Quality of Service Framework for Supporting Next Generation Mobile Services

Ramneek, Korea University of Science and Technology, Daejeon,Korea. E-mail:ramneek@kisti.re.kr

Patrick Hosein, The University of the West Indies, Trinidad and Tobago.E-mail:patrick.hosein@sta.uwi.edu

Wonjun Choi, Korea University of Science and Technology, Daejeon,Korea.E-mail:cwj@ust.ac.kr

Woojin Seok*, Korea Institute of Science and Technology Information, Daejeon, Korea. E-mail:wjseok@kisti.re.kr

Abstract---

Background/Objectives: Wireless and cellular networking technologies as well as the corresponding mobile devices have evolved significantly over the past few years. Hence their use in future internet services such as mobile cloud computing, high performance mobile computing, 3D imaging and hologram service, the Internet of Things and other real-time services has emerged as a key area of research.

Methods/Statistical analysis: Such applications involve the processing and dissemination of a large amount of data among the mobile users. Therefore, to support such applications, a high throughput, low latency, reliable network with low packet loss rates is needed. As mobile data traffic has been increasing rapidly, serving such applications with stringent QoS requirements is a challenging task.

Findings: Therefore there is a need for strict QoS guarantees to efficiently support such applications over existing wireless networks. Such applications not only require ultra high data transfer speeds and low latency, but also seamless connectivity across multiple devices as well as high reliability and fault tolerance. There is also a need for flexible, QoS-based pricing as this can directly influence user behavior and so be used to manage resources more efficiently.

Improvements/Applications: In this paper we discuss various requirements and challenges for ensuring QoS for supporting advanced mobile applications and services over existing wireless networks. We also propose a QoS framework for supporting these services over existing and future wireless networks and briefly discuss how it can be implemented using dynamic pricing, admission control, congestion control and optimal resource allocation.

Keywords---QoS, Mobile Cloud Computing, High Performance Computing, Pricing, Cellular Networks.

*Corresponding Author

Special Issue on "Engineering and Bio Science"

I. INTRODUCTION

Wireless networks have evolved to support resource intensive applications such as interactive multimedia and video streaming as well as traditional services such as email, web and voice, over the same network infrastructure. The number of mobile devices and the overall mobile data traffic has been increasing exponentially over the past few years and it is expected to increase even more rapidly in the future[1]. The total mobile data traffic is expected to increase up to 30.6 exabytes per month by 2020. The main contributing factors include advancement in network technologies capable of providing high data rates and QoS guarantees, affordability of smart mobile devices including cell phones, tablets, etc., and the introduction of advanced data intensive applications. The cellular networks have been evolving providing higher performance and capabilities when compared to first and second generation networks. The Third Generation of wireless networks (3G) had been optimized to enable better connectivity and mobile broadband services. The Fourth Generation Networks (4G), including LTE and LTE-A is further capable of providing more capacity and higher data speeds for enhanced mobile broadband experience. The future Fifth Generation wireless systems (5G) will further revolutionize mobile communications. It is not only focused on enhancing the mobile broadband experience (higher capacity and data rates), but also on providing low latency, high reliability and massive machine to machine (M2M) communications[2].

Hence the use of advanced wireless mobile networks and mobile devices in future internet frameworks such as mobile cloud computing, high performance mobile computing, 3D imaging and hologram service, Internet of Things (IoT), and other real-time services has emerged as a key area of research interest. Such applications are characterized by high processing power, data intensive computations, high data storage and access rates, high data transfer speeds, reliability and accuracy.

Hence, for supporting advanced applications that require stringent QoS requirements over existing wireless mobile networks there is a need for a QoS-based framework, and dynamic pricing plans based on it, to maximize resource utilization while providing high user satisfaction. In this paper we discuss various requirements and challenges for ensuring QoS for supporting advanced mobile applications and services. In addition, we discuss how proper charging and resource allocation can be used to balance between resource consumption and the QoS provided. The paper is organized as follows: Section 2 describes some advanced mobile applications and their QoS requirements; Section 3 provides a brief overview of the QoS framework and discuss how QoS requirements of different services can be satisfied, followed by the conclusion and acknowledgement in Section 4 and 5 respectively.

II. ADVANCED MOBILE SERVICES

Due to the evolution of wireless technologies, many revolutionary applications and services will be developed for mobile devices in the next few years. These applications will combine the benefits of smart mobile devices with high computation power and memory, and advanced networks such as 5G. Some of the potential applications and their QoS requirements are discussed below:

Mobile Cloud Computing (MCC)

Cloud computing enables on-demand access to a number of computing and storage resources provisioned as a service to the end users. Cloud eco-systems have become a popular choice for scientific research collaborations where cloud services can be deployed by different service providers, at distributed locations, using different middleware software stacks thereby creating a heterogeneous eco-system. MCC integrates cloud computing and mobile computing, and hence it is important to overcome various challenges related to the environment (e.g., scalability, availability, heterogeneity) and performance (e.g., throughput and delay constraints, etc.). Hence network performance and QoS assurance plays a vital role in MCC. A lot of work has been done to enhance the performance and efficiency of dedicated cloud networks and collaborative research networks [3] in order to meet the demands of large scale resource intensive cloud applications. However, due to a rapid increase in the adoption and use of mobile devices, provisioning stringent QoS for supporting MCC over existing wireless mobile networks is still a challenging issue.

High Performance Computing on Mobile

High performance computing over mobile devices involves the integration of advanced wireless networks, smart mobile devices with high computational power, and advanced applications and services. Mobile devices offer high computational capabilities, including processing power, memory, energy efficiency, at relatively low cost. Hence they have become a popular choice for accessing HPC facilities even when users are mobile. Many firms,including Intel and NVIDIA [4], have been focusing on the optimization of mobile platforms for supporting advanced mobile services. In addition, some applications have also been developed to access HPC facilities via mobile devices[5]. From the mobile communication aspect, a high level of QoS support (high throughput and reliability with low latency) is needed to support data intensive workloads of HPC applications including parallel processing. Significant work has been done to optimize the performance of HPC cluster networks and data-centres but supporting such applications over mobile devices is still a challenging issue. Exisiting 3G and 4G networks and the future 5G networks are capable of provisioning high data rates and increased capacity and hence have the capability of supporting such HPC services. However, there is a need for an adaptive QoS framework[6] to overcome the challenges imposed by the intrinsic properties of mobile networks including interference, channel fading, mobility, etc.

3D Imaging and Hologram Service

Existing 3G and 4G networks are already capable of supporting high definition (HD) video calls and voice calls over LTE (VoLTE). In the next few years services such as 4K-UHD and 8K-UHD, offering 4 times and 8 times the resolution of full HD respectively, as well as 3D imaging and holograms will be available over mobile devices. With the ultra-high capacity, reliability and ultra-low latency that the future 5G networks are expected to provide, supporting such mobile applications will become possible. However, such applications have stringent QoS requirements. For instance, 3D holograms require huge bandwidths[7], and even its low bandwidth alternatives such as super multi-view stereoscopic images and computer generated holograms, will impose strict QoS constraints on the underlying network. According to the projection by IBM [8], cell phones and other mobile devices will be capable of providing 3D holograms, and 3D video telephony in real-time in next five years. For example, a real-time 3D hologram of a person or object, projected from the surface of the mobile devices. This will consume a large amount of bandwidth and a 3D moving image will further increase the amount of data required.

Internet of Things (IoT):

The Internet of Things can be defined as a network of physical objects with embedded communication capabilities and other features such as sensors, etc., enabling them to sense information and to interact with other objects and the environment. These objectsrange from simple devices such as cell phones, headphones, household machines (washing machine, microwave, coffee maker, etc.), wearable devices etc., to more complex ones such as monitoring sensors implanted in human bodies, automobile sensors, etc. Currently, IoT is being applied in many applications such as smart homes, smart cities, smart grids, traffic management, connected health, etc. In the future, when massive connectivity will be possible through the 5G infrastructure, IoT can be applied to other areas such as video surveillance, remote monitoring and control, etc. Such applications require massive connectivity and, although throughputs tend to be relatively small, the signalling required to support potentially billions of devices will be challenging. So in this case the signalling network infrastructure will be of concern.

Remote Surgical Procedures

Another potential application is the use of teleoperation in the field of medicine[9]. By using this technology, the surgeons can perform specialised medical surgeries and procedures remotely. This allows the experts to apply their expert knowledge remotely, without the need for on-site presence at the time of surgeries. The experts can see the images of the patient and remotely control the robot through their computer. Although such applications are still in their nascent state, their widespread use is expected with the advancement of technology and the communication networks.In addition, sensory information and arifical intelligence can further help to enhance the quality and precision of the remote operations. In order to carry on the remote operations accurately, there is a need of a high quality communication connectivity between the remote control station and the machines being operated. The current solutions either make the use of wired connections of wi-fi to implrmrnt the last hop of the communication link. Although these provide low latency and higher reliability, the cost of installation and maintainance is high. Hence, the cellular networks offer a number of benefits in terms of wider coverage and low cost of installation and management. However, such applications require low latenct and jitter with minimum bit rate guarantee. These applications involve HD video transmission and haptic interactions and hence such links require low latency and low jitter , while some packet loss can still be tolerated.

Other Applications

Some of the other applications include autonomous driving (vehicle to vehicle communication), industrial applications such as remotely controlling industrial operations using teleoperation, large scale immersive virtual reality services, big-data based intelligent services, etc.

III. QoS Framework

In this section, we describe, in brief, the QoS-based scheduling framework and discuss how QoS constraints for different applications can be satisfied. Note that we only focus on the bearer channels (i.e. the connections for the actual data) and not on the signalling aspects. However, as we noted above, some applications such as those for IoT, will require improvements in the Signalling aspects of the wireless network. As discussed in the previous section, advanced mobile applications have stringent QoS requirements. Hence there is need of an adaptive QoS control framework to ensure that the stringent QoS requirements of such applications are satisfied.

If we consider the shared resource allocation problem, as defined in[10], the utility function can be defined in terms of one or more QoS parameters such as throughput, latency, etc. The utility function can be defined in terms of throughput as follow:

$$maximize \quad F(\vec{r}) \equiv \sum_{i=1}^{k} (UT_i(r_i) \ (1)$$

$$subject\ to \quad \sum_{i=1}^{k} r_i < C \ (2)$$

$$and \quad r_i \geq \quad rmin_i \qquad (3)$$

$$over \quad r_i \geq 0, \ 1 \leq i \leq k. \qquad (4)$$

where,

k = the number of active users competing for the channel,

$UT_i(r_i)$ =Utility function of user i experiencing an average throughput r_i ,

C = the total channel capacity,

$rmin_i$ = the minimum throughput for user i ,

Since the cellular networks are dynamic in nature in terms of varying channel conditions, user mobility, intereference, channel fading, etc., the optimal solution can be found using the dual ascent method. If we consider the utility function as a linear function of the average throughput r, then it can be defined as

$$U(r) = \alpha r \qquad (5)$$

for some constant α.

In this case, the scheduler will pick the user with the maximum achievable bit rate. This results in an increased average sector throughput,however there is no consideration given to QoS and fairness. Now let us consider the traditional proportional fair algorithm (Describe what this is in more detail), In this case the Utility function can be defined in terms of throughput r as

$$U(r) = \log(r)(6)$$

Here some degree of fairness is ensured as the users in bad radio conditions are also served when their throughput drops significantly. However, this comes at the cost of reduced sector throughput. Hence, the utility function should be defined to guarantee the QoS as per the application requirements. One way of doing this is by including some Barrier function with the utility function[11]. Barrier functionscan help to penalize the movement into areas where the QoS constraint is violated.

For instance, the QoS demands of high performance mobile computing will impose throughput, delay, jitter and Packet Loss Rate (PLR) constraints on the underlying wireless networks. Hence there is a need to provide QoS guarantees for each such parameter. In the case of throughput and delay, the utility functions can be defined using the concept of a barrier function to enforce various constraints and the PLR can be maintained by using the Hybrid Automatic Repeat Request (H-ARQ) retransmission mechanism. H-ARQ is a combination of the automatic repeat request error control mechanism and the high-rate forward error correction coding. If a transferred block is received correctly, an ACK is sent. If no ACK is received or a NACK is received, then the transmitted block will be sent again. Two commonly used methods of H-ARQ for high speed data networks include chase combining and incremental redundancy.

For throughput, a lower bound on the achieved throughout (r_{min}) is required, and hence, a suitable Barrier function can be chosen such that the priority function value increases rapidly as the throughput approaches the minimum throughout threshold. Similarly Barrier functions can be defined in terms of delay such that the priority function increases rapidly as the delay approaches the maximum delay threshold[11]. Similarly, suitable barrier functions can be defined for different QoS constraints to guarantee QoS depending on the application requirements, as shown in Figure1. Advanced mobile services and applications can be classified based on the QoS constraints that are critical for each of them.

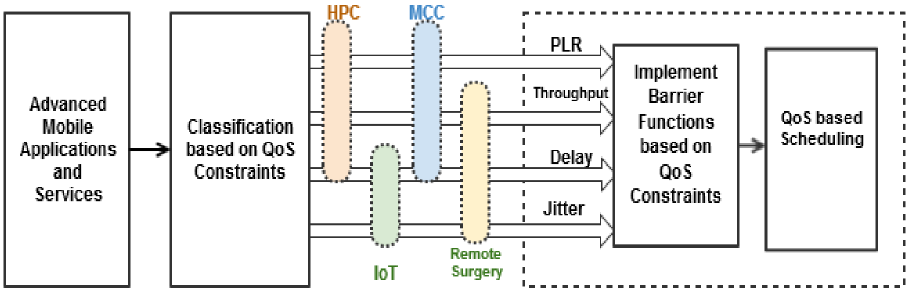
Advanced Mobile Applications and Services
Classification based on QoS Constraints
HPC
MCC
IoT
Remote Surgery
PLR
Throughput
Delay
Jitter
Implement Barrier Functions based on QoS Constraints
QoS based Scheduling

Figure 1: QoS Based Scheduling Framework

QoS-Aware Pricing

In the previous section we described the QoS framework and discussed how the constraints on different performance attributes can be satisfied for supporting advanced mobile applications and services over existing mobile networks. As pricing is an important part of the QoS control framework, we now discuss how proper pricing can be used to maintain a balance between user quality requirements, service efficiency and operator revenue.

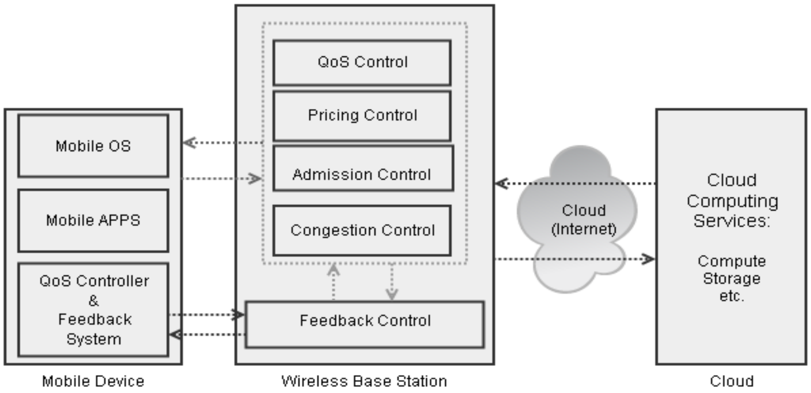

Figure 2: QoS Framework for High Performance Mobile Computing

We propose the use of tiered data plans, integrated with appropriate congestion management and admission control. We propose various data plans: A gold plan devised for advanced application users, and silver (QoS based) and bronze (best-effort) plans for normal users. The per bit price charged will be higher for the QoS plan as compared to the best effort plan, and the total price paid will depend on the QoS level as well as the monthly data cap.

In the case of the gold plan subscribers, high QoS (high throughput, low delay and jitter, and low PLR, depending on the application requirements) can be guaranteed, along with a high data allowance. In this case, there will be strict QoS guarantees and hence a higher cost per bit. When the monthly data cap is exhausted, the user can still achieve high QoS but will be charged an even higher cost per bit.

Apart from the differential pricing as described above, management of QoS in the case of congestion must also be addressed. Different congestion detection mechanisms can be used to monitor the load on the network. For instance, as proposed in[12], the average value of the priority over all users, (i.e., the average utility function gradient) can be calculated to estimate the current network load. If this exceeds some threshold then various actions can be taken (e.g., block new connection requests or reduce QoS constraints etc.). In addition, to avoid excessive charges when a user is not using a high performance computing application, they can switch to a lower QoS level on demand.

Next we provide an example of how the QoS based scheduling and pricing frameworks can be integrated to satisfy the QoS requirements of the advanced application users while optimizing the operator's revenue. The QoS control framework for Mobile Cloud Computing application and its various components are shown in Figure 2.

The QoS controller and feedback system provide the QoS information to the wireless base station at periodic intervals of time. The feedback controller at the base station collects the information received from various mobile nodes, analyzes the QoS information, and passes it on to the QoS control functions for dynamically adjusting the resources in order to meet the QoS requirements of the connections.

IV. Conclusion

Due to the evolution of wireless data networks, there has been an increasing demand for better Quality of Service and performance. Hence there is a need for network operators to adopt dynamic QoS and pricing strategies in order to support high performance mobile computing workloads. Such applications not only require ultra high data transfer speeds and low latency, but also seamless connectivity across multiple devices, reliability and fault tolerance. In this paper, we discussed various requirements and challenges and proposed a framework for addressing these requirements. The proposed method helps the network operator to manage the network resources more efficiently while maintaining a balance between user quality requirements and operator revenue. In addition, by providing features such as dynamic quality variation, the approach allows the end user to use the network services in a more flexible manner.

Acknowledgement

This work has been supported by the research projects: Computational Engineering Advanced Research (K-17-L01-C02-S04) of Korea Institute of Science and Technology Information (KISTI), Daejeon, Korea.

References

[1] Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update2015-2020. http://www.cisco.com. Date accessed:01/03/2016

[2] Balazs Bertenyi, 3gpp system standards heading into 5g era. http://www.3gpp.org. Date accessed: 20/02/2016

[3] P. Zhang and Z. Yan.A QoS-aware system for mobile cloud computing. 2011 IEEE International Conference on Cloud Computing and Intelligence Systems, Beijing, 2011, pp. 518-522.

[4] NVIDIA:Mobile phones, tablets and HPD (cloud)-Stream Computing. https://streamcomputing.eu/blog/2012-05-12/nvidia-mobile-phones-tablets-and-hpc-cloud/, May 2012.Date accessed: 01/04/2016

[5] C. V. Deepu, N. Kurkure, P. Dinde, A. Das, A. Gupta and G. Misra. E-Onama: Mobile high performance computing for engineering research. 2013 Third International Conference on Innovative Computing Technology (INTECH), London, 2013, pp. 532-536.

[6] Ramneek, P. Hosein, W. Choi and W. Seok.Quality of service support for high performance computing on mobile devices. 2016 International Conference on High Performance Computing & Simulation

(HPCS), Innsbruck, 2016, pp. 213-217.

[7] SKT Telecom's View on 5G Vision, Architecture, Technology and Spectrum. SKT Telecom 5G White Paper, http://www.sktelecom.com. Date accessed:20/08/2016

[8] Holograms on cell phones coming in five years, IBM predicts. http://www.computerworld.com/article/2512036/emerging-technology/holograms-on-cell-phones-coming-in-five-years--ibm-predicts.html. Date accessed:20/08/2016

[9] Industrial remote operation: 5G rises to the challenge - Ericsson. November 2015 Available:https://www.ericsson.com/thecompany/our_publications/ericsson_technology_review/archive/industrial_remote_operation. Date accessed:20/08/2016

[10] F. Kelly, Charging and rate control for elastic traffic. European Trans. On Telecommunications, vol. 8, pp. 33-37, 1997.

[11] P. A. Hosein, QoS control for WCDMA high speed packet data. 4th International Workshop on Mobile and Wireless Communications Network, 2002, pp. 169-173.

[12] Ramneek, P. Hosein and W. Seok. 1Load metric for QoS-enabled cellular networks and its possible use in pricing strategies. 2014 IEEE Symposium on Wireless Technology and Applications (ISWTA), Kota Kinabalu, 2014, pp. 30-35.