundamental transformation of business support systems (BSS). At Ericsson, we believe that the best way to achieve this is by forging BSS and artificial intelligence (AI) together to create truly AI-native BSS.

LARS ANGELIN,
JOHAN SILVANDER

Although AI is of obvious benefit in terms of business optimization, and has been used in all sorts of businesses for decades, AI and BSS have never been integrated into one efficient system.

Examples of areas in which AI is already used in conjunction with BSS software include customer retention, chatbots, revenue and cost predictions, customer analysis, customer experience management, customer yield optimization, automation, process reengineering, simulations, quality improvements, and fraud and anomaly detection.

AI capabilities enable improved business decision dynamics and better decision precision, resulting in better business performance and agility. Virtually all business activities can and will benefit from AI, and as 5G and the IoT continue to expand, the number of use cases will only continue to grow. The challenge we face at present is that the learning, insight-building and reasoning capabilities of AI in today's telco BSS are not as strong as they need to be to cope with emerging use cases.

For the most part, AI capabilities today are simply bolted onto telco BSS one by one. But this is inefficient in terms of life-cycle costs, because the BSS must be repeatedly upgraded to benefit from the AI algorithms. Further, as they are separate systems, the BSS information must be transformed to fit AI systems, and vice versa. A much more efficient alternative is AI-native BSS – that is, BSS with intrinsic AI capabilities where the AI logic is a natural part of BSS logic in terms of both design and operation. This approach results in a system that can handle more complex business situations, generating more optimized business outcomes.

BSS evolution drivers

The main growth opportunities for communication service providers (CSPs) within the next decade are 5G and the IoT, with an estimated annual value of approximately USD 600 billion [1, 2]. To capitalize on this opportunity, CSPs must be able to support a marketplace with an ecosystem of many actors that have their own business models, where each actor may be both supplier and customer to other actors. Business support complexity increases dramatically in this environment. Current telco BSS, which can only support a single enterprise shop with a few business models, are simply not up to the task.

Surveys show that a clear majority of the telecommunications industry actors expect AI to have significant business impact in the coming five years, affecting both the top and bottom lines. They also expect AI to bring a significant competitive advantage to the enterprises using them, growing proportionally with AI usage. Analysts predict that enterprises will invest in AI competence, AI maturity and in organizational AI capabilities [3, 4], despite the costs [5].

There are essentially three main forces driving the combined AI-BSS evolution [1]. Firstly, business agility is a highly valued BSS property since the business itself evolves and new business opportunities emerge. AI plays a key role in both identifying opportunities and in shaping the new business models to pursue them. Secondly, the maturity of the communications industry is driving ever lower business transaction costs, as demonstrated by existing platform players like Amazon and Alibaba. In light of this, AI-supported process automation and reengineering are the tools

ARTIFICIAL INTELLIGENCE

Artificial intelligence (AI) depends on software algorithms. At Ericsson, we use the term AI in its widest sense, including several subfields such as machine learning, representation learning and deep learning. AI-related areas such as natural-language processing, automated reasoning, multiagent systems, symbolic learning, knowledge representation, intelligent tutoring systems and high-level computer vision are also included [8].

Terms and abbreviations

AI – Artificial Intelligence | **BSS** – Business Support Systems | **CRISP-DM** – Cross-Industry Process for Data Mining | **CSP** – Communication Service Provider | **IoT** – Internet of Things

AI-NATIVE BUSINESS SUPPORT SYSTEMS ✱

> A BUSINESS INTENT STATES THE DESIRED OR OPTIMAL OUTCOME OF A GIVEN SITUATION

of choice. Finally, the cloud provides an ideal foundation for continuous introduction of AI and BSS marketplace capabilities. This is because the cloud offers deployment flexibility, elastic scaling and a micro-service architecture that enables a more fine-grained separation of concerns and componentization with loose coupling.

Key challenges to adding AI to BSS

Introducing AI into existing BSS is not straightforward. Some challenges, such as data acquisition, data piping and training algorithms, are obvious and well known. Others, however, are less so. One example of a less obvious challenge is the fact that the interpretation of the AI results requires business competence; another is that monetization requires both retraining within the organization and redesigning of the existing set of business rules and processes and system reengineering [5].

Traditional, non-AI-native BSS are divided into component silos – such as customer relationship management, catalogs, billing and order management – each with their own information. This arrangement contradicts AI efficiency and dynamics enablement, which require an open, pan-BSS information and rules view. When an AI capability is added to this environment, it is treated as an add-on, requiring both AI and BSS system competence, information transformation and in many cases partial system re-implementation or reconfiguration. BSS performance issues such as latency and scaling may arise.

Many business situations are multifaceted, have many root causes and may include both gains and risks. In many cases, a main business intent (also known as a KPI) must be broken down into a combination of subintents. These will be based on many data sets and algorithms, and then be stitched together by a super-algorithm to deliver the main business intent. There is also a risk of lost business control, as a high-level intent may affect many of the lower-level business rules and processes. This effect is considerably smaller when an intent is introduced at lower levels, but in those cases there will be less business gain. The uniqueness of an intent and its context means there is little opportunity for reuse or experience building.

Ericsson believes that a new BSS architecture style that incorporates AI-native properties – including data-centric, learning loop, intent and event-driven logic, business rule hierarchy, and support for strategic, tactical and operational levels – is a much more efficient way to integrate AI with BSS. We fully agree with the view that future BSS and AI will be inseparably linked and must mature together [4, 5].

Introducing intents to BSS

An enterprise is a hierarchical or line-of-command structure in which business rules at the top steer, align and control the activities and behaviors further down in the structure. Business rules also steer all behavior in BSS, in pursuit of the goal of creating and maintaining a successful business. The step between a business rule and a business intent is very small; just a small shift of perspective.

A business rule is static and states what to do in a given situation, while a business intent states the desired or optimal outcome of a given situation – that is, interpreting what the stakeholder's interest is and trying to deliver as close to it as possible.

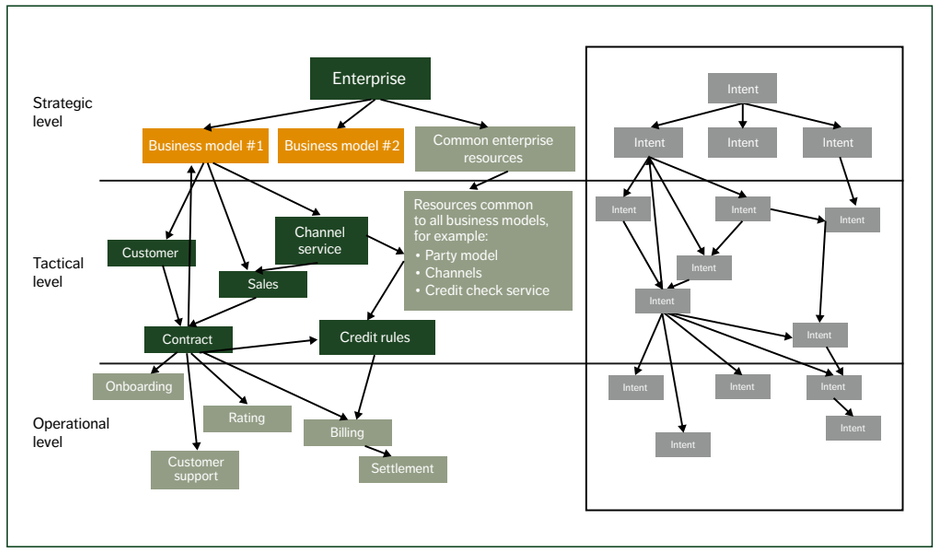


Figure 1 The business intent hierarchy mirrors the business rule hierarchy

Business intents can handle complex and dynamic situations and allow for feedback and comparison of actual and desired outcomes, enabling learning and knowledge building [1].

Higher-level intents in BSS are often expressed as business rules or KPIs. Intents are found at all levels of the business hierarchy, supporting both top- and bottom-line outcomes. An intent can range in complexity from an 'atomic intent' to an 'algorithm of intents' that combines a set of subintents. The term 'atomic intent' refers to the simplest possible intent structure, such as "our company will run a prepaid business model." Note that an atomic intent on a strategic level is likely to fan out into several intents on a lower level. While intents can be formulated for both human and machine consumption, they must be stated in a declarative format to facilitate automation in BSS.

Figure 1 illustrates the structure of the business intent hierarchy, which is designed to mirror the business rule hierarchy. An enterprise must have at least three different intent levels – strategic, tactical and operational [5] – all of which are the responsibility of the BSS. In software terms, these levels are equivalent to requirements, design and implementation, and execution.

	Strategic level	Tactical level	Operational level
Recurrence	One-off or few	Enough to learn	Very many
Human interaction	Yes	Yes	No
Learning/reasoning	Reasoning	Learning/reasoning	Execution
Feedback	Limited	Yes, key element	Large volumes
Main constraint	Quality	Quality	Time
Clarity of data use	Undetermined	Limited to own data	Deterministic and limited BSS set
Context, sources and data volumes	Many, external and large volumes	BSS	BSS and inference thereof
Data size	Large	Large but limited	Optimized for the intent
Data types and time series	All types	Many but limited	Few and limited to BSS origin

Figure 2 AI usage patterns are different at strategic, tactical and operational levels

The business strategy level owns and formulates the enterprise's top intents. For the sake of simplicity, Figure 1 breaks down only one business model to all three levels, but it is important to note that an average-size operator runs several different business models. The business tactic level is responsible for designing and implementing the intents of each of these business models at the operational level. This is normally achieved by breaking down the strategic intent into smaller, digestible parts, such as subintents with associated rules, information and processes. The business operation level – the core of traditional BSS – is responsible for executing to deliver the intents. This level requires further automation to meet ownership and business transaction cost requirements. All three levels benefit from AI support but have different usage patterns and characteristics, as shown in *Figure 2*. The differences most worth noting are in terms of repetitions, context, exploration and data characteristics.

The OODA loop

The OODA loop [6] was initially developed in the 1970s as an in-combat decision tool of the U.S. Air Force. OODA stands for observe, orient, decide and act. Many of the OODA loop's basic concepts are found in today's software agent systems. The version

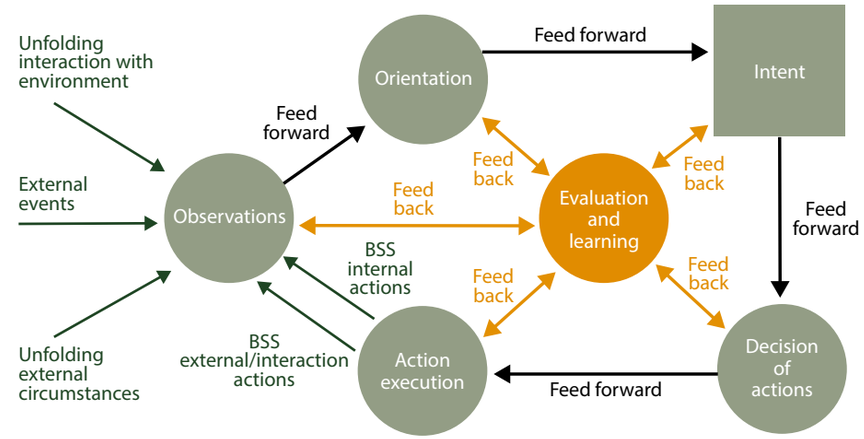


Figure 3 The OODA loop, modified to enable learning and intents

shown in *Figure 3* is complemented with explicit intent and learning capability.

The OODA loop idea is quite simple. First, observe or detect changes, events, stimuli or other things that happen in the context of interest, including internal states. Observations can be single or unfolding events, and they can be simple or complex in structure. Depending on the data, AI is often needed to interpret observations. Typical observations in BSS could be the availability of a customer's usage record or the arrival of a potential customer to a web shop.

The orientation step consists of aggregating and analyzing the observations that form the basis for the decisions. Analyzing the individual observations and aggregating them into a complete situation description requires multilevel AI support. The orientation step in BSS should enrich the observation with customer data as much as possible. This data enables the BSS to select the correct rating and charging parameters for a particular customer when their usage record becomes available, for example, or to conclude that a visitor to a web shop is looking for a new phone but seems to be price sensitive.

To clearly separate the intent from the decision of action that fulfills the intent, we have added intent to our modified OODA loop. We achieved this by

dividing the traditional OODA decision step into two distinct process steps: intent and decision. Intents in BSS are statements of the desired business outcome in a given situation. In the case of invoicing, this would mean ensuring that the rates/charges on the invoice are in accordance with the customer's contract. In the case of a price-sensitive potential customer visiting a web shop to look for a low-priced phone, the intent would be to convince that individual to buy a phone in the medium-price range rather than choosing the least-expensive option.

The decision is limited by the inventory of available possible actions. A selection is made from the available array of actions that best matches the intent. It is also possible to enrich the decision with simulations to predict the action outcome. The action is simply the execution of the decision. In BSS, actions are carried out by business processes and they result in business outcomes.

Examples of BSS decisions (and resulting actions) would include the decision to apply the standard rate/charge process in the case of a new customer usage record, or the decision to show a price-sensitive web shop visitor not only low-priced phones but also medium-priced models that have received excellent customer ratings.

Evaluation, learning and feedback are essential to build a system with optimal performance that can adapt to both business and enterprise-external changes. An optimal system uses the process of orientation, intents and decisions to continuously compare and evaluate both business outcomes and its own capabilities. Adaption may require new or additional higher-level analysis, algorithm redesign and algorithm retraining. Sometimes, it is enough to have good in-operations learning, such as a continuous algorithm retraining capability. An example of learning in BSS could be reaching the conclusion that when dealing with price-conscious web shop visitors, better outcomes can be achieved by showing a mix of low-price and medium-price phones with good ratings, as opposed to including the expensive phones as well.

The OODA loop can have various depths of reasoning, from deterministic to deep learning – that

> EVALUATION, LEARNING AND FEEDBACK ARE ESSENTIAL TO BUILD A SYSTEM WITH OPTIMAL PERFORMANCE

is, the same OODA-loop engine can be used to observe, decide and select the proper actions for all event types, regardless of complexity. It can also be used recursively to build layered structures with arbitrary depth – that is, it can support multilayered business processes and interactions.

It is critical that the people working in AI-enabled processes are able to understand the reasoning behind AI-generated results. All of the steps in our modified OODA loop can be understood and executed by both humans and machines – which makes it possible to work together in the most efficient way possible based on the particular circumstances of the organization.

Use case: reducing manual handling of invoices

An invoicing-related use case provides a good illustration of how AI adds value to BSS. In this scenario, a telecom operator notices an increase in the number of invoices that require manual handling, which is costly for the company. The executive team initiates a strategic project to address the issue.

The objective at strategic level is twofold: to establish facts on the invoice situation and to state a future invoice strategy – that is, to set an intent – regarding invoice handling and cost. The strategy team pools information about known challenges in invoicing and gathers external data for benchmarking. With the help of classification and statistical analysis algorithms, the following strategic-level intents for invoices are established by the strategic project:

- manual handling of less than 1 percent of all invoices
- average handling cost per invoice of less than USD 1.

This use of two dimensions of intent ensures a sound business balance, helping to avoid potential pitfalls, such as the possibility of reaching 0.0001 percent of manually handled invoices at an average cost of USD 100 per invoice.

Once the strategic intent has been set, work on the tactical level can begin. The primary challenge at this stage is to understand and classify all the reasons why some invoices require manual handling while others don't, and to select the optimal AI algorithms that can deliver results in line with the strategic intents. The tactical level begins by defining an efficient subintent structure and estimating each subintent's yield to the strategic intent, in order to select the most valuable ones. Then, for each subintent, it classifies possible AI algorithms to identify the best ones. Important considerations include:

- the information requirement and the complexity
- the volatility of the constituent knowledge components in the problem, which in turn determines whether machine learning is enough or if it must be combined with machine reasoning or deep learning to create deep enough or adaptable algorithms
- feasibility, effort and automation level in business operations
- cost estimations, implementation, operation and support.

In this type of invoicing use case, it makes sense to introduce customized communication patterns that vary according to customer character type. Therefore, part of the work at the tactical level involves defining three distinct customer personas – angry, regular and docile complainers, for example – and customize anomaly invoice messages for each of them. The next step is to test these different messages on a small portion (1-3 percent) of the customer population to find the right message for each customer persona to ensure that their complaints are resolved without escalation to manual handling. Selecting only a small fraction of the total customer population reduces the business risk.

Finding the necessary knowledge components is an iterative task that requires AI support and access to relevant information. The tactical level is also responsible for the AI algorithm life cycle including design, implementation and operational launch. Further, it is responsible for stating the necessary changes to BSS, so it can both calculate according to the AI algorithms and automatically execute the new behavior in the invoice functionality.

The tactical-level subintents for the invoicing use case are:

- invoice input correctness: higher than 99.999 percent
- invoice anomaly statistics and predictions at both group and individual level: anomaly type, costs, volumes, services and dates
- customer persona classification into three levels (angry, regular and docile complainers) with less than 1 percent error
- customized message success rate above 90 percent.

The tactical-level changes to BSS in terms of new rules, information and processes are:

- calculate invoice anomalies and recheck invoice input if anomaly probability is higher than 15 percent
- determine customer persona complaint classification with continuous learning capability
- customize the message success rate to achieve continuous learning capability
- instruct customer support team to "cut it short" in cases with no anomaly and with angry
- calculate and expose subintent and intent outcome, along with their projections and variance.

The methodology of the tactical level is similar to that of AI-supported CRISP-DM [7]. Once the tactical level steps have been completed, the role of the operational level is simply to execute the algorithms and the new BSS logic with as much automation as possible. After a short training period, the CSP can sort out the invoices that require

manual handling, identify the root causes, classify customers that systematically complain, and select the most efficient customized message. The error rate and the cost per invoice are initially quite high but decrease rapidly below the strategically stated intent as the invoice communication is tuned.

Implications

The invoicing use case makes it clear that BSS must have an omnipresent AI ability to be able to support strategic and tactical investigations, as well as having the agility in operations to change behavior to accommodate new algorithms, rules, information and processes. The inclusion of business models – that is, the grouping of rules, information and processes tuned to work together to deliver business outcomes for specific business situations – is also critical in the evolution of BSS. The intent structures must mirror the business model structures and their life cycles.

AI-native BSS require an expansion of the business logic elements – rules, information/objects and processes – to include intents and events. The business logic elements, often hidden inside applications, must be externalized to support the conversion of AI findings to automated BSS behavior. The business information must be structured in an ontology and made available to all business support users and applications, AI systems included, into a business information lake.

AI-native BSS must support both run-time and business-design-time. Traditional BSS are primarily built with few configuration options for run-time, resulting in less agility [5]. While it is true that AI can help a business evolve in running BSS (for example in continuous development and operations mode), there is no avoiding the fact that this requires a BSS architecture that is at least partially new.

Conclusion

It is widely recognized that 5G and the IoT represent the main growth opportunities for communication service providers (CSPs) in the coming decade. To support emerging use cases in these areas, CSPs require business support systems (BSS) that can handle complex business situations and optimize outcomes with minimal manual intervention. Artificial intelligence (AI) is the obvious answer, but introducing it into existing BSS is problematic for a number of reasons. Instead, Ericsson recommends an architectural change to traditional BSS to create AI-native BSS. Most significantly, this evolution requires the inclusion of an enterprise's strategic, tactical and operational levels in the BSS, together with the introduction of two new business logic elements (intents and events). One of the key differences between traditional BSS and AI-native BSS is the fact that AI-native BSS enable the various applications within the BSS to share business information with each other in an efficient and secure manner – a critical capability in the emerging 5G-IoT world.

Further reading

- **Ericsson, The dawn of machine intelligence, available at:** *https://www.ericsson.com/en/news/2017/9/the-dawn-of-machine-intelligence*
- **Ericsson, Zero-touch could herald a new era in service provider customer interaction, available at:** *https://www.ericsson.com/en/press-releases/2018/5/ericsson-zero-touch-could-herald-a-new-era-in-service-provider-customer-interaction*

THE AUTHORS

Lars Angelin
◆ is an expert in BSS within Business Area Digital Services at Ericsson. He has more than 30 years of experience in the areas of

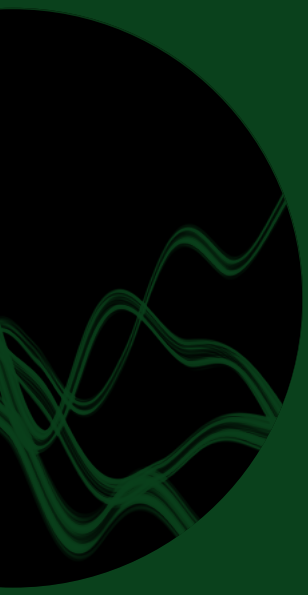
concept development, architecture and strategies within the telco and education industries. Angelin joined Ericsson in 1996 as a research engineer, and in 2003 he moved to a position as concept developer in the M2M and OSS/BSS areas. Since 2006 he has focused on BSS – specifically business support, enterprise architectures and the software architectures to implement BSS systems. He holds an M.Sc. in engineering physics, a Tech. Licentiate in tele-traffic theory from Lund Institute of Technology in Sweden, and an honorary Ph.D. from Blekinge Institute of Technology in Sweden.

Johan Silvander
◆ is a senior specialist in information management who has worked at Ericsson for more than 20 years. Over the years, his work has focused on the areas of OSS and BSS in a variety of different roles, including serving as a member of core architecture teams, working as a designer, taking technical responsibility for integration and installation projects, and being a test leader. He holds a Tech. Licentiate in computer science from Blekinge Institute of Technology, where he is currently pursuing a Ph.D.

The authors would like to thank Jörg Niemöller for his contributions to this article.

References

1. **TM Forum, Open Digital Architecture, 2018, available at:** *https://www.tmforum.org/resources/whitepapers/open-digital-architecture/*
2. **Ericsson, Unlocking 5G's revenue potential: a roadmap for operators (press release), February 26, 2018, available at:** *https://www.ericsson.com/en/press-releases/2018/2/unlocking-5gs-revenue-potential-a-roadmap-for-operators*
3. **Forbes, How Artificial Intelligence Is Revolutionizing Business In 2017, September 10, 2017, Louis Columbus, available at:** *https://www.forbes.com/sites/louiscolumbus/2017/09/10/how-artificial-intelligence-is-revolutionizing-business-in-2017/#58ebeab25463*
4. **Harvard Business Review, Artificial Intelligence for the Real World, January-February 2018, Thomas H. Davenport and Rajeev Ronanki, available at:** *https://hbr.org/2018/01/artificial-intelligence-for-the-real-world*
5. **McKinsey, Smarter analytics for banks, September 2018, Carlos Fernandez Naviera et al., available at:** *https://www.mckinsey.com/industries/financial-services/our-insights/smarter-analytics-for-banks?cid=other-eml-alt-mip-mck-oth-1810&hlkid=d3be9327efb84eccb44a1d8d391d0d8f&hctky=2669978&hdpid=ddc7fd1c-3815-4675-ae1a-f53e67d88452*
6. **OODA loop definition available at:** *https://en.wikipedia.org/wiki/OODA_loop*
7. **CRISP-DM definition available at:** *https://en.wikipedia.org/wiki/Cross-industry_standard_process_for_data_mining*



ISSN 0014-0171
284 23-3337 | Uen

Ericsson
SE-164 83 Stockholm, Sweden
Phone: +46 10 719 0000