ETSI White Paper No. 17

Next Generation Protocols – Market Drivers and Key Scenarios

First edition – May 2016

ISBN No. 979-10-92620-11-5

Editors: Andy Sutton, Richard Li

ETSI (European Telecommunications Standards Institute)
06921 Sophia Antipolis CEDEX, France
Tel +33 4 92 94 42 00
info@etsi.org
www.etsi.org

About the authors

Andy Sutton

Editor and Rapporteur, BT & 5GIC (University of Surrey) andy.sutton@ee.co.uk

Richard Li

Editor, Chief Scientist, Huawei renwei.li@huawei.com

Padma Pillay Esnault

Contributor, Distinguished Engineer, Huawei Technologies, padma@huawei.com

Gerry Foster

Contributor, Chief Architect, University of Surrey, g.foster@SURREY.AC.UK

Lin Han

Contributor, Principal Engineer, Huawei Technologies, lin.han@huawei.com

Eduard Grasa Gras

Contributor, RINA research line leader, Fundacio i2CAT, eduard.grasa@I2CAT.NET

Mark Shepherd

Contributor, Director, Tencastle Limited, mark@tencastle.com

Jiang Sheng

Contributor, Senior Research Scientist, Huawei Technologies, jiangsheng@huawei.com

Kiran Makhijani

Contributor, Senior Research Scientist, Huawei Technologies, kiran.makhijani@huawei.com

John Grant

Contributor, Partner, Nine Tiles, Cambridge, j@ninetiles.com

Mojan Mohajer

Contributor, Senior Manager & S/W Architect, U-Blox, mojan.mohajer@U-BLOX.COM

Diego R Lopez

Contributor, Senior Technical Expert, Telefonica, diego.r.lopez@telefonica.com

John Day

Contributor, Senior Lecturer, Boston University, jeanjour@comcast.net

Yue, Chenhao (philips), Yinyue (Michael)

Contributors, Senior Research Scientists, Huawei Technologies

Contents

List of Tables

List of Figures

Executive Summary

Many existing communications systems have adopted the TCP/IP protocol suite for networking and inter-networking, but increasingly find that these protocols do not meet their demands as well as they were expected to. Over time, there have been incremental improvements, often targeted at specific issues which they did not always adequately resolve. ETSI ISG NGP has been set up to consider, in the light of experience with the IP-based protocols used in fixed and wireless networks, what would be the best protocol architecture for the next generation of communication systems. Its vision is a much more efficient system that is far more attentive to user demand and responsiveness - whether “the user” is a human or billions of things - both for access to services on the Internet and for conveying signals such as audio, video and tactile feedback.

Standards bodies such as IETF and 3GPP have been the driving force behind the success of the Internet and its integration with mobile communications; however, such organizations tend to solve problems in a segmented manner, focusing on a specific protocol layer or service requirement. NGP, on the other hand, emphasizes a holistic approach and broader scope across various aspects of the current network functions and operations.

In this evolution, we will need to be mindful that during the existence of the TCP/IP protocol suite a mobile revolution has happened such that at least one end of any communications path across the network is mobile in the majority of cases. In modern converged mobile networks, multiple access technologies are often used simultaneously by the same user. There was a time when the cellular radio access technology incurred more delay than the core network, but with significant advances in recent years, now the end to end delay is roughly equal for both access and core. 5G research aims to further reduce the radio interface and access network delay such that other factors will contribute the bulk of the latency budget in a future 5G network. New network and protocol architectures will help reduce the overall end to end latency and enable new products and services. This applies equally to fixed networks; however, the 5G timeline is an interesting opportunity for implementing NGP and interworking with TCP/IP as a first step.

There is a need for ETSI NGP to stimulate the development of a network protocol architecture that is fit for the next generation of communications systems and meets the throughput, delay and mobility requirements of our growing service needs. This white paper reviews the fundamental limitations of TCP/IP and associated protocols (such as 3GPP GTP) to enable a wider understanding and appreciation of the issues and gives a preliminary indication of ways in which the new protocols might address them.

The driving vision of ETSI NGP is a considerably more efficient and gradually evolving Internet that is far more attentive to service and traffic demands while enhancing efficiency and lowering the total cost of ownership for network operators. Technology requirements from different market segments, such as IoT, high and ultra-high definition video, and 5G networks and services, offer initial scenarios where the next generation of protocols should significantly simplify solutions. NGP assimilates a diverse set of requirements from these sectors/market segments as well as from a range of different network operations within the global ICT sector. Case studies in LTE-mobile networks, industry 4.0 and multiple Packet Data Network gateways in 4G are used as reference frameworks to highlight the benefits of state-of-the-art NGP solutions.

ETSI ISG NGP is a forum for all interested parties to be able to contribute to its overall goals by sharing research and results from trials and developments in such a way that a wider audience can be informed of the future challenges and discuss potential solutions.

Introduction

Most modern digital communications systems implement the TCP/IP protocol suite as a means of providing network connectivity and transmission of applications and services for end users. TCP/IP has been in use for many years and over that period has evolved many times to address particular issues at particular levels within the protocol hierarchy. Examples include the introduction of IPv6 to manage address exhaustion which had occurred with IPv4 and the introduction of HTTPS to enhance security and privacy.

It is the view of many within the telecommunications and networking community that it's now necessary to fundamentally review the future direction of TCP/IP and to question how information communications protocols should evolve to meet the needs of the 21st century.

The TCP/IP protocol suite has undoubtedly enabled the evolution of connected computing and many other developments since its invention; however, by evolving communications and networking protocols in a timely manner, we can ensure the on-going success of our increasingly connected world. The telecommunications industry has reached a point where forward leaps in the technology of the local access networks (such as LTE-A, G.Fast, DOCSIS 3.1 and 5G) will not deliver their full potential unless, in parallel, the entire information communications protocol stacks evolve more holistically.

ETSI ISG NGP aims to review the future landscape of Internet Protocols, identify and document future requirements, and trigger coordinated follow up activities. The prize is to remove the anchor drag of historic sub-optimal IP protocol stacks and allow all the next generation networks to inter-work in a way that accelerates a post-2020 connected world unencumbered by past developments.

Scope

The scope of this ETSI White Paper is to review the drivers for an evolution in networking protocols and associated network architectures for future information communications networking protocols. It provides some background and context along with exploring the rationale behind the work of ETSI NGP. The paper reviews each fundamental issue and explains the limitations to enable a wider understanding and appreciation of the issues to be addressed. Additionally, the paper explores the requirements of 5G and other communications systems and how they could be realized far more efficiently through a profound transformation of network protocols, thereby offering improved user plane throughput, reduced latency, reduced energy consumption in both end-devices and networks, and optimized congestion management. The paper also discusses issues associated with inefficient overlay tunnels which are used to manage requirements such as mobility, device and network management, and security.

Objectives of ISG NGP

ETSI ISG NGP identifies the requirements for next generation communications protocols and, where appropriate, network architectures, from all interested user and industry groups. The goal is to formulate a series of ETSI Group Specifications which focus on documenting the state of the art in Future Internet research and proposals; recap what issues those proposals try to resolve and which requirements they try to meet; and identify where the shortcomings exist or are likely to emerge. The scope also includes describing the requirements of evolving access network technologies and identifying which communities appear to be already working on such new requirements and could be important stakeholders to engage or coordinate with and identify any gaps.

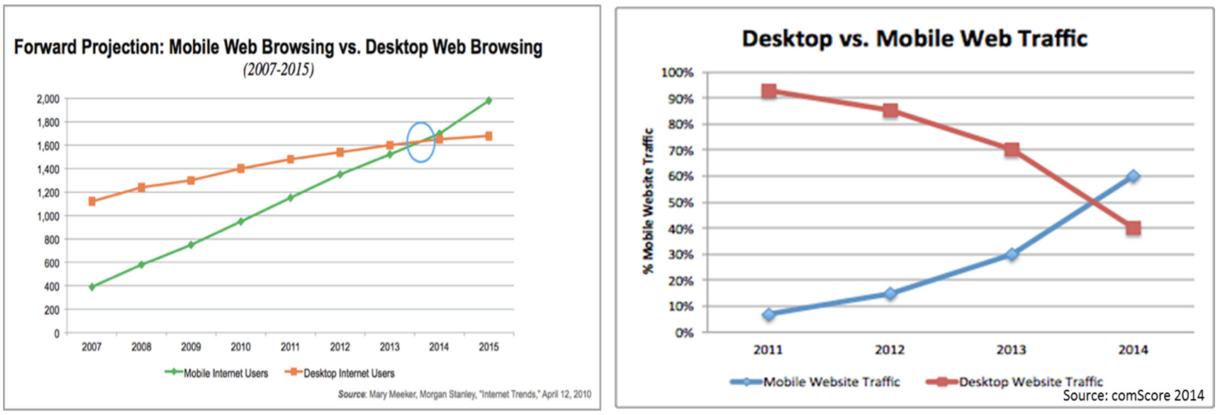
Topics of interest include the following:

- Addressing in the Internet architecture
- Mobility based scenarios in mobile networks
- Security, identity, location and authorization
- High throughput transport with low latency
- Requirements from the Internet of Things
- Requirements from ultra-low latency content distribution use cases from different sectors (e.g. automotive, high definition video)
- Requirements from network operators (e.g. challenges with end-to-end encrypted content)
- Requirements from e-commerce
- Requirements for increased energy efficiency within the global ICT sector

The deliverables will include a summary of relevant technologies, architectures and protocols under research, including an assessment of their maturity and practicality for start of implementation in the 2020 timeframe. Requirements from key technology areas (e.g. 5G and IoT) along with a business case or cases (e.g. the cost benefits of increasing the efficiency of applications and protocols over cellular radio spectrum) will justify the consequential need for follow-on standardization effort in other organizations. Additionally, a set of goals and an action plan to engage other standards groups, such as IETF, IEEE, 3GPP etc., will be developed so that parallel and concerted standardization action can take place in the most appropriate standards group.

Market Drivers

The instantaneous communication and hyper-connectivity in the Internet age has transformed the way our society interacts. The following characteristics are driving fundamental changes in the current Internet.


Forward Projection: Mobile Web Browsing vs. Desktop Web Browsing

(2007-2015)

Source: Mary Meeker, Morgan Stanley, "Internet Trends," April 12, 2010

Desktop vs. Mobile Web Traffic

Source: comScore 2014

Figure 1: Mobile vs. Desktop - (a) number of users and (b) Internet traffic volume

- **Mobility Growth**: Communication behaviour is swiftly shifting from PC based fixed computing to smartphone and tablet based mobile computing. Mobile data traffic grew 4,000-fold over the past 10 years and almost 400-million-fold over the past 15 years [1]. According to Figure 1, mobile network usage (in terms of both users and traffic) exceeded desktop usage in 2014 and will continue to grow.
- **IoT of all kinds**: It is further projected that movable devices, such as driverless cars, wearable sensors and mobile robots will play bigger roles in our connected daily lives. The traditional Internet was not engineered to provide access to network devices that are not stationary, or frequently change location, or add or drop connections.
- **Scale of Connected Things**: Every day more and more physical or virtualized end-nodes are being added to the Internet. With the deployments of IoT projected to increase significantly, these numbers present a daunting challenge to stability and scalability of the Internet by demanding more IP addresses and space in routeing tables.
- **High Throughput Super Media**: Streaming media dominates the Internet traffic today. The transmission rates for emerging video technologies such as 4K, 8K ultra-high definition (UHD), Augmented Reality (AR) and Virtual Reality (VR) will stress the network throughput for a consistent immersive experience.

Technical innovations in sectors such as medicine, agriculture, manufacturing, etc., driven by computational power and digital tool advancements, will give rise to newer categories of application ecosystems. For example, industrial and infrastructure automation will generate a diverse set of sensors and IoT devices of massive scale that must be interconnected, while other scenarios, such as transportation tracking systems, will have high mobility requirements but less variation in sensors. In contrast, a connected self-driving vehicle requires dynamic connections across a high number of vehicles and multiple sensors with efficient mobility functions. Already in these few use cases, different kinds of communication patterns can be seen, each with a specific demand for latency, mobility, bandwidth and energy sensitivity.

The prevailing network protocols are single-dimensional, built for interconnectivity of topology-centric fixed devices; moving an endpoint involves setting up a new connection and changing the topology, resulting in unpredictable latency and inefficient bandwidth consumption. These challenges present an opportunity to evaluate a multi-dimensional approach that comprises mobility, scale, diverse objects and efficient transport as essential characteristics. The following discussion exposes further issues and limitations with the current architecture.

Network Addressing - Primitive to Progressive

Similar to the postal service, the fundamental purpose of communication networks is to deliver packets from a sender to a receiver. In networks, the "receiver address" could indicate both a network device that is the destination of the packet and the "receiver's name", a logical entity that is expected to receive the packet. The ***logical entity*** could represent a device, computer process, protocol, user, application, etc. In the IP-based Internet of today, the IP address (v4 or v6) is used to identify both *address* and *name* of sender or receiver; i.e. the '*address*' is a direct reference but the 'name' is implicit through other pieces of information within the IP header, such as port numbers and protocol types.

The IP address is used to identify two different properties.

1. address of the source and destination (normally a device)
2. logical entity that actually sends or receives the information. This logical entity is associated with a physical device but may be different from the device itself.

With a single layer implying both destination and identity of the information, the OSI [2] model gets extremely simplified; the explosive growth of the Internet itself is a proof of its success. However, it must also be acknowledged that this simple IP addressing system has saturated its own limits, and in its current form the IP-based Internet finds it difficult to meet the expectations of the new generation of applications.

IP Address Challenges

The dual semantics or ambiguity of an IP address is neither effective nor sufficient to satisfy the requirements of near-future scenarios as summarized below:

1. **Scalability**: The growth of the Internet has led to an increased number of multi-homed sites and a reduced amount of aggregation of IP prefixes. As a result, BGP (Border Gateway Protocol) routeing tables keep growing at steady rates (Figure 2) [3] over the past decade and are projected to be saturated soon.
2. **Mobility**: A topology-centric, provider-assigned IP address does not perform well for a moving endpoint and results in non-trivial handoff schemes. This is by far the biggest challenge to resolve in the IP address based framework.
3. **Security**: An IP session, characterized by port numbers and protocol type, itself lacks intrinsic security and is a fundamental flaw in the IP stack, causing session hijacks and denial of service attacks.
4. **Non-IP devices**: Many non-IP IoT devices connect to the Internet through dedicated network gateways, which translate non-IP flows to IP flows and vice versa. Not all IoT devices support IP, and applications are generally IP-based servers. In order to efficiently and uniformly communicate with both IP and non-IP devices, a newer flexible addressing model is required.

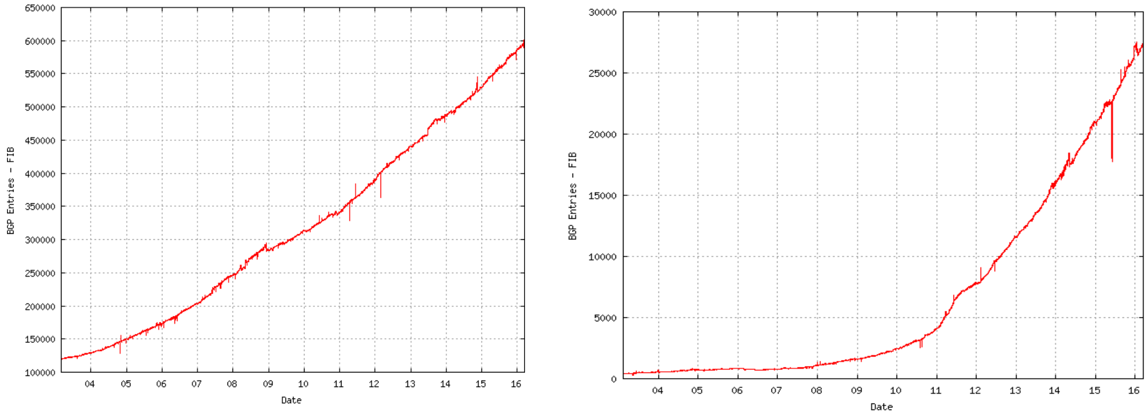


Figure 2: BGP Table sizes (Left IPV4, Right IPV6) source: potaroo

Different Perspectives and Ongoing Investigations on Addressing Systems

Driven by the challenges relating to Internet growth and content distribution, the need to revisit network architectures was already felt in the first decade of the 21st century. Various prominent technical groups such as the Internet Engineering Task Force (IETF), National Science Foundation (NSF), ITU-T and ISO/IEC JTC1/SC6 launched several investigations to study newer architectures.

ISO/IEC JTC1/SC6 in their series of Technical Reports TR 29181 also documented the problems with current networks and the requirements for future networks. TR 29181-1 [4] provides an overview, while TR 29181-3 [5] outlines an architecture which satisfies the requirements that were identified, which are broadly similar to the requirements that have been identified for 5G. In this architecture, packet destinations are identified in control plane messages rather than in user plane packet headers, thus allowing a wide variety of different forms of address to be used.

The Locator/ID Separation Protocol (LISP) [23] was one of the early proposals that came out of discussions at IETF. LISP separates the address through location and logical name through ID. The original protocol represented both location and ID through IP address formats but non-IP formatted IDs are under discussion [6].

While LISP concentrated on extending the IP address space through overlays, another IRTF research group Information Centric Networking (ICN) [7] is investigating direct support of the use of uniquely named location independent data as a core Internet principle. It specifically identifies content as the last entity of communication which is a non-IP object.

Table 1: Limitations with Current Addressing of End entities

Limitations	Expectations
The last logical entity of communication	Not just IP address format Framework for mobility
Format of the Identities	Broader abstraction, flexible representation
Scale	Manageable core, Connected objects

With an objective to solve modern communication address challenges, a redesign of the current network layer (IP) must be evaluated to meet the expectations summarized in Table 1.

Table 2: Analysing Clean Slate Vs Evolvable Addressing in Network Architectures

Options	Challenges	Benefits	Examples
Clean slate (New non IP Network Layer)	Interworking and migration from current architectures.	Comprehensive solutions to overcome limitations identified in IP	ICN
Overlay Solution	Currently ID limited to IP	Fits in today's architecture	LISP
Evolvable Network Layer	Impact on Transport Layer	Easier mobility management Easier authentication Simpler cross-silo applications	-

The inclusion of flexible address formats provides better abstraction and representation of different entities that may be logical or physical, in whole or in part, and require connectivity. With a pragmatic view of not limiting the last logical entity of communication to IP addresses only (as is the case in the TCP/IP stack), a new layer called the ***ID layer*** may be imagined as a possibility.

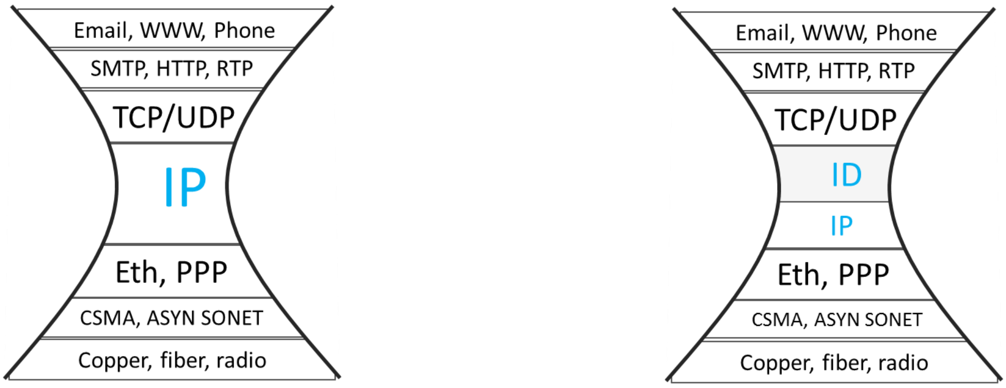


Figure 3: (i) Traditional Network layer, (ii) Evolvable ID Aware Solution Layer

A comparison of such ***Evolvable ID layer*** on top of current network layer (IP) is illustrated in Figure 3 (i) and (ii). While *Evolvable ID layer* could save vast investments in current Internet infrastructure, setting off faster adoption of new technologies, ETSI NGP could study ID awareness further with respect to

a) An efficient ID aware data plane stack.

b) The management of the ID itself in terms of its scope, assignment, allocation and lifetime.

c) ID aware routeing protocols.

d) The adoption of ID in transport layer protocols, such as TCP or UDP (changes to support both IP and non-IP IDs).

Ubiquitous and Continuous Connectivity - Mobility

Mobile devices are increasingly the preferred platform to connect to the Internet. This trend is accelerating further as more and more services and applications are progressively designed with emphasis on handhelds, tablets, and human-friendly voice & touch interfaces. The connectivity requirements vary across different mobility-aware applications, and it is not possible to represent them in traditional network protocol suites. For example, it is anticipated that 5G will support connectivity for everything, everywhere and at any time, by maintaining user sessions across different radio or fixed access technologies. This is not possible today and with a view of realizing 5G, mobility is particularly challenging for certain scenarios with strict latency and bandwidth constraints as shown in Table 3.

Table 3: 5G Requirements and projected applications

Use cases	Attributes	Requirements
High resolution media	Speed	1-10Gbps
4k-8k video	Bandwidth	10 times of 4G
Augmented Reality, Virtual Reality, Tactile Network	Latency	Sub millisecond (< 1ms)
Wearables, home and industrial networks, etc.	Scale	Billions of devices
Seamless user experience for critical applications	Session continuity	Ultra-reliability
Ubiquitous connectivity for everyone everywhere – connected vehicles, remote healthcare	Heterogeneous environments	Multi-access technologies such as Wi-Fi, radio, and fixed

5G Networks Mobility Management and Session Continuity

According to the GSMA intelligence report *Understanding 5G Perspectives* [8], the newly enabled applications in 5G: Virtual Reality, Augmented Reality and Tactile Internet sit in the top right-hand zone of the chart (Figure 4). That zone represents a network resource of either low latency, or high bandwidth, or both. While other services identified in the GSMA chart are already viable through legacy networks, newer services that rely on 5G are inherently mobile and dynamic and are expected to work seamlessly across heterogeneous networks. Historically, mobility support has been network centric with operators electing a “disconnect and re-attach” approach to session management for moving devices. This has to change for applications in 5G and beyond, with session persistence supported in the network and a mobility-aware stack in end-points.

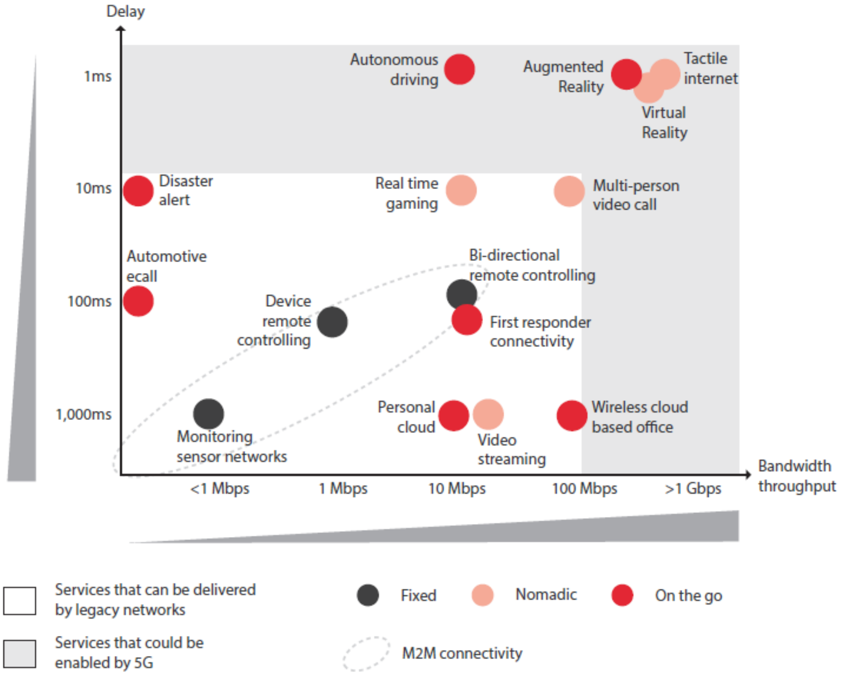


Figure 4: Bandwidth and Latency constraints of 5G applications (source: GSMA Intelligence 2014)

Evolution of Mobility

Mobility is best described through participating entities. Referring to Mobile IP nomenclature, in mobility management, there is a moving entity called a mobile node (MN) connected to a remote correspondent node (CN). A home agent (HA) typically is a network gateway through which an MN acquires a home IP address. A foreign agent (FA) is a new location which assigns a new IP address. In the cellular networks context, there is a mobility anchor point, which is the closest network device to which an MN is attached. The mobility is achieved when MN and CN remain connected, as an MN moves from HA to FA network or location.

The mobility management functionality is difficult to plan and implement because

a) the path packets take when an MN moves is not always optimal;

b) most sessions are 5-tuple TCP based connections, and a change in either source or destination IP address is considered a different connection; and

c) buffering schemes through mobile anchor nodes cannot be deployed predictably.

There are various mobility solutions available in both cellular and fixed IP networks since the early 1990s; however, each approach is less than ideal for 5G applications. Mobility solutions and corresponding limitations are listed in

Table 4. Note that for a few solutions in the table, technical short comings may not be obvious but a lack of market adoption over a long period leaves the technology as experimental.

Table 4: IP based mobility solutions and limitations

SDO	Solution	Limitations	Market Proven
IETF	Mobile IPv4[1]	Handover latency, signalling overheads in transition, suboptimal triangular routeing, limited QOS	
IETF	Mobile IPv6[2]	Handover latency, limited awareness of heterogeneity, requires kernel changes, limited security of c/o address	
3GPP	3G/GTP[3]	Tunnel re-creation on move, limited service continuity across GGSN	Yes
3GPP	4G/LTE/GTP[4]	Service continuity is limited within a P-GW	Yes
IETF	Proxy Mobile IPv6 (PMIPv6)[5]	Only for local mobility; service continuity limited within one LMA domain	Yes
IETF	Distributed Mobility Mgmt (DMM)[6]	Triangular routeing for on-going sessions. Optimized for new sessions only. Mobile node needs multiple IP addresses for service continuity crossing DMM-GW. Not a mature solution yet without an RFC	
	MobilityFirst	Experimental, part of academic research program under NSF.	
IETF	LISP[7]	Experimental, ongoing trials through beta-network, waiting for multi-vendor market adoption.	

1: RFC 5944[24], 2: RFC 6725 [25], 3: 3GPP TS 29.060 V6.9.0 [26], 4: 3GPP TS 29.274 V8.1.0 [27] 5: RFC 5213 [28], 6: IETF-DMM [29], RFC 7429[30], 7: RFC 6830 [23]

Holistic Approach to Mobility

To provide a seamless mobility function, as end-users or services move, the connectivity must remain intact. Mobility should be considered with two different view-points. The first, ***Session continuity,*** implies that the IP address is preserved for the lifetime of the data session. The second, ***Service continuity,*** implies an uninterrupted “at par” experience of a service, including the cases where the IP address and/or anchor point changes. Session continuity ensures service continuity automatically. Simply put, a comprehensive mobility solution has three primary characteristics:

1. ***Session Continuity***: seamless connections when the communication endpoint moves.
2. ***Multi Access***: allows mobility across heterogeneous access methods, e.g. when the communication end-point moves from WiFi to an LTE/5G network.
3. ***Optimal Routeing***: allows session hand-offs between two locations through an optimal path.

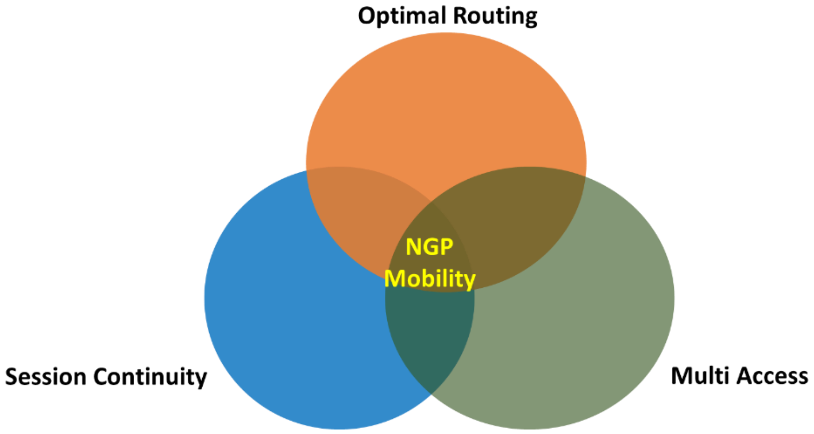


Figure 5: Three Criteria of Mobility

An NGP mobility solution study will cover all three characteristics of mobility shown in Figure 5. Session continuity is a direct requirement for the new connected era, and it is equally important to unify solutions across mobile and fixed networks. With the evolved network layer discussed earlier, a user can move during the lifetime of sessions with ease because the connections are specified by service and identity tuples and not by the IP address.

Security

Security requirements and expectations are changing. If new protocols are to be adopted, they will be driven by the business benefits perceived from 5G use cases. This implies that the security requirements must be supportive of 5G scenarios. The conclusion is that NGP security must perform two functions. First, today’s security challenges must be addressed in an increasingly efficient way to support new bandwidth and latency requirements. But also, new concerns and use cases must be accommodated without compromising core security principles. These new concerns include:

Role in critical infrastructure

It is already the case that communications networks are being added to the list of components which are considered “Critical National Infrastructure” (CNI). However, this underestimates the role they will play. Networks will be vital to almost all of the existing components of CNI. Also, 5G networks will create and underpin entirely new components of CNI, such as Tactile Internet or Remote Monitors in health care and Vehicle Communications in transport. Availability and reliability requirements will be critical to the success of an NGP.

Privacy concerns

There is an increasing public debate about the importance of user privacy and about who should or should not have access to user data. The debate asks which organizations (public or private) are trusted to hold user information and what they are entitled to do with it. This document is not a philosophical or legal one and does not attempt to draw any conclusions from this debate. Instead, the technical requirement is that NGP protocols shall support a range of outcomes from this public debate, i.e. the technology must facilitate privacy where this is required, and access to data where this is required. This implies that security may need to be handled differently depending on the component and information type in question (Personal, Network, Provider, Content, Government, Financial, Utility Control & V2x (…CNI), Personal-IoT).

Internet of Things

These use cases impact security partly through scale: the number of devices to be authenticated will be an order of magnitude larger than at present. The devices will typically be built-in, i.e. without human access (cars, meters, sensors), making it impractical to physically swap identity or security modules. Low-power IoT will have a significant impact: many traditional security techniques require considerable bandwidth (e.g. for handshaking even if not for traffic delivery). Also, connected IoT devices potentially provide a way of bypassing security measures (such as firewalls) protecting other equipment on the same network.

Network optimization

Where security components have been "bolted-on" on top of existing protocols, there can often be inefficiencies with additional proxies and protocol layers removed and re-applied. Such inefficiencies could defeat the core benefits (bandwidth/energy) which form the heart of what NGP is trying to achieve. Network optimization is only possible where carriers are able to understand key meta-data from the traffic they are conveying. A carefully thought-through approach to confidentiality can enable operators to see the information they need without exposing excess information.

Virtualization and isolation

Many infrastructures running future protocols will be based on virtualized architectures. From a security point of view, the key consequence is that it will not be possible to rely on physical separation and therefore there should be an underlying assumption that data at rest and in transit will be visible to other actors (for example, hypervisors will have access to the memory of functions they are hosting; also, network attacks may mean that there will be many compromised components running in the same environment as sensitive functions). Security-critical functions such as key negotiation or key storage will need to be built based on effective, strong isolation, e.g. enforced through hardware roots of trust. As the network gets more virtual, we need a system that not only supports users' secure access and privacy for their data, but also formally logs all instantiations in terms of What When Where Why and Who so that it is possible to audit and trace issues in the first instance and, as we evolve, write security algorithms to monitor malicious behaviour.

Limitations in the TCP/IP protocol suite

The design and use of the TCP/IP protocol suite leads to many reasons behind the high cost and low effectiveness [9] of the current Internet security architecture. Firstly, the TCP/IP protocol suite doesn't identify where security functions such as authentication, access control, integrity, confidentiality, or auditing should belong. Instead, many later or more recent enhancements to protocols (DNSSEC, BGPsec, Transport Layer Security, TCP authentication header, OSPF authentication, etc.) contribute to a large overhead and complexity due to segmented security. It is noted in [10] that as systems become more complex security gets worse. IPsec attempted to secure the network layer, but it either protects only the IP payload or turns IP into a connection-oriented solution (by tunnelling encrypted IP packets as the payload of unprotected IP). Thus, the security functions should be inherent to the architecture.

Second, the Internet addressing architecture lacks naming of content, in particular the lack of application names makes the Internet more vulnerable. The IP layer exposes source and destination addresses to applications, which provides rogue members with the ammunition to mount attacks targeting the network infrastructure or applications connected to the Internet. As an example, a script with 4 lines of code running in a conventional host attacked thousands of printers connected to the Internet simply by 'walking' the IP address space [11]. Therefore, any viable future-proof network architecture must provide support for application names and provide isolation of addresses and name spaces.

Lastly, the TCP/IP based Internet does not define "scope", therefore, it is not clear who connects to whom. Millions of transactions or message exchanges are carried out over a general-purpose Internet without clear scoping semantics. Alternatively, many mechanisms enforce 'scope of connectivity' through firewalls, Access Control Lists (ACLs), VLANs or Virtual Private Networks (VPNs) to name a few, but these are not coordinated together. A carefully thought-through approach to *'scope'* and *'confidentiality'* can enable operators to see the only information they need without exposing other aspects of user data.

Ultra-Low Latency and High Throughput Transport

Transmission Control Protocol (TCP) is the most widely used reliable transport protocol in the Internet. TCP throughput, which is the rate that data is successfully delivered over a TCP connection, is an important metric to measure the quality of a network connection. TCP employs slow start and congestion avoidance mechanisms to ensure reliability and fairness, which in turn limits TCP throughput. It is believed that the TCP throughput can be computed by the formula below,

$$\text{TCP Throughput} \leq \min\left(\text{BW}, \frac{\text{WindowSize}}{\text{RTT}}, \frac{\text{MSS}}{\text{RTT}} \times \frac{\text{C}}{\sqrt{\rho}}\right),$$

where RTT is the Round Trip Time, ρ is the Packet Loss Ratio (number of lost packets divided by number of received packets), BW is the bandwidth, and C is a constant whose value is different for different TCP algorithms. It can be seen that TCP throughput is mainly determined by RTT and packet loss ratio.

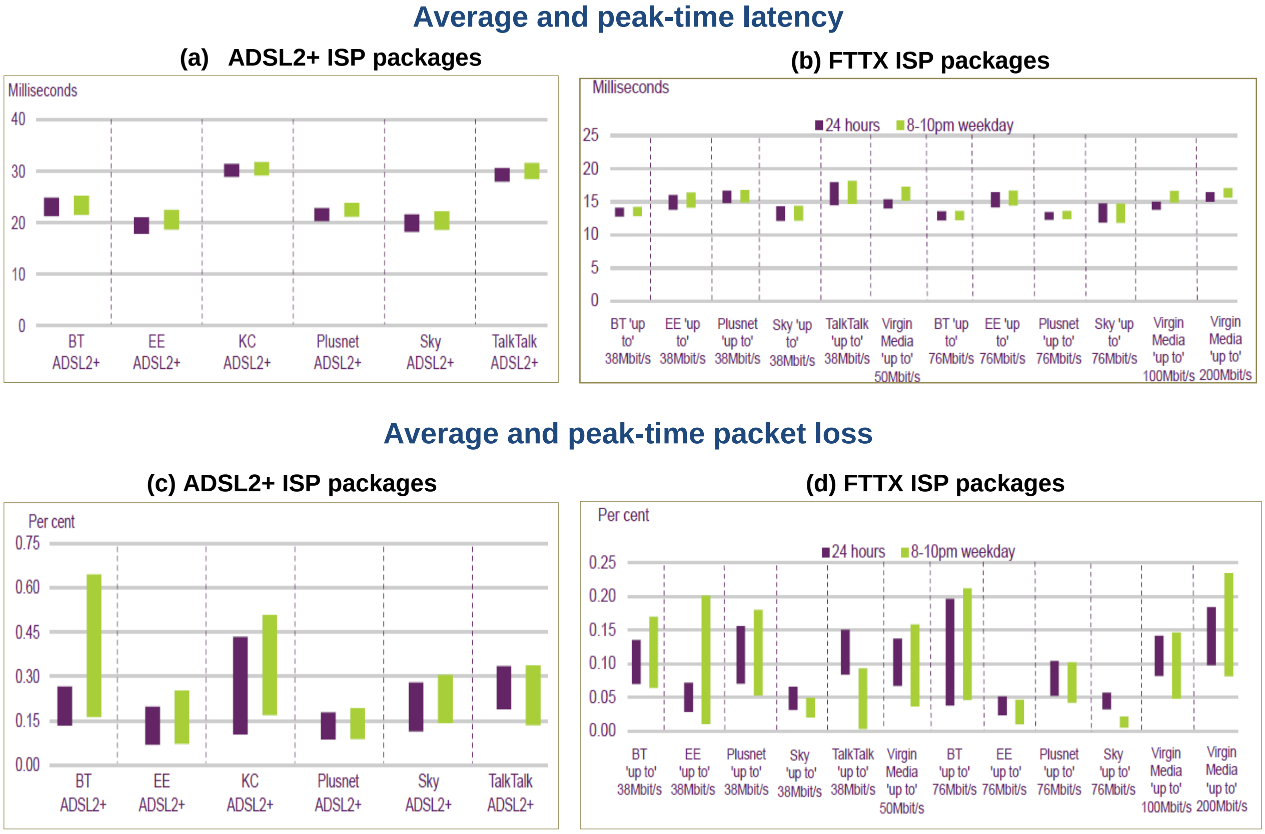
That is, both longer round trip times and higher packet loss ratios result in low TCP throughput which cannot exploit the full capacity or bandwidth of the available media. By minimizing packet loss and RTT, not only are retransmissions prevented but bandwidth utilization is maximized by transmitting more packets to the peer end-point.

Minimizing RTT delays

Typically, RTT delays can be reduced by placing the streaming media servers closer to the users because the content travels over shorter distances and fewer hops. Usually, a scheduling algorithm determines the closest Content Delivery Network (CDN) server that can service a user's request, however, this does not always guarantee optimal performance. Consider a case where actual content is not available on the selected server and the request is redirected to another content server, then the determination of the different CDN adds to the initial latency. In another case, the selected content server may be overloaded and cannot process user requests predictably. Each of these factors represents a CDN cache hit problem, because the user request has to be redirected to a different server which makes the RTT unpredictable and limits the achievable throughput of TCP. Next Generation transport should be able to determine RTT delays more accurately and possibly investigate decentralized content distribution solutions where content is placed much closer to the user.

Link-specific packet loss and latencies

That the TCP throughput discussed earlier varies over different media and access networks (e.g. FTTx, DSL), because packet loss ratios and latencies depend on link types, is supported by the fixed-line broadband performance study [12]. According to this study (Figure 6), the top two charts below show that the latency for FTTx (b) based access network (13ms to 17ms round trip time) is generally lower than the ADSL (a) based access networks (21ms to 31ms). The bottom two charts show that packet loss ratios for FTTX (d) (0.02% - 0.19%) are much lower than for ADSL (d) (0.06% - 0.61%) accesses.


Average and peak-time latency

(a) ADSL2+ ISP packages

(b) FTTX ISP packages

Average and peak-time packet loss

(c) ADSL2+ ISP packages

(d) FTTX ISP packages

Figure 6: FTTX and DSL Latency (RTT) and Packet loss ratios

Thus the impact of the type of link must be accounted for in the computation of estimated throughput. Even with a better access network such as FTTx, the achievable throughput of existing TCP cannot meet the bandwidth requirements of 8K video streaming transmission.

A Case for High Performance Networks

Internet traffic will be dominated by access to dynamic, ever changing live content and context, a transition from the static file or web page access behaviours of just a few years ago. It influences traditional traffic patterns from sudden burst and best-effort service and compels the need for consistent high bandwidth and high quality of experience.

Immersive technologies such as Augmented Reality (AR) and Virtual Reality (VR) further increase the demand for a high performance network. Transmission of a 360-degree AR/VR live stream requires a high bandwidth. Furthermore, interactive media transmission demands a deterministic end-to-end latency which also means that large jitter is unacceptable. The transport layer needs to factor in external parameters and network environment beyond RTT delay and packet loss ratio. Accordingly, newer solutions are needed to maximize bandwidth usage, deterministic computations for RTT delay and packet loss with the goal of enhancing network performance predictably.

Self-Organizing Control & Management Planes

The success of the Internet has made IP-based networks bigger and more complex. The large scale of the networks and diverse set of network operations are extremely demanding for human based management.

Growth in Network Complexity

Transition from physical to cloud-based infrastructures, the sheer volume of active network devices, and the impact of vast geographies, has led to more frequent network changes. A network administrator needs to modify configurations often and in timely manner. On the other hand, manual verification and validation processes are usually slow, painstaking and still error prone. It is reported that most network problems (above 95%) are caused by human misconfiguration.

In the current IP based network systems, only routeing functions may be considered as autonomic. Even that requires manual provisioning of peer neighbours, route policies and other attributes to achieve the desired effect. It results in a rigid network traffic management. Although several network management tools can automate repeatable work through scripts, the overall network operations, control and management functions still require human intelligence and experts with in-depth knowledge of all aspects.

Background of self-configuring and self-optimizing networks (SON)

The SON concept [13] first emerged in release 8 of the 3GPP architecture and is being evaluated much more widely in other communications circles as well. The motivation behind SON was cost and complexity reduction of the OSS operations through automation of optimization of many RF, controller and antenna parameters (in 3G) or base station, MME and eNB (in LTE). There are three important functional components in SON described as self-healing, self-configuring and self-organizing.

In 3G and LTE SON deployment architectures, there are two variants called centralized SON (C-SON) and distributed SON (D-SON). C-SON solutions are based on a central network management system; the optimization algorithm operates above the base stations with a wide area scope and can potentially orchestrate multi-vendor and multi-technology equipment in a radio network. C-SON usually exhibits slower response due to its centralized nature and derives inputs from the RANs to adjust base station configuration. In contrast, D-SON solutions have optimizations that operate at a base station or cell level with a narrow geographical scope. Due to their reduced scope, D-SONs usually respond very fast to changes in network conditions, deriving their input from RAN control or trace information. On the downside, a D-SON architecture is generally vendor-specific and does not operate well where equipment from different vendors comes into play; therefore, a hybrid approach called H-SON is adopted in which C-SON, acting as a supervisory layer, coordinates among multiple instances of D-SONs across a much broader scope and scale.

It is envisioned that a potential expansion and applicability of these concepts to fully automated orchestration of fixed network systems is quite possible and will benefit network manageability.

Autonomous Network operations and control

SON concepts are quite actively deployed but are limited to radio networks. With most mobile networks transformed to an IP based EPS core, many other components in the mobile core as well as the fixed network may benefit from similar orchestration methods. In effect, any NGP-enabled network system must be able to maintain and manage itself autonomously, meeting the variable set of requirements from the network users. In particular, a higher degree of intelligence is expected through:

- **Self-organization**: In order to make complex network systems more manageable they must organize themselves through smaller and simpler agents. A large number of such agents have well defined tasks and interact with each other to make complex systems manageable and robust.
- **Self-configuring, self-optimizing**: In massive scale systems, it is extremely difficult for a system administrator to tune parameters for the best possible network performance, due to lack of knowledge or time. A self-configuring module can smartly learn about its network and optimize the configurations without human intervention.
- **Self-protection, Self-healing**: Similarly, a self-defending module will protect against known security threats while self-healing modules will be able to automatically manage and respond to risks, for instance by isolating compromised or faulty nodes and auto-uploading software patches.

The aim for self-organizing control and management is to make the network adapt to unpredictable changes dynamically, while hiding intrinsic complexity from network operators and end users. Both protocol-oriented closed-loop and machine learning mechanisms, which may also require support from new protocols, should be studied and investigated for the next generation of protocols.

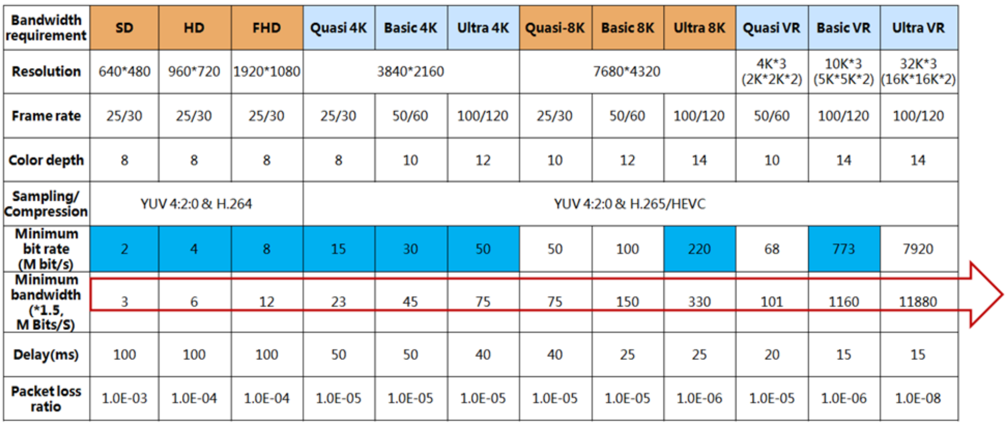
Key Scenarios for the Next Generation Protocols

Requirements from Video and Content Distribution

ITU-R [14] has developed new standards for Ultra-high-definition (UHD) video involving 4K UHD (2160p) and 8K UHD (4320p) video formats. At 3840 x 2160 resolution, 4K UHD offers superior quality video, which is four times better than Full High Definition (FHD or 1080p with 1920 x 1080 resolution). A case for high-throughput transport is most justified in the context of UHD video content distribution. In order to stream media content with higher-resolution over IP, greater network bandwidth is needed.

Table 5 summarizes the numerical values of various attributes for different video formats. For 4K video streaming 45 Mbps is the minimum bandwidth needed; this is four times the requirement for FHD, while 8K (at 150 Mbps) requires 12 times the FHD bandwidth. Even with significantly improved (50% bit-rate savings) video compression in H.265 [15] compared with H.264, the desired bandwidth grows sharply with higher resolution video transmission. In fact, the future bandwidth demands for emerging VR techniques are up to 1Gbps, much higher than 4K or 8K UHD streams.

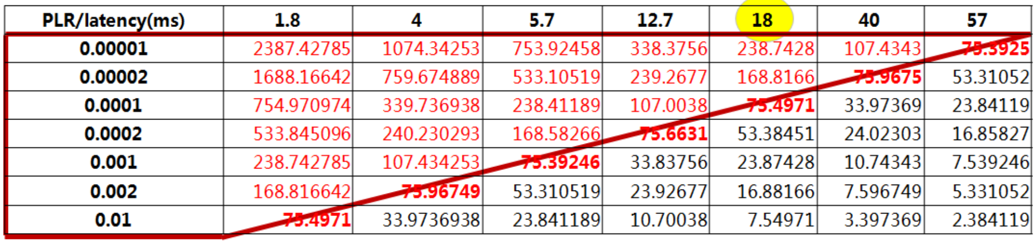
Table 5: Bandwidth Computations for different video formats



Bandwidth requirement	SD	HD	FHD	Quasi 4K	Basic 4K	Ultra 4K	Quasi-8K	Basic 8K	Ultra 8K	Quasi VR	Basic VR	Ultra VR
Resolution	640*480	960*720	1920*1080	3840*2160			7680*4320			4K*3 (2K*2K*2)	10K*3 (5K*5K*2)	32K*3 (16K*16K*2)
Frame rate	25/30	25/30	25/30	25/30	50/60	100/120	25/30	50/60	100/120	50/60	100/120	100/120
Color depth	8	8	8	8	10	12	10	12	14	10	14	14
Sampling/ Compression	YUV 4:2:0 & H.264			YUV 4:2:0 & H.265/HEVC								
Minimum bit rate (M bit/s)	2	4	8	15	30	50	50	100	220	68	773	7920
Minimum bandwidth (*1.5, M Bits/S)	3	6	12	23	45	75	75	150	330	101	1160	11880
Delay(ms)	100	100	100	50	50	40	40	25	25	20	15	15
Packet loss ratio	1.0E-03	1.0E-04	1.0E-04	1.0E-05	1.0E-05	1.0E-05	1.0E-05	1.0E-05	1.0E-06	1.0E-05	1.0E-06	1.0E-08

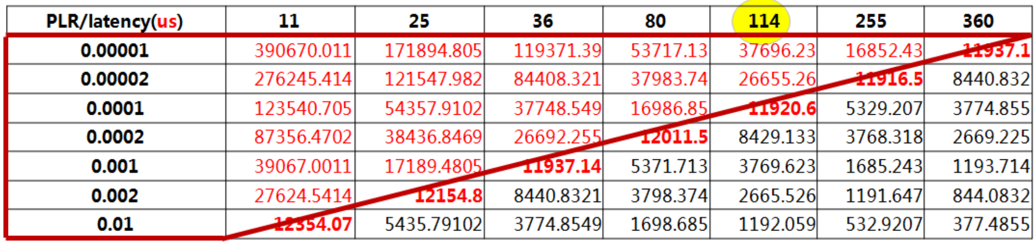
Existing video streaming is still mainly over TCP as the transport layer protocol. In Table 6, TCP throughput is computed for different values of packet loss ratio (PLR) and latency. As we move to the right (increased latency) and bottom (higher PLR), the throughput falls drastically.

Table 6: PLR and Latency Impact on TCP Throughput



PLR/latency(ms)	1.8	4	5.7	12.7	18	40	57
0.00001	2387.42785	1074.34253	753.92458	338.3756	238.7428	107.4343	**75.3925**
0.00002	1688.16642	759.674889	533.10519	239.2677	168.8166	**75.9675**	53.31052
0.0001	754.970974	339.736938	238.41189	107.0038	**75.4971**	33.97369	23.84119
0.0002	533.845096	240.230293	168.58266	**75.6631**	53.38451	24.02303	16.85827
0.001	238.742785	107.434253	**75.39246**	33.83756	23.87428	10.74343	7.539246
0.002	168.816642	**75.96749**	53.310519	23.92677	16.88166	7.596749	5.331052
0.01	**75.4971**	33.9736938	23.841189	10.70038	7.54971	3.397369	2.384119

∘ ∘ ∘



PLR/latency(us)	11	25	36	80	114	255	360
0.00001	390670.011	171894.805	119371.39	53717.13	37696.23	16852.43	**11937.1**
0.00002	276245.414	121547.982	84408.321	37983.74	26655.26	**11916.5**	8440.832
0.0001	123540.705	54357.9102	37748.549	16986.85	**11920.6**	5329.207	3774.855
0.0002	87356.4702	38436.8469	26692.255	**12011.5**	8429.133	3768.318	2669.225
0.001	39067.0011	17189.4805	**11937.14**	5371.713	3769.623	1685.243	1193.714
0.002	27624.5414	**12154.8**	8440.8321	3798.374	2665.526	1191.647	844.0832
0.01	**12354.07**	5435.79102	3774.8549	1698.685	1192.059	532.9207	377.4855

To emphasize the significance of transport throughput from the above table, a Quasi 8K UHD (at 75Mbps) can afford a delay of 18mS, assuming a packet loss ratio of only 1/10000. Whereas with same packet loss ratio, an Ultra VR at 1.1Gbps throughput can only afford a delay of 0.114mS. The mandatory bandwidth requirements for the immersive visual experience of emerging UHD techniques make a strong case for a high performance transport layer and the need to study alternative ways to achieve high throughput with lower dependency on RTT**.**

Live Media Streaming

Live transmission of signals such as audio, video, or tactile feedback, as part of a conversation or control loop, has very different requirements from "streaming" of broadcast or recorded content. The signal is converted, in real time, to a stream of digital values which are conveyed across the system in a stream of packets. At the receiving end, the values are converted back into a copy of the original signal, delayed by a fixed amount which is the end-to-end latency. Most applications require this latency to be much less than is tolerable for the streaming of content.

A major contributor to the end-to-end latency is the time packets spend waiting for onward transmission in queues in network switches. In many IP networks, the queuing time is not well-defined, so some packets will arrive too late for the data they contain to be used; waiting longer for the packets to arrive increases the end-to-end latency. Other packets will be lost because of buffer overflow in the switches. Forward error correction can be used to mitigate the losses, but will increase the latency. Use of protocols such as TCP, which retransmit lost packets, will increase it even more. To guarantee a maximum latency requires reservation of resources along the path the packets will follow; this process can also guarantee that no packets will be lost for lack of buffer space.

Tactile Internet applications need the end-to-end latency to be no more than about 1 ms, and some audio applications need it to be less than about 10 ms. To achieve a low enough queuing time may require fine-grained scheduling, synchronized along the path. In that case packets can be identified by their time of arrival rather than by any information in the header.

Requirements from the Internet of Things

The projected forecast of Internet of Things (IoT) numbers by the year 2020, will be up to 21 billion devices [16] generating about 44 zeta bytes of data. IoT connected objects are not conventional communication end points with standard network interfaces. These objects may be composed of 'things' such as sensors (for measurements and monitoring), actuators (receive instructions to augment processes) or devices (set of sensors and actuators) that are constrained by power, bandwidth, memory and connectivity range. They often require wireless connections because wired connectivity is not economically viable or geographically feasible. Because of these unique attributes and different device types, connecting IoT objects is very different from building traditional networks. The challenges include addressing and discovery of the objects in the network and connecting them to applications in the cloud (or private network) at a massive scale. In addition, they also require ad-hoc, less chatty, and secure routeing protocols in a resource constrained environment.

Presently, industry focus is to unify IoT communications through IP as a network layer. It continues to grow through several alliances and consortia such as IP Smart Objects (IPSO), Industrial Internet Consortium (IIC) and Threads, and the most widely accepted protocols are based on IETF 6LoWPAN [17] over IEEE 802.15.4 [33] media access networks. 6LoWPAN uses addressing mingled into the media layer; base specification RFC 6282 [31] provides header compression and RFC 6775 [32] provides neighbour discovery optimizations for low power communication.

Arguably, the IP based IoT stack has support through different alliances, but it is also true that vendor specific IoT device communication protocols continue to exist, which is why many IoT gateways are being introduced in the networks that translate between the IP and non-IP worlds. 3GPP IoT cellular network developments are not related to 802.15.4 [33] MAC layer, so impose no requirement to use traditional transport. In fact, within the GSMA (www.gsma.com) and the Open Mobile Alliance (OMA, openmobilealliance.org) mobile communities, OMA Lightweight Machine to Machine (LWM2M) [34] simply prefers to directly use Constrained Application Protocol (CoAP) [35] to minimize bits carried over the air-interface.

Considering that IoT based applications are still at an early stage, an IP based addressing scheme may seem restrictive for innovation, and next generation protocols should take the opportunity to identify, discover and connect IoT objects in a simple manner without requiring too many IoT gateways. Some of the goals for IoT can be summarized in Table 7.

Table 7: Requirements for IoT

Requirements	Current Challenges	Goals
Non-IP IoT connection to IP based applications	Number of IoT Gateways for translating non-IP to IP are growing	Carry device ID as is without needing IP based translation towards application
Communicate with each IoT object autonomously	Solutions for non-IP are vendor specific and not standard; cross-silo application connectivity is not possible without IP	Easier to register, discover, and authorize things based on ID.
Massive Scale of IoT	Resource burden on the network state	Minimize forwarding table size
Support for cellular based IoT	IoT stack predominantly assumes lightweight IP stack	Flexible and interoperable IoT stack for different media (802.15.4, cellular IoT)

The present-day lack of interoperability remains an unresolved hurdle to meet the diverse scenarios and over-whelming IoT scale of 2020. On the basis of the IoT requirements mentioned above, ETSI NGP will investigate means to provide universal IoT device communication (IP or non-IP) and an architecture to support large numbers of interconnected, complex event-driven IoT systems.

Requirements from Network Operators

Mobile network operators require the ability to optimize their network architecture to align with existing and new services. To do this effectively requires the ability to implement solutions such as edge computing at suitable locations and in a multi-vendor environment, ideally in a dynamic way able to address the changing requirements of users as close as possible to real time. The existing LTE protocol architecture as illustrated in Figure 8, highlights the challenges with managing tunnels associated with mobility and security, which makes an integration with a third party edge computing platform quite challenging and therefore expensive to implement, not to mention the barriers to properly addressing network plasticity requirements.

An additional challenge faced by mobile network operators is the need to optimize the use of radio frequency spectrum. Spectrum is a finite resource, and therefore it's vital that throughput be maximized by ensuing that a high ratio of application bits per second per Hertz of radio spectrum is achieved. The interaction between TCP/IP protocols, in particular TCP, and radio network protocols can be a challenge with a small packet error loss rate at the radio level serving to reduce effective throughput at the IP layer considerably, due to TCP behaviours. Mobile operators have solved this lack of adaptation by means of different optimization techniques, mostly applied at the TCP layer, which in most cases rely on some degree of access to packet headers and very limited access to some kind of payloads, like request URLs. The envisaged applications, their required data rates and emerging trends like pervasive encryption, impact directly on these practices, so it becomes necessary to explore adaptation mechanisms suitable for the radio access networks.

Network operators need next generation protocols to support virtualization along the lines of NFV/SDN today, in order to support rapid soft reconfiguration of their networks in line with evolving usage over time as well as dynamic current usage. Operators also want to be able to have the ability to choose how they host their networks, themselves as tenants or entirely as a managed service, and NFV/SDN-like capabilities are essential to do this.

Requirements from eCommerce

eCommerce represents a growing sector and huge commercial opportunity. It's vital that users can access retail websites with the minimum of delay to ensure they remain engaged and complete the transaction; a slow response often results in users clicking off and looking elsewhere for products. In addition to response time, security is a major consideration which is constraining growth in this sector.

The US government census on ecommerce (https://www.census.gov) reported $1.2 trillion sales. It shows 14% growth in 2014-2015 and at the same time $113bn were lost in cybercrime (https://csis.org/). There are several consequences to security breaches [17] both to companies and to consumers; they include loss of intellectual property, direct financial losses, sensitive business and personal information leaks, service disruptions, reduced trust, and reputational damages. The severity of these problems exists in cloud based models as well. To improve profits and build trusted eCommerce, cyber-attacks have to be minimized through comprehensively protecting websites, consumer identities and the active transactions. Almost all NGP security requirements discussed earlier apply here.

In another report from the Internet Statistics Compendium [19], it was shown that users do not stick around to complete shopping if a webpage loads slowly ($1.73bn lost). The report also showed that if a page does not load in 10 seconds, 40% of the users will abandon the website. Ten seconds seems like a long time in the context of DSL, LTE and 5G access speed and bandwidths.

eCommerce is closely tied to the application layer in the OSI model, and in the following subsection a few risks and issues are exposed that indicate the scope of the challenges to be studied within ETSI NGP.

Auditing/Management

Without proper e-commerce auditing, insecure actions of users and administrators are difficult to monitor, and without good management systems and tools the operation of the platform itself can be inefficient and hence costly in terms of OPEX. Neither auditing nor management systems are essential to provide an e-commerce platform, but without them security and OPEX tends to get costlier. Better auditing and management/standards are required for e-commerce to work with next generation protocols if we are to improve the productivity of the market and reduce the losses due to poor accounting/auditing.

Security

E-commerce security is not good enough for today’s world and costs main economies dearly in terms of losses to cyber-crime. At the basic level, automated logout is not universally supported, and SSL certificates based site validation doesn’t always verify the certificates through online trusted bodies. Most sites do not have the capability to use online trusted bodies. Future protocols need to include automated validation and site checking.

Site Access Speed

All the top 500 sites in US are now 16% slower than last year due to increased demand (2015). Many ecommerce applications still use legacy web publishing sites which require “enhancements” for security, auditing and efficient validation payment authentication controls. These platforms add workarounds for many issues with the TCP/IP suite through web accelerators. Site accesses are also affected by complex page redirections to many other sites and even self-redirection at times.

HTTP has proven a simple and efficient single server protocol for requesting information. However, requests become progressively slower as other facets of complex web pages are introduced that involve multi-server linking and extra DNS lookups per page, resulting in long page request delays. Quick UDP Internet Connections (QUIC) and HTTP2 have gone some way towards adding controls to manage transport options such as MP-TCP (multipath TCP, IETF) and some other facets of the download, but a richer meta-data based HTTP protocol with standardized tags would improve complex e-commerce site speed considerably. This is supported by current research which suggests that improvements of up to 30% are achievable with meta-data enabled versions of HTTP.

Compression

Header compression and image compression are underused, under the misapprehension that they are slow, but on an average for rich sites it actually helps with transport and is faster to the user (the smartphone is a powerful device now, but many sites have not moved on to suit).

Table 8: Areas of focus for eCommerce

Requirements	Current Challenges	Goals
Auditing/ management	Inefficiency in ecommerce transactions due to lack of sufficient information	protocols need to be more supportive and aware of access and operation of e-commerce sites with built-in efficient extensions to provide features such as user and payment authentication, site validation, auditing, and login control
Security	SSL certificates not always verified, user access not fully secure	Enforced and automated validation of SSL certificates Effective login management
Site Access Speed	multiple DNS lookups from the site pages	Optimize information retrieval
	Complex redirects	provide meta-data capabilities in new management protocols to reduce/optimize the overall number of redirects
Compression	Faster handsets not fully utilized. Media unaware	provide controls for compression over radio based access links for both header and image transfers of complex pages

In summary, next generation protocols should seek to provide better HTTP management through optimization, proxy support and page meta-data awareness suitable for eCommerce web site accesses.

Requirements for Increased Energy Efficiency Within the Global ICT Sector

The global ICT sector is a major consumer of electrical power and the current trends predict increased consumption as demand for communications and Internet based services continues to grow. A study [20] performed by the Seventh Framework Programme (FP7) on power consumption for the time period 2007-2012 is based on three main ICT categories: communications networks, personal computers and data centres. It estimated that the yearly growth of all three individual ICT categories (10%, 5% and 4% respectively) is higher than the growth of worldwide electricity consumption in the same time frame. It's vital that the sector find ways to minimize overall power consumption to reduce harmful emissions and lower operating costs. To this end, minimizing the processing power required to manage end to end communications has a role to play.

Requirements for Reducing Network Complexity

Often, complexity in networks gradually increases as a side effect of addition of new features on top of existing ones. Many inline network functions are processing–intensive and exist as result of problems identified in real deployments such as security and traffic profiling. These complex functions are mandatory, have to be implemented in-line and do not scale well, which makes end to end traffic performance difficult to predict and optimize. It adds cost to the network in terms of hardware processing requirements, increased energy requirements and increases in the overall complexity of the network. Typical network complexities in current communications systems (Table 9) include ***control and data protocol processing*** (involving header inspections, compression, management, and translation overheads, in both control plane and data plane) and ***Performance Management*** (complexity in congestion being handled on a per-layer or per-protocol basis, it is not easy to accurately view the scenario). Besides, traffic is also subjected to many network functions such as security, NAT, DPI, etc. that incur processing overheads.

Table 9: Network Complexity Considerations

Aspects	Challenges
Data plane	Optimizing headers in general and specifically for over-air transmission
Control plane	Complex interactions between different protocols and layers of network stack
Network functions	Providing services such as DPI, NAT state management, and security are processing intensive activities

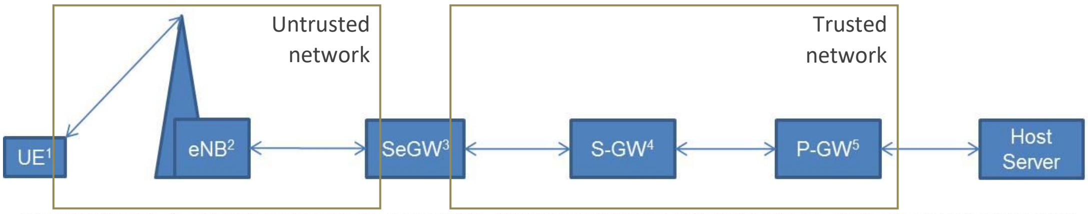
New network protocols designed by ETSI NGP should be such that they offer low handling impact, across all complexity aspects (e.g. processing, memory, ease of distribution and so on) to make resource consumption efficient, end to end traffic streamlined according to its profile, and with minimized OPEX and CAPEX.

Case Studies

LTE Mobile Network Case Study

As means of a case study let's explore a typical mobile operator's LTE network architecture. 3GPP defined the initial LTE network architecture as part of Release 8, published in 2009; this defines the main network nodes along with their functions and interfaces. Due to the distribution of all RAN functionality to the LTE base station, known as an evolved Node B (eNB), the encryption which is available on the radio interface is terminated at the base station and, therefore, 3GPP recommends the use of IPSec between the eNB and core network site (trusted environment) to provide both authentication of connected devices (to prevent a man in the middle attack) and encryption of traffic flows (to prevent eavesdropping).

Note: 3GPP specifies an IP transport network layer which is typically implemented over a Carrier Ethernet transmission link.



1: User Equipment - 2: evolved NodeB (LTE radio base-station) - 3: IP Security gateway - 4: Serving Gateway - 5: Packet data-network Gateway

Figure 7: High-level LTE network architecture

Inflexible Security Tunnels in Radio Access Networks

The introduction of IPSec results in a security tunnel which runs between the eNB [21], of which there are often many thousands distributed throughout the network to provide the cellular radio coverage, and the mobile core network sites, of which there are typically some tens. These numbers do vary based on size of the geography to be covered. As an example, EE/BT in the UK has approximately 19,000 base station sites connected to 18 core network sites. The IPSec tunnel is by design an end to end secure environment which can't be 'tapped into' and as such restricts the access to the control, user and management planes running between the eNB and core network; this makes features such as local breakout, optimization and content caching at the edge (often referred to as MEC or Mobile Edge Computing), a particularly complex and expensive integration.

In addition to the IPSec tunnel, a second tunnel is implemented to support mobility; this is specified by 3GPP and follows the principles established for GPRS in the late 1990s. GPRS tunnelling protocol (GTP) encapsulates the user plane data and operates between the eNB and Packet Data Network Gateway (P-GW). The P-GW is typically located deep in the mobile operator's core network, close to transmission links for external network access, private APN's and the public Internet.

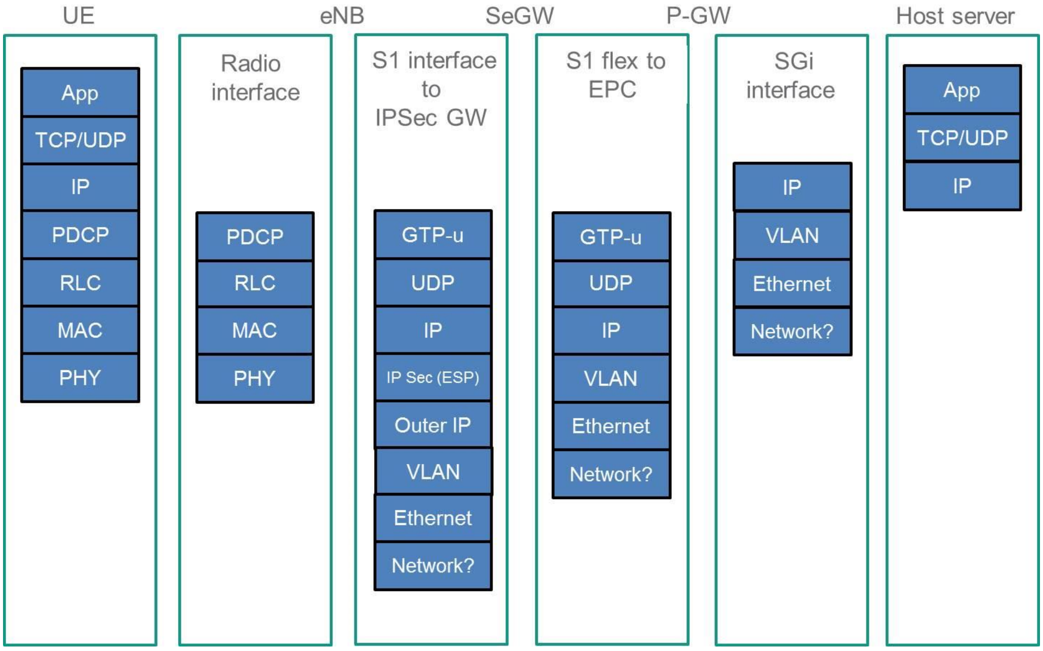


Figure 8: Typical LTE network protocol architecture for S1 interface user plane traffic

Ideally, a new protocol architecture will remove the need for overlay tunnels such as IPSec and GTP (shown in Figure 8) by integrating the functionality of security and mobility within the native protocol architecture.

Connected Objects – Industry 4.0

Industry 4.0 (a term first used in 2011) [22] delivers the future of the manufacturing industry by combining smart objects and digital systems capable of autonomously exchanging information, triggering actions and controlling each other independently. The goal is to create flexible and resource-efficient production systems in manufacturing and to integrate industry, business, and internal processes through computer based systems. With IoT at the core, Industry 4.0 is an ideal scenario for the study of new network communication protocols that serve to realize the Smart Factory. A generalized Smart Factory case study is envisioned as below:

A typical manufacturing plant in a Smart Factory collects and distributes vast amounts of device status data, video, and control parameters across dozens of buildings over a wide-area production facility. Data transmissions across the facility must be delivered in real-time, securely and uninterrupted.

The technical challenges begin at the lowest level, where each object or manufacturing part needs to have communication capability. It then requires flexible mechanisms to connect everything through an integrated wired and wireless solution for increased bandwidth, reliable Wide-Area Network (WAN) coverage, flexible networking and high security. Intelligent manufacturing demands seamless connection of large numbers of people, application systems, intelligent machines, and sensors from on-site and remote locations.

Once connectivity is established, the second level of challenge involves the ability to leverage a range of services — such as environmental and energy monitoring, remote broadband access of intelligent terminal devices and sensors, High-Definition (HD) video surveillance, and emergency response.

The data generated from installed smart meters, sensors and HD cameras offers high-end production services that include repair, fault location, emergency rescue, etc. These services are complex and require collaboration across teams, seamlessly connecting application systems and intelligent machines, to implement mobile management and energy conservation with their intelligent, innovative manufacturing, information, and communications systems.

IP based protocols are not always suitable for IoT; different types of IoT are integrated through network controllers or translation IoT gateways. This makes applications vendor specific, complex and expensive, in terms of both cost and maintenance. To realize a Smart Factory solution, the new protocol framework should integrate (1) a uniform and flexible connectivity to all the connected devices, (2) low latency IoT gateways with minimal network overheads, and (3) better intelligence and autonomous decision making systems. This should be done through auto-discovery, self-monitoring and self-configuring of connected devices.

Multiple PDN Gateways in Mobile Networks

5G promises ubiquitous connectivity over heterogeneous accesses. A direct consequence of this is increased density in the regional mobile core network which is typically served by a single PDN gateway (P-GW). As the subscribers and devices in a mobile core grow, the scale of the centralized PDN gateway may not be manageable at one place and it has been suggested that multiple distributed PGWs per mobile core should be used in future. Today, a user experiences seamless mobility within an operator’s regional mobile network and this will be impacted if PGWs are distributed (effectively forming islands of smaller networks in the core). Maintaining session continuity across multiple PGWs becomes a new problem that may be addressed through NGP based mobility solutions.

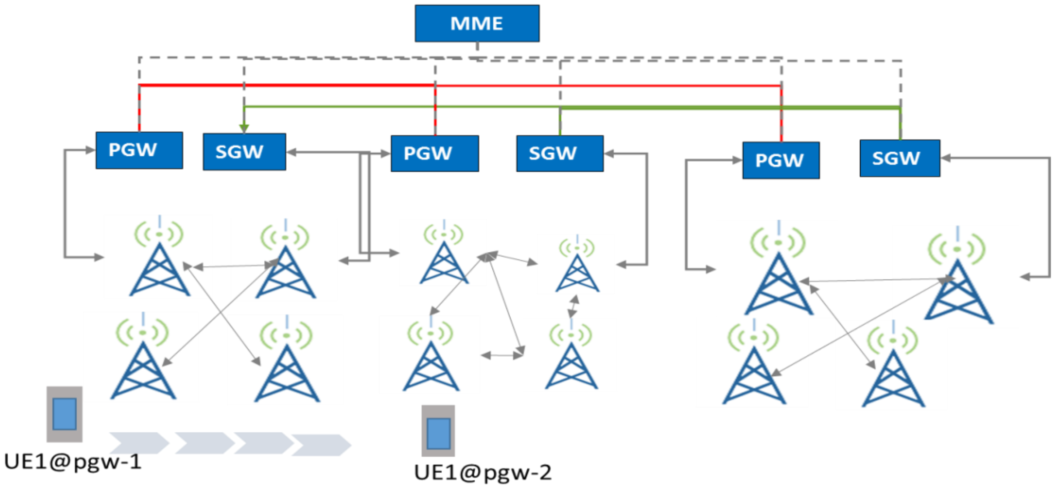


Figure 9: Multiple P-GW Scenario

Expected Deliverables and Future Outlook

With the introduction of any new networking technology that supports many other technologies, the impact is likely to be significant. As such, the introduction of a new NGP architecture needs to be managed very carefully in order to ensure success within a reasonable time period. This section of the white paper proposes a phased introduction of the ETSI NGP outputs.

Next Steps

To this end the NGP team have proposed a three-step approach to the ETSI NGP initiative to deliver a set of key milestones over the next few years, as follows:

a) **Scenarios and use cases**: Define the scenarios for use of NGP and map them to well defined existing use cases (wherever possible), clearly identifying the key issues we want to solve as compared to today’s experience and identifying the key requirements to enable us to move forward from these issues

b) **Architecture and Recommendations**: Produce a protocol architecture template for NGP, in the process making recommendations to industry and standards bodies. A close collaboration with 3GPP and IETF will be necessary.

c) **Near term strategy:** Suggest how and where NGP could be introduced into early versions of next generation communications systems. This stage best demonstrates early system gains within, say, the next 5 years while still readily supporting existing services and some new Next Generation services that cannot be supported today (ideally) or are difficult to support.

Future Outlook

Part of the underlying problem lies in the lack of generality of IP-based solutions; part in the flaws embedded in its design. To move from the prototype research Internet to the current Internet the TCP/IP protocol suite has been heavily patched to reach a scale it had not been designed to reach and to be applied to environments it was not designed to support. However, instead of re-considering the fundamental design decisions that caused the problems, the TCP/IP protocol suite has grown through the addition of new protocols, keeping its core structure essentially unchanged. Whenever there is a new requirement or a new application to be supported, a new protocol is designed, implemented and deployed. Each individual protocol increases the complexity of the whole suite and, since it is often designed in isolation, interacts with other protocols in ways that are hard to predict. Essentially, the IP protocol suite is like a huge computer program that has never been refactored in 40 years: the base structure has never been updated or consolidated. As such, it suffers from a huge technical debt that all of us – network designers, operators and users – are paying in different ways (e.g. the cost of building and operating networks, the price and performance of network services available to users and hazards generated due to the unexpected unavailability of a service assumed “universal”).

This is our present, but doesn’t need to be our future. ETSI NGP thinks there is another path forward. One in which we can consolidate all the lessons learned during the 40-50 years of computer networking theory and practice, as well as experience with other kinds of digital media, in a new network architecture that is really fit for its purpose. The goal is a consolidated architecture and protocol suite upon which all types of networks are based: Current IP-based technologies are not fit for this purpose.

ETSI NGP envisions a network architecture that captures all the invariants of the problem of internetworking between digital systems and materializes them in flexible protocol building blocks that can be used in heterogeneous network environments. In this vision, accommodating new application requirements or physical media would not require the development of new protocols, but only the addition of plug-ins to the generic building blocks. These plug-ins would optimize the generic protocols for different specific environments, allowing for a flexible solution whose complexity can still be bounded. The NGP common structure will also provide constructs to enable an efficient support of mobility and multi-homing without the need of special protocols, a clear security model enforced across all layers of the architecture and a consistent approach towards delivering predictable performance guarantees over multiple networks. The commonality provided by the NGP architecture will greatly simplify network management, turning the problem of managing layers of different protocols – one for each function – to a problem of managing a common structure with different plug-ins.

ETSI NGP’s vision can only be achieved by re-architecting the protocol stack; just continuing with incremental improvements to the TCP/IP protocol suite would mean even greater complexity and inability to meet demands of new applications and technologies. In order to envision the future, one must break with the past but in order to move to the future one must interoperate with the past. ETSI NGP doesn’t assume any flag days, nor any instantaneous replacement of existing protocols. Even assuming the NGP architecture was already specified and its benefits proved in theory, it still has to be implemented in commercial products and deployed in the field. Moreover, existing assets must be amortized. Last but not least, it is not the goal of ETSI NGP to take over the role of the multiple standard bodies that are currently producing computer-networking standards. Therefore, how to start moving to the future, and what role should the ETSI NGP play?

Clearly, the goal of ETSI NGP is not to drive an academic exercise, but to provide answers to the networking industry's problems. To be incrementally deployable in the networks of today, NGP solutions must interoperate with the TCP/IP protocol suite at different points, initially through NGP proxy devices. Implementation of NGP is more likely to be viable in scenarios that (1) are challenging for the traditional TCP/IP protocols and (2) require the deployment of new equipment. In this regard, 5G radio networks are an ideal setting for NGP deployments.

Conclusion

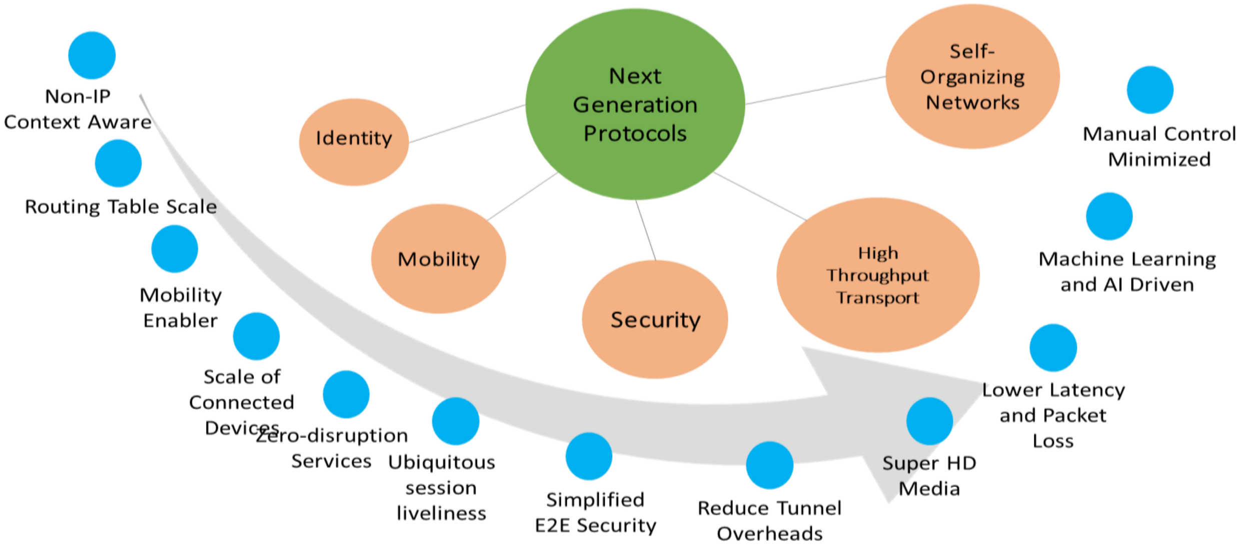


Figure 10: Main Goals and Focus Areas of NGP

The NGP initiative is to drive the development of a Next Generation Network architecture that will be based upon requirements from many use cases and deployment scenarios. ETSI ISGNGP aligns its goals to make the end to end Internet more efficient, taking into consideration the requirements of different technology sectors as discussed in earlier sections of this white paper. The vision, focus areas and goals that ETSI NGP has set itself are summarized in Figure 10.

In this document we've highlighted concerns and challenges from different market segments that will have a significant impact on existing network infrastructures. NGP is conscious of the fact that for industry-wide acceptance, the investment in global communication and corresponding OPEX and CAPEX, though not a technological barrier, is an important consideration. In this regard, the forum values close cooperation with interested parties to produce recommendations that best fit sustainable business models for the whole ecosystem and solve the problems discussed in this paper.

Re-architecting the protocol stack is an effort that can potentially impact all stakeholders in the computer networking community. Fully achieving ETSI ISG NGP's vision requires buy-in from a significant majority of the industry, especially through collaboration with established SDOs. Therefore, it is not the intention of this ISG to undertake this mission alone – it would never work. But, in order to obtain industry buy-in, first ISG NGP has to demonstrate that achieving its vision is possible: this is the ISG's mission. In order to do so it plans to work in two parallel tracks: one is to identify and specify the elements of a network architecture which meets the demands of the 21st century; the other is to design, implement and deploy an early adoption scenario in which the NGP solution is deployed in the field interoperating with existing protocols. We expect that the combination of these two tracks will provide the right incentives for current SDOs to engage in a dialogue with the ISG in order to explore how the NGP architecture can be gradually introduced in different types of networks, gradually moving to a truly converged world with simple, scalable, secure and robust networks delivering a rich set of networking services to distributed applications.

References

[1] Cisco VNI Forecast: Global Mobile Data Traffic Forecast Update, 2015–2020 http://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/mobile-white-paper-c11-520862.html

[2] OSI Conceptual Model, https://www.iso.org/obp/ui/#iso:std:iso-iec:10731:ed-1:v1:en

[3] BGP Routing Table Analysis Reports: http://bgp.potaroo.net/

[4] ISO/IEC TR 29181-1 Future Network Problem Statement and Requirements -- Part 1: Overall aspects, ISO, Geneva: http://standards.iso.org/ittf/PubliclyAvailableStandards/c057480_ISO_IEC_TR_29181-1_2012.zip

[5] ISO/IEC TR 29181-3 Future Network: Problem Statement and Requirements – Part 3: Switching and Routing, ISO, Geneva: http://www.iso.org/iso/catalogue_detail.htm?csnumber=57485

[6] Draft LCAF: LISP Canonical Address Format, https://tools.ietf.org/html/draft-ietf-lisp-lcaf-12

[7] ICN: Information Centric Networking: https://irtf.org/icnrg

[8] Understanding 5G: Perspectives on future technological advancements in mobile, December 2014. https://gsmaintelligence.com/research/?file=141208-5g.pdf

[9] B. S. Center on International Security, "Risk nexus: Overcome by cyber risks? Economic benefits and costs of alternate cyber futures," Zurich Insurance Group, Tech. Rep., 2015.

[10] B. Schneier, "A plea for simplicity: You can't secure what you don't understand," Information Security, 1999.

[11] Motherboard; "A Hacker made 'Thousands' of Internet-connected printers spit out racist flyers" Available online at: http://motherboard.vice.com/read/hacker-weev-made-thousands-of-internet-connected-printers-spit-out-racist-flyers

[12] UK fixed-line broadband performance, November 2015, http://stakeholders.ofcom.org.uk/market-data-research/other/telecoms-research/broadband-speeds/UK-home-broadband-performance-Nov-15/

[13] 3GPP TS 32.500: "Telecommunication management; Self-Organizing Networks (SON); Concepts and requirements", http://www.3gpp.org/DynaReport/32500.htm

[14] ITU-R Standards for Ultra-high-definition broadcast, http://www.itu.int/en/ITU-D/Regional-Presence/Europe/Documents/EBU_UHDTV%20ITU-D%20Seminar%20Budapest_2014.pdf

[15] ITU-T H.265 - High efficiency video coding: http://www.itu.int/rec/T-REC-H.265-201504-I

[16] Research Report Gartner Symposium 2015, http://www.gartner.com/newsroom/id/3165317

[17] 6lowpan: http://www.ipso-alliance.org/wp-content/media/6lowpan.pdf

[18] Economic Impact of Cybercrime, https://csis.org/files/publication/60396rpt_cybercrime-cost_0713_ph4_0.pdf

[19] Internet Statistics Compendium, https://econsultancy.com/reports/uk-internet-statistics-compendium

[20] Assessment of power consumption in ICT, http://cordis.europa.eu/docs/projects/cnect/0/257740/080/deliverables/001-trendd16finalwp1report.pdf

[21] 3GPP TS 33.210: 3G security; Network Domain Security (NDS); IP network layer security http://www.3gpp.org/DynaReport/33210.htm

[22] EPRS briefing Industry 4.0: http://www.europarl.europa.eu/RegData/etudes/BRIE/2015/568337/EPRS_BRI(2015)568337_EN.pdf

[23] IETF RFC 6830: The Locator/ID Separation Protocol (LISP): https://tools.ietf.org/html/rfc6830

[24] IETF RFC 5944: IP Mobility Support for IPv4, Revised: https://tools.ietf.org/html/rfc5944

[25] IETF RFC 6725: DNS Security (DNSSEC) DNSKEY Algorithm IANA Registry Updates: https://tools.ietf.org/html/rfc6725

[26] 3GPP TS 29.060 V6.9.0: General Packet Radio Service (GPRS); GPRS Tunnelling Protocol (GTP) across the Gn and Gp interface: http://www.3gpp.org/DynaReport/29060.htm

[27] 3GPP TS 29.274 V8.1.0: 3GPP Evolved Packet System (EPS); Evolved General Packet Radio Service (GPRS) Tunnelling Protocol for Control plane (GTPv2-C); Stage 3: http://www.3gpp.org/DynaReport/29274.htm

[28] IETF RFC 5213: Proxy Mobile IpV6: https://tools.ietf.org/html/rfc5213

[29] IETF-DMM: Protocol for Forwarding Policy Configuration (FPC) in DMM: https://tools.ietf.org/html/draft-ietf-dmm-fpc-cpdp-03

[30] IETF RFC 7429: Distributed Mobility Management: Current Practices and Gap Analysis https://tools.ietf.org/html/rfc7429

[31] IETF RFC 6282: Compression Format for IPv6 Datagrams over IEEE 802.15.4-Based Networks https://tools.ietf.org/html/rfc6282

[32] IETF RFC 6775: Neighbor Discovery Optimization for IPv6 over Low-Power Wireless Personal Area Networks (6LoWPANs) https://tools.ietf.org/html/rfc6775

[33] IEEE 802.15.4: IEEE Computer Society, "IEEE Std. 802.15.4"

[34] Open Mobile Alliance OMA Lightweight M2M: http://technical.openmobilealliance.org/Technical/technical-information/release-program/current-releases/oma-lightweightm2m-v1-0

[35] IETF RFC 7252: Constrained Application Protocol (CoAP): https://tools.ietf.org/html/rfc7252

ETSI (European Telecommunications Standards Institute)
06921 Sophia Antipolis CEDEX, France
Tel +33 4 92 94 42 00
info@etsi.org
www.etsi.org

This White Paper is issued for information only. It does not constitute an official or agreed position of ETSI, nor of its Members. The views expressed are entirely those of the author(s).

ETSI declines all responsibility for any errors and any loss or damage resulting from use of the contents of this White Paper.

ETSI also declines responsibility for any infringement of any third party's Intellectual Property Rights (IPR), but will be pleased to acknowledge any IPR and correct any infringement of which it is advised.