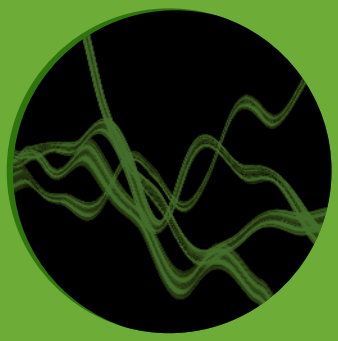
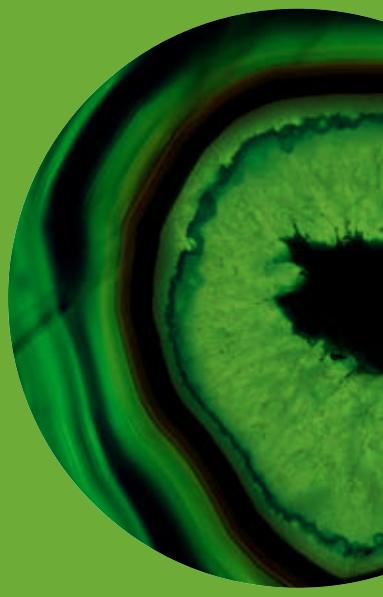
CHARTING THE FUTURE OF INNOVATION | VOLUME 94 I 2017-01

ERICSSON TECHNOLOGY Review





SUSTAINING TRUST IN A DATA-DRIVEN SOCIETY

5G AND FIXED WIRELESS ACCESS

INSIGHTS FROM CUSTOMER EXPERIENCE AWARENESS

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 on the potential, practicalities, and benefits of a wide range of technical developments, and help provide an 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

Additionally, content can be accessed on the Ericsson Technology Insights app, which is available for Android and iOS devices. The download links can be found on the Ericsson Technology Review website.

PUBLISHER
Ulf Ewaldsson

EDITOR
Tanis Bestland (Sitrus)
tanis.bestland@sitrus.com

FEATURE ARTICLE
Sustaining legitimacy and trust in a data-driven society by Stefan Larsson (Lund University Internet Institute)

ART DIRECTOR
Kajsa Dahlberg (Sitrus)

PRODUCTION LEADER
Fay Scafe (Sitrus)

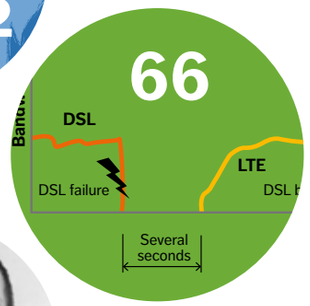
LAYOUT
Jade Birke, Carola Pilarz (Sitrus)

ILLUSTRATIONS
Claes-Göran Anderson , Sitrus Ukraine, Michèle Harland (Harland Creations)

CHIEF SUBEDITOR
Birgitte van den Muyzenberg (Sitrus)

SUBEDITORS
Paul Eade and Ian Nicholson (Sitrus)

ISSN: 0014-0171
Volume: 94, 2017



DELIVERING **VALUE** IN THE **DIGITAL AGE**

■ **THE WORLD CONTINUES TO CHANGE** at breakneck speed. Regardless of which industry we work in, we must all stay alert to an ever-wider sphere of knowledge and information to make sure we don't miss the boat on "the next big thing". To that end, this issue of Ericsson Technology Review includes five articles that present a broad range of ideas and solutions that boost capacity, efficiency and security, along with one about the latest technology to help you better understand the needs of your users.

This issue also features a guest article by Stefan Larsson, Associate Professor in technology and social change at Lund University Internet Institute (see page 42). Larsson is an expert on digital socio-legal change, including issues of trust, consumption, traceability and privacy. He points out that while user trust is heavily based on their perception of the technological security of a solution or service, "it is also fundamentally dependent on social norms and values such as privacy, legitimacy and perceived fairness in the collection and handling of individual information." He argues persuasively that the long-term success of the digital economy will be dependent on consistently high levels of both technological and sociological trust among users, and explains his view of what service providers must do to preserve them.

Every new generation of mobile technology brings with it fresh hope for true network convergence – the ability to provide both mobile and fixed broadband access via the same technology and the same infrastructure. Today, the hope is that 5G will prove to be the technology that will help us achieve that goal, by enabling both high-speed mobile broadband and fixed wireless access (FWA) on a massive scale. While FWA has previously been proven to work in 4G or even 3G scenarios, the huge capacity improvements in 5G are set to enable a proliferation of FWA solutions for both small and medium-sized enterprise and residential applications around the world. The article on page 50 explores the FWA opportunity for 5G and presents a solid 5G FWA use case, as well as a variety of possible FWA transport solutions. Together with all the benefits it will bring in the areas of mobile broadband and the IoT, its potential to be used as an FWA enabler makes the case for 5G stronger than ever.

No matter how efficient, robust and reliable a network is, the amount of bandwidth available to users is ultimately determined by the length of the last mile – the final segment of the network that physically delivers the service to users. As a result of the massive growth in popularity of bandwidth-consuming online activities in recent years, many service providers are struggling to meet consumers' ever-expanding bandwidth requirements. While the obvious solution would be to shorten the length of the last mile by deploying new street cabinets and fiber lines, doing so is not always feasible in terms of time and cost. In these cases, access aggregation is a viable alternative. The article on page 66 presents a Multipath TCP-based solution that we believe is the most efficient and effective option for access aggregation. The use of carrier-grade Multipath TCP proxies makes it possible to increase bandwidth, improve reliability and achieve seamless connectivity without the need to introduce Multipath TCP in end devices or internet servers.

Just a few years ago, encrypted web traffic was the exception rather than the rule, but revelations about pervasive surveillance and an increase in cyber-attacks have prompted many service providers to switch to communication using HTTPS to reduce the vulnerability of data carried over networks. As a result, 70 percent of today's web traffic is encrypted, and HTTPS will soon become the default protocol. While the move to HTTPS offers benefits in terms of protecting data, it also presents some challenges for operators and content providers – particularly when it comes to caching. The article on page 8 presents a solution that enables optimal placement of cache stores within a network, and traffic management in the RAN, minimizing backhaul traffic and reducing latency. As such, the proposed solution provides operators with a new business opportunity in the context of an all-encrypted web: to offer content providers optimally-placed caches for secure content delivery.

In the near future, IoT systems will need to support extremely large-scale field applications made up of an enormous diversity of connected things. Since each of these things needs to be identified in multiple domains, a solid understanding of identity management is essential to the development of IoT security solutions that are both flexible and robust. There are currently several ways to manage identities across domains. Finding the optimal one for a particular application depends on the relationship(s) between the domains, the domain-specific identity data, and the systems and technologies available. The article on page 32 presents an overview of the key concepts in identity management, and explains how they can be applied in the IoT environment to achieve optimal levels of efficiency, usability and security. In particular, it highlights how technologies like Generic Bootstrapping Architecture and specific identity management systems for the IoT can substantially reduce the complexity of identity management.

A deeper understanding of the multifaceted user experience is critical for any service provider to succeed in an increasingly competitive space. Objective QoS, while still important, is no longer enough to satisfy users in the constantly changing digital environment they live in today. For them, high service quality is simply a given. What they're looking for are partners that can help them navigate and utilize technologies and capabilities in the emerging Internet of Smart Everything as smoothly as possible, providing them with seamless services and intuitive spaces for interaction. Meeting the ever-evolving expectations of today's users requires a high level of customer experience awareness. The article on page 20 explains the concept of customer experience awareness and demonstrates how next-generation customer experience management analytics tools can be used to generate actionable business insights that enable service providers to continuously improve the user experience.

As always, I hope you find the contents of the magazine relevant and inspiring. All of the articles included here are also available online at www.ericsson.com/ericsson-technology-review, through the Ericsson Technology Insights app and on SlideShare.

ULF EWALDSSON

SENIOR VICE PRESIDENT, CHIEF STRATEGY AND TECHNOLOGY OFFICER AND HEAD OF GROUP FUNCTION STRATEGY AND TECHNOLOGY

> A DEEPER UNDERSTANDING OF THE MULTIFACETED USER EXPERIENCE IS CRITICAL FOR ANY SERVICE PROVIDER TO SUCCEED IN AN INCREASINGLY COMPETITIVE SPACE.

Blind cache

A SOLUTION TO CONTENT DELIVERY CHALLENGES IN AN ALL-ENCRYPED WEB

GÖRAN A.P. ERIKSSON, JOHN MATTSSON, NILO MITRA, ZAHEDUZZAMAN SARKER

As the internet has shifted from an informal network for information sharing among universities and scientific institutes to a fundament of modern commerce, the traffic it carries has also shifted from short message exchanges to massive files encrypted for protection. At one time, the use of the secure communication protocol HTTPS tended to be limited to applications such as internet banking and online shopping. By the end of 2016, however, 70 percent of web traffic will use HTTPS [1]. Encryption is likely to become mandatory everywhere as the rise in pervasive surveillance has changed public perception on privacy and internet security, which has in turn prompted internet content providers to adopt HTTPS to protect consumers' content consumption data. The shift to pervasive encryption brings many benefits, but it also presents significant challenges for network service providers (such as mobile network operators and internet service providers) – particularly when it comes to caching content in their networks.

CONTENT DELIVERY networks (CDNs) improve delivery efficiency by replicating popular content like video streams on a cache serving users that are networked geographically nearby. The widespread adoption of CDNs has reduced latency and the amount of traffic carried by backbone networks. Existing CDN technology requires content providers to delegate their valuable content and expose traffic to the CDN provider, compromising end user privacy and security while revealing valuable business information.

By adopting HTTPS, the content delivery process becomes more robust, and both user security and privacy improve. However, the use of end-to-end encryption takes away the ability for the network service provider to use transparent inline caches to serve previously requested content, thus increasing backbone traffic as all requests for content have to be forwarded to and served by the content provider. Large content providers can sidestep this issue by placing content on their own edge servers, but this approach is costly and increases system complexity. Smaller content providers need another solution so that they can maintain low-latency content delivery and at the same time protect consumer privacy and ensure security. This need presents a business opportunity for network service providers to offer a better way to cache content that preserves the security and privacy of content providers and consumers.

To overcome the caching challenge inherent in HTTPS, Ericsson is collaborating with internet companies that have expertise in this area. Together we are exploring a solution we call blind cache, which is also referred to as out-of-band cache in industry discussions.

> THE USE OF END-TO-END ENCRYPTION TAKES AWAY THE ABILITY FOR THE NETWORK SERVICE PROVIDER TO USE TRANSPARENT INLINE CACHES TO SERVE PREVIOUSLY REQUESTED CONTENT

Encryption everywhere

There are very strong indications that the web is moving quickly toward HTTPS everywhere, or at least almost everywhere. This transition is driven by industry trends such as browser vendors providing HTTPS as the default option, warnings for non-HTTPS connections, and simpler and cheaper certificates and certificate management capabilities for smaller websites.

As *Figure 1* shows, the use of HTTPS has been rising since 2012. Further, it is greater in mobile networks than in fixed ones [1], owing to the growth of video mobile apps, which tend to use HTTPS by default. From a standardization perspective, security and privacy are the primary factors that need to be taken into consideration in developing

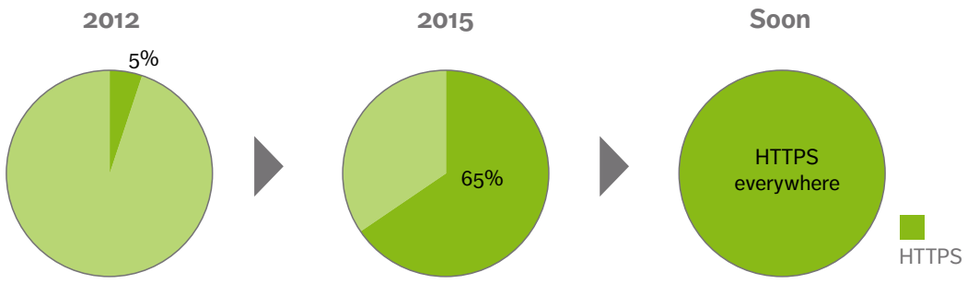


Figure 1
Evolution of HTTPS usage share in mobile networks

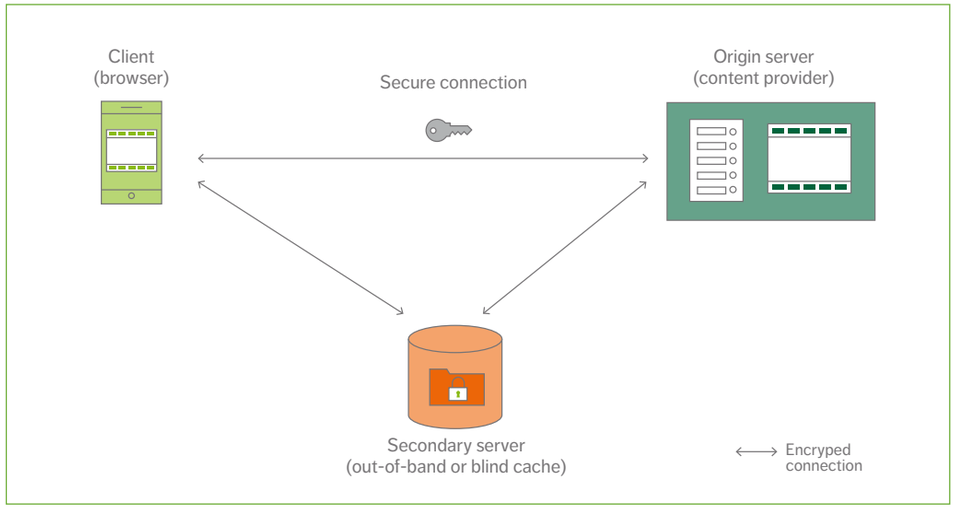


Figure 2
Blind cache architecture

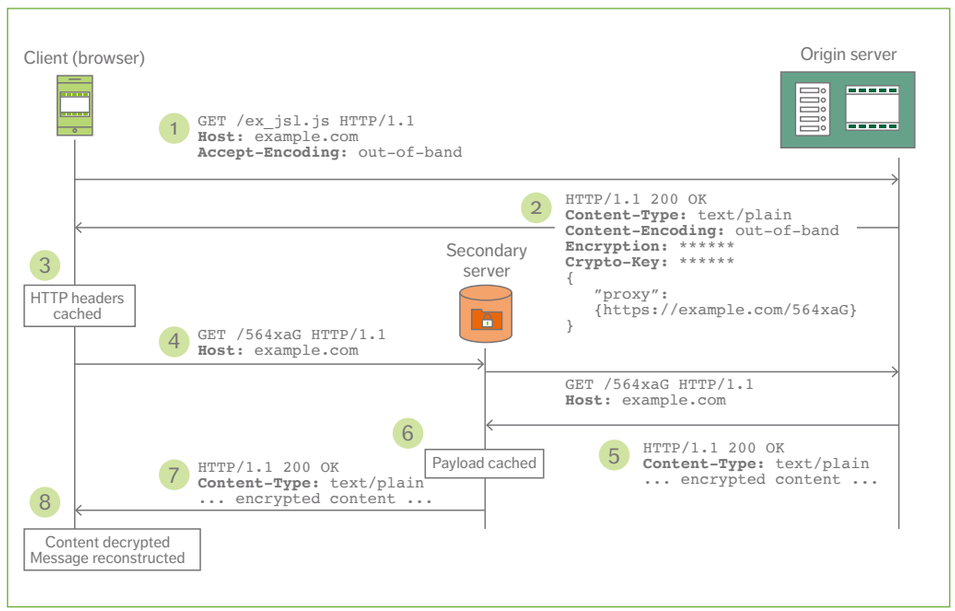


Figure 3
Blind cache call flow

the blind cache solution. Content providers, on the other hand, are primarily driven by the desire to secure end-to-end delivery, to protect ownership of valuable analytics data, and to protect against issues caused by network intermediaries such as ad injectors and application layer firewalls.

The web platform is transforming rapidly and the trend towards HTTPS is just one component of the ongoing evolution of the content delivery stack. On the client side, the industry is moving towards a new encrypted transport protocol evolving from QUIC [2], while content providers and CDNs are migrating to HTTP/2 [3] where the standard option is to mandate TLS or QUIC. The future content stack is HTTPS – HTTP or HTTP/2 over QUIC or TCP/TLS, so that a third party cannot read, alter, delete, insert, or replay data in any way.

Whether they like it or not, network service providers will have to adapt to HTTPS for all their traffic in the near future. As this requires preparation, it is time to start planning for their role in an all-encrypted web.

> “WHETHER THEY LIKE IT OR NOT, NETWORK SERVICE PROVIDERS WILL HAVE TO ADAPT TO HTTPS FOR ALL THEIR TRAFFIC IN THE NEAR FUTURE”

Solution concept

The philosophy of the blind cache solution is rooted in the concept of awareness; that all principal parties involved in a communication are explicitly aware of, and authorize the presence of, any intermediary participating in the data exchange. Adhering to this philosophy overcomes one of the main concerns about current, non-HTTPS, practice, in which the legitimate use of an intermediary such as an inline cache is made without the knowledge or consent of the user, and is indistinguishable from the actions of attackers. The target solution needs to ensure that service providers, such as a content provider, do not lose control of their content or other information flowing between client and server – a current drawback when content is served via a third party such as a CDN.

The blind cache solution [4] meets these requirements by creating a three-way relationship among the relevant actors, as shown in *Figure 2*.

It allows a content provider to deliver content faster by utilizing the support functions of external secondary servers for caching purposes. The solution places a requirement on the secondary server that any content cached on it remains encrypted and tamper-proof. It allows the content provider to decide if a secondary server should be used for a particular resource, such as an image, a JavaScript library or a set of video segments.

There are three scenarios available to a content provider for such delegated caches, with decreasing levels of trust.

Case 1 – edge origin

In this case, the secondary server hosting the blind cache belongs to the content provider. It is under the administrative and legal control of the content provider and is thus similar in character to the origin server.

Case 2 – CDN

In this case, the secondary server hosting the blind cache belongs to a third party – such as a CDN provider or a network service provider – with which the content provider has business and service level agreements.

Case 3 – proxy cache

In this case, the secondary server hosting the blind cache is hosted by any party known to the device, but does not require a business relationship with the content provider.

In all three cases, the content provider remains in charge of deciding what, if any, content to serve via blind caches. Since the caches are blind to the content they serve and unable to modify it, it is more likely that content providers will have

> THE CONTENT PROVIDED TO A BLIND CACHE CAN BE ENCRYPTED, AND THE INTEGRITY OF THE CONTENT IS PROTECTED BY APPLYING SUITABLE TECHNIQUES

the reassurance they need to use them. The blind cache solution also allows the content provider to delegate caching and serving of sensitive content to an untrusted server (or to a server on an untrusted site or cloud platform), or use a CDN provider that it might not have selected previously.

Case 1 is interesting given the possibilities for distributed cloud computing in mobile networks, which will lower the price point for a content provider operating an edge server remotely on a site and/or cloud platform under the control of a third party. An example of this is a mobile network operator that offers a cloud execution platform at a local central office site.

Case 2 enables a mobile network provider or a global CDN operator to provide deep edge-caching infrastructure.

Case 3 allows a client with a configured proxy – which is almost always the setup for enterprise users – to indicate its presence to the content provider. The content provider can then decide whether or not to allow the proxy to serve content on its behalf – restoring the efficiency gains of proxies in the service delivery chain, while preserving the security of the client-server communication.

It should also be noted that if the situation requires, the content provider could decide to have a virtual edge server serving only one user, a private cache. This could, for instance, be motivated when delivering a large file such as a software update to a particular enterprise customer, as an alternative to reserving a VPN connection to secure the timely delivery from a faraway origin server site.

Solution design

The message format used in the proposed technical solution is JSON, and the call flow is illustrated in *Figure 3*. The blind cache solution introduces the new parameter out-of-band [5] in the Accept-Encoding HTTP request-header field. Through this new parameter, a client indicates if it can (step 1 of Figure 3) handle HTTP responses where the payload is retrieved out-of-band (that is, from another server) separately from the main response. If the blind cache solution were to become well established, browsers would likely implement this feature and add the value by default in all requests.

If the content provider makes use of a blind cache, the HTTPS response from the origin server (2) will include the same out-of-band value in the Content-Encoding header field, informing the client that a cache will be used to deliver the payload of the response. The URL of the cache is included in the message body. Multiple URL may be provided, the first one linking to the primary cache with subsequent addresses linking to backups if the primary cache is inaccessible. The solution introduces two additional HTTP headers, Encryption and Crypto-Key, which contain the information for the client to decrypt the payload after retrieval, which the client stores (3).

In the next step, the client retrieves the payload from the URL provided (4). If the content does not exist in the cache, which may be the case when content is requested for the first time, the cache retrieves it from the origin server at the content provider (5), and caches it (6) for future requests. The content is sent back to the client (7) which can then decrypt the requested content (8) by applying the key acquired in step 3 to the encrypted payload obtained in step 7.

In deployment scenarios like edge origin and CDN (cases 1 and 2), content providers might want to pre-populate blind caches with content that is likely to be popular.

In addition to the call flow shown in Figure 3, the use of a client-selected proxy cache (case 3) can be achieved through the use of a proposed new HTTP header field with the working name BC [6]. If present, BC indicates that the client is connected to a proxy cache that it is willing to use to retrieve content. If the content provider returns a response including the out-of-band value in the Content-Encoding header field, it accepts the use of that proxy cache (with which it has no relationship whatsoever) and is willing to delegate the handling of certain content to it.

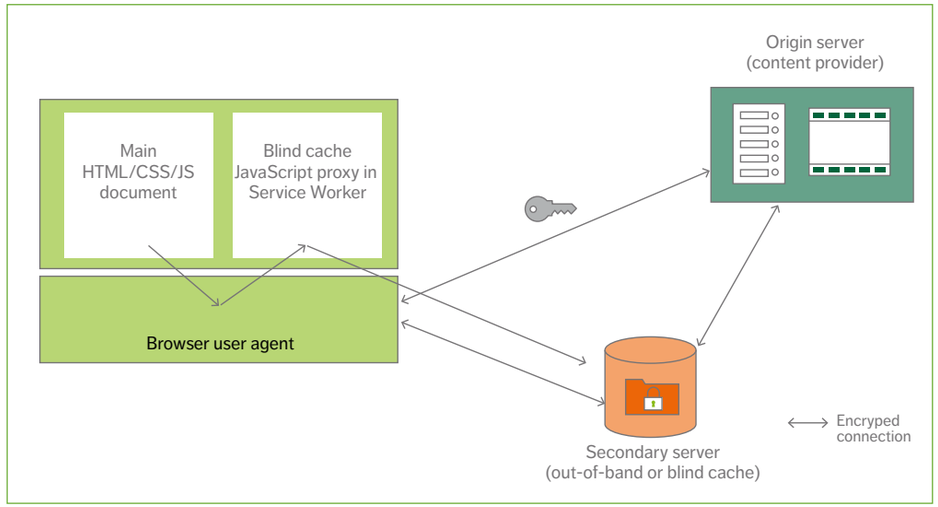


Figure 4
Overview of a browser-based client test environment

The content provided to a blind cache can be encrypted, and the integrity of the content is protected by applying suitable techniques. One such technique ensures that resources at the origin cannot be inferred from delegated resources stored in a cache, and another prevents a cache from pretending to be a client and querying the origin, which could lead to discovery of the actual resources at the origin server [7].

Efficiency gains can be achieved through the use of a resource map that contains meta-information that the origin server can provide to the client. The resource map includes meta-information for all the delegated sub-resources – such as scripts, images, and video clips – that the client needs to create a complete representation of the set of cached resources. If the client and the origin server both support HTTP/2, the meta-information can be pushed to the client directly using the server push mechanism. Subsequent individual client requests for sub-resources can be redirected to the blind cache, which not only reduces latency for the actual request, but also decreases network traffic and processing costs at the origin server.

Tried and tested

Initial lab experiments conducted by Ericsson Research show that the use of a blind cache results in significant gains for serving a typical web page when the blind cache is pre-populated with the sub-resources or when the client is provided with the resource map – compared with retrieving the same content directly from the origin over a secure connection. As might be expected, the gain is greater when latency between the client and the origin server is longer.

To assess the HTTP extensions for blind cache, Ericsson Research developed a browser-based experimental testbed that is shown in *Figure 4*. The test environment uses the service worker mechanism [8] to implement a JavaScript-based

The tests require the client to fetch a test webpage directly from the origin server and the same content from the blind cache. The page load time is used as the key performance indicator (KPI) in the test results.

The blind cache can be **primed** (that is, it already has webpage resources), or **non-primed** (it does not have webpage resources and the resources are pushed from the origin server).

The client can also be in two states: **configured** – the Service Worker knows where to fetch contents, or **non-configured** – the resource map is not yet known.

The tests were run simulating different RTT values in the path between (1) the client and the origin server, and (2) the blind cache and the origin server.

Scenario A
RTT between the client and the server = 200 ms-300 ms
RTT between the client and blind cache = 40 ms
RTT between the blind cache and the origin server = 100 ms

Scenario B
RTT between the client and the server = 200 ms
RTT between the client and the blind cache = 40 ms
RTT between the blind cache and the origin server = 100 ms-200 ms

The results show that given high RTT between the client and the origin server, the proposed solution architecture will still be able to improve the user experience by a substantial margin (a page load time improvement of up to 30 percent). It also shows that the different delay between the cache and the origin server does affect the overall performance. The extra overhead required in terms of the number of extra bytes exchanged and extra request generated is also low compared with the gain in the responsiveness in page load.

The testbed helped identify different issues in the prototype and important features, which contributed to the evolution of the solution architecture and protocol design. The quest for more data and results continues; additional content types including video and delay scenarios will be added in the future.

Overhead due to use of BC

	% of extra bytes exchanged	Extra request generated
Primed, non-configured	3.03	3
Primed, configured	0.85	None
Primed, configured, all content via cache	1.29	None

Scenario A compared with end-to-end TLS

Cache primed?	Client configured?	All content via cache?	Page load time efficiency
RTT = 200 ms			
Yes	Yes	No	+27%
Yes	No	No	+11%
Yes	Yes	Yes	+38%
RTT = 300 ms			
Yes	Yes	No	+30%
Yes	No	No	+13%
Yes	Yes	Yes	+45%

Scenario B compared with end-to-end TLS

Cache primed?	Client configured?	All content via cache?	Page load time efficiency
RTT = 160 ms			
Yes	Yes	No	+41%
Yes	No	No	+14%
Yes	Yes	Yes	+39%
RTT = 200 ms			
Yes	Yes	No	+42%
Yes	No	No	+12%
Yes	Yes	Yes	+41%

proxy in the user agent that intercepts the HTTPS request from the main page and retrieves the resource map from the origin server as well as the origin payload from the secondary server.

When a response is received from the secondary server, the payload is decrypted and its integrity verified using the key provided by the origin server. Once the payload has been decrypted, subsequent requests for the webpage content are provided using the standard APIs in the Service Worker.

The content used in our tests included typical web resources, such as images, text, and DASH-segmented video.

Given the basic assumption that the latency between the client and the cache is lower than latency between the client and the origin server, tests were carried out to evaluate the proposed design for RTT – latency – in the various connections between the client, origin server, and secondary server (cache).

One potential disadvantage of a blind cache solution is the extra RTT required to retrieve out-of-band encoding meta-information from the origin followed by the content from the secondary server hosting the blind cache. To avoid this, the origin server responds to the client with out-of-band encoding information for a set of resources and not only the requested one. The tests were carried out with a view to improving the implementation of the protocol extensions, and to verify the assumption that a blind cache placed in close proximity to the client actually provides the desired benefits.

> THE BLIND CACHE SOLUTION IS A SIGNIFICANT STEP TOWARD ENABLING CONTENT PROVIDERS TO LEVERAGE DEEPLY DISTRIBUTED EDGE CACHES WHILE MAINTAINING CONTROL OVER THEIR CONTENT AND ITS USE

Figure 5 is an example of a test scenario with different RTT measurements between client-origin, origin-cache, and cache-client in our testbed. In this particular scenario the page load time at the client improves by about 20 percent when fetching the web resources from the secondary server rather than fetching these directly from the origin server over HTTPS, thus confirming the benefit of using a resource map. As the client-origin RTT increases while other factors remain unchanged, the improvement in page load times rises.

A cloud of caches

The blind cache solution establishes the basic concept of placing and accessing content in a delegated cache in a secure manner. The next step is to explore possible ways to improve the efficiency of delivery, potentially using new deployment modes. *Figure 6* shows one area Ericsson Research is investigating: a set – or cloud – of caches.

In this hierarchy of caches, the top-level cache might belong to the same administrative domain as the origin server – the content provider, in other words. With the proper cache topology knowledge, the top-level cache can act like an HTTPS request router, redirecting requests for sub-resources to appropriate secondary caches based on a number of parameters, such as predictive traffic load, client location and topology related costs such as transport and CDN usage.

◀ PERFORMANCE TESTBED RESULTS

Key assumptions for the testbed:
a) Low bandwidth and high latency between the client and the origin server.
b) High bandwidth and low latency between the client and the blind cache.
c) The client and the blind cache may have the same access network characteristics towards the origin server.

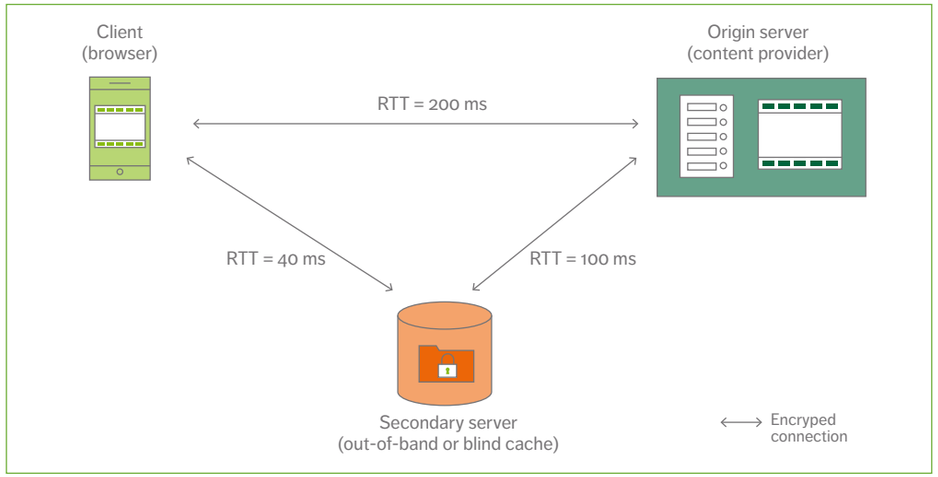


Figure 5
Example of a test with some sample latencies

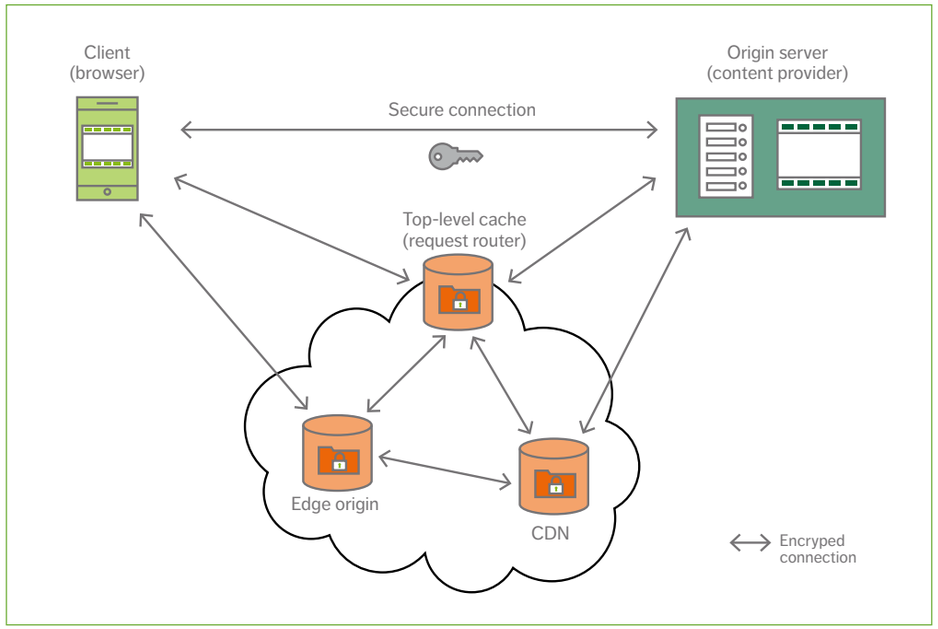


Figure 6
A set of caches

In Figure 6 the edge origin cache may belong to the administrative domain (the self-delegation case) of the content provider. The content provider could deploy such caches deep in a mobile operator's network – at a local central office site or a remote data center site, for example – leveraging shorter RTT and avoiding internet peering costs.

If the edge origin caches are provided by third parties for volume deployments in, for example, hotels, or to support Wi-Fi hotspots, the cost of provisioning certificates to secure the channel between the client and the caches can be reduced by provisioning the cache with inexpensive, self-certifying certificates such as those offered by the Let's Encrypt organization [9] and provisioned using the ACME protocol [10].

The benefits of using blind caches

The blind cache solution is a significant step toward enabling content providers to leverage deeply distributed edge caches while maintaining control over their content and its use. The solution provides network service providers with a business opportunity in an all-encrypted web, as it enables them to provide optimal placement of caches within their networks, as well as, for mobile operators, traffic management in the RAN that reduces latency and minimizes backhaul traffic. The solution will be particularly useful in 5G and LTE-U scenarios, where the need for caches will intensify – in residential gateways, hotspots, vehicles, and transportation systems – due to the dramatic increase in applications based on video traffic, which is projected to experience a 45 percent compound annual growth between now and 2021 [11]. ✱

Terms and abbreviations
ACME – Automated Certificate Management Environment | **API** – application programming interface | **CDN** – content delivery network | **DASH** – Dynamic Adaptive Streaming over HTTP | **HTTP/2** – Hypertext Transfer Protocol Version 2 | **IETF** – Internet Engineering Task Force| **JSON** – JavaScript Object Notation | **LTE-U** – LTE in unlicensed spectrum | **QUIC** – Quick UDP Internet Connection | **RTT** – round-trip time | **TLS** – Transport Layer Security | **W3C** – World Wide Web Consortium

References

1. **Sandvine, 2016, report, Global Internet Phenomena Spotlight – Encrypted Internet Traffic, available at:** *https://www.sandvine.com/trends/encryption.html*
2. **IETF draft, July 2016, QUIC- a UDP based Secure and Reliable Transport for HTTP/2, Work in Progress, available at:** *https://datatracker.ietf.org/doc/draft-hamilton-early-deployment-quic*
3. **IETF RFC 7540, 2015, Hypertext Transfer Protocol Version 2 (HTTP/2), available at:** *https://tools.ietf.org/html/rfc7540*
4. **IETF draft, June 2016, An Architecture for Secure Content Delegation using HTTP, Work in Progress, available at:** *https://tools.ietf.org/html/draft-thomson-http-scd*
5. **IETF draft, July 2016, 'Out-Of-Band' Content Coding for HTTP, Work in Progress, available at:** *https://tools.ietf.org/html/draft-reschke-http-oob-encoding*
6. **IETF draft, March 2016, Caching Secure HTTP Content using Blind Caches, Work in Progress, available at:** *https://tools.ietf.org/html/draft-thomson-http-bc*
7. **IETF draft, June 2016, Encrypted Content-Encoding for HTTP, Work in Progress, available at:** *https://tools.ietf.org/html/draft-ietf-httpbis-encryption-encoding*
8. **W3C, 2015, specification, Service Workers, available at:** *https://www.w3.org/TR/service-workers/*
9. **Let's Encrypt, available at:** *https://letsencrypt.org/*
10. **IETF draft, July 2016, Automatic Certificate Management Environment (ACME), Work in Progress, available at:** *https://tools.ietf.org/html/draft-ietf-acme-acme*
11. **Ericsson, 2016, Mobility Report, available at:** *https://www.ericsson.com/res/docs/2016/ericsson-mobility-report-2016.pdf*

THE AUTHORS

Göran A.P. Eriksson

◆ joined Ericsson Research in 2000 and is an expert in communication services. He has worked on web technologies on the device and server side, ranging from SOA and composition engines to web protocols such as SIP, RTCWEB and HTTP to client side browser APIs – in particular WebRTC. He holds an M.Sc. in engineering physics from KTH Royal Institute of Technology, Sweden.

John Mattsson

◆ joined Ericsson Research in 2007 and is a senior researcher. In 3GPP, he has had a great influence on the work being done on IMS security and algorithm profiling. He coordinates Ericsson's security work in the IETF, and is currently working on cryptography as well as transport and application layer security. He holds an M.Sc. in engineering physics from KTH Royal Institute of Technology, Sweden, and an M.Sc. in business administration and economics from Stockholm University.

Nilo Mitra

◆ joined Ericsson in 1998 after 15 years at AT&T Bell Laboratories. He is an expert standardization architect for media solutions at Business Unit Media, where he coordinates Ericsson's participation in various media-related

standards organizations. He has participated in many standardization fora including HbbTV, Open IPTV Forum, W3C, OASIS, WS-I, OMG, ITU-T, ISO, OMA and Streaming Video Alliance, having held leadership roles in several of these. Nilo holds a Ph.D. in theoretical physics from Columbia University, US.

Zaheduzzaman Sarker

◆ joined Ericsson Research in 2007 and is a senior researcher in the services, media and network feature domains. During this period, he has been involved in work from prototype development

to standardization, and from protocol design to radio network simulations (4G). His special areas of interest are real-time video communication, web technologies and performance measurements. He works with transport and application layer protocols in IETF, and is currently serving as Ericsson IETF coordinator. He holds an M.Sc. in computer science and engineering from Luleå University of Technology, Sweden.

GENERATING ACTIONABLE INSIGHTS FROM

customer experience

AWARENESS

In today's ICT marketplace almost all networks provide a high degree of objectively measurable quality. As a result, quality alone is no longer sufficient to distinguish a service provider from the competition and ensure customer loyalty. Instead, customer experience awareness has emerged as one of the most important business enablers for service providers, by helping them understand the opinions, needs and motivations of users. New enabling technologies for data analytics in the customer experience awareness field can provide richly detailed and actionable insights for business optimization.

JÖRG NIEMÖLLER, NINA WASHINGTON, GEORGE SARMONIKAS

THE INTERNET OF THINGS (IoT) offers users many great opportunities and simplifies many facets of life, but for some, its all-encompassing nature can be overwhelming and alienating. Service providers that recognize this risk have the opportunity to differentiate and create environments where each individual user feels comfortable and can rely on their intuition. Doing so successfully, however, requires a high degree of customer experience awareness.

In the not so distant past, objective QoS was the central concern for service providers. The idea was that a high degree of QoS enabled by an excellent infrastructure would lead to a positive customer experience with a high degree of satisfaction and loyalty. This was reflected in business metrics such as churn rates and users' propensity to call customer care.

While that logic still holds true to a certain extent, there are many factors that it fails to take into account on an individual user level. The fact is, individual users are never truly objective, and a subjective individual user might not always feel satisfied – even when experiencing good service. Understanding why this is the case is essential to developing customer experience awareness and gaining the insights required to make the right decisions to actively manage the user's perception.

There are several ways to go about developing a higher level of customer experience awareness, including training, goal setting and optimization of the organizational structure. In many cases, a customer experience management (CEM) system will also play a key role.

A CEM system is a business assurance tool that monitors and actively controls the impact that the user's perception of a product, brand or service has on the business result of a service provider. The central figure in CEM thinking is the individual user: a human being with personal opinions based on subjective perceptions, who is embedded within a particular social environment. The user's perception is based on their individual expectations. The moods, feelings and specific context of every user play a major role in developing their opinions and attitudes, and ultimately determining their actions and behaviors. Just as users can never be objective in forming their perception of their service providers, they are frequently inconsistent in their actions.

A good CEM system requires a holistic understanding that goes all the way down to an individual level, with a broad range of contextual information about the user taken into consideration. The resulting insights become actionable if they are combined with business processes at the individual user's level that make it possible to personalize their experience.

The path to actionable insights

A CEM system provides insights that can be used as the basis for decision making and action taking that will help optimize business results. These insights are typically expressed in scores and indicators that quantify a particular aspect of users and their experience. For example, the Net Promoter Score (NPS) quantifies in a single number the user's general willingness to promote the service provider, which is an indirect expression of their level of satisfaction and loyalty. The NPS measures user perception of the overall performance of the service provider, and has become a very useful tool for raising awareness of customer experience within an organization.

> "THE MOODS, FEELINGS AND SPECIFIC CONTEXT OF EVERY USER PLAY A MAJOR ROLE IN DEVELOPING THEIR OPINIONS AND ATTITUDES, AND ULTIMATELY DETERMINING THEIR ACTIONS AND BEHAVIORS"

CEM systems are specifically designed for particular business optimization use cases, generating a variety of use-case-specific scores and indicators as their primary output. In network operations, for example, any substandard user experience is detected automatically with low latency from performance metrics and brought to the attention of support technicians for further analysis and rapid response. This contributes to overall business optimization, as problems are solved quickly, hence limiting their effects as well as the number of users who are exposed to them.

Every business process or optimization use case that would benefit from customer experience

Terms and abbreviations

BSS – business support systems | **CEM** – customer experience management | **ELI** –experience level index | **IOT** – Internet of Things | **MOS** – Mean opinion score | **NPS** – Net Promoter Score | **OSS** – operations support systems | **S-KPI** – service KPI

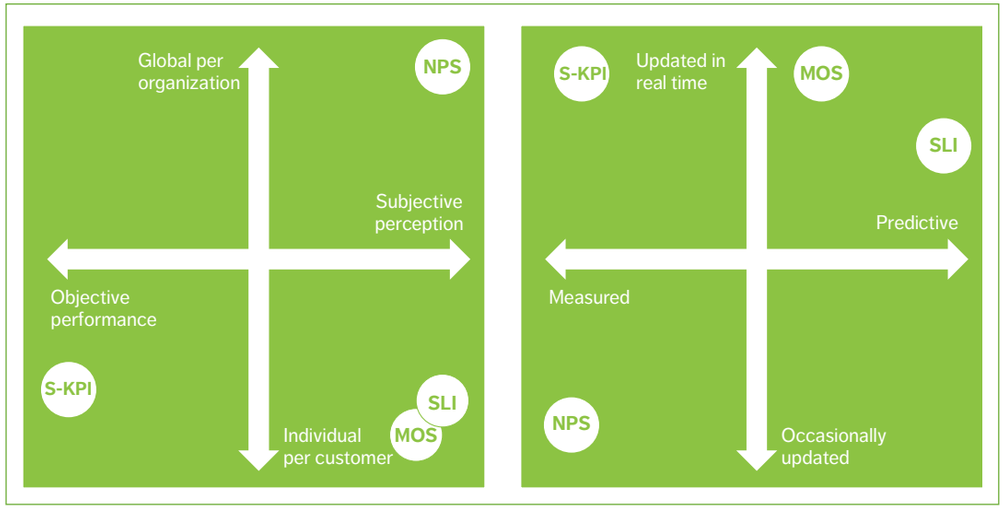


Figure 1 Properties of scores

awareness will have its own particular requirements with respect to scores. Aside from the main subject that a particular score expresses, the following characteristics are relevant:

- **Scope:** Does the score reflect an insight at the individual user level, for a group of users, or for an entire organization?
- **Outreach:** How many users are included?
- **Subjectivity:** Does the insight reflect an objective fact or a subjective perception?
- **Predictive:** Is the insight directly measured or the result of a predictive model?
- **Latency:** How quickly does the score need to reflect an experience?
- **Frequency:** How often is an update of the score needed?

These use-case-specific requirements are directly reflected in the way a score is obtained and implemented. A service KPI (S-KPI) and similar low-latency, high-outreach measurements are needed for swift corrective action to be taken. Most of them are objective metrics measuring technical performance, and they make it possible to distinguish individual users. The characteristics of the KPI are reached with considerable effort in terms of the efficient handling of a real-time input data stream. The raw data comes from an extensive distributed probe network, and is correlated and processed in near real time.

Surveys and studies that approach the user directly and ask for feedback have a completely different technical profile from the S-KPI – that is, the characteristics of these scores are quite different. The NPS is a prominent example, as it is a direct measurement that typically has high latency, with significant time intervals between distinct measurements. Furthermore, the outreach is not very high, with only a few percent of the user base included in every measurement activity. In *Figure 1*, the characteristics of four different types of scores are compared – S-KPI, NPS, service level index (SLI) and mean opinion score (MOS).

Measuring subjective user perception

The ability to understand the user on an individual and personal level has become one of the most promising areas within CEM. Nevertheless, it is one of the most challenging tasks due to the complexity of the user as an individual. Understanding the user as an individual means gaining an accurate idea of their level of satisfaction as well as their expectation, behavior, loyalty and intention. These factors are determined and influenced by personal properties such as sentiment, perception, experience and need. The links between these properties are complex, and can vary significantly depending on personal context, such as the social environment and timing of experiences. Furthermore, experiences that are totally unrelated to the service provider can have a major impact, because they influence many factors from general priority setting to momentary moods.

It is possible to measure these factors by asking the user to complete a survey. Getting consistent and accurate answers requires smart ways of asking questions, however. For example, NPS surveys ask users if they would recommend their service provider.

This is a proxy question used to indirectly measure satisfaction and loyalty by triggering more accurate responses than asking users directly about their level of satisfaction.

Survey-based methods fail completely in use cases that require agile action to be taken, however, because low latency in terms of insight availability and outreach to the entire user base are prerequisites for taking personalization actions. Consequently, while perfectly suited to clarifying customer experience performance at an organizational level, the NPS is of no use in personalizing the treatment of users.

The individual's indication of their level of satisfaction provides a valuable insight for many use cases in marketing and service operations. It helps in the selection of recipients in marketing campaigns, for example, and can support product planning. The individual insight level allows for the customization of service offers tailored to single users. Low latency and frequent updates make it possible to adapt the personalized offer dynamically according to changes in the individual user's needs.

Ericsson has introduced the SLI as a satisfaction score that meets the technical properties of use cases that require agile action [1]. Reached using a predictive model, it is a personal score available for every user. It is updated frequently, and with low latency after a user activity. Experience-related S-KPIs that originate from the network probe infrastructure are used as input. Psychological factors relating to subjective perception form the basis for interpreting the objectively measured experience expressed by the KPIs, and indicate a subjective level of satisfaction. The following psychological effects have been identified:

> UNDERSTANDING THE USER AS AN INDIVIDUAL MEANS GAINING AN ACCURATE IDEA OF THEIR LEVEL OF SATISFACTION AS WELL AS THEIR EXPECTATION, BEHAVIOR, LOYALTY AND INTENTION

- Perception is individual, so different models are needed for different users.
- A negative experience has more impact than a positive one on the overall level of user satisfaction.
- Surprising experiences are more significant than less surprising ones.
- The user forgets experiences as time passes.
- The more significant the experience, the longer it will be relevant.
- The context of each experience event may affect the user's perception.

The SLI model is a psychology-based hypothesis that represents these factors mathematically. The model is trained using survey-based reference data. In this way, it is calibrated to how service and network experiences are perceived by a particular user base. The combination of psychological research with state-of-the-art machine learning algorithms is the central innovation that makes it possible to master the complexity of individual user perception. The insights provided by the SLI are designed for direct consumption in decision-making and action-taking

processes, contributing to increased personalization of the user experience.

In short, the SLI and the NPS both provide insights into subjective satisfaction, but they target different sets of use cases and therefore take different approaches to reach the technical use case requirements. The SLI introduces the ability to act on an individual user level and ultimately improve the NPS, which is an important indicator of business success.

Understanding the experience journey

In all of their interactions with a service provider, users undergo an experience journey: a sequence of experience events they are involved in over time. These events can be direct interactions such as visiting a point of sales, calling customer care or simply using a service. But even events that do not involve the service provider directly are considered part of the experience journey. For example, the user might share their opinion of a service with other users on social media platforms. The experience journey is a powerful conceptual tool for continuous monitoring and categorization of experiences. A prominent and widely used example is TM Forum's Experience Lifecycle Model [2], which defines 22 phases to categorize the individual experience events in the journey. This model, illustrated in *Figure 2*, was recently updated under Ericsson's leadership to cover digital service providers across domains and user experience in the Internet of Smart Everything.

There are always two distinct roles in the context of a model for experience journeys: the observed user, and the service provider that is using the model for structured recording of observations with rich details about the user.

It is notable that the user's experience while consuming the service is just one of 22 phases in the journey. It begins before the user becomes a customer, with the realization that there is a need for a service or that current services are no longer a good fit. The first few phases of the journey determine interest and loyalty, and a good understanding of these can be highly relevant, allowing a service provider to approach users at the right time with the right message. The Experience Lifecycle Model clearly illustrates the significant paradigm shift from network-operationcentric CEM to taking a holistic view of the user.

Each phase in the journey and every experience event is connected to a rich set of contextual details and metrics. In service usage, the S-KPIS are related to the usage event. For customer care, respective KPIS determine the details of an interaction between the user and the service provider. All of these experience-event KPIS provide rich input to analytics. For example, the SLI model utilizes service-usage events to determine user satisfaction.

The Experience Lifecycle Model reveals, however, that the SLI is missing experiences from other phases of the journey, which also contribute to user satisfaction. The psychological scoring model of the SLI is designed to utilize further KPIS as long as they correlate with user satisfaction. This means the scoring can easily be extended to further parts of the journey for higher prediction accuracy. Ericsson calls the resulting score the experience level index (ELI) [1].

Subjective perception scoring

- The NPS is a business performance benchmark for the service provider that measures customer experience. It is based on surveys in which a number of users are asked if they would recommend the service provider to friends, colleagues and family.
- The SLI predicts a user's current level of satisfaction by interpreting observations about their service usage and the delivered service quality. An analytics model evaluates the observations and delivers a score. The SLI is frequently updated and available for every user.
- The ELI expands the SLI model to include additional phases of the user journey.

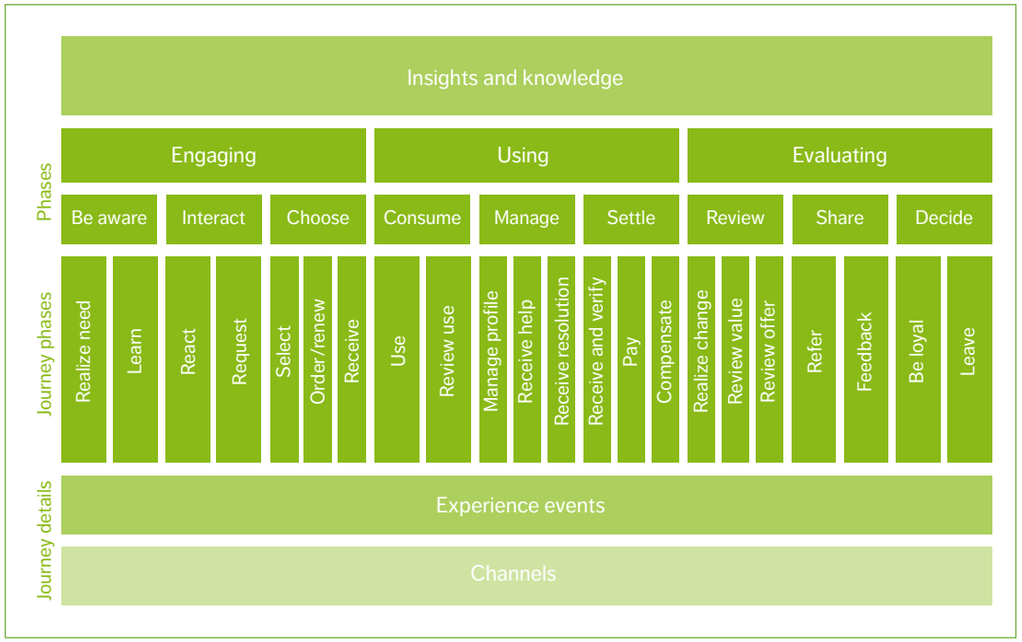


Figure 2
Experience Lifecycle Model

With respect to the user journey, interaction between the user and the service provider can occur in various phases and through many different channels. The diversity of these touchpoints means a user can start an interaction on one channel and then continue later on a different one. For example, the user might first call customer care to get information about a service offer, and later resume this dialogue at a point of sales. The overall experience is not only determined by each individual interaction, but by the entire journey. In this respect, continuity and consistency across the individual touchpoints are highly important for ensuring a good user experience.

The concept of a seamless and consistent journey experience is referred to as "omni-channel" by TM Forum [3]. The technical solution for reaching it consists of a consolidated operations support systems/business support systems (OSS/BSS) backend with consistent data and detailed handling of each user's communication history. This is another example in which delivering a good user experience requires a high degree of personalization.

From IoT to Internet of Smart Everything

The combination of the IoT and the exponentially increasing smartness of all manner of things is driving us into a new phase of digitalization known as the Internet of Smart Everything. In this emerging reality, the user is part of an infrastructure along with a large number of physical things, services and social media platforms. Things can be understood as autonomous entities with built-in intelligence that represent physical entities in the digital context. At the same time, new smart services are launched every day, and social media platforms and other interaction channels continue to grow. Familiar offline services – public transportation and utilities such as electricity and water, for example – are also increasingly presented and managed

through digital channels. Terms like "smart city", "smart cars" and "smart meters" refer to this ongoing transformation.

The Internet of Smart Everything constitutes a vital new dimension in CEM. The number of digital interactions with the user is set to increase dramatically, resulting in a new level of complexity in the user journey.

In the IoT environment, users experience a presentation layer through which more or less smart things and related services interact with them. The things present information, request decisions and learn from the interaction. A good user experience is an effortless and intuitive one without unnecessary interactions. Users should always be aware of what they can do, how they can reach their goals efficiently, and what consequences an action will have. All of this should be enabled with the right level of relevant information available at the right time.

It is important to recognize that users are different in their abilities to cope with and accept this new environment. The digital natives of the 21st century will interact more intuitively than many others. Personalization makes it possible to customize the entire experience for each individual.

CEM suppliers follow IoT developments closely, and will launch new capabilities to manage new types of interaction experiences. New and extended scores will help to capture experiences and facilitate personalization. The analytics backend will incorporate new KPIs that are a more accurate reflection of user experience. And new and improved algorithms will process all available data into ever better recommendations for action taking.

While the experience of users when interacting and using smart things is obviously paramount, it is also true that smart things themselves act like users when interacting with each other or with human users. The smart thing decides autonomously to use services or to communicate. It has requirements that need to be satisfied by the services and infrastructure that it utilizes. As a result, the same tools used to determine and actively manage the experience of a person can be applied to the experience of a smart thing.

The analytics infrastructure

In order to provide a high level of customer experience awareness in a vast and complex digital world, a CEM system needs to support four types of analytics in a flexible analytics backend:

- **Descriptive analytics** determines information about the current situation and presents it in a way that makes it possible to capture the essential insight easily and decide on the appropriate action.
- **Diagnostic analytics** goes one step further, and finds causalities in the data.
- **Predictive analytics** learns from current and historical references to detect trends and anticipate situations before they occur, enabling early countermeasures to be taken to mitigate problems before they become significant.
- **Prescriptive analytics** directly proposes the best actions to be taken.

All of these approaches depend on access to raw data coming from various sources. The CEM system must be able to support many different systems with a great variety of interfaces and protocols for data access, together with respective information models. Typical sources of data are customer relationship management systems, billing and other BSS, network probes, IoT service enablement and any type of existing data warehouse. The results of user surveys can also be included in the analysis.

In early processing, the raw data is filtered, aggregated and correlated in order to maximize relevance. Basic statistical methods are deployed and S-KPIs are calculated in this step, creating the basis for more elaborate analytics. This phase includes a redundancy and irrelevance reduction, which is particularly important for input data streams, where the sheer volume of incoming data that also needs processing with low latency is a challenge in itself.

This infrastructure is ideal for use cases where the system is expected to deliver the basis for agile actions and countermeasures. A typical example is network operation, which needs to react rapidly on detected

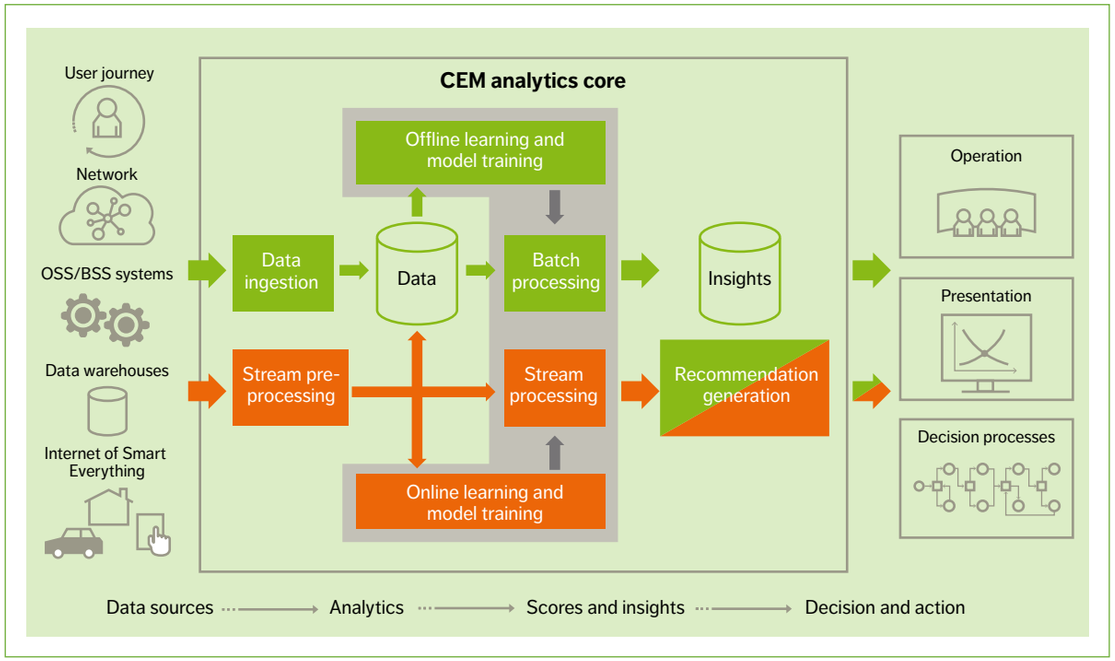


Figure 3 Logistics in a warehouse

service quality degradation in order to contain the effects of the issue. For the analytics backend, this implies the need for a scalable stream processing infrastructure that is able to process high volumes of data. Scalable rules engines can be applied to filter out interesting insights from the stream. In-memory data handling is another essential component in this context.

In other use cases, the insights will be based on a similarly huge amount of available data, but low latency is not required. Batch processing components based on MapReduce techniques, for example, constitute an essential enabler for this type of use case. Access to comprehensive historical data is often important for generating these insights.

Prescriptive analytics will support the service provider's technical and business experts in making their operational decisions. The domain experts will be able to automatize their decisions and action processes for more agile reaction and implementation of change. These expert systems are based on techniques for knowledge management and artificial intelligence. Rules engines are able to generate straightforward recommendations based on previous analytics insights. Ontologies and semantic models give meaning and context to data and analytics results. They make it possible, for example, to inform the recommendation system about business-level goals and strategies. Pairing this with machine reasoning

techniques leads to recommendations that are dynamically aligned with business-level concerns.

The four types of analytics act as enabling toolsets that allow a data scientist to set up analytics algorithms with the right technical properties for a given use case. Ericsson's state-of-the-art architecture is highly scalable for this purpose in terms of processing and data handling capabilities. Its particular strength is in low-latency processing of network-based streams of metrics, the integration of diverse data sources and information models, and the incorporation of results into subsequent decision-making and action-taking processes throughout diverse OSS/BSS tools.

The analytics infrastructure of our CEM solutions has become a common base for Ericsson's entire digital business systems portfolio. This is an important step to reach a consolidated offering where a consistent experience for users across several touchpoints along their journeys is relatively easy to achieve.

The end-to-end analytics path is outlined in *Figure 3*, which distinguishes between the stream processing track – where low latency and real-time processing are enabled – and the batch processing track – where analytics on more static data is performed. Raw data enters the system on the left at preprocessing, which interfaces with a great variety of sources. The incoming data is immediately used in stream processing for low latency insights and stored in a distributed in-memory database to make it easily available for further analytics processing.

Training analytics models via machine learning, and using them in actual scoring and insight generation, are two distinct functions in the analytics workflow. Learning and model adaptation cycles can, however, be highly dynamic processes with continuous adaptation of the way that insights are generated.

The insights generated in analytics are kept available in a database for subsequent internal or external usage. Recommendation systems usually operate on the analytics insights. Decision making and action taking are typically distributed throughout the various business and operational level management and planning systems. Examples of business level systems and processes that profit from analytics insights are campaign and revenue management, as well as investment planning.

An essential component of future CEM solutions will be recommendation systems that are use-case aware and able to directly propose an action that would optimize a user experience. These systems would, for example, recommend when to start a marketing campaign for upsell and propose which users to include, or recommend investments into infrastructure upgrades to optimize the user experience.

These examples show the ever closer integration of CEM into business level processes and activities. Future recommendation systems will be aware of the service provider's business goals and preferred strategies to support staff in decision making. The resulting change in scores will then be used to track progress and verify the success of the actions taken.

The customer-centric organization

Many service providers find they need to adapt their business processes to make use of the new insights they receive from CEM. These process adaptations can be expected to produce significant improvements in business results. They are, however, only the first step in the process of becoming a truly customer-centric organization.

A customer-centric organization is characterized by four externally visible key abilities:

1 **Being attentive.**
No requests are left unanswered. The goal is to be perceived as flexible and available. The user feels every effort is being made to accommodate their personal needs with a customized solution.

2 **Being proactive.**
The service provider anticipates customer concerns and addresses them early before they become an issue. In this respect, it is important to choose the right time and channel for interaction.

3 **Being consistent.**
For a good experience, the dialogue with customers across channels needs to be seamless. The user expects the service provider to have a record of all previous contact so that they don't have to repeat themselves or receive conflicting information from different touchpoints. Information needs to be consistent, flawless and immediate.

4 **Being adaptive.**
Services and products are adapted continuously to meet customer needs. The customer feels their needs are fulfilled and experiences any changes as improvements.

These abilities crosscut a service provider's entire organization, demanding a high degree of engagement from every employee. This can be achieved through goal setting and training, but it also requires the right organizational structure and internal interfaces.

CEM helps identify both shortcomings within an organization and any necessary corrective actions. Furthermore, CEM insights enable each unit within the service provider's organization to master their tasks with continuous awareness of the customer experience impact. Customer-experience-related benchmarks are used as major success criteria for a unit. They can also facilitate investment decisions and indicate the return on investment of actions and changes.

Conclusion

Customer experience management is the practice of continuously managing and improving an organization's customer touchpoints and interactions. In an increasingly vast and complex digital world, there is a clear need for service providers to understand the customer experience – not only from the perspective of network and service performance, but also, to recognize each user as a subjective individual. Customer experience awareness has the potential to act as a key differentiator in the ICT industry, helping service providers transform into the kind of customercentric organizations that can offer a high level of personalization. By combining technology, strategies and resources, any service provider has the chance to use customer experience awareness thinking – and CEM systems in particular – to significantly improve customer satisfaction and loyalty. ❁

References:
1. Jörg Niemöller and Nina Washington, Subjective Perception Scoring, at the 19th International Conference on Innovations in Clouds, Internet and Networks (ICIN 2016), March 2016, Paris, France
2. TM Forum Best Practices, GB-995, 360 Degree view of the Customer R16.0.0, Frameworx Release 16
3. TM Forum Best Practices, GB-994, Omni-Channel Guidebook R16.0.0, Frameworx Release 16

THE AUTHORS

Jörg Niemöller

◆ is a systems manager for customer experience management and data analytics within Digital Business Systems. He has established subjective perception scoring within Ericsson's products for CEM and analytics. Niemöllerholds a Ph.D. in computer science from Tilburg University, the Netherlands,and a diploma in electrical engineering from the Technical University of Dortmund, Germany. He has worked for Ericsson since 1998 in various system management positions and at Ericsson Research.

Nina Washington

◆ joined Ericsson Research in 2012. She is an experienced services researcher focusing on psychological perspectives, studying user behavior, attitudes and values. In her current research she explores goal setting, communication and cross-cultural differences, all in interplay with technology as an enabler. She holds an M.Sc. in psychology from Lund University, Sweden, as well as an M.Sc. in positive psychology from the University of East London, UK.

George Sarmonikas

◆ joined Ericsson in 2013 after several years of working for mobile operators. He currently leads the IoT Analytics and CEM Innovation group within Digital Business Systems. Prior to this he was responsible for product management of Ericsson's CEM and analytics portfolio, including assets, for subjective experience scoring. Sarmonikas holds both an M.Sc. in communication systems and an M.Eng. in electronic engineering and computer science from Bristol University, UK, as well as a diploma in innovation with artificial intelligence from Stanford University, US.

CROSS-DOMAIN identity OF THINGS

The rapid expansion of the Internet of Things (IoT) calls for a clearer common understanding of how identities function in the digital world. The numerous domains that make up the IoT result in single entities having multiple overlapping identities. In order to operate successfully in this environment, a company must be able to manage related identities across domains in an efficient manner. To do so, it needs to determine which cross-domain identity management solution best meets its own specific requirements.

THOMAS WEIDENFELLER, CLAUDIA BAUSCH

Identity is a concept used in fields ranging from philosophy to mathematics, with a variety of definitions. Even within the fields of ICT and IoT, the interpretation of the terms identity and identity management can vary widely, depending on the specific application and particular school of thought.

For many, identity management involves nothing more than giving a thing a traceable name or number, and perhaps adding a password or a public key certificate. For others, it means applying a consistent naming scheme or using a particular protocol to provide a computer with a host name, or a system user with a convenient sign-in experience.

According to ISO/IEC 24760-1:2011, an identity is "a set of attributes related to an entity", and an entity is defined as "an item ... that has a recognizably distinct existence" [1]. These definitions are very broad, and clearly cover more than just devices and people. For example, not only is an IoT device an entity according to this definition; all of its physical and virtual components are also entities, as are all of the actors that interact with them. The definition also covers parts and groups of such items, as long as they have a recognizably distinct existence.

From this perspective, even a small IoT device consists of many entities. Not every entity needs to have one or more identities, however; nor do all established identities need to be managed throughout the complete lifetime of the device. Defining the set of identities that need to be established, and working out how to manage them, are the result of decisions made on multiple levels at different times. For example, some identities are the result of design decisions about communication technologies or hardware component selection for a particular device.

It is also important to note that an identity does not necessarily have to be unique. For example, an identity can refer to a group of devices, such as in multicasting. An entity can also have – and typically has – more than one identity.

Entities, identities and domains

The application and validity of an identity tend to be finite, and are often dictated by technical limitations. For example, a private IP address has no global meaning; it only has meaning in a private network, and cannot be used on the internet. It is also possible to limit the applicability of an identity even further by design. The resulting domain of applicability describes where an identity may be used.

A car is an example of an entity that has multiple identities that are valid in different, partly overlapping, domains. A car receives its vehicle identification number (VIN) during the manufacturing process. The VIN is used by government agencies to track the car throughout its lifetime. The VIN's domain of applicability is typically limited to administrative purposes. However, at some point the vehicle will also receive a license plate number, which is used to identify it in public. Its domain of applicability is the public realm. Both the VIN and the license plate number identify the same entity: a particular vehicle. Both should be registered to the same owner. Depending on the type of operation to be performed, a particular one of the two identities or identifiers will be used. In some cases, both might be required. However, rarely can one identity be provided in place of the other.

> "THE APPLICATION AND VALIDITY OF AN IDENTITY TEND TO BE FINITE, AND ARE OFTEN DICTATED BY TECHNICAL LIMITATIONS"

Although they are separate, identities in different domains are related. In the car example, the relevant identity management systems (IDMSs) are designed to make it possible for government authorities to find out the license plate number from the VIN, and the VIN from the license plate number. When a license plate number is issued, identity management activities affect both domains to ensure traceability.

Understanding identity management

The term identity management is defined in ISO/IEC 24760–1:2011 as "the processes and policies involved in managing the lifecycle and values, type and optional metadata of attributes in identities known in a particular domain" [1].

The car example clearly illustrates that identity

Terms and abbreviations

eUICC – embedded Universal Integrated Circuit Card | **GBA** – Generic Bootstrapping Architecture | **IDMS** – identity management system | **IIP** – identity information provider | **IoT** – Internet of Things | **LWM2M** – Lightweight M2M | **M2M** – machine-to-machine | **OAuth** – open standard for authorization | **OpenID** – open standard and decentralized authentication protocol | **SAML** – Security Assertion Markup Language | **SSO** – single sign-on | **UICC** – Universal Integrated Circuit Card | **VIN** – vehicle identification number

management is not about managing the entity itself (that is, performing operations on the entity). Rather, it is about managing "a set of attributes related to an entity" – data that describes or identifies the entity. Identity management is fundamentally a security technique – not an entity management one. As such, identity management supports the identity-based decisions [1] that must be made to ensure security.

Typical identity-based decisions that are related to security include device authentication, controlling authorizations (typical authentication, authorization and accounting functions) and the categorization of data. For example, identity-based decisions can be used to ensure that the data returned by an IoT sensor (such as a temperature measurement) is associated with the correct entity (the machine from which the temperature was taken). In general, the routing of input and output data to and from an IoT device is based on identities.

The distinction between managing an entity and managing an entity's identity is important. Managing identities can have side effects that impact the entity, but won't necessarily. For instance, an attempt to manage an entity via identity management will at best be indirect, and at worst a complete failure.

For example, in geolocation applications, an entity's location might be one of its identities. The entity might even be addressed (identified) by its location. Performing a particular identity management activity could affect the location data attribute in the identity register. But this change would have no effect on the entity's actual position. In the best-case scenario, there would be additional mechanisms in place to take the identity management data and translate it into action that would in turn affect the entity itself, such as commanding it to move to the new location. This could work if the entity was a mobile machine, but would obviously fail if it were a factory building (the worst-case scenario).

The limitations of identity management are particularly significant for IoT devices. Identity management is no substitute for proper device management; rather, the two need to work in parallel. Device lifecycle changes must be supported by identity management activities.

The identity management lifecycle

Figure 1 provides an example of the lifecycle of an identity in terms of states and state transitions. This example is a modified version of the reference lifecycle model in ISO/IEC 24760–1:2011 [1]. Other lifecycle models may also be used, depending on the specific purpose of the particular identity. An IDMS supports the creation, provisioning, maintenance and decommissioning of identities throughout the lifecycle of a particular type of identity [1] following its lifecycle model.

The lifecycle example in Figure 1 manages an identity within a specific domain. However, we know an entity can have more than one related identity within the same domain, or multiple identities spread over several domains. As a result, the requirements for a real-world IDMS extend beyond merely transitioning through the states for an identity.

The scenario involving multiple identities that is easiest to manage is when the related identities are within the same domain and under the control of the same authority. At the other end of the spectrum are scenarios in which the identities are in different domains, and are controlled by different authorities, and the relevant IDMSs are not able to communicate with each other.

Figure 2 illustrates the relationship between IDMS coupling and domains, and its impact on the relative difficulty of managing identity data. In cases where there is no communication between IDMSs, manual intervention and handling are mandatory. Such cases are therefore best avoided.

Cross-domain management architectures

Two common cross-domain identity management architectures are particularly relevant to IoT identity management. The first, shown in *Figure 3*, uses one IDMS for coordination, giving it special authority among its peers.

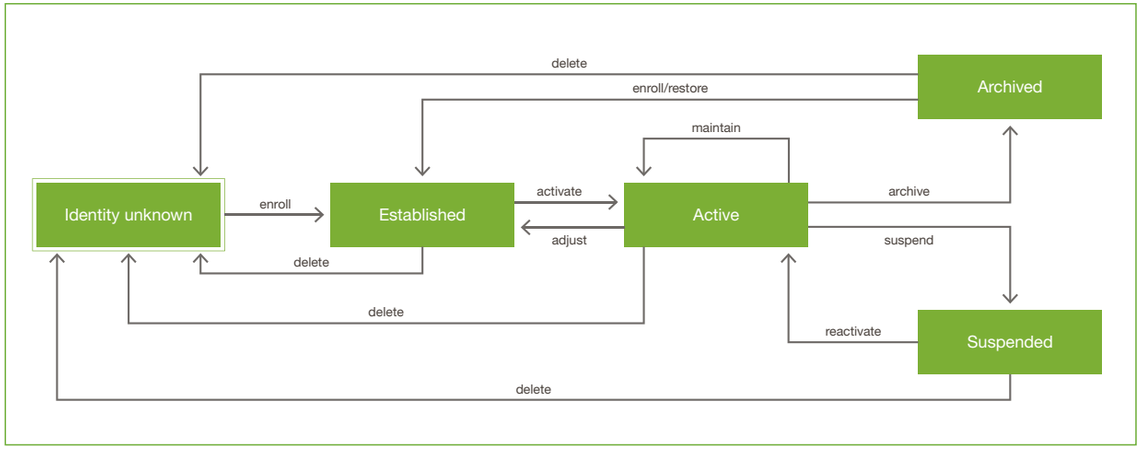


Figure 1 Example of an identity lifecycle

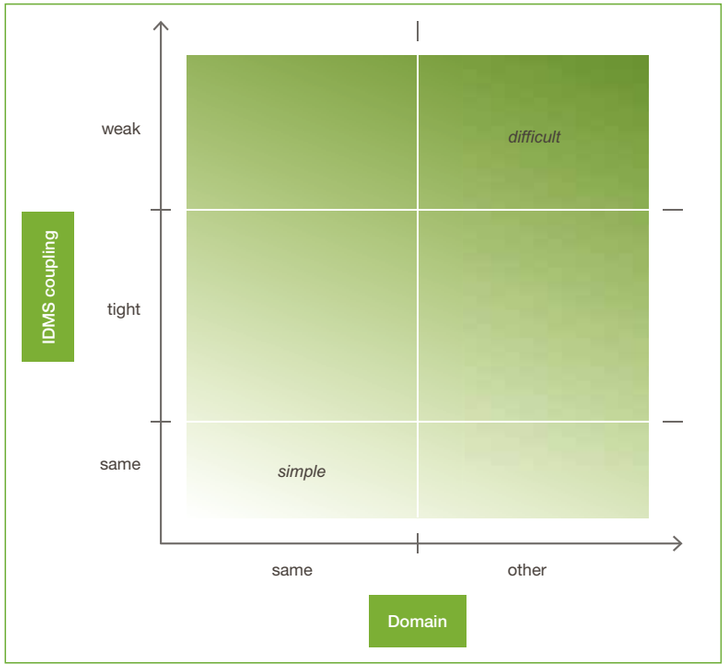


Figure 2 Relative difficulty of managing related identity data

The architecture shown in Figure 3 is similar to an architecture used in network management, in which individual element managers are each responsible for managing a particular network element, and a network management system coordinates network-wide issues above the element management layer. The architecture can be enhanced by adding hierarchy levels with intermediate coordinating IDMSs.

Figure 4 shows the second common architecture, in which the various IDMSs coordinate with other IDMSs on a peer-to-peer basis. Note that not every IDMS coordinates with every other IDMS; this depends on whether there is any need for them to coordinate, as well as technical or administrative limitations.

There are no hard and fast rules dictating which architecture is preferable. Other architectures also exist, including hybrid versions of the architectures presented in Figures 3 and 4. Practitioners need to consider their existing systems and any administrative barriers they may have, and make compromises, adapting their integrations to suit their particular circumstances. Ideally, they should establish one of the architecture options as the primary one and add diverging IDMS and management subsystems as satellite systems in isolated areas.

Techniques to build a coordinating system

There are technical and administrative issues to overcome when building a coordinating system. The technical issues begin with the communication layer. The individual IDMSs that should take part in cross-domain identity management as shown in Figures 3 and 4 need to communicate in some way – typically via the TCP/IP suite. When faced with legacy protocols on the network layer [2], an adaptation to IP should be considered. Which protocols to use on layers above the transport layer (particularly the application layer) is both a technical and an administrative decision.

Administration of cross-domain identity management includes the creation of an identity federation: "[an] agreement between two or more domains specifying how identity information will be exchanged and managed for cross-domain identification purposes." [1] The system that is subsequently built according to this agreement is typically also known as an identity federation.

Single sign-on identity federation

One highly sought-after feature when building identity federations – especially when humans are involved – is single sign-on (SSO). With SSO, the identity of an entity in one domain can be used for authentication of the same entity in another domain. The purpose of SSO is to avoid having to perform identity management in two or more domains in parallel. This is achieved by having fully automated protocols and processes in the identity federation agreement for handling the data processing and exchange between the domains.

Enterprise and cloud system architectures are good examples of how cryptography-based identity federations can be used to provide SSO services. SAML, OpenID and OAuth 2.0 (with or without additional application programming interfaces like OpenID Connect) are typical protocols used to build SSO identity federations for authentication or authorization purposes in this context. Essentially, these protocols are used to exchange trust in an identity – and by association, an entity or groups of entities – between domains.

For humans, SSO is a highly valued convenience feature that removes tasks like remembering user login credentials. But for non-human IoT entities, which connect to a rather limited number of services, the use of identities and identity-based decisions in IoT device communication does not necessarily require SSO.

A typical IoT device might, for example, make use of the following services:

- a network service that provides basic communication
- a device management service provided via the Lightweight M2M (LWM2M) management protocol [3]
- a service management service provided via LWM2M, either separate from or in cooperation with the device management service
- a payload or application service to which the IoT device delivers data and from which it receives application information.

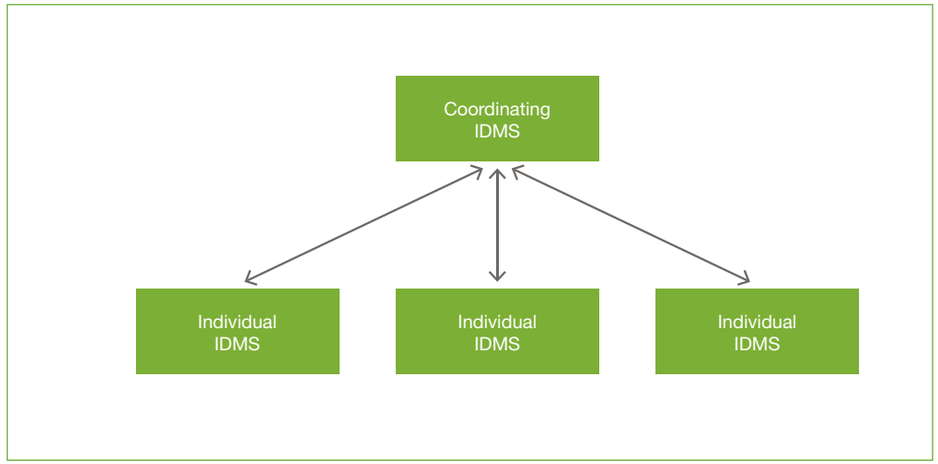


Figure 3 Centrally coordinated IDMSs

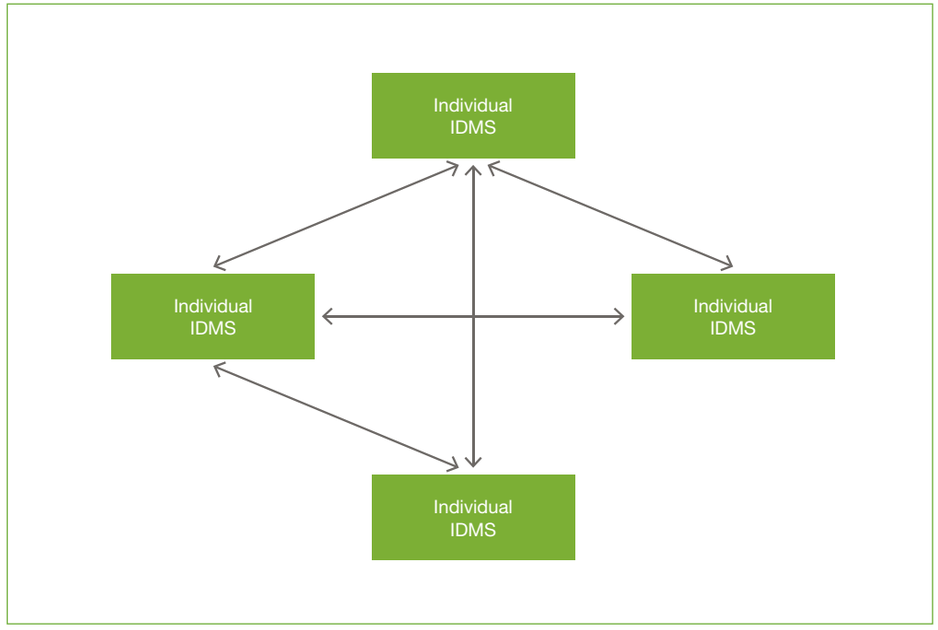


Figure 4 Peer-to-peer coordination

> THE 3GPP IDENTITY AND GBA ARE CURRENTLY ASSOCIATED WITH CELLULAR NETWORKS, [BUT] THIS TECHNOLOGY CAN ALSO BE USED FOR DEVICES CONNECTED TO A NETWORK USING OTHER, NON-3GPP TECHNOLOGIES

Since the number of services used is relatively static over the lifetime of the IoT device, and there is no human convenience advantage, an SSO-capable identity federation is not absolutely necessary in this type of case. In fact, for small IoT devices, the use of enterprise SSO protocols adds considerable overhead to the device firmware. When SSO is needed on an IoT device, lightweight SSO protocols should be considered instead.

The Generic Bootstrapping Architecture (GBA) [4] is a mobile network technology that makes it possible to reuse an identity from within the mobile network domain in other domains. Solutions based on the GBA architecture make use of mobile network subscribers' identities, associated cryptographic key material and cryptographic algorithms to establish a temporary, cryptographically-secured security association between an IoT device and a service in the application layer, for example. The security association can then be used for tasks such as authenticating the IoT device before granting access to the service. One promising realization of an identity management solution using GBA as a federation technique is a trial project for agricultural applications known as the Connected Vineyards project [5].

GBA uses well-known mobile network identity information providers (IIPs). A UICC/eUICC with a SIM application suitable for GBA is used in the IoT device, while the corresponding identity information on the mobile network side is provided by the Home Location Register/Home Subscriber Server.

Note that, although the 3GPP identity and GBA are currently associated with cellular networks, this technology can also be used for devices connected to a network using other, non-3GPP technologies. The identity credential (shared secret) and associated software may in this case be protected by hardware-specific isolation and protection mechanisms to avoid the extra cost of (e)UICC in IoT devices. GBA can also be used to extend the federation beyond SSO – for example, to provide cryptographically derived, temporary pre-shared keys to secure communication.

It is important to recognize that setting up an identity federation, for SSO purposes or otherwise, requires effort. The need to manage identities in multiple domains is replaced with the need to manage the federation. More importantly, an identity federation requires trust. An enrollment in one domain affects all federated domains, which means that improper identity proofing in one domain creates a potential security risk in all federated domains. However, in some cases – such as GBA – a mobile network operator with an established track record of managing signup and access to network services is in a good position to provide the necessary trust.

Mapping

The SSO identity federation protocols presented above all rely on sound cryptographic principles. The original identity data, including passwords and cryptographic material, are not copied between the domains. Only the trust in some identity – an identity assertion [1] – is exchanged, enabling a federated domain to authenticate an entity, and, if desired, bootstrap its own cryptographic material. This is not the only way to build an identity federation, however.

Another common way to build an identity federation is by mapping. Identity data valid in one domain is mapped to some other identity data in another domain. The mapping can be 1:1 (the data is copied as is) or with some adaptations. For example, the mapping could include adding supplementary identity data, or adding an identifier as an attribute to one's own identity data.

One way to perform mapping is to synchronize at regular intervals. At certain points in time, the contents of two or more IIPs are compared with each other. Algorithms are then used to resolve any detected discrepancies and generate a consistent state across domains.

Tracking changes is another way to perform mapping. When this method is used, each of the state transitions shown in Figure 1 is communicated to the federated IDMSs. The IDMSs then map the received event data and add the result to their identity registers. A message bus is one possible software architecture that can be used for communication and exchange of events between the IDMSs. Regardless of which of these two mapping methods is chosen, it is vital to address the issue of concurrent changes to the mapped data in the federated domains. This can be dealt with by considering one domain to be the master for particular identity data. That is, one domain always has precedence, or may even be the only domain in which the data is allowed to actively be changed. If this solution is not possible in a particular case, operational transformation techniques can be used to handle issues of concurrent changes, especially in the case of tracking changes. Three-way merge or differential synchronization are other techniques for resolving issues when tracking changes or synchronizing.

Identity management domains in the IoT

The selection criteria for identity domains in an IoT IDMS are largely technical, but they are also influenced by organizational factors and sometimes even individual preferences. Domains

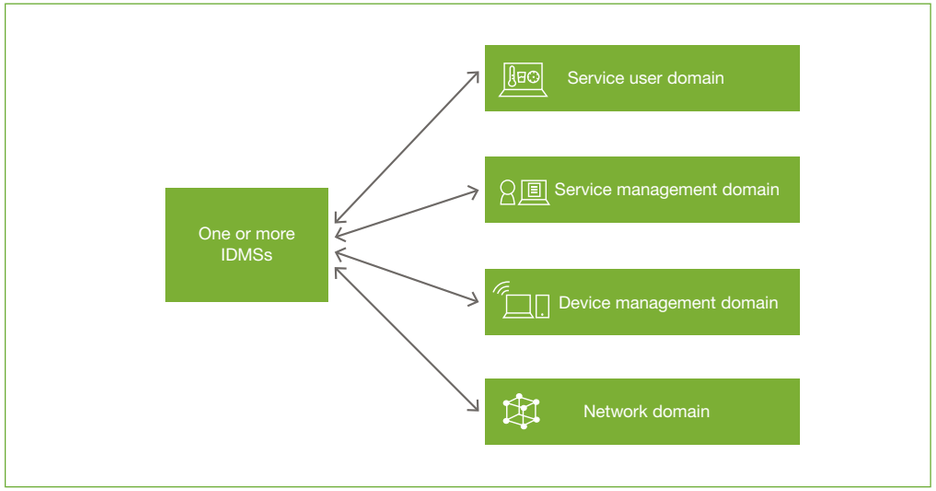


Figure 5 Four identity management domains in the IoT

can be quite small or rather broad, containing only a few or many different types of identity data.

Figure 5 illustrates four identity management domains that capture the technical and organizational properties of an IoT system at a high level:

- **service user domain** – where the IoT system is exploited for benefits. Services on top of the IoT device(s) are provided here. They supply a machine or a human with accumulated data and value-added services.
- **service management domain** – where the application(s) and/or service(s) running on the IoT device are managed, along with their association with enterprise application servers responsible for dealing with the payload data. A service delivery platform would work in this domain, for example.
- **device management domain** – where basic device functions are managed, including the device lifecycle and firmware (operating system). Services based on the LWM2M protocol would run here, for example.
- **network domain** – the "I" in IoT, where the communication happens, such as a cellular network or another type of WAN, or a LAN.

Identity management and security

There is another point that must be considered when coupling IDMSs to manage identities across domains. Identity management itself needs to be performed securely to fulfill the promise of helping to secure systems. It can only do so when identity management is performed in such a way that the managed identities are not compromised. For example, during enrollment, the right entity must be paired with the right identity. This is the most important aspect of this activity.

The basic security requirements for identity management are nearly identical to the security requirements of modern ICT systems. Both data at rest (storage) and data in motion (communication) need to be protected; and in each case, common ICT security techniques and technologies are relevant. This applies particularly to the exchange of identity information in identity federations, in those cases where identity data (such as access credentials) are simply copied or mapped from one domain to another.

There can be additional security requirements for identity management, depending on the particular domain or system, and on the system providers' level of commitment to offering a secure system. In general, the security of the management process and the security of the IDMS will have a direct impact on the trustworthiness of the managed identities.

Conclusion

With the spread of IoT systems to almost all areas of life, IoT security is set to become one of the most important technology development areas in the coming years [6]. IoT systems will need to be able to support large-scale field applications comprising a diversity of connected things. This will require massive enrollments of identities at an early stage of the device lifecycle, as well as the maintenance of those identities throughout the devices' lifetimes. The use of technologies like GBA and specific identity management systems for the IoT will substantially reduce the complexity of these activities.

It is clear that identity management systems – based on sound identity principles and intra-domain identity lifecycle models – have an important role to play in ensuring IoT security. Due to the heterogeneous setup of IoT end-to-end solutions, an IDMS that can only support one domain is not adequate for the complete identity management of IoT devices. Devices that must be identified in multiple domains need to have their identities managed across them. There are several ways to achieve this, depending on the systems and technologies available, and the relationship between the domains and the domain-specific identity data.

References

1. **International Organization for Standardization, ISO/IEC 24760-1:2011, Information technology – Security techniques – A framework for identity management – Part 1: Terminology and concepts, available at:** *http://standards.iso.org/ittf/PubliclyAvailableStandards/c057914_ISO_IEC_24760-1_2011.zip*
2. **International Organization for Standardization, ISO/IEC 7498-1:1994, Information technology – Open Systems Interconnection – Basic Reference Model: The Basic Model, available at:** *http://standards.iso.org/ittf/PubliclyAvailableStandards/s020269_ISO_IEC_7498-1_1994(E).zip*
3. **Open Mobile Alliance, Lightweight Machine to Machine Technical Specification. OMA-TS-LightweightM2M-V1_0-20160407-D. Draft Version 1.0. 07 April 2016, available at:** *http://member.openmobilealliance.org/ftp/Public_documents/DM/LightweightM2M/Permanent_documents/OMA-TS-LightweightM2M-V1_0-20160407-D.zip*
4. **3GPP, Generic Bootstrapping Architecture (GBA). 3GPP TS 33.220, available at:** *https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=2280*
5. **Ericsson, Connected Vineyards, available at:** *http://www.ericsson.com/res/docs/2015/iot-connected-vineyards.pdf*
6. **Gartner, Gartner Identifies the Top 10 Internet of Things Technologies for 2017 and 2018, available at:** *http://www.gartner.com/newsroom/id/3221818*

THE AUTHORS

Thomas Weidenfeller

◆ is a master systems designer at Portfolio & Systems within Customer Group Industry & Society. He has more than 20 years of experience at Ericsson, starting in telecommunication management systems. Over the years, he has worked in such diverse areas as software design, systems management, mobile packet backbone design and software architectures. He is currently working on IoT security issues. He holds a degree in electrical engineering from the Cologne University of Applied Sciences (now called the Technical University of Cologne), Germany.

Claudia Bausch

◆ joined Ericsson in 1998. She holds a degree in computer science from RWTH Aachen University, Germany. Her expertise covers several areas of software design, configuration management and project management. She is currently working as senior systems designer at Portfolio & Systems on IoT studies and end-to-end solutions within the Customer Group Industry & Society.

The authors would like to acknowledge the support and inspiration they received from their colleagues Per Ståhl, Patrik Teppo, Dhruvin Patel and Gustavo Tanoni.

Guest author

DR. STEFAN LARSSON

◆ Stefan Larsson is an associate professor at Lund University Internet Institute (LUii). He holds two Ph.D.s and an LL.M., and is an expert on digital socio-legal change, including issues of trust, consumption, traceability and privacy. He is a member of the scientific board of the Swedish Consumer Agency and recently published the book *Conceptions in the Code: How Metaphors Explain Legal Challenges in Digital Times* with Oxford University Press.

For more information about his work, visit:
http://luii.lu.se/about/stefan-larsson/

SUSTAINING legitimacy & trust IN A DATA-DRIVEN SOCIETY

FEATURE ARTICLE ✱

Human-centric data is at the core of the digital economy and most consumer-targeted innovation. What we sometimes forget, however, is that the quantification of everyday human life that produces this data depends not only on technological capabilities, but also on social norms and user values.

TRUST IS A VITAL determining factor influencing users' decisions to adopt innovations and sign up for new services – particularly those that they know will generate data for the service provider. While user trust is heavily based on their perception of the technological security of a solution or service, it is also fundamentally dependent on social norms and values such as privacy, legitimacy and perceived fairness in the collection and handling of individual information. The long-term success of the digital economy is dependent on consistently high levels of both technological and sociological trust among users. In light of this, it is of utmost importance that service providers consider the implications of social norms and user values in the service design process.

Human-centric big data

A large proportion of ICT innovation today is driven by the collection and analysis of human-centric data – a key component of the big data phenomenon. In some cases, human-centric data is collected by a company from the users of its current services. In other cases, a company may have purchased the data from another company to gain a better understanding of a new target group, for example. No matter how it is sourced, data is collected, analyzed and traded on a continuous basis, acting as a backbone for wide-ranging products and services: from health to consumer goods and services to urban planning.

The implications of this trend extend far beyond mere digitalization in terms of communication and infrastructure. Several scholars have argued that the growing strength of social networks is causing society to become not only "digitized" but increasingly "datafied" – with profound effects on how we read, write, consume, use credit, pay taxes and educate ourselves [1, 2].

This large-scale quantification of human activities has occurred within a very short period of time. Just a few years ago, it was much more difficult to gather human-centric data and use it for service development or commodification. But now, whenever we use the internet or carry a smartphone that is connected to it, we are tracked, logged, analyzed and predicted in a variety of ways: by way of web cookies, search engines, social media, e-mail and online purchases, as well as various types of sensors (including RFID tags and GPS-enabled devices such as cameras, smartphones and wearables). Offline purchase history is another useful resource, which can be administered through loyalty cards and club memberships, for example.

All of this information relating to our activities is not only used by the organizations that collect it; it is also exchanged by numerous commercial and governmental players for a whole variety of reasons. Beside this, there are companies known as data brokers that specialize in collecting and trading consumer data that is often at least partly collected from public sources. Such data collection and trading activities rarely involve a human observer who actually monitors the data points. They rather tend to be handled by an automated, quantitative and ubiquitous storage system built into the infrastructure – in the widest sense of the word – itself [cf. 3].

Some social scientists claim that this trend represents one of the most far-reaching social changes of the past 50 years [cf. 4]. As a result, these data-driven and technology-mediated practices are increasingly gaining the attention of scholars in various disciplines, particularly as they relate to privacy, but also in a variety of critical perspectives on transparency and algorithmic accountability [5, 6], big data ethics [7], behavioral and traditional discrimination [8, 9] or other consequences of a data-driven "platform society" [10].

Web cookies and the black box society

Web cookies are among the tools being used by companies such as Google, Facebook and traditional media houses to create extensive data retention infrastructures. The 2015 update of the Web Privacy Census revealed that a user who visits the world's 100 most popular websites receives more than 6,000 web cookies, which are stored on their computer [11]. Furthermore, it found that Google tracking infrastructure is on 92 of the top 100 most popular websites and on 923 of the top 1,000 websites, which contributes to making Google the world's most powerful information manager, with a central place in the modern information economy.

Similarly, a 2015 study by the Norwegian data protection authority Datatilsynet showed which parties were present when visiting the front page of six Norwegian newspapers [12]. The report noted that between 100 and 200 web cookies were placed on any computer being used to visit these home pages, that information about the visitor's IP address was sent to 356 servers, and that an average of 46 third parties were "present" during each visit. However, none of the six newspapers provided their audience with any information relating to the presence of this large selection of third-party companies.

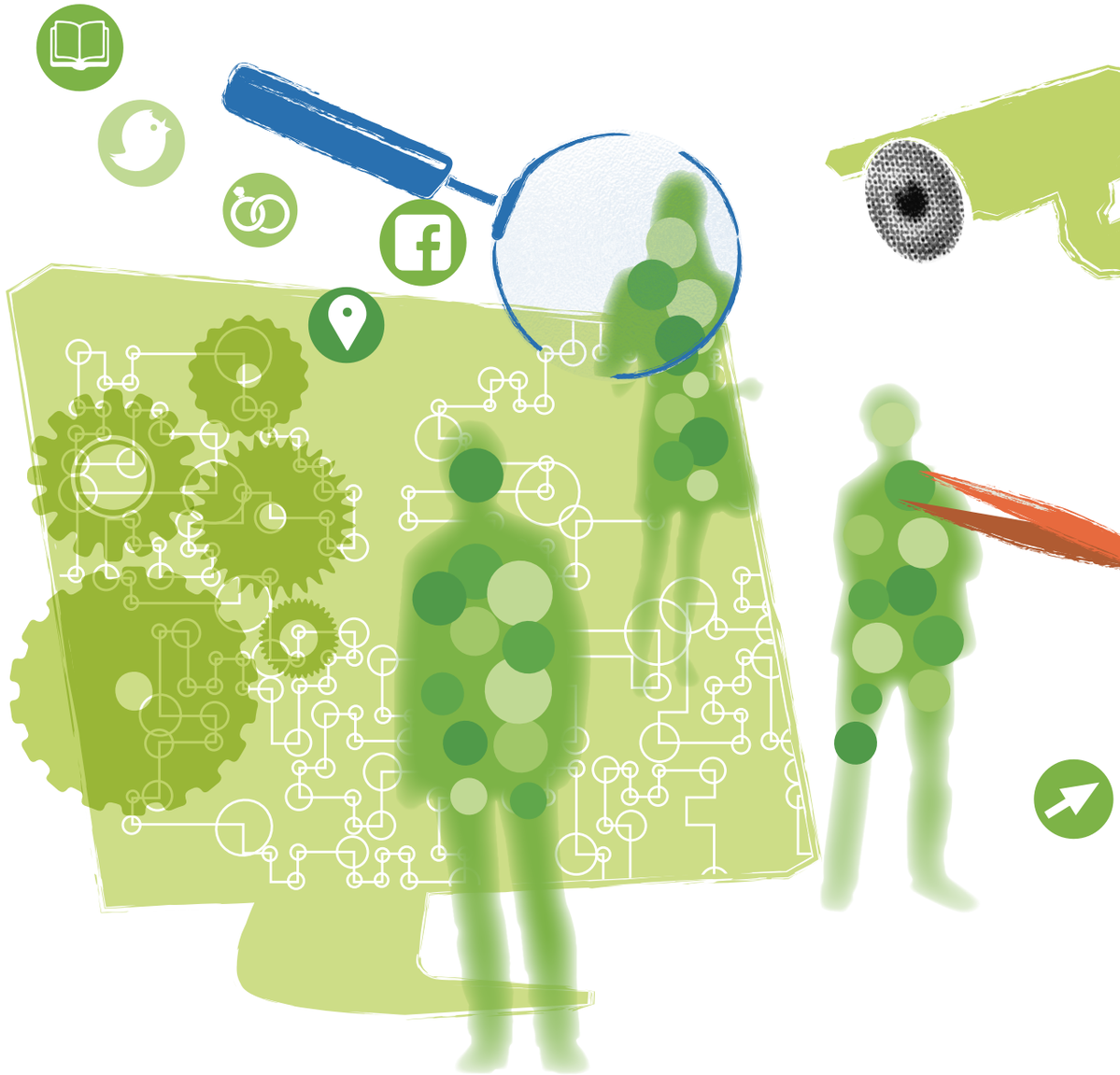
> "SEVERAL SURVEYS CARRIED OUT IN RECENT YEARS HAVE REVEALED THAT USERS ARE BECOMING INCREASINGLY CONCERNED ABOUT THEIR LACK OF CONTROL OVER THE USE AND DISSEMINATION OF THEIR PERSONAL DATA"

The use of web cookies in this manner contributes to the creation of what has been dubbed the "black box society" [13], where users are unable to make informed decisions when choosing services. Any attempt to find the services that are the most privacy friendly is doomed to fail because users are kept largely in the dark.

While advertising companies are the key players in this arena, theirs is far from the only segment that sees the benefits of individually targeted data-gathering practices. The ongoing introduction of innovative analytical methods adds to the importance of the data, including the shift from descriptive to predictive analytics [14].

Growing concerns over lack of control

Several surveys carried out in recent years have revealed that users are becoming increasingly concerned about their lack of control over the use



and dissemination of their personal data. They are particularly worried about having no control over their internet-generated personal data, and the possibility of it being used in ways other than those they originally intended when sharing it [15, 16]. Many people are concerned about the capability of third parties such as advertisers and other commercial entities to access their personal information [16, 17, 18, 19].

A clear majority of internet and online platform users in the European Commission's Special Eurobarometer 2016 expressed their discomfort over the fact that online platforms use information about their internet activities and personal data to tailor advertisements or content to their interests [20]. Further, according to the EU Commission in 2015, only 22 percent of Europeans fully trust companies such as search engines, social networking sites and e-mail services, and as many as 72 percent of internet users are worried about being asked for too much personal data online [21]. In a survey conducted by the Pew Research Center in 2014, as many as 91 percent of US users who took part in the study felt they had lost control over the ways in which their personal details are collected and used by companies [16].

Data collection and handling is clearly fueling many users' growing sense of distrust in service and goods providers. This is naturally a great cause for concern since access to user data is a key enabler of the digital economy. At a certain point, the users' increasing unease could have a damaging effect on service usage levels, and serious repercussions with respect to the digital economy as a whole.

Information overload

Meanwhile, just as the lack of consumer control, and in a sense, the shortage of available information, are problematic, there are indications that the exact opposite – information overload – is also presenting a problem. The information overload in question is specifically related to user agreements, privacy policies and cookie usage. Online user agreements do not appear to be particularly effective in terms of enabling informed user choices. Critics argue that this kind of "privacy self-management" does not provide meaningful control and that there is a need to move beyond relying too heavily on it [22].

In relation to a study on consent practices on Facebook, media scholar and digital sociologist Anja Bechmann posits that "the consent culture of the internet has turned into a blind non-informed consent culture" [23, p. 21]. User agreements often constitute little more than an alibi for providing data-driven businesses with access to user data. The validity of this kind of agreement is consequently questionable.

The trouble with these agreements is that they tend to be too long, too numerous and too obscure. The result is that most users don't read them carefully and are therefore not fully aware of what they are agreeing to when they sign them. For example, a study that tracked the internet browsing behavior of 48,000 monthly visitors to the websites of 90 online software companies found that only one or two of every 1,000 retail software shoppers accessed the license agreement, and that most of those who did access it read no more than a small portion. The conclusion in that study was that the limiting factor in becoming informed thus seemed not to be the cost of accessing license terms but reading and comprehending them [24, cf. 25]. Arguably, the sheer amount of lengthy license agreements that even an average user of digital services agrees to constitutes a sort of information overload. For example, Norway's consumer ombudsman Forbrukerrådet recently conducted a study that involved reading the terms and conditions of all the apps on an average smartphone. Reading them was found to take 31 hours and 49 minutes [26].

Media researcher Helen Nissenbaum has pointed out that the obscurity of the agreements may serve a purpose: if they were written more clearly, they would likely be far less readily accepted [27]. In a recent study, the privacy policies of 75 companies that track behavior in digital contexts were reviewed, and the researchers found that many of them lacked important consumer-relevant management information, particularly with respect to the collection and use of sensitive information, the tracking of personally identifiable data and companies' relationships to third parties [28]. In the short term, a fuzzy and extensive privacy policy appears to be a helpful tool in the data-gathering race. But will there be a price to pay in the long run?

The privacy paradox and acceptance creep

In many cases, there is a significant gap between a service provider's commercial data practices and the normative

> PERCEPTIONS OF PRIVACY AND SOCIAL NORMS RELATING TO COMMERCIAL USE OF INDIVIDUAL DATA CHANGE DYNAMICALLY OVER TIME DUE TO SOCIO-TECHNOLOGICAL SHIFTS IN GENERAL, AND IMPROVED SERVICES IN PARTICULAR

preferences of many – or even most – of its users. Yet research shows that many users often continue to use services that can be very intrusive, while at the same time stating that they are concerned about data being collected when they use products and services online [cf. 23]. Other studies demonstrate that many individuals have not made any major changes to their data sharing or privacy practices in recent years, despite their concerns regarding online data collection [17, 29, 30, 31]. In our behavior, we tend to "accept the cost of free," as noted by competition law scholars Ariel Ezrachi and Maurice Stucke [8, p. 28].

US consumption researchers have put this "privacy paradox" down to consumers' sense of resignation toward the use of their personal data [32]. In the case of loyalty cards, studies show that although consumers do not necessarily feel satisfied with receiving discounts as a trade-off for sharing their personal data, they feel resigned about the situation rather than driven to address the imbalance.

Are these all signs that we are experiencing a phenomenon that legal scholars Mark Burdon and Paul Harpur [33] call "acceptance creep," with massive data collection practices becoming normalized among users? If so, does the acceptance creep merely point to a sense of resignation (too many choices, too much information – resistance is futile) or to the beginning of a fundamental shift in social norms (perceptions) regarding data and privacy?

The answer likely contains a little bit of both. Perceptions of privacy and social norms relating to commercial use of individual data change dynamically over time due to socio-technological shifts in general, and improved services in particular. But the current gap between the stated norms of users and the data practices of service providers is very clear. A great deal of the commercial data collection and handling that is taking place at present is simply not perceived as legitimate. Figuring out how to handle users' normative and behavioral preferences and navigating the "non-informed" consent culture is a major challenge for service designers in a data-driven digital economy.

Ethical implications of information asymmetry

The emergence of big data has added to the information asymmetry between customers and the companies in the insurance, airline and hotel industries and other traditional markets – an effect that is further amplified by the advent of predictive analytics [cf. 8]. This raises several questions about service development and design in terms of how the more qualitative aspects of humanity might be incorporated into all of this quantification. The first question relates to balancing powers on the markets, which in most cases would mean empowering the consumers who are often in the dark with regard to how their data is being collected, analyzed and traded. One way of doing this would be to increase transparency about data practices; another would be to redesign the legal and structural protective measures to better protect weaker parties that have provided "non-informed consent" from being taken advantage of by service providers.

A somewhat more complex question that needs to be addressed is the extent to which users' values and cultures should be considered when designing large-scale automated systems and algorithms. This is related to the ethical and moral questions

that may arise as an outcome of quantification and automation of a particular kind. Concerns like this have only begun to be conceptualized and discussed – one example being a recent report from the committees of the IEEE Global Initiative for Ethical Considerations in Artificial Intelligence and Autonomous Systems [34].

We can already see a growing tendency among market players such as insurance brokers, money lenders and health institutions to base interest rates, insurance costs and payment plans on detailed, big-data-based analyses of individuals. This could also concern predictive analytics of future health, income and life expectancy. What are the potential risks and repercussions of this kind of development from an ethical and normative point of view? More specifically, how should we understand and govern complex (and often proprietary) algorithms and machine-learning processes that may produce troublesome consequences from a social, legal, democratic or other perspective? These are questions that both the public and private sectors need to address urgently.

Commercial practices and the lagging law

One of the key challenges met when regulating the use of human-centric data is that the use of

such information has already become so integral to innovation at a time when both lawmakers and private individuals are still largely unaware of how it is collected and used. From a legal point of view, the challenge is arguably largely the result of a lack of knowledge of growing data practices and their outcomes, but is also of a conceptual kind: how should new practices and phenomena be understood and governed? Law is inevitably path dependent in that it is reliant on past notions or past social and technological conditions when regulating contemporary challenges. The result, according to emerging socio-legal research in the field, is a sort of path-dependent renegotiation of traditional concepts for the regulation of new phenomena [35]. For example, should Facebook be liable for content in that same way as a traditional news outlet when mediating news for its 1.79 billion monthly active users (as of the third quarter of 2016)? Should Uber be regarded as a taxi company and an employer, and be taxed accordingly in each of the more than 60 countries it operates in?

Given that contemporary digital innovation is often disruptive (creating new markets and value networks, and displacing established firms, products and partnerships), the development of new services and products tends to be carried out iteratively. In light of all of this, the fact that the law is lagging is therefore not surprising or strange. Nonetheless, it is vital that we continuously strive to close the gap – particularly in the face of new conceptual dilemmas that involve data-driven innovation, legitimacy and trust.

Conclusion

From a legal point of view, I think regulators must develop a more critical perspective and a better understanding of how to manage data-driven and algorithm-controlled processes as well as data analyses. They must continuously improve their ability to recognize when consumers need protection and empowerment, and strive for transparency with regard to how new technologies work and what kinds of regulations we need to ensure that future developments are in users' best interest.

Ultimately, the continued success and future development of the digital economy will depend on our ability to strike a balance between the interests of individuals, commercial players and governments when it comes to data collection and usage. While regulation in the form of laws such as the Swedish Personal Data Act (Personuppgiftslagen or PuL) and the EU's General Data Protection Regulation (GDPR) will continue to play an important role, the pace of technological development is likely to continue to leave lawmakers playing catch-up.

It is therefore crucial for the private sector to take a proactive approach to addressing normative and ethical questions as part of the service design and development process. Otherwise, there is a significant risk that consumers' trust in digital services will decline in the mid to long term. A low level of trust in new features, services and devices could substantially reduce their potential scalability, and consequently have a negative impact on the digital economy as a whole. ✱

References

1. **Kitchin, R.** *(2014) The Data Revolution. Big data, open data, data infrastructures & their consequences,* **SAGE**
2. **Mayer-Schönberger & Cukier (2013)** *Big Data – A Revolution That Will Transform How We Live, Work, and Think, Boston and New York:* **Eamon Dolan/ Houghton Mifflin Harcourt**
3. **Andrejevic, M. (2013)** *Infoglut. How too Much Information is Changing the Way We Think and Know. New York, NY:* **Routledge**
4. **Rule, J.B. (2012) "Needs" for surveillance and the movement to protect privacy. In K. Ball, K. D. Haggerty & D. Lyon (eds.)** *Routledge handbook of surveillance studies* **(pp. 64-71). Abingdon, Oxon: Routledge**
5. **Rosenblat, A., Kneese, T. & Boyd, D. (2014) "Algorithmic Accountability." A workshop primer produced for The Social, Cultural & Ethical Dimensions of "Big Data" March 17, 2014, NY**
6. **Kitchin, R. & Lauriault, T.P. (2014) Towards critical data studies: Charting and unpacking data assemblages and their work,** *The Programmable City Working Paper 2*
7. **Richards, N.M. & King, J.H. (2014) Big Data Ethics, 49 Wake Forest Law Review, 393-432**
8. **Ezrachi, A. & Stucke, M.E. (2016)** *Virtual Competition. The Promise and Perils of the Algorithm-Driven Economy.* **Harvard University Press**
9. **Datta, A., Tschantz, M.C., Datta, A. (2015) Automated Experiments on Ad Privacy Settings – A Tale of Opacity, Choice, and Discrimination.** *Proceedings on Privacy Enhancing Technologies.* **1: 92–112, DOI: 10.1515/ popets-2015-0007**
10. **Andersson Schwarz, J. (2016) "Platform logic: The need for an interdisciplinary approach to the platform-based economy", paper presented at** *IPP2016*: **The Platform Society, Oxford Internet Institute**
11. **Altaweel, I., Good, N. & Hoofnagle, C. (2015) Web privacy census.** *Technology Science*
12. **Datatilsynet (2015) The Great Data Race.** *How commercial utilization of personal data challenges privacy*
13. **Pasquale, F. (2015)** *The Black Box Society. The Secret Algorithms That Control Money and Information*, **Harvard University Press**
14. **Siegel, E. (2016) Predictive Analytics:** *The Power to Predict Who Will Click, Buy, or Die.* **Wiley**
15. **Lilley, S., Grodzinsky, F.S. & Gumbus, A. (2012)** *Revealing the commercialized and compliant Facebook user.* **Journal of information, communication and ethics in society, 10(2): 82-92**
16. **Pew (2014)** *Public Perceptions of Privacy and Security in the Post-Snowden Era.* **Pew Research Center**
17. **Findahl, O. (2014)** *Svenskarna och Internet 2014.* **Göteborg: .SE**
18. **Kshetri, N. (2014)** *Big data's impact on privacy, security and consumer welfare.* **Telecommunications Policy, 38(11)**
19. **Narayanaswamy, R. & McGrath, L. (2014)** *A Holistic Study of Privacy in Social Networking Sites*, **Academy of Information and Management Sciences Journal, 17(1): 71-85**
20. **Special Eurobarometer 447 (2016) Online Platforms**
21. **COM (2015) 192 final.** *A Digital Single Market Strategy for Europe*
22. **Solove, D.J. (2013)** *Privacy Self-Management and the Consent Dilemma.* **126 Harvard Law Review 1880-1903**
23. **Bechmann, A. (2014)** *Non-informed consent cultures: Privacy policies and app contracts on Facebook.* **Journal of Media Business Studies, 11(1): 21-38**
24. **Bakos, Y., Marotta-Wurgler, F. & Trossen, D.R. (2014)** *Does Anyone Read the Fine Print? Consumer Attention to Standard Form Contracts.* **Journal of Legal Studies, 43(1): 9-40**
25. **McDonald, A.M. & Cranor, L.F. (2008)** *The Cost of Reading Privacy Policies.* **Journal of Law and Policy for the Information Society**
26. **Forbrukerrådet (24 May 2016)** *"250,000 words of app terms and conditions"* **http://www.forbrukerradet.no/ side/250000-words-of-app-terms-and-conditions/**
27. **Nissenbaum, H. (2011)** *A contextual approach to privacy online,* **140 DAEDALUS 32-48**
28. **Cranor, L.F., Hoke, C., Leon, P.G. & Au, A. (2014)** *Are They Worth Reading? An In-Depth Analysis of Online Advertising Companies' Privacy Policies,* **TPRC Conference Paper**
29. **Christensen, M. and Jansson, A. (2015)** *Complicit surveillance, interveillance, and the question of cosmopolitanism: Toward a phenomenological understanding of mediatization.* **New Media & Society, 17(9): 1473-1491**
30. **Light, B. & McGrath, K. (2010)** *Ethics and Social Networking Sites: a disclosive analysis of Facebook.* **Information, technology and people, 23(4): 290-311**
31. **Martin, S., Rainie, L., & Madden, M. (2015)** *Americans Privacy Strategies Post-Snowden.* **Pew Research Center**
32. **Turow, J., Hennessy, M., Draper, N. (2015)** *The Tradeoff Fallacy: How Marketers Are Misrepresenting American Consumers and Opening Them Up to Exploitation.* **Research report, Annenberg School of Communication, University of Pennsylvania**
33. **Burdon, M. & Harpur, P. (2014)** *Re-conceptualizing Privacy and Discrimination in an Age of Talent Analytics,* **University of New South Wales Law Journal 37(2): 679-712**
34. **The IEEE Global Initiative for Ethical Considerations in Artificial Intelligence and Autonomous Systems (2016)** *Ethically Aligned Design: A Vision For Prioritizing Wellbeing With Artificial Intelligence And Autonomous Systems*, **Version 1. IEEE. http://standards. ieee.org/develop/indconn/ec/ autonomous_systems.html**
35. **Larsson, S. (2017)** *Conceptions in the Code: How Metaphors Explain Legal Challenges in Digital Times.* **Oxford University Press**

fixed wireless access

ON A MASSIVE SCALE WITH 5G

The promise of ubiquitous fixed wireless access (FWA) looms large with every new generation of wireless or mobile technology, and 5G is no exception. Indeed, one of the 5G use cases currently gaining momentum in the industry is FWA for both small and medium-sized enterprise (SME) and residential applications.

KIM LARAQUI,
SIBEL TOMBAZ,
ANDERS FURUSKÄR,
BJÖRN SKUBIC,
ALA NAZARI,
ELMAR TROJER

FWA is a concept for providing broadband service to homes and SMEs that is particularly attractive in cases where there is no infrastructure in place to deliver wired broadband via copper, fiber or hybrid solutions. It can also be used when the existing infrastructure is not able to provide sufficient service. With 5G due to provide 10 to 100 times more capacity than 4G, it has the potential to enable cost-efficient FWA solutions on a massive scale.

Already today, in LTE with 40MHz of bandwidth, there is often a working business case for FWA as an add-on improvement to mobile broadband (MBB) and it only becomes stronger as LTE evolves. The further evolution toward 5G has the potential to take FWA to a whole new level. This is because 5G offers unprecedented technology options that make it possible to use larger chunks of radio spectrum and provide consumers with benefits like low latency (1ms) and major capacity improvements. Many of these options are relevant to the evolution of 4G as well.

Compared with fiber-to-the-home (FTTH) and other wireline solutions, FWA offers a variety of benefits including significantly lower rollout costs, rapid service rollout and lower opex. This is because the bulk of the costs and most of the complexity involved in fixed access deployments are associated with the last mile: the portion of the network that reaches the user premises.

FWA also offers an opportunity to double the impact of a 5G deployment by addressing the two prominent 5G use cases – MBB and fixed wireless – simultaneously. The 5G beams that serve mobile users outdoors during the daytime can be redirected to an FWA terminal when people return home in the evening, thereby strengthening the case for 5G deployment and its outlook as an affordable and sustainable technology.

5G-based FWA is expected to enable robust services with sustainable rates that are high enough to meet the foreseeable needs for home use well into the future. It is also poised to offer peak cell rates that few fixed technologies will be able to match without very costly investments into deep-fiber fixed access infrastructure deployments.

In many situations, FWA – based on 3G, 4G or 5G – may be the only feasible broadband access (BBA) option, particularly in rural areas and emerging markets with limited fixed BBA infrastructures,

> 5G-BASED FWA IS EXPECTED TO ENABLE ROBUST SERVICES WITH SUSTAINABLE RATES THAT ARE HIGH ENOUGH TO MEET THE FORESEEABLE NEEDS FOR HOME USE WELL INTO THE FUTURE

which comprise the majority of homes around the globe. For example, although more than one-third of all households in developing countries have internet access, only about 20 percent of that access is provided through fixed broadband [1]. 5G in lower frequency bands – such as 3.5GHz – opens up for much higher capacities in the realm of 3GPP radio access as a residential broadband technology.

5G FWA could also be used to boost existing fixed BBA in dense urban deployments to achieve higher peak rates and thereby meet increasing bandwidth and latency requirements without having to make comprehensive upgrades to the

Terms and abbreviations
BBA – broadband access | **BF** – beamforming | **CO** – Central Office | **CPE** – customer premises equipment | **CPRI** – Common Public Radio Interface | **CWDM** – coarse wavelength division multiplexing | **FDP** – Fiber Distribution Point | **FSO** – free-space optics | **FTTH** – fiber-to-the-home | **FWA** – fixed wireless access | **GE-PON** – Gigabit Ethernet Passive Optical Network | **IoT** – Internet of Things | **ISD** – inter-site distance | **ISP** – internet service provider | **MAC** – Media/Medium Access Control | **MBB** – mobile broadband | **MNO** – mobile network operator | **MU-MIMO** – multi-user multiple-input, multiple-output | **MVNO** – Mobile Virtual Network Operator | **MW** – microwave | **NGCO** – next generation central office | **NG-PON2** – Next Generation Passive Optical Network 2 | **NR** – New Radio | **ODN** – optical distribution network | **OTN** – optical transport network | **P2P** – peer-to-peer/point-to-point | **PHY** – physical layer | **RBU** – Radio Baseband Unit | **RRU** – remote radio unit | **SME** – small and medium-sized enterprise | **SNR** – signal-to-noise ratio | **STA** – Station | **TWDM-PON** – time and wavelength division multiplexed passive optical network | **UE** – user equipment | **WDM-PON** – wavelength division multiplexed passive optical network | **WR-WDM-PON** – wavelength routed WDM-PON | **WS-WDM-PON** – wavelength selective WDM-PON | **XG-PON** – 10-gigabit-capable passive optical network

physical infrastructure. In particular, 5G FWA appears capable of addressing the bandwidth saturation issue caused by the high demand for typical residential services such as IPTV. The very low latency of 5G access is also a potential key enabler for future applications.

Thanks to high-efficiency data compression techniques, variable bitrate video and adaptive bitrate streaming, 5G FWA is well positioned to become a leading media distribution technology. High-efficiency data compression techniques allow for the delivery of high-resolution video with less bandwidth, while variable bitrate video enables the transport of more video streams using less bandwidth than constant bitrate. Finally, adaptive bitrate streaming is a technique that enables the best possible multimedia viewing experience, as it adapts automatically to any changes in network conditions (such as fluctuations in available bandwidth).

5G and the FWA opportunity

To enable higher user data rates and greater system capacity, 5G radio will make use of new and often higher frequency bands. The most prominent band options currently under consideration are 3.5GHz, 28GHz, 37GHz and 39GHz, in addition to the bands used for legacy cellular technologies.

A technique called beamforming makes it easier to provide coverage at high frequencies. Massive beamforming at high frequencies creates narrow beams that can be redirected easily as required. The signals from multiple user terminals can be multiplexed simultaneously on the same frequency resource in different beams. This is often referred to as multi-user multiple-input, multiple-output (MU-MIMO).

The possibility of using high-gain antennas on the terminal side (both indoors and outdoors) makes the higher frequencies more useful. The more static channels simplify beamforming and MU-MIMO user pairing.

FWA works by using wireless technologies (such as 5G) to connect a base station or wireless access point to a special kind of user terminal referred to as a fixed wireless terminal (FWT), which then provides backhauling services for customer premises equipment (CPE). Sometimes the FWT is integrated with the CPE in the same enclosure. But most often the FWT is installed in one place (fixed on a particular premises) in proximity to an outdoor antenna, and it is rarely moved.

Figure 1 outlines a few alternatives for possible 5G FWA deployments. The placement of the RBS in relation to other nodes depends on the frequency it operates on – the higher the frequency, the shorter the reach of radio links from the RBS. The mainly indoor entities that provide consumer connectivity are orange, and the corresponding outdoor entities are green.

Since 5G will support multi-access networks, it will be possible to deploy FWA as a complement to existing fixed BBA to boost peak rates for the home CPEs [2]. It is also increasingly evident that 5G FWA will be able to offer very attractive services that can compete with high-capacity fixed solutions.

FWA is steadily becoming a more sustainable alternative to fixed BBA due to the ongoing incorporation of more spectrum, beamforming, advancements in terminals, optimization of media distribution, virtualization of RAN and core, and other forms of technological progress in the 4G/5G arena. *Figure 2* is a schematic illustration of production cost per subscriber as a function of traffic per subscriber, depicting how 5G technologies can add value for consumers in an FWA scenario. (Note that the numbers on the x and y axes are representative – the actual numbers depend on many factors that may vary in different parts of the world.)

One obvious advantage of 5G FWA is its ability to support very high peak rates without requiring dedicated fixed facilities for each consumer. In fixed networks, the fiber or copper plant needs to be physically dimensioned for each consumer's fixed rate. Upgrading existing fixed plants is typically a slow and costly process, not least due to deployment costs and rights of way. By 2020, BBA speeds in excess of 100Mbps are expected to be available in less than 10 percent of all residential connections worldwide [3]. In

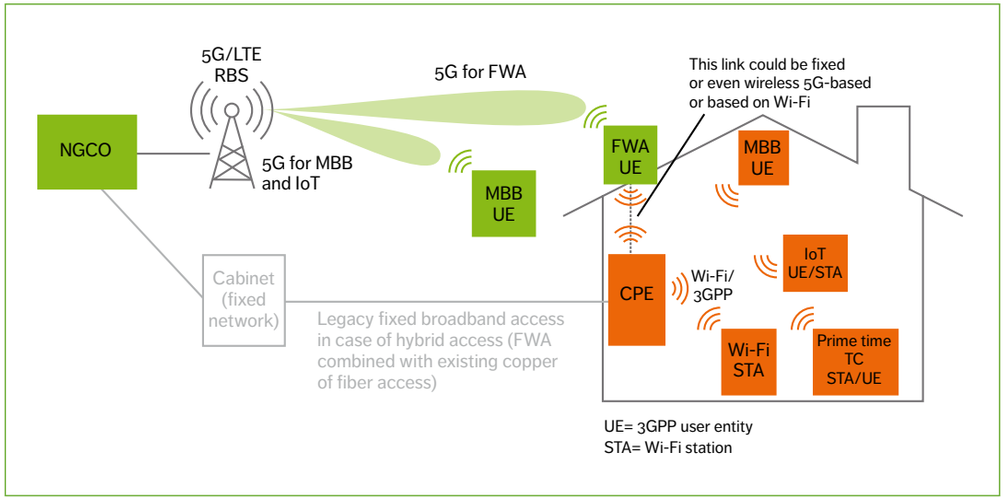


Figure 1 Examples of FWA deployment alternatives

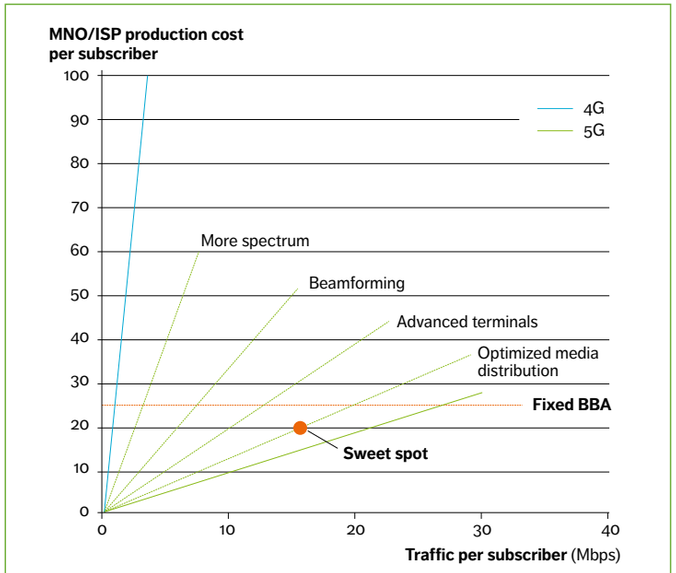


Figure 2 Production cost per subscriber as a function of traffic per subscriber

Parameter	Value
Base station transmit power	35dBm
CPE transmit power	30dBm
Channel bandwidth	200MHz
Operating frequency	28GHz
Duplex mode	TDD with 57 percent downlink allocation

Figure 3 Simulation assumptions

fact, in 2020, almost two-thirds of all broadband connections will still provide peak rates below 25Mbps. Such numbers suggest that there are many opportunities around the world for 5G to complement existing solutions or create new ones, as our use case below also suggests.

The 5G-based FWA use case

We have studied a range of different scenarios to assess the system performance that is achievable in fixed wireless use cases. Our findings show that coverage and overall performance largely depend on which frequency band is used, the environment or terrain the system operates in and the placement of the terminal antenna.

One of the key scenarios we have studied is a suburban environment with 1,000 households per square kilometer. Twenty-five percent of the households use a 4K UHD video service (video on demand or linear) that requires a download speed of at least 15Mbps for uninterrupted playout of basic 4K video streams. To support this demand, a network is deployed with base stations on utility poles that are 6m tall. Terminal antennas are placed outdoors – often on rooftops or walls – as well as indoors. The buildings are 4-10m tall, and there are trees in the area that reach heights of 5-15m and attenuate the signal. In our study, the buildings and trees were represented on a three-dimensional digital map, and a ray-tracing technique was used to create a model showing their impact on radio propagation; this included diffractions, reflections, path loss owing to the effect of foliage, and building penetration loss.

The system design was based on a preliminary 5G New Radio (NR) concept operating at 28GHz with a bandwidth of 200MHz, utilizing beamforming and MU-MIMO and enabled by a base station antenna array of 8x12 cross-pole elements. We made the conservative assumption that terminal antennas were omni-directional with a 10dBi gain. Two-layer MIMO was used for each user. A summary of the simulation assumptions is presented in *Figure 3*.

Figure 4 shows a map of the environment where each user was assigned a color based on the data rates they received at a low traffic load. A majority of the users enjoyed data rates in excess

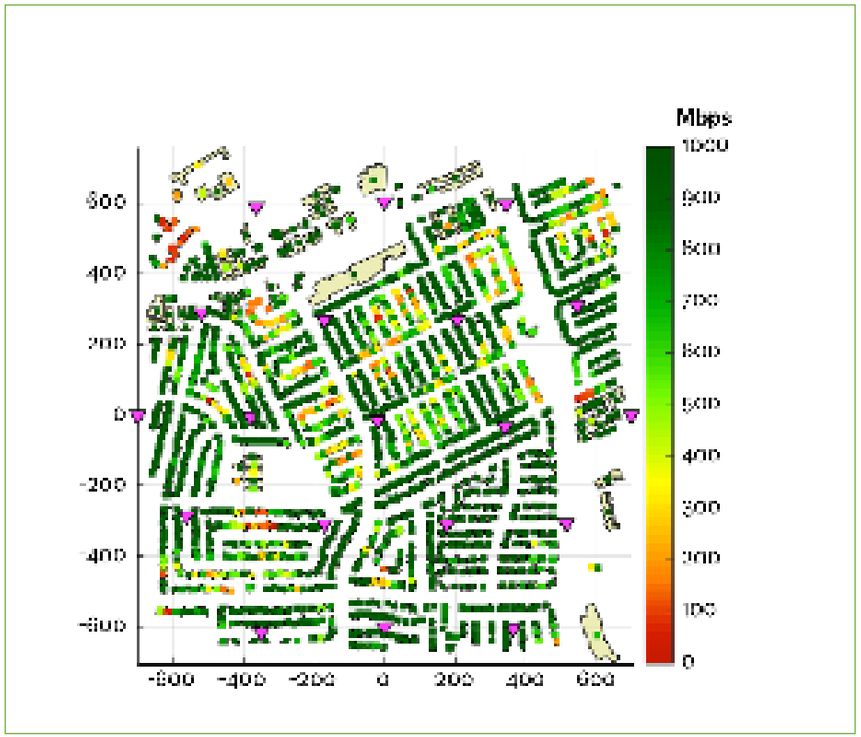


Figure 4 Throughput map of suburban area at low load

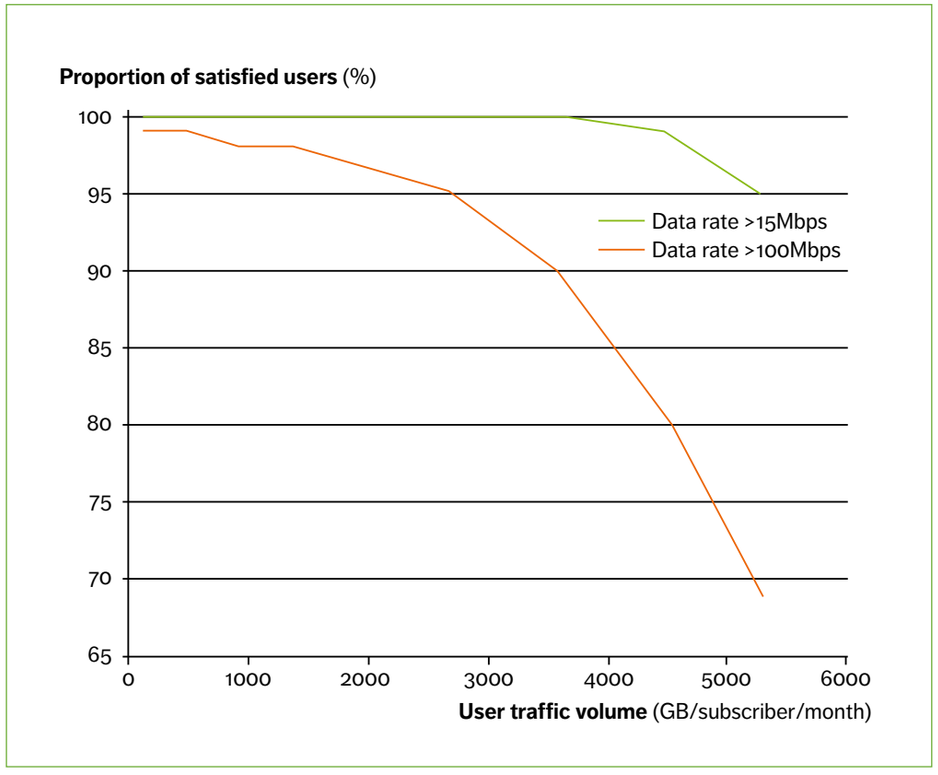


Figure 5 Proportion of satisfied users as a function of traffic

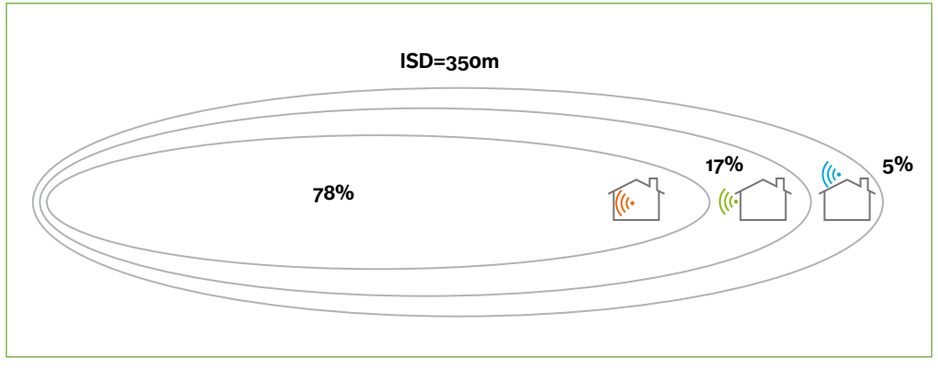


Figure 6 Breakdown of indoor, wall-mounted and rooftop antennas for an ISD of 350m

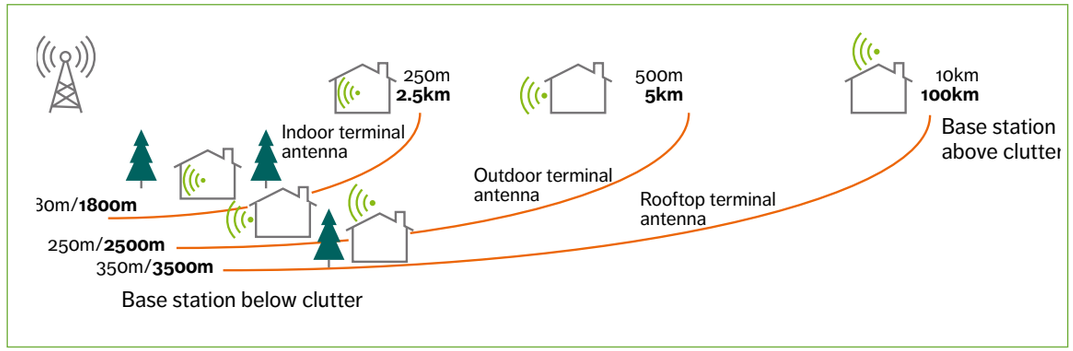


Figure 7 Key findings on FWA feasibility for different scenarios and frequencies (3.5GHz cell ranges appear in bold text and 28GHz cell ranges in plain text)

of 800Mbps. Only 11 percent of them had a data rate of below 400Mbps and all of them had the targeted 15Mbps.

When the traffic load increases in this scenario, user data rates decrease owing to greater interference and queuing. *Figure 5* shows the proportion of users with data rates exceeding the required 15Mbps and 100Mbps, depending on the traffic load. The maximum traffic load in this scenario is 5200GB per month per subscription, with 95 percent of the users benefiting from a data rate that exceeds the targeted 15Mbps. This is equal to 1Gbps per site. Sixty nine percent of the users enjoy a data rate of more than 100Mbps at this high load level.

The results shown in Figure 5 are based on the use of rooftop antennas. However, it is possible to achieve similar results with wall-mounted and indoor antennas, but only for smaller cells, due to diffraction and indoor penetration losses.

Our research indicates that the optimal 5G FWA solution is likely to include hybrid terminal antenna placement, where only the users who are furthest from the base station use rooftop antennas, while those closer to it use outdoor wall-mounted or indoor alternatives. The percentage of households using each of the three antenna placement variations is illustrated in *Figure 6*. The results show that for an inter-site distance (ISD) of 350m, 78 percent of the households can use indoor antennas (usually integrated into CPEs), whereas the rest should rely on either outdoor wall-mounted antennas (17 percent) or rooftop-mounted antennas (5 percent) to achieve better propagation conditions.

We made similar analyses for other combinations of frequency bands, environments and terminal antenna placements. *Figure 7* provides a summary of a selection of these, focusing on the 3.5GHz and 28GHz frequency bands. While these two bands are by no means the only frequency options for FWA, they are good examples of low and high frequency FWA solutions that can provide insight in terms of the feasibility and usability of FWA for different applications and services.

The key performance findings for both 3.5GHz and 28GHz are encouraging. As expected, 3.5GHz provides very good mobile coverage, allowing longer reach compared with the 28GHz band. Although the available bandwidth is smaller

> THE ABILITY TO USE SOFTWARE TO DYNAMICALLY CONFIGURE CORE AND SERVICE NETWORKS ALSO PLAYS AN IMPORTANT ROLE IN GENERATING THE FLEXIBILITY REQUIRED TO ENABLE THE DEPLOYMENT OF FWA SOLUTIONS ON A TRULY CONVERGED MOBILE AND WIRELINE HARDWARE INFRASTRUCTURE

at 3.5GHz, the use of massive beamforming and MU-MIMO provides very high cell spectral efficiency, making this band a great candidate for delivering video services. Moreover, this band can be used to deliver basic home broadband connectivity – as an outside-in MBB available to indoor users, for example – which would make it easier to realize the vision of connecting the billions of unconnected people in rural and remote areas.

With 28GHz, the achievable cell ranges are much lower owing to worse propagation conditions, and they strongly depend on the environment or terrain in which the system operates. The most important factors affecting the feasible ranges are:

- the placement of the terminal antenna
- the height and density of the trees and buildings
- the height of the base station antenna placement.

Since the 28GHz band is more sensitive to building penetration and diffraction losses, rooftop placement of the terminals provides the largest range due to higher line-of-sight probability between the terminals and the base station. Use of outdoor wall-mounted and indoor terminals reduces the ranges significantly.

It is critical to consider the propagation effect of foliage on the cell ranges at 28GHz. Deploying the base station antennas at a height greater than that of the tallest trees in the area significantly boosts the cell ranges. In terms of capacity, the availability of larger bandwidth and the possibility of utilizing a large number of antennas for massive beamforming enables very large cell capacity at 28GHz. These factors make the 28GHz band suitable for fixed wireless service in dense suburban and urban areas.

Several options for FWA transport

FWA poses new challenges in providing cell site connectivity. Compared with conventional macro deployments, FWA may require 10 times more cells and cell site connections, putting significant strain on the backhaul network. As shown in Figure 7, the required ISD varies from several hundred meters to a few kilometers depending on the actual 5G radio deployment.

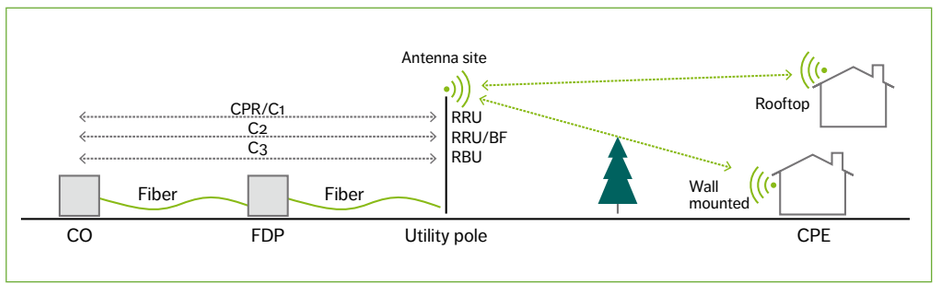


Figure 8 Schematic figure of FWA showing CPE, antenna site, FDP and CO

RAN split	Parameters that determine capacity (antenna configuration, user traffic and so on)	Required transport capacity per site
CPRI (4G)	Bandwidth, SNR (quantization, reach), number of antennas, traffic load, overhead	40-100Gbps
C1 – Evolved CPRI Any form of digital time-domain radio carrier and beamforming weight representation in 5G, including compression	Bandwidth, SNR (quantization, reach), number of antennas, traffic load, overhead	40Gbps
C2 – Split-PHY (lower stack split) Digital frequency domain radio signal and weight representation	Bandwidth, SNR, number of layers, traffic load, overhead	10-25Gbps
C3 – MAC-PHY split (higher stack split) Digital MAC bearer and state representation	Bandwidth, SNR, number of layers, traffic load	5-10Gbps

Figure 9 Table showing interface capacities of different RAN splits

With 5G, several functional splits have been proposed to enable new scenarios for the deployment of RAN functions across sites [4]. Three of these – C1, C2 and C3 – are shown in *Figure 8*. The specific requirements on the transport network depend on how the RAN is deployed and which interfaces are carried over the transport network.

Generally, FWA deployment requires the use of advanced array antennas to support MU-MIMO and beamforming for the required capacities and peak rates of residential access. This in turn determines the interface capacities of different RAN splits, as shown in *Figure 9*.

In the FWA use case we have presented here,

each cell site (utility pole) serves approximately 25 customers, resulting in user data bandwidth requirements per utility pole of 5Gbps at peak rate and 1Gbps sustainable rate. For lower splits such as CPRI, the antenna configuration in FWA would require very high transport bitrates, which is not feasible in the access segment. Instead, splits higher up in the layers are more likely (MAC-PHY, for example) where transport bitrates can be kept below 10G per site.

Compared with fixed access systems, the fan-out requirements for FWA transport solutions are lower, while requirements on capacity are higher. The requirements on latency and jitter are also more stringent. The densification of cell sites required by FWA means that the transport solutions may need to provide connectivity to 10 times as many sites as in today's mobile deployments. This is still just a fraction of the number of connections needed in fixed wireline access deployments, though.

The choice of optimal transport solution depends on factors such as available copper/fiber infrastructure and site structure. A range of possible transport solutions (both optical and wireless) that could support FWA are listed in *Figure 10*.

Enabling technologies

The 5G FWA concept will only become stronger and more flexible as a result of the family of enabling technologies that is currently being incorporated into implementations of various portions of 5G systems, including:

- 5G and 4G RAN [4], which will increase deployment flexibility and network scalability – necessary for meeting a wide variety of coming performance requirements
- core networks [5], with a focus on software-defined networking and virtualization to provide elastic connectivity
- next generation central offices (NGCOs) [6], which will be able to provide the necessary facilities (such as mini data centers) required to meet fixed and mobile service and infrastructure convergence challenges.

Taken together, these new attributes of next generation mobile networks will provide a very potent toolbox to meet future FWA needs.

A split 5G and 4G RAN architecture is of particular significance when building FWA solutions because it makes it possible to place functions (including those in the RAN) dynamically across the access network to fulfill various needs. Functional node types execute on pools of hardware with both special and general purpose processors. This provides the necessary flexibility to adapt networks to future capacity, latency and other needs, such as supporting future virtual and augmented reality applications in homes.

The ability to use software to dynamically configure core and service networks also plays an important role in generating the flexibility required to enable the deployment of FWA solutions on a truly converged mobile and wireline hardware infrastructure. This means that features, functions and operational capabilities conceived for mobile networks can also be used for FWA where appropriate. Solutions like blind cache [7] for optimized content delivery, models for network sharing, and unbundling via Mobile Virtual Network Operators (MNVOA) and other approaches are just a few examples.

Conclusion

Key technology enablers such as beamforming and new frequency bands, in combination with advances in mobile back and front hauling, network virtualization and network programmability, are strengthening the FWA concept significantly. While the exact characteristics of any FWA deployment are case specific, our research suggests that 5G-based FWA is definitely an option to fulfill the advanced future service requirements of the homes and SMEs of tomorrow in many types of environments around the globe.

With 5G, we have the opportunity to achieve true network convergence, since the same technology and indeed the same infrastructure can be used to provide next generation MBB, IoT and FWA. ✱

	Solution	Advantages and disadvantages
Optical systems	P2P fiber (grey optics)	⊕ Low-cost optics and support for high capacity and low latency ⊖ Requires fiber-rich deployment
	TWDM-PON (such as XG-PON, NG-PON2, GE-PON)	⊕ Low-cost potential and potential system reuse between FWA and FTTH clients ⊖ Limited capacity (≤10G) and limited low latency support, limiting possible RAN deployment options (functional splits, RAN coordination) and RAN services (low latency services)
	WDM-PON (such as WS-WDM-PON, WR-WDM-PON)	⊕ Dedicated solution for RAN transport where optical distribution network (ODN) deployment can be tailored for desired RAN deployment ⊖ Limited reuse of potentially existing FTTH infrastructure and potential issues for future migration of customers (individuals or groups) to FTTH, which then requires a separate ODN. Low fan-out of typical scenarios limits need for dense WDM (CWDM is sufficient)
	P2P WDM overlay (such as NG-PON2)	⊕ Reuse of potentially existing fiber plant for providing P2P connections for mobile transport. Support for high capacity and low latency ⊖ High costs and footprint associated with ODN filters
Active systems	Ethernet (such as for CPRI over Ethernet), OTN	⊕ Reuse of existing infrastructure suitable for active network deployment ⊖ Deployment options (RAN splits) practically limited by deployed active equipment (capacity and protocol support)
Wireless systems	In-band wireless (5G, LTE)	⊕ Low-cost deployment ⊖ Spectrum is shared between access and transport (less overall capacity or more spectrum needed)
	Out-of-band wireless (MW, FSO and so on)	⊕ Low-cost deployment compared with fiber but more effort needed compared with in-band ⊖ Dependent on solution or spectrum, whether or not licensed spectrum is needed, sensitivity to weather conditions and so on

Figure 10 Possible transport solutions for FWA

Further reading

- **Ericsson 5G PlugIns, Enabling the evolution:** *https://www.ericsson.com/networks/offerings/5g-plug-ins*
- **The Connected Building – Microwave to and between buildings :** *https://www.ericsson.com/spotlight/industries/our-industries/real-estate/microwave-connected-buildings*
- **State and future of the mobile networks:** *https://www.ericsson.com/mobility-report/state-and-future-of-the-mobile-networks*
- **Wireless backhaul in future heterogeneous networks:** *https://www.ericsson.com/news/141114-wireless-backhaul-in-future-heterogeneous-networks_244099435_c?fromDate=2014-01-01&categoryFilter=ericsson_review_1270673222_c&toDate=2014-12-31*

References:

1. **ITU-T, ICT Facts & Figures, 2015, available at:** *https://www.itu.int/en/ITU-D/Statistics/Documents/facts/ICTFactsFigures2015.pdf*
2. **Ericsson Technology Review, Bolstering the last mile with multipath TCP, Robert Skog, Dinand Roeland, Jaume Rius i Riu, Uwe Horn, Michael Eriksson, October 2016, available at:** *https://www.ericsson.com/thecompany/our_publications/ericsson_technology_review/archive/bolstering-the-last-mile-with-multipath-tcp*
3. **Cisco, white paper, The Zettabyte Era - Trends and Analysis, June 2016 , available at:** *http://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/vni-hyperconnectivity-wp.html*
4. **Ericsson Technology Review, 4G/5G RAN architecture: how a split can make the difference, Erik Westerberg, July 2016, available at:** *https://www.ericsson.com/thecompany/our_publications/ericsson_technology_review/archive/4g-5g-ran-architecture-how-a-split-makes-a-difference*
5. **Ericsson Technology Review, A vision of the 5G core: flexibility for new business opportunities, Henrik Basilier, Lars Frid, Göran Hall, Gunnar Nilsson, Dinand Roeland, Göran Rune, and Martin Stuempert, February 2016, available at:** *https://www.ericsson.com/thecompany/our_publications/ericsson_technology_review/archive/5g-core-vision*
6. **Ericsson Technology Review, The central office of the ICT era: agile, smart and autonomous, Nail Kavak, Andrew Wilkinson, John Larkins, Sunil Patil, and Bob Frazier, May 2016, available at:** *www.ericsson.com/thecompany/our_publications/ericsson_technology_review/archive/central-office-of-the-ict-era*
7. **Ericsson Technology Review, Blind cache: a solution to content delivery challenges in an all-encrypted web, Göran A.P Eriksson, John Mattsson, Nilo Mitra, Zaheduzzaman Sarker, August 2016, available at:** *https://www.ericsson.com/thecompany/our_publications/ericsson_technology_review/archive/5g-core-vision*

THE AUTHORS

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 has an M.Sc. in electrical engineering and a Ph.D. in radio communications systems, both from KTH Royal Institute of Technology in Stockholm, Sweden.

Kim Laraqui

◆ joined Ericsson in 2008 as a senior customer solution manager, and is currently principal researcher on transport and access solutions for heterogeneous networks. Previously, he was a senior consultant on network solutions, design, auditing, deployment and operations for mobile and fixed operators worldwide. He holds an M.Sc. in computer science and engineering from KTH Royal Institute of Technology in Stockholm, Sweden.

Sibel Tombaz

◆ joined Ericsson in 2014 and currently works as a senior researcher in the Wireless Access Networks department at Ericsson Research. Her work mainly focuses on 5G concept development, 5G use cases and scenarios, and RAN energy performance. She holds an M.Sc. in electrical and electronics engineering from Istanbul University, Turkey, and a Ph.D. in communication systems from KTH Royal Institute of Technology in Stockholm, Sweden.

Ala Nazari

◆ is a media delivery expert within Business Unit Media. He joined Ericsson in 1998 as a specialist in datacom, working with GPRS, 3G and IP transport. More recently, he has worked as a senior solution architect and engagement manager. Prior to joining Ericsson, he spent several years at Televerket Radio working with mobile and fixed BBA and transport. He holds an M.Sc. in computer science from Uppsala University in Sweden.

Björn Skubic

◆ is a senior researcher in networking technologies who is currently driving activities in 5G transport. He joined Ericsson in 2008 and has worked in several areas including optical transport, energy efficiency and fixed access. He holds a Ph.D. in physics from Uppsala University, Sweden.

Elmar Trojer

◆ joined Ericsson in 2005 and is currently working as a principal researcher on fronthaul networking technologies for 5G. He has been active on both fixed and MBB technologies such as VDSL2, G.fast, GPON and mobile backhaul technologies for 4G small cells. He holds a Ph.D. in electrical engineering from the Vienna University of Technology, and an MBA from the University of Vienna, Austria.

BOLSTERING the last mile WITH MULTIPATH TCP

The rapid uptake of bandwidth-consuming services such as video on demand and linear TV has many service providers struggling to keep pace with ever increasing bandwidth demands. The problem is particularly acute on the last mile: the segment of the network that delivers broadband services to users' homes and workplaces.

ROBERT SKOG, DINAND ROELAND, JAUME RIUS I RIU, UWE HORN, MICHAEL ERIKSSON

AS AN ALTERNATIVE to building out the physical communications infrastructure – which in some geographical areas may be too costly or time consuming – Ericsson proposes an access aggregation solution based on Multipath TCP. Our solution consists of a carrier-grade Multipath TCP proxy that allows the use of Multipath TCP across access networks without the need to introduce it in end devices or internet servers.

The last mile is the part of the telecommunications network that physically reaches user premises, either by wireless technology (cellular networks) or wireline technology such as cable, fiber or digital subscriber line (DSL). The achievable data rates for each of these access technologies vary, but in many cases the bandwidth depends on the distance between the access termination point in the service provider network and the device in the user premises. This means that no matter how fast the service is up to the access termination point, the users who are farthest away from it will experience significantly slower service than the ones who are closer.

For example, although the MOST recently standardized DSL technologies allow bitrates of up to 1Gbps, most subscribers today are still getting less than 20Mbps. The reason for this is the dependency between the achievable bitrate and the length of the copper line connecting a household to the DSL access multiplexer (DSLAM). As *Figure 1* shows, if the distance between the user premises and the DSLAM exceeds 2km, DSL speed falls quickly below 20Mbps. The obvious solution is to reduce the length of the last mile. If the copper line distance can be reduced to less than 250m, new technologies and standards such as vectoring and G.fast will allow bitrates of about 1Gbps. However, reducing the copper line distance is costly because it requires the deployment of more street cabinets connected by fiber lines to the backbone network. To get around this, some fixed broadband service providers have started to launch offerings that combine DSL with LTE as a cheaper way to boost the bitrate for DSL customers than deploying more fiber-connected DSLAM street cabinets.

Similarly, LTE/wi-fi aggregation is useful as a booster for mobile phones. Some operators have started deploying solutions that combine Wi-Fi and LTE accesses in areas such as shopping malls and big event venues as a means to increase user capacity while at the same time offloading their cellular network traffic to the fixed networks when possible.

> MULTIPATH TCP CAN BE USED ACROSS ALL KINDS OF ACCESS NETWORKS, PROVIDING A RICH TOOLKIT THAT SUPPORTS ACCESS AGGREGATION FOR USE CASES SUCH AS BANDWIDTH AGGREGATION, RELIABILITY AND SEAMLESS CONNECTIVITY

Technologies for access aggregation

Many standardized aggregation technologies only support use cases in which links using the same access type are aggregated. This is known as bonding, and examples include the bonding of several Ethernet links, or of two DSL access links. Notable exceptions are IP Flow Mobility and multiple-access PDN connectivity – both defined by 3GPP – which are able to support aggregation of multiple access types [1]. However, these two technologies have gained little traction because

Terms and abbreviations
ACK – ACKnowledgment | **CCA** – Congestion Control Algorithm | **CPE** – customer premises equipment | **CPU** – central processing unit | **DPDK** – Data Plane Development Kit | **DSL** – digital subscriber line | **DSLAM** – DSL access multiplexer | **IETF** – Internet Engineering Task Force | **MFDN** – Media First Delivery Node | **RNA** – Radio Network Aware | **RTT** – round-trip time | **TCP RNA** – TCP Radio Network Aware | **VDSL** – Very high-speed DSL

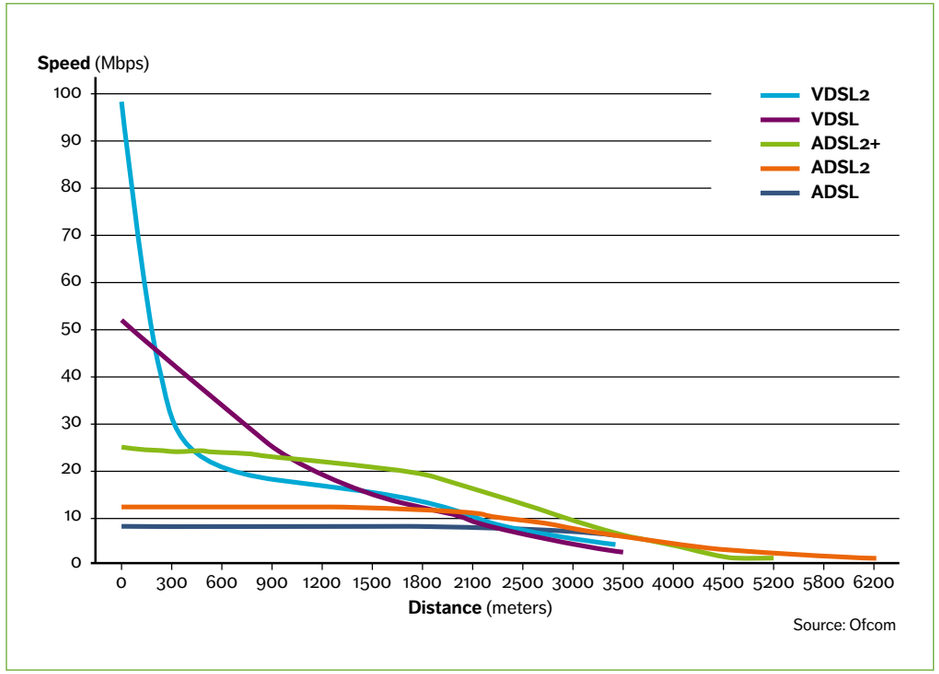


Figure 1 Speed versus copper line length between user premises and the DSLAM for the most widely deployed DSL technologies

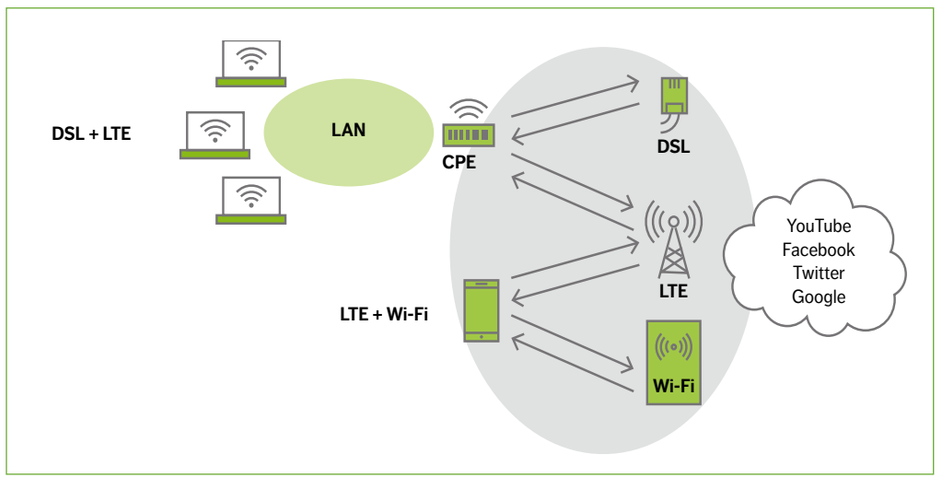


Figure 2 Examples of access aggregation enabled by Multipath TCP

their introduction on mobile devices would require a significant implementation effort, and even the apps running on them would require modifications.

Multipath TCP as specified by the IETF [2] can be deployed in existing networks more easily than other alternatives because it is an evolution of TCP [3] – the most widely used protocol in the internet today. This guarantees interoperability between equipment from different vendors. Like TCP, Multipath TCP works on top of IP. Since IP is the foundation of all internet protocols, Multipath TCP can be used across all kinds of access networks, providing a rich toolkit that supports access aggregation for use cases such as bandwidth aggregation, reliability and seamless connectivity. In addition, there is an open source reference implementation for Multipath TCP that is continuously developed and improved by a large community of developers [4].

Figure 2 shows two access aggregation scenarios enabled by Multipath TCP. The first scenario shows DSL/LTE aggregation, where an existing DSL connection is combined with LTE. If the DSL link provides 12Mbps and the LTE link provides 8Mbps, the aggregated bandwidth that can be obtained via Multipath TCP is roughly 20Mbps.

The second scenario shows LTE/Wi-Fi aggregation, which functions according to the same principle. Together with a mobile device manufacturer, Ericsson has performed successful field trials in public LTE and Wi-Fi networks using commercially available mobile devices. Only the firmware was modified to support Multipath TCP.

> BANDWIDTH AGGREGATION REFERS TO THE ABILITY OF MULTIPATH TCP TO COMBINE THE BANDWIDTH OF SEVERAL LINKS INTO ONE LOGICAL CONNECTION

Although the benefits of Multipath TCP are often presented in the context of two different access networks, there is no limit in Multipath TCP that would prevent the use of three, four or more access networks. The access networks could even be operated by different service providers, which is an additional benefit for use cases aiming for improved resiliency.

Aggregating bandwidth

Bandwidth aggregation refers to the ability of Multipath TCP to combine the bandwidth of several links into one logical connection. *Figure 3* shows

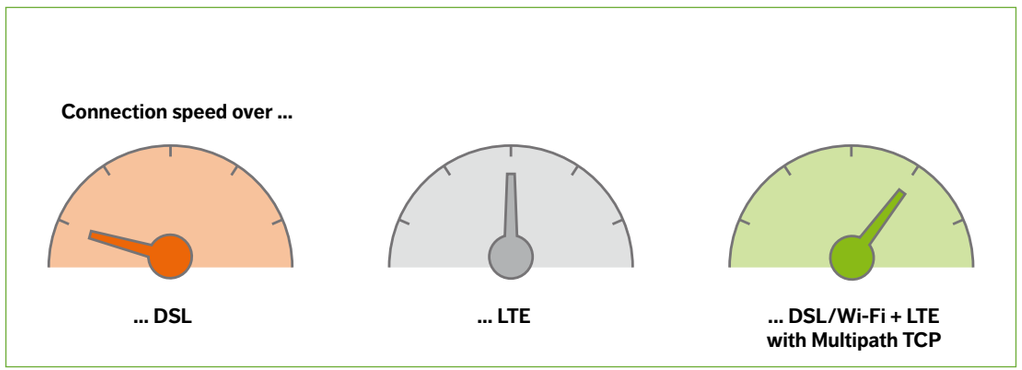


Figure 3 DSL and LTE bandwidth aggregation with Multipath TCP

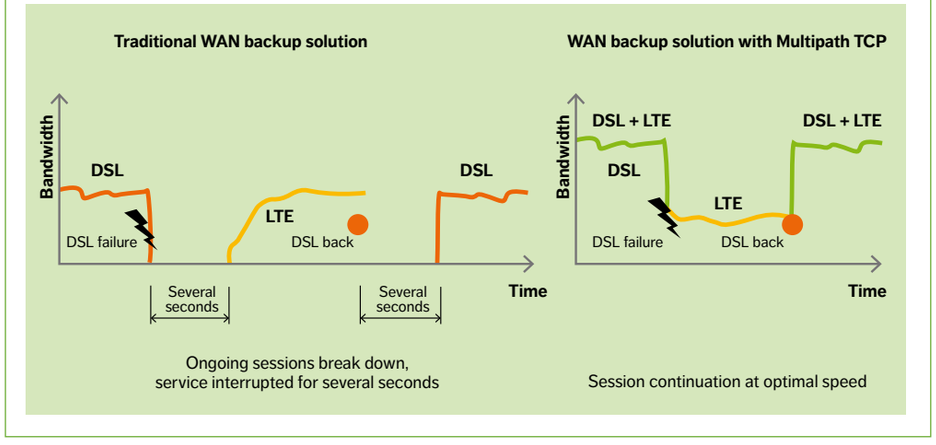


Figure 4 Improved connection resiliency with Multipath TCP

an example of how Multipath TCP adds together the bandwidth of DSL and LTE. This is equally valid for the LTE + Wi-Fi scenario depicted in the bottom part of Figure 2.

The bandwidth aggregation features of Multipath TCP apply to both downlink and uplink directions. As a result, Multipath TCP also helps to improve uplink speeds, which are only a fraction of the downlink speed in existing (asymmetric) DSL consumer services. For instance, the uplink speed over a 6Mbps asymmetric DSL connection is usually below 1Mbps. Aggregating DSL and LTE makes it possible to boost the uplink speed to 10Mbps and more.

Examples of services that would benefit from the Multipath TCP bandwidth aggregation are:

- A user watching HDTV (high definition TV) over a DSL access connection that is not capable of providing enough bandwidth – Multipath TCP can be used to schedule surplus traffic over LTE (particularly useful for the downlink).
- A user uploading documents or photos to a server – when the DSL uplink capacity is exceeded, Multipath TCP can add LTE capacity for quicker upload.

Improving reliability

In the context of access aggregation, reliability refers to the ability to maintain data exchange within a session, even if one or several access links become unavailable. *Figure 4* compares the behavior of a traditional WAN backup solution with that of a solution based on Multipath TCP. Traditional solutions cannot react quickly to the disappearance and reappearance of access links. Whenever a link disappears, sessions break and need to be reestablished, which can lead to data loss and the need for human intervention.

Multipath TCP is able to react more quickly to access links disappearing and reappearing. And as long as at least one access link is up and running, a Multipath TCP enabled session will continue without interruption – albeit at a lower bitrate. Likewise, if an access link reappears, the bitrate goes up. The connection always runs at an optimal speed in relation to the availability of the links involved.

Achieving seamless connectivity

The concept of seamless connectivity is related to reliability, referring more specifically to the ability of Multipath TCP to switch from one access to another without having any impact on the application. A typical use case would be a session started over Wi-Fi. If the mobile device leaves Wi-Fi coverage and enters mobile broadband coverage, the session will break and need to be reestablished. This can be quite annoying and time consuming for the user, especially if two-factor authentication is involved. With Multipath TCP, the session does not get interrupted due to the change of access.

Changing from one access to another can also be triggered by service provider policies. For example, a service provider could have a policy to use LTE by default, but move some traffic to Wi-Fi when there is good coverage and available capacity. Or, alternatively, the service provider could set a policy where Wi-Fi is used by default and LTE is used to provide wide-area coverage. In all cases, the use of Multipath TCP prevents sessions from being interrupted if and when access systems change.

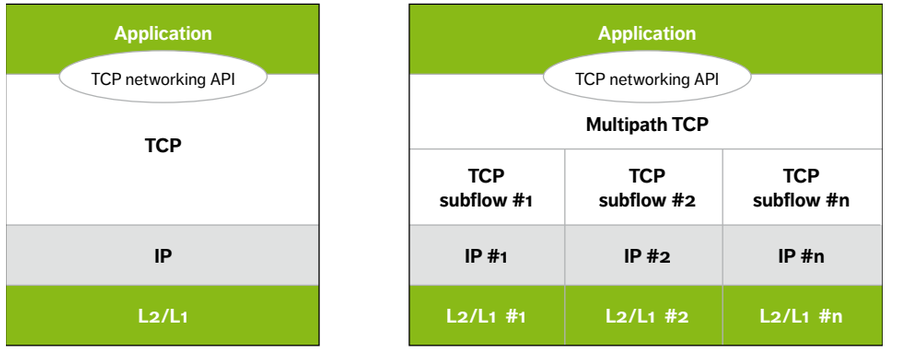


Figure 5 Protocol stack for TCP and Multipath TCP

How Multipath TCP works

TCP [3] is one of the main protocols in the IP suite, providing a reliable means of communication between two endpoints. Once a TCP connection has been set up, both endpoints can send a data stream to each other. TCP is designed to cope with data that is damaged, lost, duplicated or delivered out of order. Furthermore, it provides a means to perform flow control. Upon receiving data, the receiver sends an acknowledgment (ACK) back to the sender. Such an ACK contains a "window," which indicates the maximum number of bytes the sender is allowed to transmit before receiving further permission. This way, the receiver controls the amount of data transferred by the sender. Finally, the receipt or non-receipt of ACKs guides the TCP Congestion Control Algorithm (CCA) to determine the pace at which data may be sent.

Today, many endpoints have multiple data communication interfaces and therefore multiple IP addresses. For example, a laptop is often equipped with both a wired and a wireless interface, and a smartphone often has the capability to use multiple wireless communication technologies. Using regular TCP, these devices are capable of establishing multiple simultaneous TCP connections, with each connection tied to one specific IP interface. In other words, each TCP connection is bound to a single path defined by the IP addresses of the connection's endpoints. Note,

however, that a path is defined here in terms of endpoint identifiers; it is not the same as the route that individual packets take on their way from one endpoint to the other.

Multipath TCP [2] is a set of extensions to standard TCP that allows connections to use multiple paths simultaneously. Multiple regular TCP connections, also known as subflows, are aggregated into a single Multipath TCP connection. *Figure 5* compares the protocols stack of regular TCP with that of Multipath TCP.

In regular TCP, an application initiates communication by opening a connection via an application programming interface (API) provided by the operating system. The TCP layer communicates in its turn with the IP layer. In Multipath TCP, the TCP layer has been extended. Upwards, the Multipath TCP layer exposes an interface that is perceived as regular TCP by the application. Downwards, the Multipath TCP layer may set up multiple regular TCP connections. These may be bound to different IP layers. In *Figure 5*, the host is equipped with multiple data communication interfaces. Each one is associated with its own IP address. The Multipath TCP layer aggregates the multiple TCP connections into a single Multipath TCP connection. The application does not need to be aware of which protocol stack is used.

Figure 6 shows an example of how a Multipath TCP connection can be established. It starts with the setup of a first subflow (steps 2-4). These steps consist of a three-way handshake, similar to the process in regular TCP. The only difference for Multipath TCP is that an MP_CAPABLE option is used in the TCP header. With this option, the device indicates to its peer that it is Multipath TCP capable and wants to use it (step 2). If the peer is also able to use Multipath TCP, it replies with a similar capability indication (step 3). As part of the three-way handshake, the endpoints also exchange security keys. After setting up the first subflow, both endpoints can exchange data over the connection (steps 6-7).

Once a Multipath TCP connection has been established, each endpoint may initiate the setup of an additional subflow. In the example shown in *Figure 6*, the device has two network interfaces. Each interface is associated with its own IP address. Here, the device takes the initiative to establish a second subflow via its second interface. Again, a three-way handshake is used to achieve this. But this time the option MP_JOIN is used to indicate that this is a new subflow that is to be joined to an existing Multipath TCP connection. A token (step 9), derived from the earlier received key (step 3), is used to correctly bind the two subflows. Additional authentication information is also exchanged to ensure the authenticity of both endpoints.

Once the new subflow has been established, both endpoints can use it to send and receive data. In our example, the device sends data to its peer (step 14). Note that the device needs to take an active decision regarding which subflow to use (step 13). How this decision is made is not defined in the standard, which gives the designer the freedom to implement the scheduling policy that is most appropriate for each case.

Subflows may come and go for various reasons, such as connectivity problems. To ensure reliable, in-order delivery to the application, Multipath TCP uses a data sequence number that is carried in a Data Sequence Signal option (steps 6-7 and 14-15). Aside from ensuring in-order delivery, this number can be used in combination with the sequence numbers used by regular TCP at subflow level to execute retransmissions on different subflows, if needed. Multipath TCP can also

User space

In computer design, a distinction is made between kernel space and user space. Kernel space is where the operating system code runs – hardware device drivers, memory management and protocol stacks, for example. User space is where ordinary programs run. In designing our Multipath TCP solution, we chose to place a protocol stack (MPTCP) in user space rather than in kernel space. This results in faster packet processing, because packets don't need to travel from kernel space to user space. Instead, they go directly from the hardware interface to user space.

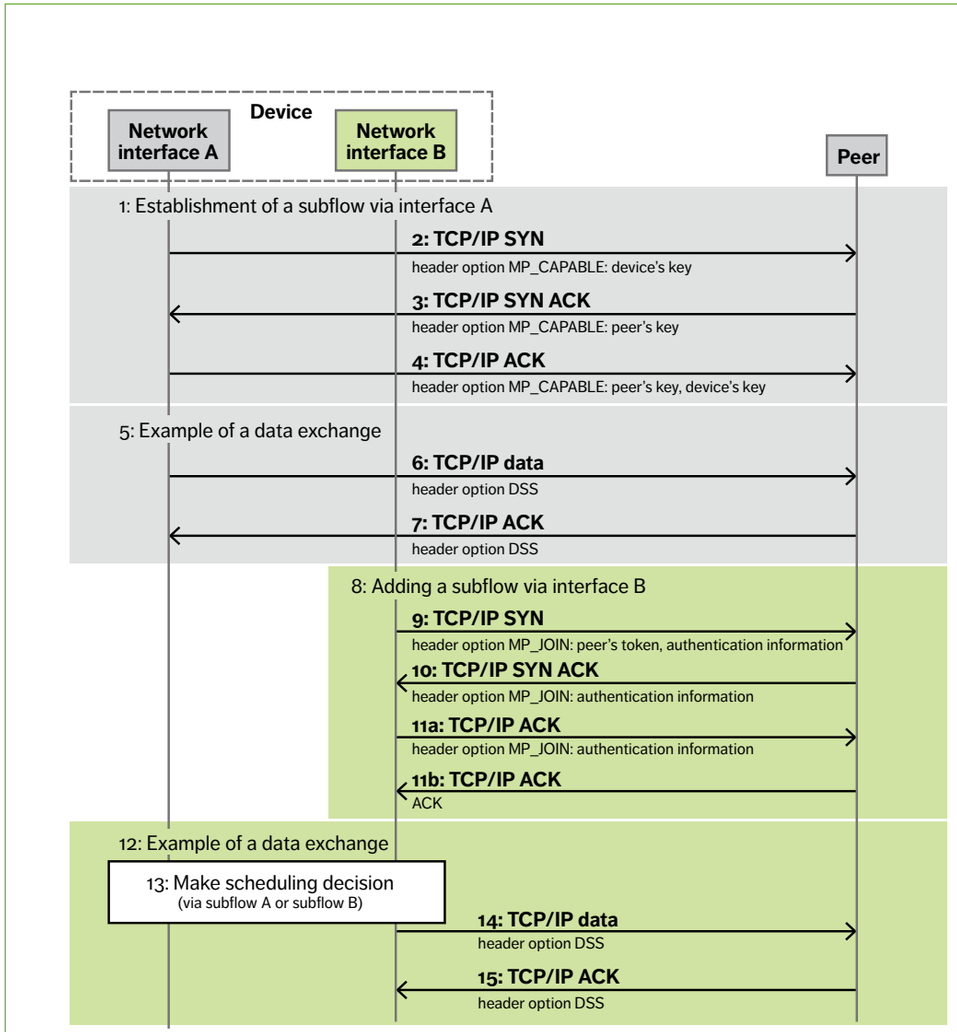


Figure 6 Establishment of a Multipath TCP connection

> ERICSSON IS PARTNERING WITH CPE VENDORS AND CHIPSET MANUFACTURERS SUCH AS INTEL TO ENSURE EFFICIENT IMPLEMENTATION OF THE MULTIPATH TCP CPE PROXY

synchronize congestion control over subflows in order to avoid unfairness to single-path users [5].

An additional benefit of Multipath TCP is that it can be introduced incrementally. In particular, if the receiver of the first subflow's TCP SYN does not support Multipath TCP, it will simply discard the capability option. It will reply with a TCP SYN ACK, but without adding the MP_CAPABLE option, and the connection will be made with standard TCP.

The proxy-based approach to Multipath TCP access aggregation

Proxies make it possible to achieve the benefits of Multipath TCP for access aggregation without requiring Multipath TCP support in all end devices and internet servers. An additional benefit of proxies is that they give the service provider control over the scheduling of the traffic. In this way, service providers can ensure that the available access alternatives are used in the most efficient and cost-effective way. The use of proxies has already been recognized by the industry, and work has been done and published by the Broadband Forum defining the architecture [6]. Ericsson is contributing actively to this work.

Figure 7 provides a high-level overview of the proxy-based approach to Multipath TCP access aggregation. There are two proxies involved: a network proxy and a customer premises equipment (CPE) proxy. The network proxy is located in the service provider's network and converts TCP sessions from internet servers into Multipath TCP sessions that operate across multiple access networks. Similarly, the CPE proxy converts a Multipath TCP session with the network proxy back into a TCP session.

End devices with built-in Multipath TCP support could also connect directly to the network proxy. There are already some smartphones on the market with built-in Multipath TCP support that can be used to aggregate LTE and Wi-Fi. Ericsson has run tests that prove the feasibility of this setup in public LTE and Wi-Fi networks.

The proxies can be used to enhance standard Multipath TCP via additional traffic-steering capabilities that are optimized for the specific application scenario. For instance, a service provider might want to ensure that the DSL pipe is filled first before using the scarcer LTE bandwidth. This traffic-steering approach is often referred to as a cheapest-link-first policy. Service providers might also want to define policies to prevent or allow the use of heterogeneous access for specific services, or to force selected services to use only one of the available access links. All of this is possible with Multipath TCP, as the IETF standard does not prescribe a specific traffic-steering method.

In an implementation, the optional CPE proxy will be integrated in a CPE such as a home or office router. This setup can be used in a residential or enterprise setting, and when it is in place, all devices connecting to the router will receive a faster and more reliable internet connection. Traffic steering can also be applied at the CPE proxy level to control the traffic in the uplink direction.

Ericsson is partnering with CPE vendors and chipset manufacturers such as Intel to ensure efficient implementation of the Multipath TCP CPE proxy. We also offer a reference design and a test lab environment for CPE vendors.

Carrier-grade Multipath TCP proxy implementation

One important requirement for a Multipath TCP proxy in the service provider network is the ability to support a high-performance, carrier-grade IP solution for traffic aggregation. *Figure 8* illustrates how Ericsson's solution can be used as a Multipath TCP network proxy, which can be deployed in either a virtualized or non-virtualized environment.

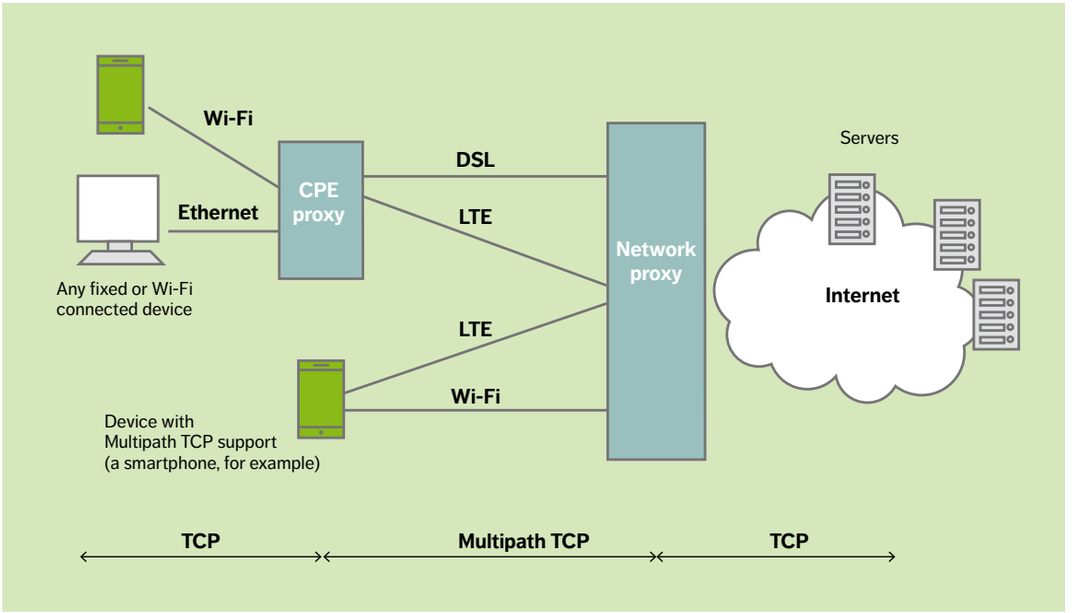


Figure 7 Proxy-based approach for Multipath TCP access aggregation

All components – including Multipath TCP functionality – are implemented in user space[7] to meet the capacity requirements. The TCP traffic can be accessed directly from hardware using a Data Plane Development Kit (DPDK)[8]. The packet distribution function is responsible for sending traffic to the Multipath TCP protocol stack, located in the user space on one or several central processing unit (CPU) cores.

The Ericsson solution implements Multipath TCP functionality as specified by the IETF [2], combined with a specifically designed TCP CCA called TCP RNA (Radio Network Aware). TCP RNA is designed to utilize the mobile ran in an optimal way, and solves the equations for the correct congestion window by using measurements of the speed of the arriving TCP ACKs in conjunction with reactions of lost TCP segments. The benefits of TCP RNA are:

- maximum utilization of available bandwidth for both uplink and downlink
- reduced retransmissions using traffic shaping
- controllable latency
- avoiding bufferbloat.

This solution is highly configurable and can be tailored to support multiple Multipath TCP use cases per access network. The traffic-steering settings are policy driven. One configuration example is to send Multipath TCP traffic on one preferred subflow, such as the DSL link. When the DSL link has reached its limit, any surplus Multipath TCP traffic will be sent on another subflow – most commonly the LTE link.

Another configuration example aims to optimize radio usage on a system-wide level. If Multipath TCP traffic is sharing radio spectrum with other non-Multipath TCP

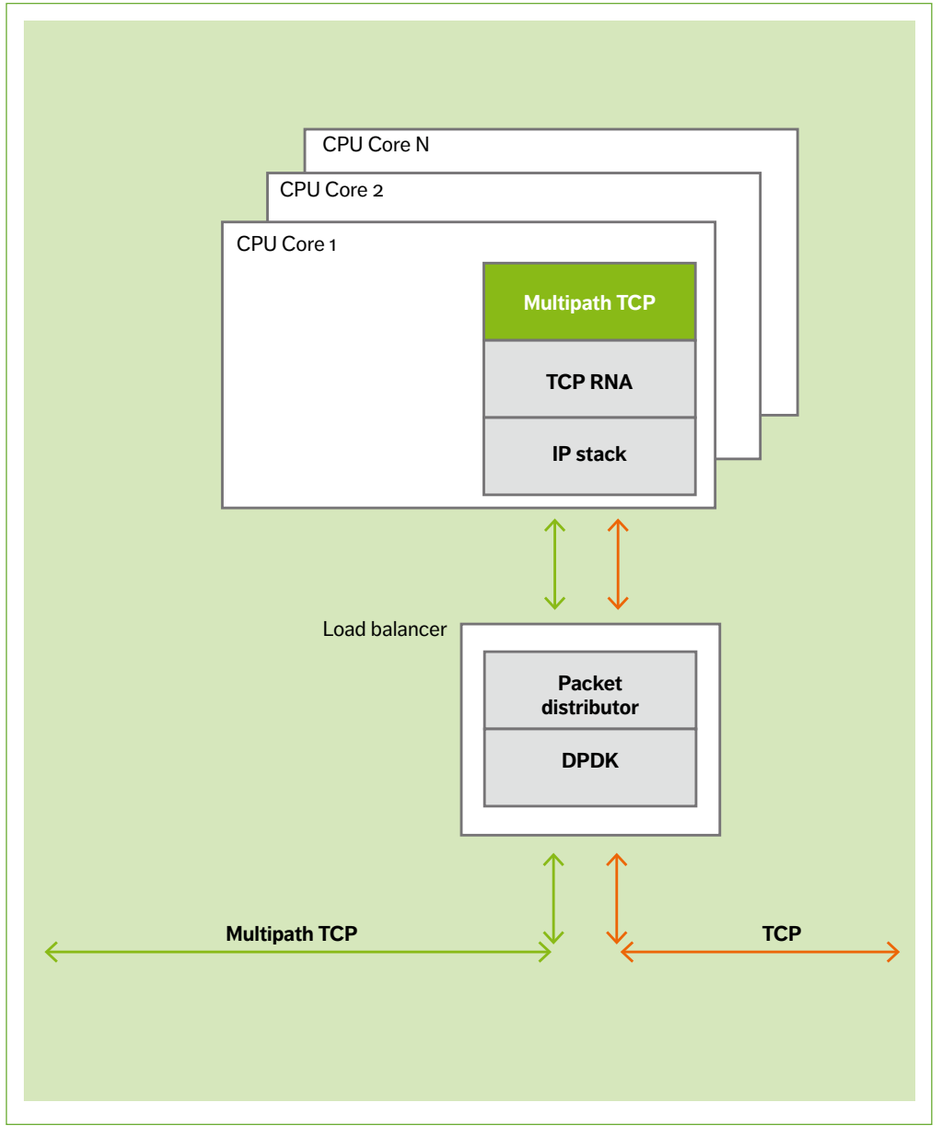


Figure 8 Multipath TCP network proxy

traffic – from LTE-only mobile phones, for example – it might be preferable to avoid excessive use of the LTE link from Multipath TCP traffic. This can be achieved by configuring the TCP RNA for the LTE link to behave like background delivery. The result is that Multipath TCP traffic will back off when TCP RNA detects that the cell is congested, in favor of LTE-only traffic.

At times, it might be desirable to configure Multipath TCP for maximum throughput – when combining LTE with Wi-Fi access for fast file download, for example. In such a scenario, the solution can be configured to use round-trip-time-based (RTT-based) traffic steering. Such traffic steering is achieved by sending data over the subflow with the lowest RTT. If that link reaches its capacity limit and there is more data to send, the rest of the data is sent over the other subflow. If one subflow can handle all the data, only the link with the lowest RTT will be used.

Conclusion

Access aggregation is a viable option for service providers to boost bandwidth across the last mile in areas where it is too costly to increase the capacity of legacy access. Typical access aggregation scenarios are the combination of DSL with LTE or the combination of LTE with Wi-Fi. Multipath TCP, as specified by the IETF, is ideal for access aggregation in the last mile, as it is able to boost bandwidth significantly, while simultaneously increasing reliability and ensuring seamless connectivity.

Multipath TCP comes as a set of extensions to standard TCP. It leverages all of the benefits of TCP such as fairness, flow control and reliability, as well as allowing the use of multiple paths through a network simultaneously. Multipath TCP proxies allow service providers to use Multipath TCP for access aggregation without the need for end devices and internet servers to be aware of it.

Ericsson has created a Multipath TCP proxy that is tailored to the specific needs of service providers. It is carrier-grade, optimized for high traffic throughput and allows service providers to implement traffic-steering policies for the use of available access networks in the most cost-effective and efficient way. ✱

References:

1. 3GPP TS 23.402, Architecture enhancements for non-3GPP accesses, available at: www.3gpp.org/DynaReport/23402.htm
2. IETF RFC 6824, TCP Extensions for Multipath Operation with Multiple Addresses, available at: https://tools.ietf.org/html/rfc6824
3. IETF RFC 793, Transmission Control Protocol, available at: https://tools.ietf.org/html/rfc793
4. Linux Kernel Multipath TCP Project, available at: http://www.multipath-tcp.org/
5. IETF RFC 6356, Coupled Congestion Control for Multipath Transport Protocols, available at: https://tools.ietf.org/html/rfc6356
6. Broadband Forum, Hybrid Access Broadband Network Architecture (TR-348), available at: https://www.broadband-forum.org/technical/download/TR-348.pdf
7. Jonathan Corbet, Alessandro Rubini, Greg Kroah-Hartman, Linux Device Drivers, 3rd Edition. Nutshell Handbooks, 2005.
8. DPDK – Data Plane Development Kit, available at: https://en.wikipedia.org/wiki/Data_Plane_Development_Kit

THE AUTHORS

Robert Skog

◆ is a senior expert in the field of media delivery. After earning an M.Sc. in electrical engineering from KTH Royal Institute of Technology in Stockholm in 1989, he joined Ericsson's two-year trainee program for system engineers. Since

then, he has mainly worked in the service layer and media delivery areas, with everything from the first WAP solutions to today's advanced media delivery solutions. In 2005, Skog won Ericsson's prestigious Inventor of the Year Award.

Dinand Roeland

◆ joined Ericsson in 2000 as a systems manager for core network products. At Ericsson Research since 2007, he is currently a senior specialist in core network architectures and features. He has been a key contributor to the

standardization of multi-access support in the GPP EPC architecture, especially in Wi-Fi. Roeland holds an M.Sc. cum laude in computer architecture from the University of Groningen, the Netherlands.

Jaume Rius i Riu

◆ joined Ericsson in 2004 and has been principal

researcher in connectivity architectures at Ericsson Research since 2014. His work focuses mainly on the standardization of fixed-mobile convergence networking technologies (hybrid access, mobile transport, converged policy control, IPv6, network controlled Wi-Fi and so on) and proof of concept development. Rius i Riu holds an M.Sc. in physics from UAB Autonomic University of Barcelona and a Ph.D. in experimental physics from KTH Royal Institute of Technology in Stockholm.

Uwe Horn

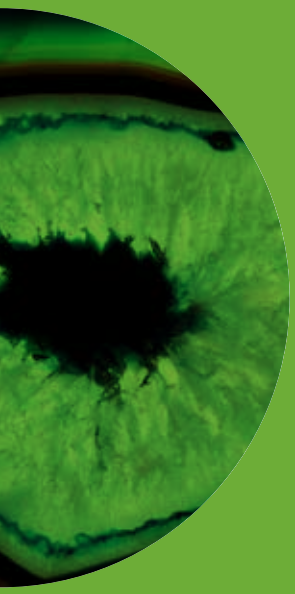
◆ is a solutions director within Ericsson's Global

Customer Unit Vodafone. He has worked in the telecommunications industry for more than 15 years and held various positions in R&D, consulting, marketing and sales. For the past 10 years, he has worked closely with Tier-1 service providers to develop new solutions based on the latest technologies. Horn holds a Ph.D. in telecommunication

engineering from Friedrich-Alexander-Universität Erlangen-Nürnberg in Germany and a diploma in computer science from the University of Bonn, Germany.

Michael Eriksson

◆ is a senior researcher at Ericsson Research. During most of his more than 20 years with Ericsson, his research has focused on the areas of computer science and networking. His current focus is on the design and implementation of advanced networking prototypes. Eriksson holds an M.Sc. in electrical engineering from KTH Royal Institute of Technology in Stockholm.



ISSN 0014-0171
284 23-3303 | Uen
Edita Bobergs, Stockholm

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