NGMN 5G P1
Requirements & Architecture
Work Stream End-to-End Architecture

Edge Computing

by NGMN Alliance

Version: **Date:** **Document Type:** **Confidentiality Class:** **Authorised Recipients:** (for CR documents only)	**1.0** **14th September 2016** **Final Deliverable (approved)** **P - Public**

Project:	**5G**
Editor / Submitter:	**Adrian Neal (Vodafone)**
Contributors:	**NGMN P1 WS1 E2E Architecture team**
Approved by / Date:	**NGMN Board, 10th October 2016**

The information contained in this document represents the current view held by NGMN Ltd. on the issues discussed as of the date of publication. This document is provided "as is" with no warranties whatsoever including any warranty of merchantability, non-infringement, or fitness for any particular purpose. All liability (including liability for infringement of any property rights) relating to the use of information in this document is disclaimed. No license, express or implied, to any intellectual property rights are granted herein. This document is distributed for informational purposes only and is subject to change without notice. Readers should not design products based on this document.

Abstract: Short introduction and purpose of document

This document describes further details of Edge Computing and identifies use cases, including those from the NGMN White Paper and 3GPP SA1 SMARTER study for 5G, where placement of compute functionality closer to or at the edge of the network could form part of the solution.

Document History

Date	Version	Author	Changes
16th February 2016	0.0.0	Adrian Neal	First draft
14th March 2016	0.1.0	Adrian Neal	Addition of 2 use cases
15th March 2016	0.1.1	Adrian Neal	Deletion of erroneous reference
16th March 2016	0.1.2	Sebastian Speicher	Text for deletion removed, alignment of other text
29th March 2016	0.1.3	Adrian Neal	Agreements from 16th March conference call. Removal/change of solution specific text in Sections 5.3.1.1 and 5.3.2.1,
7th April 2016	0.2.0	Adrian Neal	Inclusion of agreed text from 31st March conference call.
22nd April 2016	0.3.0	Adrian Neal	Inclusion of agreed text from 20th April conference call.
12th May 2016	0.4.0.	Adrian Neal	Inclusion of text agreed from the 11th May conference call.
13th June 2016	0.4.1	Adrian Neal	Rapporteur cleanup as agreed on 8th June conference call.
18th August 2016	0.4.2	Sebastian Thalanany, Adrian Neal	Text revisions as agreed on 3rd and 17th August conference calls
24th August 2016	0.4.3	Adrian Neal	Inclusion of text in Section 5.1.3
7th September 2016	0.4.4	Steve Tsang Kwong U	Inclusion of text in Section 6.2.1
14th September 2016	0.4.5	Steve Tsang Kwong U	Editorial and formatting updates
3rd October 2016	0.4.6	Steve Tsang Kwong U	Updated text to address based on feedback during board approval

Table of Contents

1 INTRODUCTION

In October 2015 it was proposed to the NGMN Board that the NGMN 5G project should include work on Edge Computing in its scope. In particular the following considerations were given;

- Edge Computing infrastructure and applications could even be introduced into the market before 5G is ready,
- Interesting use cases for mobile and for fixed access are anticipated
- In the initial renditions Edge Computing nationwide coverage is not expected.
- Edge computing is expected to leverage 5G radio capabilities to facilitate ultra-low-latency bounds to meet service specific demands.
- Edge computing is expected to significantly enhance the attractiveness of 5G service propositions in geographically relevant manner.

Edge Computing is therefore a framework that is relevant for deployment over 4G and WiFi accesses, as markets develop, but also as a key enabler for new markets and potential new revenue streams in 5G. The proposal for Edge Computing was unanimously approved.

This document identifies known use cases from the NGMN White Paper [1] and 3GPP's 5G "SMARTER" studies [2] for which Edge Computing could form part of the solution, and presents a rationale for that conclusion in each case. It also proposes high level architecture and service requirements.

2 REFERENCES

NGMN 5G White Paper v1.0. February 2015. [1]

3GPP TR22.891v14.0.0. Study on New Services and Markets Technology Enablers. [2]

Recommendations for NGMN KPIs and Requirements for 5G. June 2016. [3]

3GPP TS22.185v14.1.0. Service requirements for V2X services. June 2016. [4]

NGMN Perspectives on Vertical Industries and Implications for 5G v2.0 September 2016 [5]

3 PROBLEM STATEMENT

Standards Development Organisations and Industry Forums have been engaged in gathering use cases and developing solutions for various 5G service scenarios, where Edge Computing could provide part or all of the solution. ETSI ISG NFV is working on related virtualization and management/orchestration aspects; ETSI ISG Mobile Edge Computing is working on the requirements, architecture and open APIs for the mobile version. 3GPP is working on its next generation (5G) system which has new stricter requirements for latency and delivery of content.

Edge computing also provides the benefits of cloud computing at the edge of the network, to enhance the user experience. With this background, no single entity has a higher level view of the requirements for Edge Computing, which could provide guidance for its relevance and applicability in 5G, and how Edge Computing might impact the individual or joint work of Standards Development Organisations and Industry Forums.

4 THE CONCEPT OF EDGE COMPUTING

4.1 Edge Computing concept

Edge Computing places services enabled by IT and cloud-computing capabilities at the network edge, e.g. within the Access Network or at the edge of the Core Network, thus much closer to end-users.

The access edge offers a service environment with ultra-low latency and high-bandwidth as well as direct access to real-time network information (such as access network conditions, user location, network load, etc.) enabling applications to be context enriched and hence embellished to offer context-related services for the user or for the network provider (for network performance optimisation).

Edge Computing allows local content, services and applications to be accelerated, increasing their responsiveness and performance. The user experience can be enriched through efficient network and service operations, harmonised with known access network conditions.

Operators can open the network edge to third-party partners, allowing them to rapidly deploy innovative applications and services towards their own and roaming subscribers, enterprises and other vertical segments.

In essence, Edge Computing could form part of the solution for a variety of 5G use cases, especially where rapid delivery of content only of relevance to a defined geographic area, or low latency is needed.

4.2 Definitions

Affective computing. Systems and devices that can recognize, interpret, process, and simulate human affects, interpreting the emotional state of humans and adapting their response to them.

Haptic content Content which recreates the sense of touch by applying forces, vibrations or motions to the user

Proprioceptive media. Media which informs the user of their body position in space.

Vestibular media Media which affects the users sense of balance and spatial orientation.

5 USE CASES WHICH COULD BE SATISFIED USING EDGE COMPUTING

5.1 Use cases from the NGMN White Paper [1]

Additional information and potential latency values which may be relevant to some of the following use cases can be found in [3]. Additional use cases (such as smart airport AR/VR, collaborative gaming, New Media Experience) that may also benefit of Edge computing can be found in [5].

5.1.1 Low latency use cases

5.1.1.1 Tactile Internet

Tactile interaction is where humans will wirelessly control real and virtual objects. It typically requires a tactile control signal and audio and/or visual feedback. The user, interacting with the tactile environment, does not perceive any difference between local and remote content. Real-time reaction is expected to be sub-millisecond. The current proposed requirement in 3GPP SA1 for this use case is that the 3GPP system shall support very low latency in the region of 1ms.

Rationale for proposing Edge Computing: Ultra low latency. Edge Computing would also be an enabler for affective computing services, with human-to-machine interfaces, requiring adequate response times, in terms of haptic content, where a latency alignment among visual, vestibular, and proprioceptive media is pivotal for an attractive user-experience, and a mitigation of cyber-sickness.

5.1.1.2 Automated Traffic Control and Driving

The vehicular LTE communication applications in 3GPP TS22.185 [4] contain the following four different types:

- Vehicle-to-Vehicle (V2V)
- Vehicle-to-Infrastructure (V2I)
- Vehicle-to-Network (V2N)
- Vehicle-to-Pedestrian (V2P)

In V2I scenarios the UE supporting V2I applications transmits messages containing V2I application information to a roadside unit which may be an eNodeB
In V2N scenarios the UE supporting V2N applications communicates with an application server supporting V2N applications. The application server and its location are currently out of 3GPP scope.
The NGMN White Paper predicts that lower latency than that required for LTE applications will be necessary for 5G. Commuting capabilities will require an ultra-low end-to-end latency for some warning signals, traffic flow optimisation and higher data rates to share video information between cars and infrastructure.

Rationale for proposing Edge Computing: Low latency for V2I. (V2V likely not to be Edge Computing, Edge Server unlikely to be vehicular. On the other hand D2D services may represent a category of services, where one device may be a host for another, in a V2V or a non-V2V scenario).
Note: On the other hand in the case of an autonomous vehicle, additional attributes of Edge Computing with mobility may be relevant.

5.1.1.3 Collaborative Robots: A Control Network for Robots

In order to enable these applications with completely diverse tasks in different environments, it will be essential to provide very low latency and high reliability. For many robotics scenarios in manufacturing a round-trip reaction time close to 1ms is anticipated [1], based on the type of human-machine service content and the related QoE (Quality of Experience) guarantees.

Rationale for proposing Edge Computing: Very low latency, localised content, hosted in secure enterprise environment(s), could even be on unlicensed spectrum or fixed access.

5.1.1.4 Pervasive Video

Customers will use video broadly in their everyday workflow. Examples include data delivery for optical head-mounted displays, collaboration in 3D cyber-real offices or operating rooms (with both physical and virtual presence) and customers' support by hologram services. The number of concurrently active connections, combined with the performance required (data rate and the end-to-end latency) will present a challenging situation. For such services low end-to-end latency is required. As the video element is a key component of the user's interaction with the service it should also be delivered in a suitably responsive near real-time fashion.

Rationale for proposing Edge Computing : Low latency.

5.1.2 Localised content use cases

5.1.2.1 Smart Office

Services that need high-speed execution of bandwidth-intensive applications, processing of a vast amount of data in a cloud, and instant communication by video. Ultra-high traffic volume, and for some applications latency, are the main challenges applicable for this use case.

Rationale for proposing Edge Computing: Low latency, “instant” video, likely to be in a secure enterprise environment/localised content.

5.1.2.2 HD Video/Photo Sharing in Stadium/Open-Air Gathering

This use case is characterised by a high connection density and potentially temporary use (e.g., in a stadium, concert, or other events). Several hundred thousand users per km^2 may be served, possibly integrating physical and virtual information such as score, information on athletes or musicians, etc., during the event. People can watch high definition (HD) playback video, share live video or post HD photos to social networks. These applications will require a combination of ultra-high connection density, high date rate and low latency. They could be scaled to known subscriber numbers within the stadium/gathering and help to preserve WAN capacity due to local delivery of the localised content with a high, scalable data rate. Mobile or fixed access could be used.

Rationale for proposing Edge Computing: Localised content, low latency.

5.1.2.3 Remote Object Manipulation: Remote Surgery

The technology necessary for providing the correct control and feedback for the surgeon entails very strict requirements in terms of latency, reliability and security.

Rationale for proposing Edge Computing: Hosted in secure enterprise (hospital) environment, ultra-low latency.

5.1.2.4 Local Broadcast-like Services

Local services will be active at a cell (compound) level with a reach of for example 1 to 20 km. Typical scenarios include stadium services, advertisements, voucher delivery, festivals, fairs, and congress/convention. Local emergency services can exploit such capabilities to search for missing people or in the prevention or response to crime (e.g. theft).

Rationale for proposing Edge Computing: Localised content hosted in a secure enterprise environment (mall, stadium), video surveillance (missing person/tracing person/suspicious package). Low latency for advertisements/vouchers (i.e. before the target walks past).

5.1.3 Performance optimisation use cases

None of the use cases in the NGMN White Paper relate to performance optimisation.

5.2 Use Cases from 3GPP SMARTER study [2]

5.2.1 Low latency use cases

5.2.1.1 Ultra-reliable communications.

Rationale for proposing Edge Computing: The low latency aspects of this use case relate to potential uses already included in the NGMN White Paper and reproduced in Section 5.1 of this document (i.e. automated traffic control and driving),

5.2.1.2 Mobile Broadband for Indoor scenario.

Rationale for proposing Edge Computing: The low latency and reliability aspects relate to potential uses already included in the NGMN WP and reproduced in Section 5.1 of this document (i.e. Smart Office).

5.2.2 Localised content use cases

5.2.2.1 Mobile Broadband for Indoor scenario.

Rationale for proposing Edge Computing: The local storage and delivery of content aspects are also relevant for potential uses already included in the NGMN WP and reproduced in Section 5.1 of this document (i.e. Smart Office).

5.2.2.2 Best connection per traffic type

As mentioned in the 4G Americas white paper: "With the advent of small cells in indoor environments such as offices, there is a need for some traffic to be routed locally while other traffic needs to access MNO or third-party services".
In this use case a user has two applications running, one voice and one video streaming application. The two applications have very different requirements, as one is generating low volume, real time traffic that needs to access MNO services, and the second requires much higher data rates and access to the closest Content Distribution Network (CDN). If the user is in the coverage area of multiple cells, the best cell for the given application should be used, so that the traffic is routed in optimal manner.

Rationale for proposing Edge Computing: Some localised or locally stored content should be delivered over the local connection to the edge server, while internet, speech, etc. can be offloaded to the WAN.

5.2.2.3 In-network and device caching.

Deploying in-network content caching at the edge is an effective way to deliver video, webpages, etc. and;

1) provides a better user experience (lower latencies and channel switching times) for the end-user,
2) allows the operators to dimension their network and backhaul more cost-effectively and
3) in some scenarios, efficiently utilize its limited radio resources.

This use case is related to the use case category #14 in Annex A of the NGMN white paper [1] and to the technology building blocks, "UE centric network" and "smart edge node", in Annex D of the paper.

Rationale for proposing Edge Computing: Local caching/low latency. The local cache could be on an edge server.

5.2.2.4 Routing path optimisation when server changes

The immersive services such as augmented reality, virtual reality, ultra-high-definition (UHD) 3D video have critical requirement on transfer bandwidth and delay.
In order to ensure good user experience, the server near to the end-user may be utilized to serve these types of services, and the operator network needs to ensure optimized data path between end-user and server to address the immersive services requirement on delay, for example, based on the terminal and server location.
Subject to the service agreement between the operator and the service provider, the 3GPP network shall enable hosting of services (including both MNO provided services and 3rd party provided services) closer to the end user to improve user experience and save backhaul resources.
The 3GPP network shall be able to support routing of data traffic to the entity hosting services closer to the end user for specific services of a UE.

The 3GPP network shall support efficient user-plane paths between a UE and the entity hosting the service closer to the end user even if the UE changes its location during communication.
The 3GPP network shall be able to support charging, QoS, and Lawful Interception (LI) for services hosted closer to the end user.

Rationale for proposing Edge Computing: The requirements above also apply to edge servers if “the entity hosting the service closer to the end user” is an edge server.

5.2.2.5 Wireless briefcase.

Rationale for proposing Edge Computing: This use case implies a distributed cloud, local content caching, and low latency. Those aspects relate to potential uses already included in the NGMN WP and reproduced in Section 5.1 of this document (i.e. Smart Office).

5.2.2.6 Cloud robotics.

Rationale for proposing Edge Computing: Low latency, compute offloading. This is similar to the Collaborative Robots use case from the NGMN White Paper reproduced in Section 5.1 of this document.

5.2.3 Performance optimisation use cases

5.2.3.1 Flexibility and scalability.

The system shall support dynamic utilization of resources (compute, network and storage resources) in more than one geographic area in order to serve the differing needs of the users in each geographic area, subject to operator policy.
Using resources (compute, network and storage resources) in more than one geographic area by the system shall be supported without requiring manual re-configuration of neighbouring nodes, without service disruption, and while avoiding additional signalling due to unnecessary UE’s re-attachments (e.g. due to loss of call state information in the network).

Rationale for proposing Edge Computing: Users of applications hosted at the edge could get service from edge servers instantiated at the eNodeB, an aggregation point in the mobile RAN, a BNG, etc. This use case requires compute resources in different geographical areas to be flexibly deployed and scaled without loss of service or call state information and that could also apply to Edge Computing resources. Edge Computing allows such resources to be placed in the optimum location to improve performance or to achieve the performance required by the Edge Computing applications themselves.

5.3 Other use cases

5.3.1 Low latency use cases

5.3.1.1 Object response-time critical applications

This use-case focuses on applications that have very bursty traffic behaviour where request response delays are very critical. In the future next generation network there are a number of use-case scenarios where such applications behaviour is identified (for example in Mobile Cloud computing, Remote Object control, Web acceleration). The motive for this use case is to optimize the response time to have an object transferred with as

short a delay as possible. The total transfer time of an object can be composed of many separate time segments, including potential processing of the object before response is sent (however the actual processing time is not considered in this use-case).
With a significantly higher air-interface throughput, the difference between a) what theoretically could be achieved and b) what can be achieved (consider todays network solutions and transport protocols) is large. An Edge Cloud computing platform located in connection to RAN can make use of local properties such as:

a) bursty traffic and instant high bitrate for transfer of large data units
b) separating the queues for current transport protocol friendly traffic and bursty instant high bitrate traffic.

The above applications may over time send fewer objects and thus generate a lower average rate. A local edge computing platform may therefore support substantially improved transfer performance compared to a traditional OTT application that is limited by an e2e requirement to comply with current transport protocol fairness.

Rationale for proposing Edge Computing: Low latency in object response time for delay critical applications.

5.3.2 Localised content use cases

The future of 5G should enable many different use-cases and for example a VPN-service could make use of the protocol transparency in RAN.

5.3.2.1 Site2site transport: Enterprise Site2site communication over Cellular based on Ethernet.

In this scenario, current Ethernet-protocol transport is used. The possible edge-computing functionality is: integration into enterprise networks, SLA-monitoring.

Rationale for proposing Edge Computing: Site local deployment with integration into enterprise networks and improved QoE. Enabling local SLA performance monitoring capabilities.

5.3.3 Performance optimisation use cases

5.3.3.1 Optimization of applications

Some applications can use up-to-date indications from the RAN to improve the end-user QoE. Examples of such adjustments include adjusting video codec parameters, adjusting TCP congestion window, etc. However, given the latencies and rapid changing load variations involved in IP-packet delivery to a centralized cloud infrastructure, the derived application feedback from radio information is often out-of-date. Moreover, in many RAN implementations such indication is not available as a service consumable by 3rd party applications. In addition, without any application based feedback the RAN has less possibility to optimize the RAN performance for locally deployed applications.

Rationale for proposing Edge Computing: The availability of an application based feedback derived from indications on network conditions (RAN and other) in the mobile edge in a timely fashion can help optimize the communication for locally deployed applications.

6 HIGH LEVEL ARCHITECTURE AND SERVICE REQUIREMENTS

6.1 Low latency scenarios

The considerations in Section 5.1.1, 5.2.1 and 5.3.1 of this document imply the following.

6.1.1 Architectural Requirements

For V2I services the network shall provide mechanisms to place and operate edge computing capabilities in roadside units at the network edge.

The network shall provide mechanisms to place and operate edge computing capabilities in vehicles for autonomous driving services.

The network shall provide mechanisms to place and operate edge computing capabilities in secure enterprise environments for robotic manufacturing applications.

The network shall provide a sufficiently high number of concurrently active connections, high data rate, and low latency to enable pervasive video and data delivery for optical head-mounted displays, collaboration in 3D cyber-real offices or operating rooms (with both physical and virtual presence) and customers' support by hologram services.

Edge computing should be possible over licensed or unlicensed spectrum, or fixed access.

6.1.2 Service Requirements

For pervasive video services the video element should be delivered in a suitably responsive near real-time fashion.

The network shall provide mechanisms to place and operate edge computing capabilities in such a way that a user of tactile services does not perceive any difference in latency between locally and remotely delivered content.

6.2 Localised content scenarios

The considerations in Section 5.1.2, 5.2.2 and 5.3.2 of this document imply the following.

6.2.1 Architectural Requirements

The network shall provide mechanisms to place content relevant to edge computing in, and deliver it from, the most relevant point of presence after the terminal equipment.

The network shall provide a mechanism to broadcast specific local content to a defined local area with very low latency.

Note: existing broadcast/multicast methods should be considered without alteration.

The network shall provide a mechanism to broadcast encrypted content from a local edge computing system

The network shall be able to determine which content should be backhauled from the edge computing system to the core network and which should be delivered locally by the edge computing system, and deliver it accordingly.

The network shall be able to cache appropriate content in the edge computing system and deliver it to the local area from there.

Subject to the service agreement between the operator and the service provider, the network shall enable hosting of services (including both MNO provided services and 3rd party provided services) closer to the end user to improve user experience and save backhaul resources.

The network shall be able to support routeing of data traffic to the edge computing system for the edge computing services the UE subscribes to.

The network shall support efficient user-plane paths between a UE and the edge computing system even if the UE changes its location during communication.

The network shall provide a mechanism to ensure that encrypted unicast, multicast, or broadcast content to be transmitted to the UE can be forwarded from one local edge computing server to another one, in order to have a seamless reception, without the loss of data on the UE due to the unavailability of appropriate key material to decrypt the content.

The network shall be able to support charging, QoS, and Lawful Interception (LI) for services hosted in the edge computing system.

6.2.2 Service Requirements

No service requirements have been identified for these scenarios.

6.3 Performance optimisation scenarios

The considerations in Section 5.1.3, 5.2.3 and 5.3.3 of this document imply the following.

6.3.1 Architectural Requirements

Users of applications hosted at the edge shall be able to receive service from edge servers wherever they are instantiated in the access or core network.

The network shall provide mechanisms to flexibly deploy and scale compute, network, and storage resources required by the edge computing system in different geographical areas without loss of service or call state information.

The network shall be able to place Edge Computing capabilities in locations which deliver the required improvement in performance, or which deliver the level of performance required by the Edge Computing applications themselves.

Use of compute, network and storage resources in more than one geographic area by the edge computing system shall be supported without requiring manual re-configuration of neighbouring nodes, without service disruption, and while avoiding additional signalling due to unnecessary UE re-attachments.

The system shall allow edge computing applications to provide information on network conditions (RAN and other) in their locality in a timely fashion in order to assist in the optimisation of communication for other applications.

6.3.2 Service Requirements

No service requirements have been identified for these scenarios.

7 ABBREVIATIONS

BNG	Broadband Network Gateway
IP	Internet Protocol
MNO	Mobile Network Operator
OTT	Over-The-Top
RAN	Radio Access Network

SLA	Service Level Agreement
TCP	Transmission Control Protocol
UE	User Equipment
WAN	Wide Area Network