A Queuing Theoretic Scheme for a QoS-Enabled 4G Wireless Access

Robert Achieng, Member, IEEE and H. Anthony Chan, Sr Member, IEEE
Department of Electrical Engineering, University of Cape Town, Rondebosch, South Africa
Email: rachieng@crg.ee.uct.ac.za, h.a.chan@ieee.org

Abstract
There will be heterogeneous wireless access networks in 4G systems, e.g., WLAN, UTRAN etc. The networks will need to support QoS provisioning and full mobility. For each connection request by a mobile user in such networks, there will be need for a scheme for selecting the best network from among the available networks. The research proposes a scheme that selects the best network based on the user’s QoS requirements. The scheme models each radio access network as a network of queuing nodes. This model then uses the network state (e.g. traffic arrival and service rates) to compute the end-to-end QoS parameter statistics (e.g., delay and throughput) of user traffic flow through each available network. We postulate that the statistics indicate the QoS capabilities of the network and can therefore be used to select the best network to serve the mobile user.

I. INTRODUCTION

There has been a phenomenal growth in wireless networking over the last two decades. This growth is attributed to rapid advances in technology, competition in the business environment, and socio-economic changes on how people live and work. Due to this growth, a new paradigm in telecoms systems known as the Next Generation Network (NGN) is emerging. A clear and fixed definition of NGN is yet to be developed, but there’s universal consensus that it will comprise the following [1]:
1) a layered architecture, consisting of functionally differentiated planes, namely the Transport plane, the Control plane, the Application plane, and the Management plane.
2) open and standardized interfaces to facilitate interworking of systems across the planes.

The Transport plane (or the Data/User plane) is responsible for the transfer of user traffic across the network. The Control plane is responsible for the control of bearer capabilities in the data plane through signaling and also the facilitation of interworking between systems. The service creation and deployment are handled by the Application plane. The Management plane performs such tasks as QoS provisioning, security, and network management. In NGN Management plane will be distributed (not localized) and cross-layered.

In NGN, the Transport plane moves data packets across the access, the edge, and the core networks. The access network will be a heterogeneous network with various wireless and wire-line technologies operating in concert to provide service to the user. A wireless access network will typically consist of mobile terminals, a wireless access point, and an access router. The edge network will be a system of routers that manage ingress into the core network through admission control and QoS provisioning. The core network will be a high-speed optical switching and transport system. The entire data plane will be packet switched. Fig.1 shows the architecture of NGN.

A Fourth Generation (4G) Wireless network is a wireless network that can provide NGN functionalities. The heterogeneous wireless access system in a 4G network will have to support QoS while simultaneously providing full mobility. For QoS support, the always best connected (ABC) concept of 4G systems implies that the access network will have to perform the following tasks:
1) Determine the best network to serve the mobile user both when it joins the network and also during handoff.
2) Determine the connection QoS requirements

![Fig.1 NGN Architecture](image-url)
In traditional networks, there is no need for a best network selection scheme because services are provided by vertically structured and independent homogeneous networks. However, for heterogeneous networks, the situation is different. At some point in time and place, a mobile user in a heterogeneous network can be served by more than one network. Thus, there should be a scheme for selecting the best network to serve the mobile user. The selection procedure should be based on the user’s QoS requirements and the access network capabilities at that particular time. Thus, there is a challenge to develop a procedure for selecting the best access network based on user QoS requirements and the QoS capabilities of the various networks.

Several QoS architectures and procedures have been proposed to address this challenge. However, some deficiencies still remain. We propose a scheme that enhances these efforts.

The paper is organized as follows: Section II gives a description of three QoS architectures and highlights their limitations. Section III presents the proposed solution by motivating and describing a queuing model of the radio access network. Section IV derives the QoS statistics for traffic in UTRAN FDD (WCDMA network). In Section V we derive the QoS statistics for traffic in WLAN. Section VI presents a method for comparing the end-to-end QoS statistics of traffic obtained in Sections IV and V, and hence select the best network. In Section VII, a possible framework for evaluating the model is given. Section VIII is a conclusion of the paper.

II. EXISTING QoS ARCHITECTURES

There have been several research initiatives to meet the 4G QoS challenge. Here, we describe three major architectures namely Fuzzy-Neural Based Approach for Joint Radio Resource Management (JRRM), QoS Broker Architecture for End-End QoS Support and the SMART QoS Architecture.

A. Fuzzy-Neural Based Approach for Joint Radio Resource Management (JRRM)

The fuzzy neural scheme uses fuzzy-neural theory to formulate QoS support mechanisms in 4G networks [2]. Radio network parameters and processes are represented by fuzzy logic variables. QoS information processing and decision making is performed using both fuzzy logic theory and artificial neural networks. Thus, the essential function of this architecture is to provision QoS in the presence of uncertainty of the access network. The major drawback of this scheme is its complexity and lack of supporting technologies.

B. QoS Broker Architecture for End-to-End QoS Support

The heterogeneous radio access network in the QoS Broker Architecture comprises the mobile terminal, the radio access networks, and access routers [3]. A new network element, the QoS Broker, is introduced. The QoS Broker provides end-to-end QoS support by interfacing between the radio access networks and the core network and also serves as the Policy Decision Point (PDP) of the access network.

This architecture provides only a signaling framework for QoS provisioning; it does not specify the QoS provisioning mechanisms. Therefore any such mechanism can be integrated into the architecture.

C. The SMART QoS Architecture

The SMART 4G QoS architecture is similar to the QoS Broker architecture in that it also only provides a QoS signaling framework. It proposes an overlay radio network dedicated to QoS signaling.

The architecture has four components [4]. The components are the multi-service user terminal, the basic access network for overlay signaling, the wireless access network for data transfer, and the common core network for QoS provisioning. The architectures mentioned describe a QoS architecture but do not provide a network selection procedure. Thus, there’s need for a solution to address these limitations.

III. QUEUING MODEL OF RADIO ACCESS NETWORK

We model the wireless access network as a network of queuing nodes. The model considers both the characteristics of the wireless channel and the type of traffic. The end-to-end QoS statistics of a traffic flow in this network of queuing nodes is then determined. We use these statistics to determine the QoS capabilities of the various access networks in order to select the best network to serve the user.

A. The Wireless Access Network as a Queuing Network

In 4G systems, there will be a number of paths for user traffic through the wireless access network. User traffic will be able flow to the edge of the network through UTRAN, WLAN, WiMAX, or any other viable wireless technology. Figure 2 shows the different paths that a traffic flow can take between the mobile user and the edge.

![Fig. 2. Possible traffic paths in the access network](image)

Fig. 3 is a queuing model of the wireless access system
shown in Fig. 2. In Fig. 3, network nodes (e.g., cellular base stations, WLAN access points, routers/switches) are represented by queues. The queues are multidimensional and are characterized by service rates \((\mu_1, \mu_2, ..., \mu_n)\), arrival rates \((\lambda_1, \lambda_2, ..., \lambda_n)\). A multidimensional queue is a queue with different classes of traffic. Each traffic class in the access network has a known (or determinable) arrival rate and service rate. The principal performance parameters of a queuing system are the queue length \((N)\), waiting time \((W)\), and throughput \((\gamma)\). These parameters are in themselves the QoS statistics of the access network modeled by the queuing system. Thus, knowledge of the arrival and service rates for each traffic class is a necessary and sufficient \((?)\) requirement for the determination of the QoS statistics of the traffic class.

### B. Traffic Arrival and Service Rates in a Wireless Network

The rate intensities \((\lambda, \mu)\) of the queuing model depend on the characteristics of the wireless channel. Different wireless technologies have different wireless channel characteristics. Thus the QoS statistics of user traffic will vary from one wireless network to another. For example, the time varying interference level in UTRAN determines the wireless channel capacity while collision rate in WLAN directly affects the throughput. Moreover, user mobility within UTRAN cellular topology induces handoff, which in turn affects the traffic arrival and service rates.

![Fig. 3. Queuing model of the radio access network.](image)

C. Effect of Type of Traffic on QoS statistics

The type of traffic (e.g., real time voice or non-real time packet data) determines both the statistical distributions of the traffic rates \((\text{i.e., arrival rates } (\lambda) \text{ and service rates } (\mu))\) and the queuing system model. We use the Poisson distribution \((\text{denoted M})\) to model traffic arrivals for both real time (RT) services like conversational voice and video and non-real time (NRT) services like web service packet data. For packet length we use exponential distribution \((\text{M})\) for RT services and general distribution \((\text{G})\) for NRT packets \([9]\).

RT services are modeled with the Erlang Loss model. On the other hand, the Erlang Delay model is used NRT services. In the Erlang Loss Model, the queuing system has no waiting capacity; thus calls that arrive when all servers are busy are cleared (lost). It is the most appropriate model for RT services that are delay sensitive. The delay model represents a system with infinite buffer capacity. Thus, calls which arrive when all servers are busy are queued for delayed service \([9]\).

To provide reliable communication to loss-intolerant but delay-insensitive data traffic, packet retransmission strategies may be used at the logical link layer of the access technology. For example, UTRAN uses Hybrid Automatic Repeat reQuest (HARQ) retransmission scheme for NRT packet data.

We use Priority Queuing (PQ) to cater for multiple services which are envisaged for future 4G systems. In this scheme, the system classifies and ranks traffic and then serves the traffic based on priorities (ranks).

D. QoS statistics in the Wireless Access Network

For each possible flow path through the access network, we determine the QoS performance metrics. The statistic will depend on the wireless access channel, the type of traffic and the queuing discipline employed. The following are the QoS metrics to be determined.

1) Blocking probability. This is the probability that either a new call or handover call will be denied service in the wireless network because all servers are busy. The Erlang Loss model is used.
2) Mean end-to-end delay: The delay that the traffic is expected to experience on traversing all the nodes of the wireless access network.
3) Call dropping probability: This is the probability that a call will be dropped in the cell because of unsuccessful handoff attempt.
4) Delay probability: The probability that a call will be queued because the system is busy (Non-real time services)
5) Throughput: End-to-end throughput considering the fact that calls are delayed at the nodes.

The following sections present an approach for determining these QoS statistics for UTRAN and WLAN.

## IV. QoS Statistics in UTRAN

The multi-access technology used in UMTS Terrestrial
Access Network (UTRAN) is Wideband Code division Multiple Access (WCDMA) [5]. The QoS statistics are affected by the time varying wireless channel capacity, user mobility, and the Hybrid Automatic Repeat Request (HARQ) error control mechanism.

A. Wireless Channel capacity

In a WCDMA cell, the number of traffic channels is given by the soft traffic capacity of the cell. In [11], WCDMA soft traffic capacity $C$ is shown to be

$$C = 1 + \left( \frac{W}{R} \right) \frac{\phi_v \phi_a}{E_b/I_o} 1 + f$$

(4.1)

where

- $W =$ system bandwidth (5 MHz for UTRAN)
- $R =$ user data rate
- $E_b/I_o =$ Bit energy to interference ratio.
- $\phi_v, \phi_a =$ voice activity gain and antenna gain respectively
- $f =$ external cell interference factor (typically 0.6)

The number of servers ($m$) in the model is then set equal to $C$.

B. User mobility

In a cellular wireless network, the traffic arrival rate to a cell is composed of two components: new call arrival rate ($\lambda_{nc}$) and handoff call arrival rate ($\lambda_{hc}$). The service rate also has two components: the service rate ($\mu_t$) and channel holding rate in the cell ($\mu_h$).

Thus [12]

$$\lambda = \lambda_{nc} + \lambda_{hc}$$

(4.2)

$$\mu = \mu_t + \mu_h$$

(4.3)

C. Hybrid Automatic Repeat Request error control mechanism.

The throughput of NRT data in UTRAN is affected by HARQ. To determine the delay and throughput in a HARQ scheme, we proceed by first computing the first moment ($\bar{X}$) and second moment ($\bar{X}^2$) of the service time distribution. Assume general distribution (G) [10]

- $P_r =$ P (Packet retransmission in a channel using HARQ)
- $Pr =$ P (An error is detected in a frame). In our analysis we shall always assume that this figure is available from network measurements.

$$P_r = P \text{ (Frame successfully transmitted after } i \text{ retransmissions})$$

$$P_i = (1 - P_r) P_r^i$$

(4.4)

$N_r =$ Mean number of retransmissions to achieve successful frame transmission

$$N_r = E(i) = \sum_{i=0}^{\infty} i P_i = \sum_{i=0}^{\infty} i (1 - P_r) P_r^i = \frac{P_r}{1 - P_r}$$

(4.5)

If $t_{frame} =$ frame transmission time, $n_{frame} =$ frame length in bits

Then

$$(1 + N_r)t_{frame} = \text{Effective frame transmission time}$$

Throughput $\gamma = \frac{n_{frame}}{(1 + N_r)t_{frame}} = \frac{1 - P_r}{\mu}$

(4.6)

where $R = 1/\mu =$ link data rate.

The QoS statistics are finally obtained by utilizing the M/M/m/m queue model for the RT services, and an M/G/m queue model for a NRT service. We also use Priority queuing, and assign priority 1 to RT services and priority 2 to NRT services. Using standard formulas from Queuing Theory [8], the QoS metrics for both types of traffic are obtained. Tables I and II gives the QoS metrics.

**TABLE I**

**QoS STATISTICS FOR REAL TIME TRAFFIC IN UTRAN**

<table>
<thead>
<tr>
<th>UTRAN QoS Parameter</th>
<th>Queuing formula</th>
</tr>
</thead>
<tbody>
<tr>
<td>Call Blocking Probability $P_B$</td>
<td>$P_B = \sum_{k=0}^{\infty} \frac{\rho^k}{k!} \frac{\mu^m}{m!} (1 - \rho^m)$</td>
</tr>
<tr>
<td>Call Dropping Probability $P_D$</td>
<td>$P_D = \sum_{k=0}^{\infty} \frac{\rho_{hc}^k}{k!} (1 - \rho_{hc}^m)$</td>
</tr>
<tr>
<td>Delay $D$</td>
<td>$D = \frac{1}{\mu} (1 - P_B)$</td>
</tr>
</tbody>
</table>

**TABLE II**

**QoS STATISTICS FOR NON REAL TIME TRAFFIC IN UTRAN**

<table>
<thead>
<tr>
<th>UTRAN QoS Parameter</th>
<th>Queuing formula</th>
</tr>
</thead>
<tbody>
<tr>
<td>Call Delay Probability $P_{D_c}$</td>
<td>$P_{D_c} = \frac{\rho_{hc}^m}{m! (1 - \rho_{hc})}$</td>
</tr>
<tr>
<td>Handover Delay Probability $P_{HD}$</td>
<td>$P_{HD} = \frac{\rho_{hc}^m}{m! (1 - \rho_{hc})}$</td>
</tr>
<tr>
<td>Delay $D$</td>
<td>$D = \frac{1}{\mu} + \sum_{i=0}^{\infty} \alpha_i X_i$</td>
</tr>
<tr>
<td>Throughput $\gamma$</td>
<td>$\gamma = \frac{1 - P_r}{\mu}$</td>
</tr>
</tbody>
</table>
V. QOS STATISTICS IN WLAN

For QoS support in WLAN, the IEEE 802.11e standard has been developed. To support traffic with different QoS requirements, a channel access mechanism known as Enhanced Distributed Coordination Function (EDCF) is implemented [6]. EDCF is an enhanced version of the Carrier Sense Multiple Access with Collision Avoidance (CSMA/CA) MAC scheme. In this scheme, traffic classes are associated with Access Categories (AC).

At each EDCF station, MAC maintains four transmit queues; each queue corresponds to an AC contending for transmission with the other queues. Whenever there’s a collision, a Virtual Contention Handler resolves the contention with the higher priority queue taking precedence. Fig. 4 illustrates the transmit queue operation of the EDCF MAC protocol.

![AC transmit queues for EDCF](image)

**Fig. 4. AC transmit queues for EDCF**

The Virtual Contention Handler is essentially a contention-based MAC protocol. Before a station transmits it must first sense that the channel is idle for a time period known as the Arbitration Inter-frame Space (AIFS), and then restrains itself from transmitting for a random length of time known as the Back off time ($b_n$). AIFS and $b_n$ depend on the AC of the traffic to be transmitted. Channel Access Delay (CAD) is the total time period before a frame is transmitted. CAD has three components: an Arbitration Inter-Frame Space (AIFS), a Distributed Inter-Frame Space (DIFS) and a Contention Window (CW$_n$). DIFS has a fixed length. AIFS and CW$_n$ have variable lengths that depend on the AC. The higher the priority of the traffic to be transmitted, the shorter the CAD. CAD is determined as follows [7]

\[
AIFS = \eta \text{DIFS} \quad (5.1)
\]

\[
AIFS = 1, 2, \ldots \quad (\eta \text{ depends on traffic class})
\]

\[
\text{CW}_n = 2^{k_0 - n - 1} - 1 \times \text{Slot length} \quad (5.2)
\]

\[
r = \text{a factor that depends on the traffic type. The higher the priority of the traffic, the lower the value of } r
\]

\[
\text{CAD} = AIFS + b_n \quad (5.3)
\]

\[
0 \leq b_n \leq \text{CW}_n \quad (5.4)
\]

At each station, a medium access function maintains a back

off timer for each contention event. The timer is initialized at $b_n$ at the beginning of each contention window; it then decrements by a slot time for each time slot the medium is sensed idle. The timer stops decrementing when a transmission is detected. It resumes when the medium is sensed idle again for an AIFS time. Once the timer decrements to zero, the medium access function begins transmission on the channel. If the transmission is successful, an acknowledgement frame is sent to the sender; otherwise a collision is presumed to have occurred. In this event the back off event counter is incremented (i.e. $n$ is incremented), a new contention window is calculated according to Eq (5.2) and a new contention event is started with $b_n$ set according to Eq. (5.4).

The following section presents a statistical analysis of the EDCF MAC protocol. Because the protocol is a contention protocol, the only relevant QoS metrics are the delay and throughput. Availability metrics like blocking probabilities do not make sense.

A. Delay and Throughput Analysis for EDCF

The EDCF MAC contention channel can be modeled as an M/G/1 queue. To perform a delay analysis for the system, we first need to determine the effective frame transmission time ($X$)

\[
X = AIFS + T_b + T_f \quad (5.5)
\]

- AIFS is a constant given by Eq 5.1
- $T_f$ is a random variable that depends on the packet length distribution. In our analysis, we take RT packet lengths to be Exponentially distributed while NRT packet lengths are Pareto distributed [9].
- $T_b$ is a random sum of random variables. Here the individual random variables are the back off times ($b_n$) within a contention window. The number of back off events ($n$) before successful transmission is itself a random variable

\[
T_b = b_0 + b_1 + \ldots + b_n \quad (5.6)
\]

From the Law of Total Expectations

\[
E(T_b) = E(b)E(n) \quad (5.7)
\]

\[
E(T_b^2) = E(b)E(n) + E(n^2)E(b) \quad (5.8)
\]

\[
E(b) = \text{Mean back off time}
\]

\[
E(n) = \text{Mean number of back off events}
\]

\[
V(b) = \text{Variance of back off time}
\]

\[
E(n^2) = \text{Second moment of back off events}
\]

The back off time $b_n$ of each of the back-off events is a random variable distributed over the interval $[0, 2^{k_0 - n - 1} - 1] = [0, k_0 2^n - 1] \quad (5.9)$

Assume that $b_n$ follows a continuous uniform distribution
over the contention window. Thus
\[
b_{k} = \tau \sum_{k=0}^{a} k \left( \frac{1}{b - a + 1} \right) = \frac{\tau b + a}{2} = \frac{\tau k_{a} 2^{n}}{2}
\] (5.10)
For \( a=0 \) and \( b=k_{a}2^{n} \)
\[P_{n} = P \left[n \text{ back offs before successful transmission}\right]
\] (5.11)
p = P \left[\text{unsuccessful transmission}\right]
\lambda = \text{traffic arrival rate}; \tau = \text{slot length}
\[E(b) = \sum_{n=0}^{\infty} n P_{n} = \tau k_{a} (1 - p) \sum_{n=0}^{\infty} 2^{n} p^{n} = \frac{\tau k_{a} (1 - p)}{1 - 2p}
\] (5.13)
For \( |2p| < 1 \)
\[E(n) = \sum_{n=0}^{\infty} n P_{n} = (1 - p) \sum_{n=0}^{\infty} np^{n-1} = \frac{p}{1 - p}
\] (5.14)
Thus
\[E(T_{a}) = \frac{\tau k_{a} p}{1 - 2p}
\] (5.15)
\[E(T_{s}) = \frac{(\tau k_{a})^{2} (1 - 3p - p^{2} + 2p^{3})}{1 + 2p)(1 - 2p)^{2}} = \frac{p^{3} + p^{4}}{(1 - p)^{4}}
\] (5.16)
Finally, we apply the Pollaczek-Kinchine (PK) formula to
the M/G/1 queue model to obtain the waiting time (W), system
time (T), and throughput (\(\gamma\)) in IEEE 802.11e system to be
\[W = \frac{\lambda}{2(1 - p)}
\] (5.19)
\[T = W + \frac{1}{\mu}
\] (5.20)
\[
\bullet \ \text{let } g(n) \text{ be the attempt rate (the expected number of packets transmitted in a slot) when there are } n \text{ backlogged nodes. Then}
\text{Attempt rate = arrival rate + retransmission rate}
\]
\[g(n) = \lambda + \frac{n}{T}
\] (5.21)
The number of attempted packets per slot in state \( n \) is
approximately a Poisson random variable of mean \( g(n) \) [8].
\[\text{P (m attempts)} = g(n)^{m} \frac{e^{-g(n)}}{m!}
\] (5.22)
\[\text{P (success)} = P(m=1) = g(n)e^{-g(n)}
\] (5.23)
The throughput is the fraction of slots with a successful transmission
\[\text{Throughput (\(\gamma\)) = P (success)} = g(n)e^{-g(n)}
\] (5.24)

VI. QUANTITATIVE DETERMINATION OF END-TO-END QoS PARAMETER STATISTICS

After determining the network availability, the expected delays and throughput for the access networks as outlined above, the Access/QoS manager then selects the best network based on these parameters. A QoS function computes a rank for each of the access networks using a ranking algorithm. The algorithm considers both the type of traffic and user preferences (e.g. power budget and network operator). For example, a user running low on battery power might prefer a network that requires minimal power budget for transmission; another user might, on the other hand, prefer an access network because it is cheaper than the other access networks etc. Essentially the QoS function assigns weights to each of the QoS metrics based on traffic type and user preferences. The rank is then computed as
\[R_{i} = c_{f_{i}} (w_{a} P_{m} + w_{d} D_{i} + w_{r} / \gamma_{i})
\] (6.1)
where
\[R_{i} = \text{Rank for network } i
\]
\(c_{f_{i}}\) = Cost function for network \( i \); this factor is set by the user.
\(w_{a}, w_{d},\) and \( w_{r} = \) weights for availability, delay and throughput respectively. The weights are set by the QoS function. \( P_{m}, D_{i},\) and \( \gamma_{i} \) are the blocking probability, delay, and throughput for network \( i \) as determined in Sections IV and V. The lower the value of \( R_{i} \), the most preferable is the network

VII. PERFORMANCE ANALYSIS OF PROPOSED SCHEME

Fig. 5 shows the test bed for evaluating the performance of the proposed scheme. We build scenarios of the wireless network (e.g. different traffic arrival rates and service rates) and then feed the scenario variables into both OPNET Modeler and formulae developed; the formulae are run in MATLAB. The results obtained from both the OPNET and MATLAB simulations are analysed and compared

VIII. CONCLUSION

The paper has shown that a QoS enabled network selection procedure in 4G wireless access networks can be realized by modeling the wireless access system as a network of queues. With the model, the end-to-end statistics of the QoS parameters of an arbitrary flow in the network can be computed and then used to select the best network. We are currently carrying out a performance evaluation of the scheme.
**Fig. 5. Evaluation testbed for proposed scheme**

**REFERENCES**


