Resource Allocation Architecture with Improved QoS for SLA in Next Generation Mobile Systems

Sanjeev Shrivastava¹, D. S. Bhilare²
¹,²SCS, DAVV, Indore, MP, India

Abstract—The Internet Engineering Task Force is working on expanding and extending the TCP/IP model to support mobility and multimedia and real-time applications on the Internet. Thus, new protocols, amendment of existing protocols have been developed. Also there has been a significant increase of interest in enhancing quality of service (QoS) constraints in mobile systems.

The proposed network system architecture and mechanisms to provide real-time services to mobile users. Mobility experiences wide variations of quality of service and to obtain mobility independent service guarantees, a mobile system needs to make advance resource reservations. In this paper Transport Layer Management System (TLMS) is introduced at Transport Layer of TCP/IP reference model. TLMS ensure user’s service level agreement (SLA) by monitoring, controlling and allocation of resources at server side with the help of application layer and network layer adaptation engine.

Keywords—Mobile Systems, Quality of service (QoS), Resource Allocation, Service Level Agreement (SLA), Transport Layer Management System (TLMS)

I. INTRODUCTION

Now-a-days, tremendous researches have been done in the field of wireless network, Internet and other network architecture for supporting Quality of Services as these networks support a wide variety of services viz. file sharing, video streaming, voice communication, etc. [1] However most of them are not suitable for mobile environment because it is susceptible to bandwidth constraint and its dynamic topology presents challenges to achieve QoS. Supporting QoS in mobile environment is really difficult as every user’s links’ bandwidth is susceptible to get affected by the neighboring users’ links’ bandwidth and on the other hand it is known that mobile users move from one location to another frequently and thus they join, leave and re-join links which presents greater difficulty to QoS routing in mobile environment.

Keeping these drawbacks in mind, continuous work is being carried out in computing technology and wireless communication networks in order to guarantee smooth connection at network and transport layer so as to facilitate seamless services to users. [2,3]

The advancement of mobile devices and applications, also the continuous increase in the number of users and the wide use of MANET in battlefields and emergency situations necessitates the provision of real-time services taking into account the Quality of Service requirements viz. bandwidth, reliability, low jitters and less end-to-end delays. As the mobile user changes locations frequently that creates the necessity of resource management for guaranteeing end-to-end services.

Recent studies have provided numerous approaches and models to facilitate resource management efficiently supporting not just global mobility of mobile users but also enabling agile service creation and access agnosticism. These studies are been done keeping in account the perspective of both client and network. Some studies have also emphasized on the usage of Multiple Stream Reservation Protocol (MSRP) over Resource Reservation Protocol (RSVP) in order to provide real-time services by securing the bandwidth necessary to cater seamless service to mobile users.

This paper considers the bandwidth reservation problems in mobile environment and the problem is addressed by several works in the literature. Further, this paper provides a detail study on bandwidth reservation at transport layer that purports on allocating resources to users based on their QoS satisfaction level and SLA.

II. RELATED WORK

Kamaljit I. Lakhtaria [4] with an aim to enhance the end-to-end quality of Next generation Networks based communications, through his work, have proposed a service aware policy-based approach that employs all layers of the environment through parameterization, profiling, negation and arbitrating mechanisms. His approach takes into account the Next Generation Networks based on quality assurance QoS and Quality of Experience (QoE). The purpose of his work is to pursue the end-to-end controllability of the quality of the multimedia next-generation network (NGN) -based communications in an environment that facilitates end user’s access agnosticism, service agility and global mobility.
Keeping in consideration the bandwidth reservation problems, a route discovery protocol is proposed which makes utilization of both hidden-terminal route and exposed terminal route while establishing a route.

Jorge M. Londoño and Azer Bestavros [9] have proposed a novel resource management approach to deliver large numbers of video-on-demand (VOD) streams through the cloud using a mixture of per-stream reservations and an aggregate reservation. This is done by allocating in an online fashion. The framework makes use of three per-stream parameters which makes it scalable and very lightweight. As a bandwidth management scheme, it facilitates the delivery of streams with less bandwidth or with higher link utilization. The experimental evaluation of their proposed work reasserts the efficiency for small buffer size and delay bounds.

Wen-Hwa Liao, Yu-Chee Tseng and Kuei-Ping Shih [10] have purported a more practical and cost-efficient TDMA based bandwidth reservation protocol, to deal with the Bandwidth Reservation problems, for QoS routing in a wireless Mobile Ad Hoc Network. Problems related to hidden-terminal and exposed-terminal is taken into consideration in their proposed protocol thereby serving as a medium to calculate bandwidth.

W. Ibrahim, J. W. Chinneck, S. Periyalwar and H. El-Sayed [11] have put forward a novel charging and resource management policy that manages network resources based on the traffic condition and allocates resources and services to customers based on their QoS profile as well as that best fits their needs. They emphasized that the proposed policy outperforms other existing policies and is capable of managing different types of applications and adapting to different traffic loads apart from generating more revenue than the rate-based policy. The policy caters the advantage to those customers who are unable to pay new increased service price by shifting them to lower QoS services.

Michael G. Kallitis, George Michailidis and Michael Devetsikiotis [12] have proposed a non-linear pricing model for next generation multiservice networks that ensures efficient resource allocation and guarantees QoS. A detailed analysis of the properties of the policy that gives information about the working as well as sensitivity of various parameters is provided by them. Standard iterative methods are used to deal with the problem of convex optimisation.

Chuanxiong Guo, Guohan Lu, Helen J. Wang, Shuang Yang, Chao Kong, Peng Sun, Wenfei Wu, and Yongguang Zhang [13] have emphasized on the importance and desirability of Virtual data Centre (VDC) than physical data centres due to its benefits and thus proposed VDC as a unit resource allocation for multiple renters in the cloud.
They also designed and provided the data centre network virtualization architecture called SecondNet for VDC support. They proclaimed on the basis of experimentation and evaluation that VDC guarantees high network utilization and SecondNet provides bandwidth assurance and elasticity.

Yu Cheng, Hai Jiang, Weihua Zhuang and Zhisheng Niu and Chuang Lin [14] have carried out their work in the area of Wireless network to determine model to enable efficient resource management in China’s 3G/4G wireless communication network as the number of mobile users are expected to grow at a much faster pace than the rest of the world. In regard to this, they proposed all-IP DiffServ model and two novel QoS techniques for efficient resource allocation.

Anup K. Talukdar, B. R. Badrinath and Arup Acharya [2] have purported on the utilization of MRSVP in place of currently used RSVP protocol due to inadequacy of the later one to guarantee undisturbed availability of services when mobile users frequently move to multiple locations during the lifetime of the connection. The proposed reservation protocol for mobile hosts aims to cater real-time services in an Integrated Services packet Network. They highlighted vivid features of MRSVP that signified its suitability for the desired objective viz. ability to make advanced reservation for mobile hosts, use of active and passive reservations to improve the network utilization, introducing new messages, etc.

Hassene Bouhouche and Sihem Guemara El Fatmi [3] have proposed QoS reservation mechanism for Mobile Ad Hoc Networks named as QSRR by taking into account resource reservation mechanism of the traffic submitted to the network. The proposed QSRR method reserves the network that has adequate resources to provide the QoS requirements of the connections. They also conducted the simulation analysis that showed the superiority of performance by QSRR in terms of delays, load and throughput.

III. PROPOSED SYSTEM

It is known that number of users accessing the Internet or mobile service during festive seasons, general holidays or some special events is larger than the usual. Due to such large demand altogether at one time, delay in or sometimes failure of service occurs which is a great cause of dissatisfaction among end-users and therefore is a matter to be dealt sincerely by service providers.

This disruption in service occurs due to request for accessing the limited bandwidth by multiple users apart from congestion caused due to bandwidth utilization by users of other competing network providers.

Fig. 1 depicts how the request of service increases tremendously during certain events viz. holidays, New Year, festivals, etc. Through SLA, network service provider and end-user enters into a contract where the user defines what type of services he requires the network operator to furnish and thus becomes the “privileged user”. Thus, during the period of high demands for network bandwidth as shown in fig. 1, privileged users are given priority by the network service providers as compared to normal users.

In fig. 2, a depiction of how bandwidth is utilized for carrying out different functions viz. texting/social networking, calling and mailing by users at different time intervals during working days of the week is done. We know that the amount of bandwidth and type of services accessed is different in different time intervals. During the morning hours i.e. between 7 am to 10 am, people generally make more number of calls as compared to texting or browsing whereas during late hours i.e. between 2 am to 7 am, the frequency of calls is negligible as compared to browsing and texting. Through SLA, network service provider, caters seamless services to its privileged user by underlining the nature of their usability of services as well as the congestion in the bandwidth utilization by several users.
Communication conducted between client and primary server according to research work is shown in fig. 3. It is a four layered structure with topmost layer being the application layer which deals with the execution of multiple applications. Subsiding to it is the transport layer which has gone under modification in a way that the client system consists of client TLMS and the primary server consists of Server TLMS. In order to perform the allocation of resources as per SLA to user, TLMS facilitates sharing of information between user and server. Any communication between client and primary server takes place through transport layer. Next to it lies the network layer which is followed by host to network layer at the bottom.

To enable seamless access to networks’ bandwidth, Client TLMS model is proposed as shown in fig. 4. It consist of three important parts i.e. Application Layer adaption engine, Network Layer adaption engine and database. The purpose of Application Layer adaption engine is to facilitate uninterrupted functioning of various applications already running or willing to run on the device as per user’s need. This adaption engine takes the decision of allowing free flow of applications to run on the device based on user’s previous history, time factor and energy availability. It also ensures high security to applications running on the device. Network Layer adaption engine checks for availability of network connection. It plays a significant role at the time of network congestion as it sends request to server TLMS for allocation of suitable bandwidth based on the SLA history of the prescribed user. All the information related to user’s previous connection and application history required for efficient allocation of bandwidth as per SLA is stored in the database which is also utilized by application layer and network layer adaption engine for carrying out their functions.

All the requests for service and applications from client TLMS are received and managed by server TLMS as shown in fig. 5. It performs numerous important functions viz. allocation of resources, managing cost and checking feasibility of service provision.
The above shown Infrastructure Management Module consists of several important sections performing different functions viz. high speed server, routers, firewalls, network traffic monitor/controller, analyzer, etc. Application Layer (AL) adaption engine receives request for running several applications, Network Layer (NL) adaption engine on receiving request of service allocate the available resource to the end-user based on SLA. It is used to set up applications i.e. to provide resource allocation as per the search result.

Searching Module primarily looks for available free bandwidth, priority and storage that can be allotted based on the requests received. It returns information on feasibility, cost and resource that can be allocated. It is connected to monitoring section and management section.

Monitoring Module checks the viable performance of running applications, generates performance parameter and traffic reports as well as keep track of total resource being used or in use. All the history related to work performed in server side along with user specific details are stored in the server side database. It continuously monitors the infrastructure and notices any changes occurring in the topology promptly.

IV. CONCLUSION

Paper described modification in transport layer working that contains transport layer management system for client and server. Two patterns of mobile application requests are discussed to cater the demand of users’ SLA. First pattern explains the request during the average working day request traffic and other pattern is related to special occasion where SLA needs special attention apart from issues related to mobility. QoS is provided to user to run various applications on mobile environment based on user’s previous history, security requirements, time-factor and battery energy availability. Resource allocation is ensured by transport layer management system of server with the help of Infrastructure Management Module and searching module.

REFERENCES


