Resource management in IP-based Radio Access Networks

H. el Allali, G. Heijenk
University of Twente
(Computer Science, TSS group)
{*allali,heijenk*}*@cs.utwente.nl*

Abstract

IP is being considered to be used in the Radio Access Network (RAN) of UMTS. It is of paramount importance to be able to provide good QoS guarantees to real time services in such an IP-based RAN. QoS in IP networks is most efficiently provided with Differentiated services (Diffserv). However, currently Diffserv mainly specifies Per Hop Behaviors (PHB). Proper mechanisms for admission control and resource reservation have not yet been defined. A new resource management concept in the IP-based RAN is needed to offer QoS guarantees to real time services. We investigate the current Diffserv mechanisms and contribute to development of a new resource management protocol. We focus on the load control algorithm [9], which is an attempt to solve the problem of admission control and resource reservation in IP-based networks. In this document we present some load control issues and propose to enhance the load control protocol with the Measurement Based Admission Control (MBAC) concept. With this enhancement the traffic load in the IP-based RAN can be estimated, since the ingress router in the network path can be notified by marking packets with the resource state information. With this knowledge, the ingress router can perform admission control to keep the IP-based RAN stable with a high utilization even in overload situations.

keywords: UMTS, RAN, admission control, resource reservation, load control, IP, QoS, Diffserv.

1. Introduction

The European Telecommunication Standard Institute (ETSI) has developed the second generation cellular system Global Systems for Mobile (GSM) in the 1980s. GSM has been standardized to offer simple voice services and basic low speed circuit switched data services. Because of the rapid increase of the number of mobile subscribers, the demand for new data services is growing as well. The current GSM system cannot offer these services. The reason for this is that the GSM service rate is bound by the small radio spectrum. To fulfill the needs for future mobile services ETSI has been standardizing UMTS, the third generation mobile system. By making use of a new spectrum scheme, Wide-band Code Division Multiple Access (WCDMA) and integrating different network technologies, UMTS is capable of offering a wide range of services. These services will be delivered by a UMTS Terrestrial Radio Access Network (UTRAN), which is based on circuit- and packet switched networks. UMTS will offer data services up to 2 Mbps and shall satisfy the demand of the mobile users.

Currently, the network entities in the UTRAN are interconnected with ATM virtual circuits. Since the UTRAN will transport both data and voice traffic at different priorities, a new transport technology is needed to simplify the O&M (Operation & Management) tasks and lower the network cost. The 3rd Generation Partnership Project (3GPP) is investigating the use of a packet switched network in order to have high link utilization and more efficient support of data traffic in the RAN. The Internet Protocol (IP) is a good candidate to deliver the UMTS services through the RAN. IP has shown its strength in the Internet as a very flexible and simple forwarding principle. By using the IP technology in the RAN, low priority traffic can be transported between high priority traffic in case we have spare resources. This transmission method results in lower transmission costs in the RAN.

There are a number of unsolved research issues related to introduction of IP in the RAN. One of these is the QoS guarantees for real-time traffic (call based traffic) in the RAN. The QoS of flows in the IP-based RAN are hard to guarantee, the reason for this is that overload situations can take place through the increase of the number of users, and the large amount of real time traffic in the RAN. This behavior results in degradation of service performance of real time traffic. To solve this problem, resources in the RAN must be managed in a proper way. Load control [9] is a simple resource reservation scheme that can be used to manage the resources in the RAN. We investigate how to enhance this

protocol so that it can be used in an IP-based RAN to overcome overload situations.

The purpose of this document is to get a clear overview of the current and future situation of wireless radio access networks. In this document we focus on some resource reservation and admission control problems in IP-based RANs. Much attention is paid to the load control algorithm. We describe the load control algorithm and point out some open issues. The rest of the document is organized as follows. In the next section we give a short overview of UMTS. In the section that follows we evaluate resource reservation methods for IP networks followed by a description of load control and some open issues. The final section summarizes our conclusion and discusses future work.

2. UMTS

Architecture:

ETSI has standardized an architecture that distinguishes the UMTS access network i.e., the UTRAN, and the UMTS core network, as can be seen in Fig. 1. The UTRAN consists of a Radio Network Subsystem (RNS) connected to the Core Network (CN) through the Iu interface. Furthermore, a RNS consists of a Radio Network Controller (RNC) and one or more Base Stations (BSs). The BSs are connected to the RNC through the Iub interface. The RNC in the UTRAN is responsible for control of the BS, and the radio link to the MS (Mobile Subscriber). The RAN also interconnects the RNCs of the Radio Network Subsystems. For each connection between a MS and the RAN one RNS is the Serving RNS. Typically, a RAN consists of a couple of RNCs and hundreds up to tens of thousands of BSs.

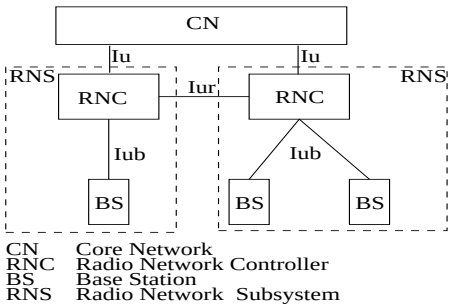

Figure 1. UMTS architecture [4].

Radio link:

UMTS uses the WCDMA scheme as radio bearer [10]. Roughly, this concept works as follows: each MS gets a unique code sequence, which is used to encode its information data signal. At the receiver-side this code of the particular MS is also known in order to decode the encoded signal. The signal reception is possible because the unique code has a small cross-correlation with other codes. Because of this the data signal of the MS can be despread, while the other users are still spread over a large bandwidth. The original data signal has a small bandwidth compared to the other spread-spectrum signals and can be filtered easily from the other code signals. The spreading principle gives WCDMA the multiple access capability.

Services:

By using the WCDMA scheme, UMTS is able to provide the mobile subscriber channels with maximum service rates of 64Kbps, 144Kbps, 384Kbps, and 2048Kbps [8]. With these service rates UMTS can deliver a range of new wireless services to the mobile user. UMTS can offer the following services, telephony like GSM, mobile access to Internet, and multi-media applications, e.g. voice over IP and video conferencing [1]. In conclusion: UMTS concept uses a common air interface to support all kind of applications and provides access to different networks (e.g. PSTN, and Internet) and different application servers.

3. UTRAN

The mobile network can be divided into two parts, namely the access network, which connects the base stations with the RNC (control unit), and the core network. Currently, the access network is built with ATM transport technology, each BS is connected via a virtual circuit to the RNC. The core network connects the RNC with other external networks, e.g. PSTN, and Internet. In the UTRAN the large number of BSs are connected with a common transmission link (tree structure). The current RAN consists of a large number of BS in different types of locations. To upgrade these RANs, bandwidth is added by introducing new transceivers. The transmission capacity needed for one base station is relatively low [7]. The transmission links that are needed to connect the RNCs are fewer and the distances between them are longer compared to the distances between one BS and another. The transmission capacity of these links is high because all the traffic from the BSs is aggregated at this point.

Due to the nature of the RAN, ATM leased lines are very expensive to upgrade and to maintain. For this reason, the telecom industry is investigating the use of IP in the RAN.

Fig. 2 shows the IP network configuration in the UTRAN. Each BS is integrated with an IP router. The BS router converts the UMTS packets into IP packets and forwards the packets to the proper outgoing interface into the IP network. The forwarding principle relies on comparing the IP destination address in the packet header with the routing table entries to route the packet to the next hop. All

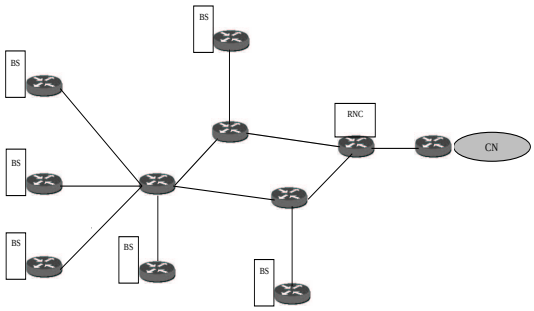

Figure 2. UTRAN based on IP service bearer.

the IP traffic from the BSs is aggregated at the core router, which interfaces with the core network. By using this transport method instead of ATM technology in the RAN we can have the following advantages [4]:

- Cost reduction of the transmission network.
- Flexible and efficient connectivity.
- Easier management by removing complex O&M operations that are required for handling the VPs/VCs.
- IP network routers can make requests to the network without making contact with the operator.
- Faster setup of calls and allocating new radio links in case of handover.
- AAL-2 signaling is not needed any more.
- Header compression is possible to have efficiency for voice connections.
- Efficient Internet access.
- IP can be built on the existing infrastructure.

IP technology in the RAN is promising but still QoS support of real time traffic in IP networks is not sufficient.

RAN requirements: The main portion of traffic carried in RAN is real-time traffic, this is why the RAN has stringent requirements on the IP technology. We list some of the RAN requirements [7]:

- Minimal complexity in RAN.
- Small end-to-end delay to offer good speech quality.
- Provide appropriate level of QoS regarding to delay, and jitter for phone calls.
- Low overhead.
- Rerouting must not impact the QoS of the existing calls.
- Good mobility performance.
- Efficient network utilization.
- Fast handover.

Today's QoS support in IP cannot meet the delay and jitter requirements for real time traffic. These delay constraints are solved in a circuit switched network but not yet in IP networks. Overprovisioning is not the best solution to this problem, since it will increase the network cost significantly because of the large distances between the BSs. The way to solve this problem in the RAN is to use resource management. Resource reservation and admission control mechanisms must be used to provide resources to the real time services in the RAN in a proper way.

4. Resource reservation in IP networks

Several methods have been developed to provide QoS to real-time traffic within the Internet. The first one is the Integrated Services (Intserv) framework, whereas the second one is the Diffserv framework. In both frameworks, admission control (AC) plays an essential role in protecting the network from overload. We give an overview of the different methods and evaluate the possible use of these concepts within a RAN environment. Note that most of the work takes place in the Internet Engineering Task Force (IETF).

Integrated Services: The Integrated Services framework provides per flow QoS using the Resource reSerVation Protocol (RSVP) [11] as a signaling protocol. RSVP can be used to allocate resources for real time traffic with high QoS constraints. The mobile user can use RSVP to request a specific QoS from the network. RSVP traverses the network and visits each intermediate router in the path that will be used for the call, and attempts to reserve resources. We can do this by sending a PATH message to the receiver to reserve the resources. The mobile user can choose the level of reserved resources by specifying the traffic characteristics in the PATH message. Every intermediate router along the path forwards the PATH message to the next hop. Then the receiver responds with RESV message to request resources for the flow. Every intermediate router in the path can reject or accept the request of the RESV message. If the request is accepted, the router sets parameters in two mechanisms, the packet classifier and the packet scheduler for that flow. The needed state information is placed in the router. The packet classifier determines the flow, a packet belongs to based on a number of packet header fields, i.e. source and destination address and port number, and protocol id and based on this, the packet scheduler makes the forwarding decisions to achieve the agreed QoS.

RSVP creates "soft states" so that every packet follows the same path. In Intserv the resources are reserved for

every individual flow. This approach cannot be realized in the core network with a large number of active flows. The intermediate routers would not be able to process the flows in the router regarding to classification and scheduling. To solve this scalability problem in the Internet Diffserv is defined by the IETF.

Differentiated services: By introducing the Differentiated services (Diffserv) framework we can allocate resources for different classes of IP traffic. Diffserv provides differential treatment to flows or aggregates of flows by using the Type of Service byte in the IP header [2]. Six bits of this byte, together called the DS field, are used for marking the differential service classes. The differentiated services define a number of Per Hop Behaviors (PHB). Each PHB corresponds to a particular forwarding treatment of the packets. The Diffserv code points are mapped to the PHBs.

Using expedited forwarding as PHB a virtual leased line can be configured in the RAN to provide real time traffic with low loss, low delay, and guaranteed service rate. However, in overload situation this method alone is not sufficient to deliver the desired QoS guarantees of real time traffic. Admission control has to be used in the edge routers to provide these QoS guarantees. Admission control admits a flow only if there are sufficient resources to provide the requested QoS for the new flow, at the same time it ensures that the QoS level of the already admitted flows is not violated.

Admission Control: To allow the network to do admission control (AC), each (mobile) user provides a traffic description relating to the resources that are requested. The "simple sum" algorithm [5] ensures that the new requested resource plus the sum of the reserved resources does not exceed the total link capacity. This algorithm is effective, but not always efficient. Standard AC results in low utilization when traffic is bursty. Tight traffic description is needed, which is very difficult due to the changing nature of real time applications.

To achieve better utilization performance in the network, Measurement Based Admission Control (MBAC) approach can be used. With MBAC we can estimate the traffic load in the network and perform admission control. Several MBAC algorithms have been developed [3]. One of them is the "measured sum" algorithm. While the "simple sum" uses requested reservations as input variable, the measured sum algorithm uses rate measurement. With the measured sum we estimate the load of the existing traffic and decide to admit or block the new call.

4.1. Discussion

Introduction of Intserv in an IP-based RAN will make the network more complex. The reason for this is that the routers must process the packets to filter the incoming flows, which is a very expensive task. The second method (Diffserv) is scalable, only the edge routers perform classification, marking, policing and shaping operations. The intermediate router performs only class filtering and scheduling. However, diffserv can only give hard guarantees with respect to one-way delay and jitter if it is supplemented by an admission control mechanism.

With standard AC it would be difficult for the IP-based RAN to fulfill the utilization requirement, since AC performance is sensitive to traffic burstiness and incidental congestion. The MBAC concept has great potential to be used because of the utilization performance.

5. Resource reservation in IP-based RANs

The best concept for an IP-based RAN is a simple and flexibly operating resource reservation algorithm. A promising mechanism is load control. This mechanism can be used to monitor the resources in the RAN and make call admission.

5.1. Load control

Load control is a new resource allocation algorithm, which can be used in IP-based RANs to monitor and control the network resources. In this section we describe one of its modes of operation, the "simple marking" algorithm. If a call is requested at the ingress router (e.g. BS or RNC), the ingress router generates a probe packet and sends it into the network. The probe packet traverses the same path as the actual traffic to probe for resources. If resources are not available in intermediate routers, the probe packet will be marked. Once the probe is marked the marking status will not be changed by the other routers along the path. At the end of the path the egress router will receive the probe packet. The egress router in turn shall forward the resource status of the path to the ingress router.

Incidental congestion can be notified to the ingress router by enabling the intermediate routers to mark the actual traffic. The egress router can detect congestion in the network path and inform the ingress router to take precautions. When the probe packet is received by the ingress router it investigates the resources of the path and decides whether to admit or block the call. If the call is accepted, the ingress interface is configured to accept the flow.

5.2. Open Issues

Load control is a new protocol that is under development. There are still unsolved issues regarding to marking strategy and probe information processing in the network nodes. We present some initial thoughts about how the load

control scheme can be enhanced. The issues are described in steps of load control actions.

Type of probe: The probe in load control can be of different forms. In all cases, the probe information is somehow included in the packet header. one of the ways to define a probe is to include the probe information in the header of the actual data packet. Another way of probing is to include the probe information in the packets carrying the signaling traffic that goes from the RNC to the BS and visa versa. The third option is to generate a probe packet (dummy packet), this probe form gives us the advantage to time stamp the payload of the probe packet. The best probe form to use is the one that has the lowest overhead and gives the best throughput performance.

Probing entity: In load control we described that the ingress router performs network probing. In the IP-based RAN it is still unclear who performs the network probing. If the BSs are connected in a tree configuration with an intermediate router, the intermediate router can perform network probing tasks. In this scenario the BS submits a request to the intermediate router to probe the next hops. We assume that there are enough resources between the BS (ingress router) and the intermediate router, this solution makes the scheme simpler but centralizes the complexity.

Measurement strategy: Probing information can be based on different types of measurements, candidates are rate, delay, and queue length measurements.

Rate measurement: We are interested in the utilization of the probing path. A way to do this is by measuring the usage rate of each interface of the intermediate router, with this method we have a measure of the resources in the path. The usage rate can be estimated by sampling the rate, this is based on periodically counting the number of bits sent trough the output interface of the router and dividing it with the sampling time.

Delay measurement: Besides the usage rate in the intermediate routers, it is also interesting to analyze the experienced packet delay in the queue (waiting time). The queue in the router is also a resource quantity. By measuring the experienced delay of the traffic in the queue, the resource availability of the path is measurable. As an alternative to packet marking, delay can be estimated by measuring the Round-Trip-Time (RTT) of dummy probe packets. This method is based on the following, the ingress router time stamps the probe packet and sends the packet trough the path. After the ingress router receives the probe packet back from the egress router, it reads out the time stamp and calculates the RTT of the path. With this information the ingress router can keep up a database with several RTT values to observe time differences between probe periods. An increase of RTT means that more packet buffering has been taken place in the path.

Queue size measurement: Another important queue quantity is the queue occupation, we can estimate the number of packets in the queue to have a measure of the resources in the intermediate router. Random Early Detection (RED) is based on this approach and is implemented in many routers. We can use this implementation to measure the average buffer size. In Fig. 3 we show the marking mode of RED. When the queue is measured we can notify

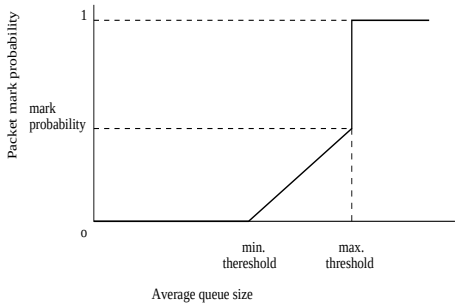

Figure 3. RED marking behavior.

the ingress router the status of the queues by marking the packets. This can be done in three different ways. The first option is to define one threshold value in RED. If the average queue size exceeds this threshold value, one bit in the probe header will be set. This is the simplest way of marking the probe packet. The second scenario is based on a number of thresholds namely, th1,...,thi. If the queue build-up exceeds one of the thresholds the probe header will be marked in the header with the corresponding threshold code. With this approach we can define i different areas of average queue sizes. With the third method we can notify the ingress router with more relevant information by relate the level of queue occupation with the level of marked packets (actual packets). This means that the amount of packets in the buffer is the same as the number of marked packets. Note that the lost packets can be seen as marked packets. With this additional information we can make better decisions at the ingress router. The threshold value is a very important factor in the RED mechanism. To achieve high utilization in the network the threshold value must be set in a proper way.

Early congestion notification: During probing, the network can exhibit incidental congestion, it is essential to notify the ingress router as soon as possible. The ingress router maintains a timer to observe the dropped probe packets to retransmit a new probe packet. In case congestion is taking place in the intermediate routers probe packets can be discarded or delayed to notify the ingress router of the congestion in a faster manner. With this

approach we do not have to feedback the probe packets during congestion and stress the path more. Proper timer setting is needed to have a stable and high utilization in the network.

The way of marking: The way of marking is also an important issue in load control. It is possible to mark packets at the front of the queue instead of at the end. The ingress router will be notified faster if overload situations occur because the packet at the front can be marked quickly during service. The front marking principle avoids additional delay due to buffering waiting time.

Reporting strategy: The most efficient manner to feedback the probe packet back to the ingress router is also still unclear. There are number of ways to solve this issue namely, by tunneling the probe packet, attach the probe information in an Internet Control Message Protocol (ICMP) messages, or generating a new acknowledgment packet.

Information processing: The received information from the egress router can be used to make admission control. The way of processing this information is still open. We can make decisions by checking one bit in the probe header or by analyzing the level of average queue size.

5.3. Discussion

The way of measurement and processing of the measured quantities is still open in the load control protocol. The performance of these methods is strongly dependent on the tuning of the measurement and processing mechanisms. Further research is needed, important research problems are: measurement accuracy of the load estimation in the queues, RED threshold setting, and load control timer setting. All this is needed to have a stable RAN with a high utilization performance. Furthermore, the decision strategy is still open in load control. Research must be done to find the way to process the measured values. The questions in the processing mechanism are as follows, when do we decide to block or admit flows and delay or drop the packets?

6. Conclusions and future work

In this document we have discussed several approaches to reserve resources in IP networks. Load control has the greatest potential to comply with the stringent requirements of the RAN, and to realize a scalable resource reservation scheme in an IP-based RAN. From [6], simulation results show that MBAC has a good utilization performance, but its performance in a RAN environment is still unclear. The MBAC approach is a starting point for our future work in load control.

We have presented some initial thoughts about the possible enhancements of load control. The tuning knobs are of great importance to achieve high utilization and good QoS performance in the IP-based RAN. The evaluation of these settings in a test environment will give us more insight which strategy to choose. Load control will be analyzed to find the best way to enhance it. We plan to implement load control in a diffserv testbed to determine its behavior and performance. Different test scenarios with mixed prioritized and best effort traffic across the network will be performed to gather statistics on the load control performance.

References

[1] *Universal Mobile Telecommunication System Strategies*, ETSI draft, 201 721 v1.1.2 edition, Feb. 2000.

[2] S. Blake, D. Black, M. Carlson, E. Davies, Z. Wang, and W. Weiss. An architecture for differentiated services. Internet RFC, RFC 2475, Dec. 1998.

[3] C. Casetti, J. Kurose, and D. Towsley. An adaptive algorithm for measurement-based admission control in integrated packet networks. In *Proceedings of Protocol for High Speed Networks workshop*, Oct. 1996.

[4] N. Drevon. IP transport in UTRAN. 3GGP draft, Working sheet of the TSG-RAN group 9, Dec. 1999.

[5] S. Floyd. Comments on Measurement-based Admission Control for Controlled-Load Services. Technical report, Lawrence Berkeley Laboratory, July 1996.

[6] S. Jamin, S. Shenker, and P. Daniz. Comparison of Measurement-based Admission Control Algorithms for controlled-load service. In *Proceedings of IEEE Infocom'97*, Apr. 1997.

[7] D. Partain, G. Karagiannis, and L. Westberg. Wireless resource issues. Internet draft, draft-partain-wireless-issues.txt, 2000.

[8] K. Salkintzis. A Survey of Mobile Data Networks. *IEEE communications surveys, vol.2, no. 3, Third quarter*, 1999.

[9] Z. Turanyi and L. Westberg. Load control of real-time traffic. Internet draft, draft-westberg-loadcntr-02.txt, Oct. 1999.

[10] H. Walke. *Mobile Radio Networks: networking and protocols*. June 1999.

[11] P. White. RSVP and Integrated Services in the Internet: a tutorial. *IEEE communications magazine*, May 1997.