Network Requirements for Latency-Critical Services in a Full Cloud Deployment

Höst, Stefan; Tärneberg, William; Ödling, Per; Kihl, Maria; Savi, Marco; Tornatore, Massimo

Published in:
SoftCom 2016

DOI:
10.1109/SOFTCOM.2016.7772160

2016

Network Requirements for Latency-Critical Services in a Full Cloud Deployment

Stefan Höst*, William Tärneberg*, Per Ödling*, Maria Kihl*, Marco Savi†, Massimo Tornatore†
* Department of Electrical and Information Technology, Lund University, Lund, Sweden
† Department of Electronics, Information, and Bioengineering, Politecnico di Milano, Milano, Italy
Emails: {stefan.host; william.tarneberg; per.odling; maria.kihl}@eit.lth.se, {marco.savi; massimo.tornatore}@polimi.it

***Abstract*—The end-point of the cloud trend is that all computational and storage resources leave the homes and move to the cloud. We can be rather certain the future eventually looks like this as IT and telecommunications services are commoditised and the end users' time, interest and skill to acquire and maintain electronics fade away. The network architecture suitable for such scenario will be analysed in terms of latency critical services to establish delay requirements and feasible localisations of the data centers. It is concluded that for cloud gaming as an example of a time-critical service, the maximum delay is limited to 20 ms. In the network architecture this corresponds to locating the data center in the Main CO, i.e. typically in the aggregation part of a metropolitan network.**

I. Introduction

In this paper we describe a scenario, which is the end-point of the cloud trend, bringing essentially all the computational power out from the user environment, and concentrated in a Data Centre (DC). There are several advantages with this, both from an operator point of view and for the user. In many cases the user does not have neither the time, skill or interest to maintain the computers in a normal home network. If the customer instead buys the service *Computer* from the DC provider, they will be guaranteed usability and security for the system. The operators, today spending noticeable time and efforts on faults origin in the home environment, can rely on functional user environments.

Cloud computing is an essential part of future network as described by both Cisco's Internet Business Solutions Group (IBSG) list of technology trends [1] and Bell lab's future networks [2]. As access networks become capable of higher data rates, using e.g. Fibre to the Home (FttH) or the next generation copper access G.fast, the networks will be able to accommodate interactive real-time content at scale. When this technology shift becomes more pronounced, it will also impact the traffic patterns over the networks.

Another clear trend listed in both [1] and [2] is that users become more mobile, both in the fixed network by using e.g. WiFi, and in the mobile networks. Currently, fixed network traffic widely exceeds mobile traffic, but this is about to change as the annual traffic growth is essentially higher for mobile networks than for fixed [3]. When performance and price differences between fixed and mobile access diminish, end users will become less conscious of how they access the Internet. To meet this trend the fixed and mobile networks need to converge into one single network, as anticipated in e.g. [4].

To support novel services and technologies the edge cloud computing paradigm has been proposed. The paradigm goes by many names such as; *Fog Computing* [5], *Telco-Cloud* [6], *Mobile Cloud* [7], [8] or *Mobile Edge Computing* [9]. Nevertheless, edge cloud computing complements the prevailing centralised infrastructure by distributing cloud computing capacity through the core and access networks. For example, a neighbourhood's required compute capacity is aggregated in a shared DC, accessible at low latency with a smaller global traffic footprint.

However, current network structures where the IP edge is located far away from the user, do not give convenient support for this type of solutions. The IP edge is the frontier in the network where IP payload is first available. In many cases it is located in the network core or at its border. From there down to the end user the content is passed through the network using tunnelling, most often encrypted. For the mobile network the IP edge is at the Packet Gateway (PGW) in the Evolved Packet Core (EPC), located in the core network, while the fixed network IP edge is normally in the Broadband Remote Access Server (BRAS), typically at the border between the aggregation and core networks, see Fig. 1.

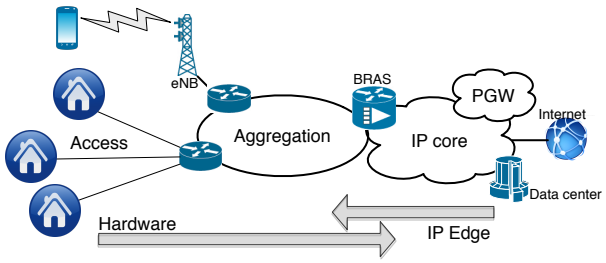

Fig. 1. Change in the network structure to meet the changes in the user behaviour in a cloud scenario.

In this paper we address the challenge of where to locate the IP edge to meet forthcoming application requirements. In other words, how far up in the network the IP edge can be located to still be considered close to the user from a latency perspective. In the study cloud gaming will be used as an example of a time critical services. Then it will is concluded that the total delay seen by the user from input to screen update should

not exceed 20 ms. Comparing with typical link delays that means the DC should preferably be located in the aggregation network, and not as high up as the core network.

The rest of the paper is organised as follows. In Section II the full cloud scenario is described in more detail. The challenge of locating the IP edge is approached in Section III by studying the network and user requirements for a very high demand cloud application, namely cloud gaming. Finally, the work is concluded.

II. The Emerging Cloud Computing Paradigm and its Abilities

The aggregate compute power in a typical household today is over-provisioned. Desktops, laptops, gaming consoles, tablets, televisions, and peripherals are left virtually unused for the better part of the day. Cloud computing introduces a new computing paradigm where compute resources are aggregated in DCs, which end users can use to run Internet services. The infrastructure is managed by the cloud operator, the capacity of which is seemingly infinite, and the end user is agnostics to where and how their content is being produced and delivered. This paradigm shift allows for user equipment to be made much simpler as the software run and maintained in a DC by the Cloud operator and the application owner, respectively.

When taking the cloud computing ideas to its end-point, essentially all computer power is moved from the User Equipment (UE) to the cloud infrastructure at the DC. What remains of the UE is mainly a terminal supporting the interfaces for the user, such as keyboard, mouse pointer and touch screens. Anything more advanced than moving the pointer on the screen is performed at a blade in the DC. Consequently all updates of the screen are performed in the cloud and sent to the user equipment as video, as in e.g. Amazon AppStream [10].

Real-time applications naturally impose strict latency requirements on the connectivity between the DC and the user. To provide the reader with a contrast, for example, cloud-based word processing services has latency demands in the same order as writing, which can be in the order of some hundreds of milli seconds. The requirements on screen updates are also fairly low rate and thus transmitting a new image once every key-stroke is not crucial. However, when considering more time critical services such as gaming or augmented reality, the requirement goes way beyond what contemporary cloud infrastructures can deliver.

The notion of employing pervasive and ubiquitous computing [11] for all our computing needs is coming to fruition. We have had real-time collaborative cloud applications such as word processing and spreadsheets for a while. Even though the more latency critical services, like Cloud gaming [12], is still in its infancy, there are examples of quite mature tests implemented, e.g. [13]. These services are pushing the boundaries of what is possible with contemporary infrastructure.

A. Network architecture for cloud applications

As stated above, today's network architectures are not designed for a full cloud deployment where the latency requirements can be crucial. With the IP edge at the BRAS for the fixed network and at the PGW for the mobile, the DC must be located above these, see Fig. 1. Hence, emerging cloud applications like mobile-offloading [14] and cloud gaming [15] are today implemented in distant centralised DCs beyond the IP edge.

For the full cloud deployment, the computer power should be placed in a DC above the IP edge. At the same time, the network trends to converge the fixed and mobile networks, and to move the IP edge further down closer to the user, are important to achieve acceptable latencies. Hence, the ideas of cloud computing where computation is relegated from the UE into a DC in the network should be reciprocated by the network operator by moving the IP edge further out towards the user, as shown in Fig. 2. In that way the user hardware can be positioned at a DC located close to the IP edge, and still comply with the requirements. In [16] the common IP edge is located at a converged functional entity called Universal Access Gateway (UAG), which plays an important role in a converged network structure. In 3GPP there are openings for positioning local PGW closer to the user by means of SIPTO [17]. In most cases the control plane (CP) for the mobile connection, e.g. assuring authentication or key exchange, does not have the same requirements and can still be located in the core network, denoted Mobile CP in the figure.

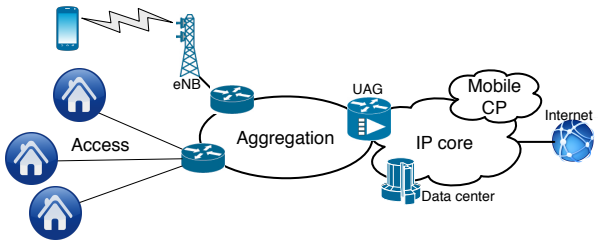

Fig. 2. Change in the network structure to meet the changes in the user behaviour in a cloud scenario. The UAG contains a local PGW and represent the IP edge for both the fixed and the mobile network.

B. Network requirements for cloud applications

Different cloud applications require different network performance metrics. For many of them there are no direct problems with the current situation, while others are more latency critical and will have issues with the latency in current networks. For example, in cloud based word processing the screen is essentially a still image and changes every time a key is pressed. This means an average of at most a couple of updates per second. It is simply a matter of asynchronously replicating the keystrokes, requiring a quite low data rate, and is relatively insensitive to delays. At the other end of the scale are latency critical real time applications like cloud gaming, augmented reality, tacile Internet or vehicular communications. In this paper cloud gaming is used as an example of latency critical services to get a reasonable bound on what delays are acceptable by the end users.

For this purpose, first it is important to make a technical distinction between on-line gaming and cloud gaming. In

on-line gaming the gaming hardware is colocated with the gamer and all computation is done locally. Somewhere on the Internet there is a Multi-Player Server (MPS) that for example aggregates the players movements, actions, and scores. This data is asynchronously transmitted to all players, which is then fed back into the game dynamics on the local hardware to reflect the current state of the game. A large delay will result in inaccurate local representation of the state.

In cloud gaming both the game mechanism and the rendering of the graphics are supposedly moved to a cloud based Game Server (GS). What remains at the user-end is a set of controllers, a receiver/transmitter, and video decoder. Cloud gaming has the potential to usher in a new era of game distribution and accessibility. Resources will be better utilised, globally, as a few shared resources can produce the equivalence of the many locally distributed, as the shared infrastructure can achieve a better level of economies of scale. Naturally, letting the user equipment be represented at a DC, there can still be connection from the DC to the MPS from here, see Fig. 3. There the UE is connected as a terminal to the GS, where all the game processing and video rendering is performed. Then the GS is connected via the Internet to the MPS where the multi-player functionality is aggregated. The time requirement on the first loop, the gaming loop from the UE to the DC, is much harder than that to the MPS.

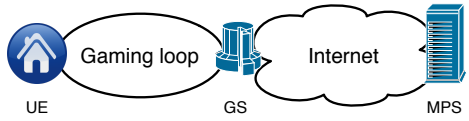

Fig. 3. A cloud gaming setup for a multi-player game.

Typically, gamers are worried about two parameters that affects the QoS for the game, namely the ping time and Frames Per Second (FPS). For online gaming the ping time is the Round Trip Time (RTT) to the MPS, and if this is too high the screen image is not corresponding to the view in the server, often referred to as lag. The movements are relatively slow and normally it is not a problem with a ping time of 70-80 ms [18]. The FPS refers to how often the screen is updated, and can thus be viewed as the sampling frequency of the game. The question of the required FPS is often debated on gaming forums. Clearly it depends on the type of game, but the measure should be between 30 and 60 Hz. Slow games like simulation games and some strategy games can cope with the lower update rate, while more time critical games like First Person Shooter Game (FPSG), Third Person Shooter Game (TPSG) and racing often require rates of up to 50 Hz or 60 Hz. Normally role playing games, like Massive Multiplayer Online Role Playing Game (MMORPG), are in the middle requiring 40-50 Hz update frequency.

III. Requirements on network delay

The latency of the gaming loop, from user input until screen update, includes the network delay, i.e. RTT, the game processing time and the video coding in the system. It is not well known in the literature how the delay in the gaming loop affects the user Quality of Experience (QoE) for cloud gaming. Jarschel *et al.* [19] present a set of QoE measurements for varying delay and packet loss. In the study users never rated a fast game better than fair with delays at 100 ms, even for average gamers. For the technology to be commonly accepted the delay must be negligible in the gaming experience. Already the fact that online gaming requires a RTT of 70-80 ms suggests the gaming loop latency for cloud gaming should be considerably lower than 100 ms.

To formulate realistic delay requirement bounds we consider the FPS as the sampling frequency of the system, i.e. the game itself. A low sampling frequency and a high delay will have the same effect on the QoE; the game will not run smooth and the reaction time will suffer. To study the sources of delay we turn our attention to the intermediate network. Fig. 4 depicts the cloud gaming network architecture. The UE, is attached to the Residential Gateway (RGW), which is connected to the Internet. For example, in mobile access the RGW is replaced by a specialised router, but since games are often played in the home, we retain the term RGW for both fixed and mobile access. The GS is located somewhere in a DC, and constitutes the equivalence of either a gaming console or computer. The GS can then be connected to an on-line server for multi-player games, as usual.

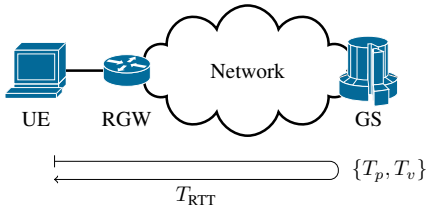

Fig. 4. Cloud gaming architecture and delay.

A significant delay in the system is the time from a user interaction to its effect is shown on screen. For the architecture in Fig. 4 that means first the signal is sent to the GS, where it is processed and the screen video updated, encoded and sent back to the UE. The time for transmission in the network is denoted by T_{RTT}. The processing time for the game is T_p and the video coding, in total at both the GS and the UE, is T_v. Thus, the total gaming loop delay is given by

$$T = T_{\text{RTT}} + T_p + T_v$$

As long as the perceived delay for the user is not dominated by the delay T, compared to the FPS, the game will run smooth and the structure can be accepted. In control theory, when applying a digital controller to an analogue system, it is a rule of thumb that the sampling time should be less than the delay [20]. If the network delay in Fig. 4 is at maximum $T_{\max} = 1/\text{FPS}$ the update will be not be delayed more than one sample. This is also the minimum delay that can be guaranteed for a sampled system. Hence, it is reasonable

to assume that the user will not be able to notice the delay, in terms of degraded QoE, if the network delay is not exceeding $T_{\max}$.

In Table I typical types of games are listed together with their requirements on FPS and what it implies in terms of maximum delay. Naturally, the classification of game types and FPS varies from game to game and should be seen as target values. Most games will work smoothly on FPS=50 Hz, even though we here have listed up to 60 Hz. Home computer screens are normally updated with 60 Hz and there is little use in exceeding this value in the update from the gaming equipment. In this work we are focused on the upper bound, as shared infrastructures always need to be scaled and engineered to satisfy highest requirements amongst all of its clients.

TABLE I
TABLE OF MAXIMUM TOLERATED DELAYS FOR DIFFERENT FPS AND TYPES OF GAMES. DELAY TIMES ARE MEASURED IN MILLISECONDS.

FPS	Type of game	$T_{\max}$	$2T_{\max}$	$4T_{\max}$
30	Simulation, building	33	67	133
40	Sport, MMORPG	25	50	100
50	TPSG, FPSG, racing	20	40	80
60	FPSG	17	33	67

The values in column $T_{\max}$ in Table I can feasibly be considered as the requirements to achieve no noticeable delay in the loop, which should be satisfactory even for skilled gamers. For average gamers, in [19] claimed as the majority, the delay can probably be set slightly higher without any considerable quality degradations. Thus, the quality should still be satisfactory even if the delay approaches two screen updates, which is reflected in the column $2T_{\max}$. However, increasing the delay even further to e.g. $4T_{\max}$ as in the table, will give delays in the order of the limit for online gaming, which will degrade the perceived quality of the game. In the continuation of this study we will assume a delay requirement of one screen update at 50 Hz, i.e. 20 ms.

A. Data Center location

Viewed from the network architecture, the requirements on delay can be translated to geographical points in the network. In Fig. 5 typical locations of key infrastructure from both the fixed and mobile networks are shown. The evolved NodeB (eNB) and Central Office (CO) are located close to the user and normally constitute the last mile access. The Main CO is typically in the aggregation network, as a central node in the metropolitan network. The Core CO and the PGW are located at the border to, or in, the core network, far away from the user. For the PGW, as part of the EPC the 4th generation packet network, there are typically a handfull locations in a country. The Core CO, Main CO, or even the CO can all be seen as candidates for hosting the IP edge, and therefore constitute reference possible locations for DC placements. Another possibility, claimed by the ETSI MEC working group [9] is to locate the DC in the eNB. To determine which of these network locations are feasible to support the latency requirements discussed in the previous section, we need to find the point where network transmission delay is not dominant, i.e. where $T_{\text{RTT}} \ll T$.

Fig. 5. Equipment and locations in the network.

To get estimations of the RTT at different positions in the network we first consider the access network. In Table II typical RTT delays are given for the most common access technologies.

TABLE II
TYPICAL RTT FOR DIFFERENT ACCESS TECHNOLOGIES.

Technology	RTT
ADSL2+/VDSL2	20 ms
G.fast	1-2 ms
FttH	1-2 ms
LTE	10 ms
5G	1-2 ms

Home Internet connections are still predominantly over copper connections, using ADSL2+ and VDSL2, but fibre solutions, FttH, are growing rapidly. Starting with the traditional copper based access methods, ADSL2+ and VDSL2, they have a built in delay of roughly 20 ms for RTT to the CO, which will dominate the game loop delay. The delay is mostly due to interleaving, set by the operator to protect for e.g. impulse noise on the cable, but also without this the latency is relatively high. The next generation copper based access, G.fast, utilising up to 250 meters of the copper loop will have a much lower delay, at 1-2 ms to the CO. The same delay is typical for a fibre access, FttH.

For gaming as well as other cloud applications, it can also be interesting to access over a mobile connection, opening for advanced games in Hand Held devices (HH). However, a typical delay for the access part in LTE is in the order of 10 ms, which is due to the RAN scheduling. However, considering the next generation mobile system, 5G, the aimed delay is similar to FttH, in the order of 1-2 ms. That means that also for LTE it is doubtful that the RTT is much less than 20 ms. The above delay estimations is provided for the average case, and does not take into account effects due to e.g. link congestion or insufficient home networking, that could also severely affect the total delay.

Summarising, all technologies for the next generation access, such as FttH, G.fast or 5G, have a delay at 1-2 ms, which can be seen as considerably lower than the stipulated 20 ms. For legacy copper technologies ADSL2+ and VDSL2, as well as for LTE, the RTT over the access part cannot be seen as negligible for the total required delay over the gaming loop.

The next step is to consider the delay in the aggregation part of the network, i.e. from eNB or CO to the Main CO. Normally this is well below 1 ms, and thus the RTT from the

UE to the Main CO is dominated by the access delay. The corresponding RTT from the CO to the Core CO is in the order of 5 ms. Even though the variation is large for different locatons, it is more doubtful that the RTT to the Main CO can be claimed much less than 20 ms. Hence, for latency critical cloud services, like cloud gaming, the DC should not be located higher in the network than the Main CO. If it is located further up, e.g. close to the Core CO or the EPC, there is a substantial risk that the user Quality of Service (QoS) will diminish.

Since the RTT difference between the CO and the Main CO is very small, it is reasonable to locate the GS at the Main CO, thus maximise the number of aggregated users for the DC, and thus the utilisation degree. The RTT requirements are harder to maintain by positioning the GS at a higher position like the Core CO, due to the extra latency in the network. It is also reasonable that the IP Edge can be located at the Main CO, making the architecture feasible.

IV. Conclusions

In most home networks the computer power is essentially unused. In this paper a full cloud scenario has ben described where the computer power is moved to DCs in the network. Interactions from the user is sent to the DC and screen updates as videos are sent back to the UE. This will increase the hardware utilisation and avoid user maintenance.

The allowed RTT in the network has been estimated by considering the time critical services cloud gaming. It has been estimated that the total latency, from input to screen update, for such service should not exceed 20 ms. With low latency access technologies like fibre (GPON), LTE or G.fast, the RTT is well below this stipulated time limit. The location of the DC can then be as far up in the network as the Main CO in the aggregation part of the metropolitan network. This is also a reasonable location in the network to locate the IP edge for the UE, which today typically is in the core network.

Acknowledgements

The research was in part funded by the EU FP7 project COMBO, grant agreement 317762, the EU H2020 project 5G-CrossHAUL, grant agreement 671598, the EIT Digital, the Swedish Research Council (VR) for the project Cloud Control under contract C0590801 and Lund Center for Control of Complex Engineering Systems (LCCC), the Excellence center ELLIIT, the EUREKA/CELTIC project GOLD nationally funded in Sweden by VINNOVA, and Mobile and Pervasive Computing Institute Lund University (MAPCI). The work was done while William Tärneberg was hosted by the University of Virginia.

References

[1] S. Taylor, "The next generation of the internet," CISCO Point of view, April 2013.

[2] M. K. Weldon, Ed., *The Future X Network: A Bell Labs Perspective*. CRC Press, 2016.

[3] CISCO, "Cisco visual networking index: Forecast and methodology, 20142019," White paper, May 2015.

[4] S. Gosselin *et al.*, "Fixed and mobile convergence: Needs and solutions," in *European Wireless 2014*, Barcelona, Spain, May 2014.

[5] F. Bonomi, R. Milito, J. Zhu, and S. Addepalli, "Fog computing and its role in the internet of things," in *Proceedings of the First Edition of the MCC Workshop on Mobile Cloud Computing*, ser. MCC '12. New York, NY, USA: ACM, 2012, pp. 13–16. [Online]. Available: http://doi.acm.org/10.1145/2342509.2342513

[6] P. Bosch, A. Duminuco, F. Pianese, and T. L. Wood, "Telco clouds and virtual telco: Consolidation, convergence, and beyond," in *Integrated Network Management (IM), 2011 IFIP/IEEE International Symposium on*. IEEE, 2011, pp. 982–988.

[7] N. Fernando, S. W. Loke, and W. Rahayu, "Mobile cloud computing: A survey," *Future Generation Computer Systems*, vol. 29, no. 1, pp. 84–106, 2013.

[8] H. T. Dinh, C. Lee, D. Niyato, and P. Wang, "A survey of mobile cloud computing: architecture, applications, and approaches," *Wireless communications and mobile computing*, vol. 13, no. 18, pp. 1587–1611, 2013.

[9] "MEC, Mobile Edge Computing, ETSI Industry specification group," Started Dec 2014.

[10] Amazon web services. [Online]. Available: https://aws.amazon.com/appstream/

[11] M. Satyanarayanan, "Pervasive computing: Vision and challenges," *Personal Communications, IEEE*, vol. 8, no. 4, pp. 10–17, 2001.

[12] C. Moreno, N. Tizon, and M. Preda, "Mobile cloud convergence in GaaS: A business model proposition," in *System Science (HICSS), 2012 45th Hawaii International Conference on*. IEEE, 2012, pp. 1344–1352.

[13] Sony Playstation Now. [Online]. Available: http://us.playstation.com/playstationnow

[14] K. Kumar, J. Liu, Y.-H. Lu, and B. Bhargava, "A survey of computation offloading for mobile systems," *Mobile Networks and Applications*, vol. 18, no. 1, pp. 129–140, 2013.

[15] R. Shea, J. Liu, E. Ngai, and Y. Cui, "Cloud gaming: architecture and performance," *Network, IEEE*, vol. 27, no. 4, pp. 16–21, 2013.

[16] COMBO project. A universal access gateway for fixed and mobile network integration. White paper, Sept. 2015. [Online]. Available: http://www.ict-combo.eu

[17] *Local IP Access and Selected IP Traffic Offload (LIPA-SIPTO)*, 3GPP TR 23.829 Std.

[18] M. Claypool and K. Claypool, "Latency and player actions in online games," *Communications of the ACM*, vol. Vol. 49, pp. 40–45, 2006.

[19] M. Jarschel, D. Schlosser, S. Scheuring, and T. Hossfeld, "An evaluation of QoE in cloud gaming based on subjective tests," in *Fifth International Conference on Innovative Mobile and Internet Services in Ubiquitous Computing*, 2011.

[20] K. J. Aström and R. M. Murray, *Feedback systems: an introduction for scientists and engineers*. Princeton university press, 2010.