Project	IEEE 802.16 Broadband Wireless Access Working Group <**http://ieee802.org/16**>
Title	Proposed Text for Evaluation Methodology and Key Criteria for P802.16m
Date Submitted	2007-03-05
Source(s)	Robert Novak, Mo-Han Fong, Peiying Zhu, Gamini Senarath, Dean Kitchener, Wen Tong, Hang Zhang, David Steer, Derek Yu, Mark Naden, Jianglei Ma, Sang-Youb Kim rnovak@nortel.com Voice: 1-613-763-7863 mhfong@nortel.com Voice: 1-613-765-8983 Nortel 3500 Carling Avenue Ottawa, On, K2H 8E9 Canada
Re:	A response to Call for Contributions, http://www.ieee802.org/16/tgm/docs/80216m-07_005r2.pdf
Abstract	Proposed text for Evaluation Methodology and Key Criteria for P802.16m - Advanced Air Interface
Purpose	For consideration and incorporation into the P802.16m Evaluation Methodology and Key Criteria document.
Notice	
Release	
Patent Policy and Procedures	

technology (or technology under patent application) might be incorporated into a draft standard being developed within the IEEE 802.16 Working Group. The Chair will disclose this notification via the IEEE 802.16 web site <http://ieee802.org/16/ipr/patents/notices>.

Proposed Text for Evaluation Methodology and Key Criteria for P802.16m

Robert Novak, Mo-Han Fong, Peiying Zhu, Gamini Senarath, Dean Kitchener, Wen Tong, Hang Zhang, David Steer, Derek Yu, Mark Naden, Jianglei Ma, Sang-Youb Kim

Nortel

1 Introduction

The scope of this System Evaluation Methodology is to develop and specify parameters and methods associated with the traffic models, and performance metrics that would serve as guidelines to aid in the evaluation and comparisons of technology proposals for IEEE 802.16m Project. Channel models suitable for the 802.16m Project are under development. It is not the intention of this document to mandate the use of this evaluation methodology in the comparisons of proposals. Proponents should provide sufficient details of simulation parameters such that it is possible for other proponents to replicate their results.

This document makes use of models and metrics defined for the evaluation methodology developed for proposals for 802.16 Relay TG [1].

1.1 Simulation overview

In this section, an example of the Simulation model is provided. Figure 1 shows the components and methodology that would serve as a baseline for the rest of this document.

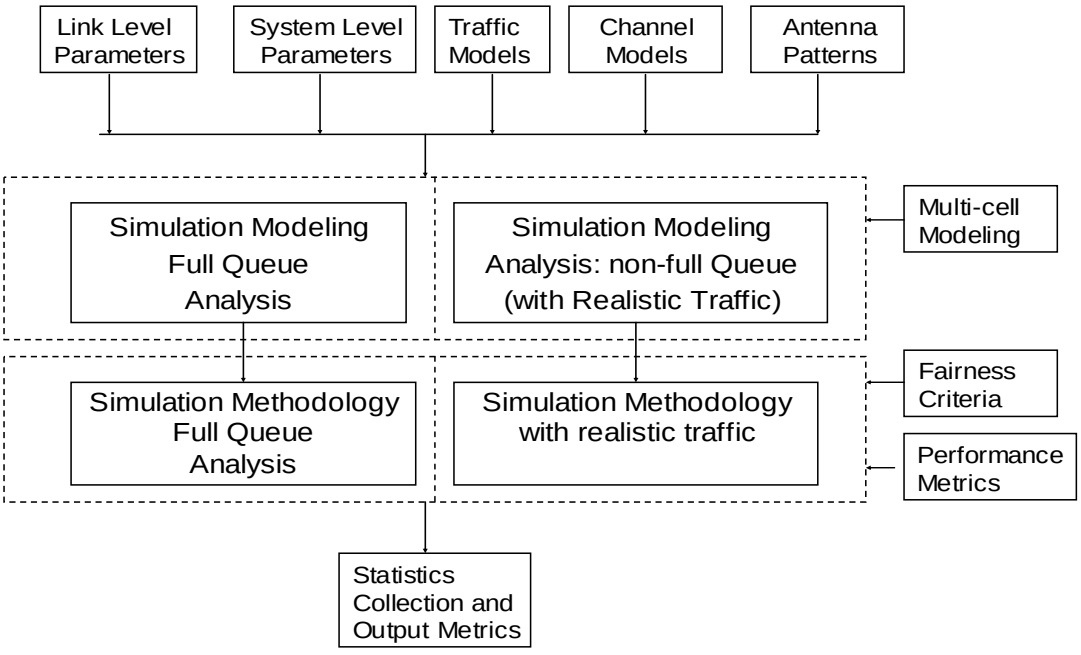


Figure 1. Simulation Components and Overall Methodology

2 Channel Models

Channel models suitable for evaluation of 802.16m system proposals needs to be developed, taking into account parameters specific to 802.16m including bandwidths, operating frequencies and modeling for MIMO configurations. Comments regarding channel model development are given in [2].

3 Traffic models

This section describes the traffic models in detail. A major objective of system simulations is to provide the operator a view of the maximum number of active users that can be supported for a given service under a specified configuration at a given coverage level. The traffic generated by a service should be accurately modeled in order to find out the performance of a system. This may be a time consuming exercise. Traffic modeling can be simplified, as explained below, by not modeling the user arrival process and assuming full queue traffic which is considered as the baseline. These two assumptions are further discussed proceeding paragraphs. Modeling non-full-queue traffic is also discussed in the next subsections.

Modeling of user arrival process: Typically all the users are not active at a given time and even the active users might not register for the same service. In order to avoid different user registration and demand models, the objective of the proposed simulation is restricted to evaluate the performance with the users who are maintaining a session with transmission activity. These can be used to determine the number of such registered users that can be supported. This document does not address the arrival process of such registered users, i.e. it does not address the statistics of subscribers that register and become active.

Full Queue model: In the full queue user traffic model, all the users in the system always have data to send or receive. In other words, there is always a constant amount of data that needs to be transferred, in contrast to bursts of data that follow an arrival process. This model allows the assessment of the spectral efficiency of the system independent of actual user traffic distribution type.

In the following sections, we will concentrate on traffic generation only for the non-full queue case. In addition, the interaction of the generated traffic with the higher layer protocol stack such as TCP is not included here. However, we will provide references to document which provide the detailed TCP transport layer implementation and its interaction with the various traffic models.

3.1 Traffic Modeling for IEEE802.16m Project Services

The required traffic models and their corresponding sections where they are defined are listed in Table 1.

Table 1. Services to be considered

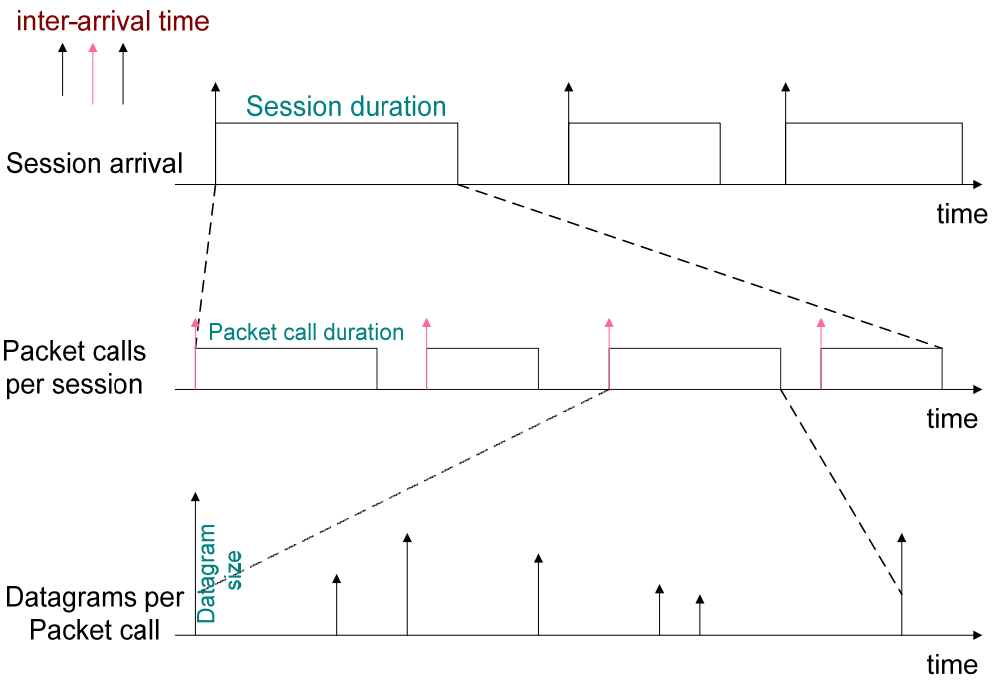
#	Application	Traffic Category	Definition
1	Full buffer		Section 3
2	HTTP (UL and DL)	Interactive	Section 3.1.1
3	FTP (UL and DL)	Best effort/ Non real-time	Section 3.1.2
4	Near Real Time (NRT) Video Streaming (UL and DL)	Streaming	Section 3.1.3
5	VoIP	Real-time	Section 3.1.4
6	Gaming (UL and DL)	Real-time	Section 3.1.5
7	Live Video	Interactive Real-time	TBD

For a simulation with HTTP, FTP and NRT video streaming traffic models, if simulation is for DL (or UL) traffic only, UL (or DL) traffic modeling (e.g. HTTP/FTP requests) can be neglected for the simplicity as the bandwidth requirements for these messages are small compared to the data traffic.

The FTP and HTTP traffic models listed in Table 1 can be generated using the bursty traffic generation model described in Figure 2. For each traffic source, the following characteristics are modeled:
Session arrival in terms of session inter-arrival time and session duration.
Packet call arrival in terms of packet call inter-arrival time and packet call duration within a session. Within a packet call, there are periods of active traffic generation and periods of no activity.
Finally, datagram inter-arrival times and datagram size within a packet call.

We consider that a single session stays from the beginning of the simulation till the end of the simulation, i.e., the whole simulation time. Therefore, packet call and datagram inter-arrival times, packet call duration and datagram size distributions for these bursty traffic models will be described in the next sections.

Figure 2. Bursty traffic generation model

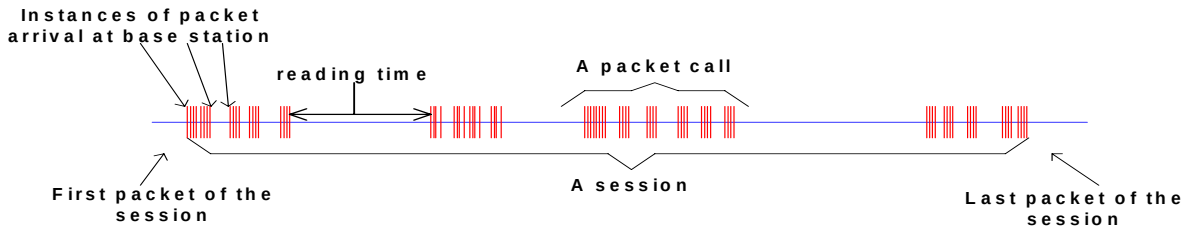


3.1.1 HTTP model (UL and DL) [3][4]

3.1.1.1 HTTP traffic model characteristics

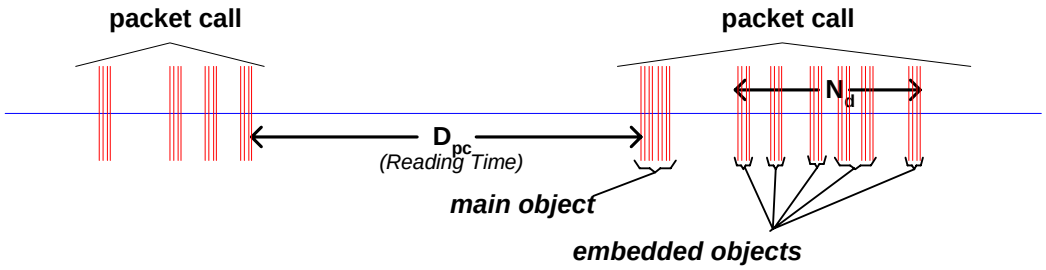
Figure 3 shows a typical web browsing session. Each session is divided into ON/OFF periods representing web-page downloads and intermediate reading times. Each web-page download is referred to as packet calls in Figure 3. During an ON period (packet call), users are requesting information. During an OFF period, user is reading/digesting the web-page.

Figure 3. Packet trace of a typical web browsing session



The activity within each packet call can be found in Figure 4. Note the similarity of the distribution for the packet calls within a session in Figure 3 and the datagram arrivals within a packet call in Figure 4. This can possibly be a result of self-similarity in web-browsing traffic.

Figure 4 Contents of a packet call



There are ON and OFF periods within a packet call. During an ON period, objects are being retrieved. Parsing time and protocol overhead are represented by the OFF periods within a packet call. During a packet call, the initial HTML page (referred to as the main object) is first downloaded. However, within the initial HTML page, there can be additional references to embedded object files such as graphics and buttons. After parsing the information on the embedded objects, the embedded objects will be loaded next as indicated in Figure 4.

3.1.1.2 HTTP traffic model parameters

The parameters for web browsing traffic are:
No of pages per session;
S_M: size of the main object in a packet call;
S_E: size of an embedded object in a packet call;
N_d: number of embedded objects in a packet call;
D_{pc}: reading time;
T_p: parsing time for main page

Table 2. HTTP Traffic Model Parameters

Component	Distribution	DL Parameters	UL Parameters	
Main object size (S_M)	Truncated Lognormal	Mean = 10710 bytes, Std. Dev = 25032 bytes, Minimum = 100 bytes; Maximum = 2Mbytes, $\sigma = 1.37, \mu = 8.35$	Mean = 9055 byes, Std. dev. = 13265 bytes, Minimum = 100 bytes, Maximum = 100Kbytes $\sigma = 1.37, \mu = 8.35$	$f_x = \frac{1}{\sqrt{2\pi}\sigma x}\exp\left[\frac{-(\ln x-\mu)^2}{2\sigma^2}\right]$ $x \geq 0$
Embedded object size (S_E)	Truncated Lognormal	Mean = 7758bytes, Std. dev. = 126168bytes, Minimum = 50bytes, Maximum = 2Mbytes $\sigma = 2.36, \mu = 6.17$	Mean = 5958bytes, Std. dev. = 11376bytes, Minimum = 50bytes, Maximum=100kbytes $\sigma = 1.69, \mu = 7.53$	$f_x = \frac{1}{\sqrt{2\pi}\sigma x}\exp\left[\frac{-(\ln x-\mu)^2}{2\sigma^2}\right]$ $x \geq 0$
Number of embedded objects per page (N_d)	Truncated Pareto	Mean = 5.64, Maximum = 53 $\alpha = 1.1, k = 2, m = 55$	Mean = 4.229, Maximum = 53 $\alpha = 1.1, k = 2, m = 55$	$f_x = \frac{\alpha k^{\alpha}}{x^{\alpha+1}}, k \leq x < m$ $f_x = \left(\frac{k}{m}\right)^{\alpha}, x = m$ Subtract k from generated random value to obtain N_d.
Reading Time (D_{pc})	Exponential	Mean = 30seconds	Mean = 30seconds	$f_x = \lambda e^{-\lambda x}, x \geq 0$ $\lambda = 0.033$
Parsing time (T_p)	Exponential	Mean = 0.13second	Mean = 0.13second	$f_x = \lambda e^{-\lambda x}, x \geq 0$ $\lambda = 7.69$

Note: when generating a random sample from a truncated distribution, discard the random sample when it is outside the valid interval and regenerate another random sample.

3.1.1.3 HTTP and TCP interactions for DL HTTP traffic

Two versions of the HTTP protocol, HTTP/1.0 and HTTP/1.1, are widely used by servers and browsers. Users shall specify 30% HTTP/1.0 and 70% HTTP/1.1 for HTTP traffic.

For people who have to model the actual interaction between HTTP traffic and the underling TCP connection, refer to 4.1.3.2, 4.2.4.3 of [3] for details.

3.1.1.4 HTTP and TCP interactions for UL HTTP traffic

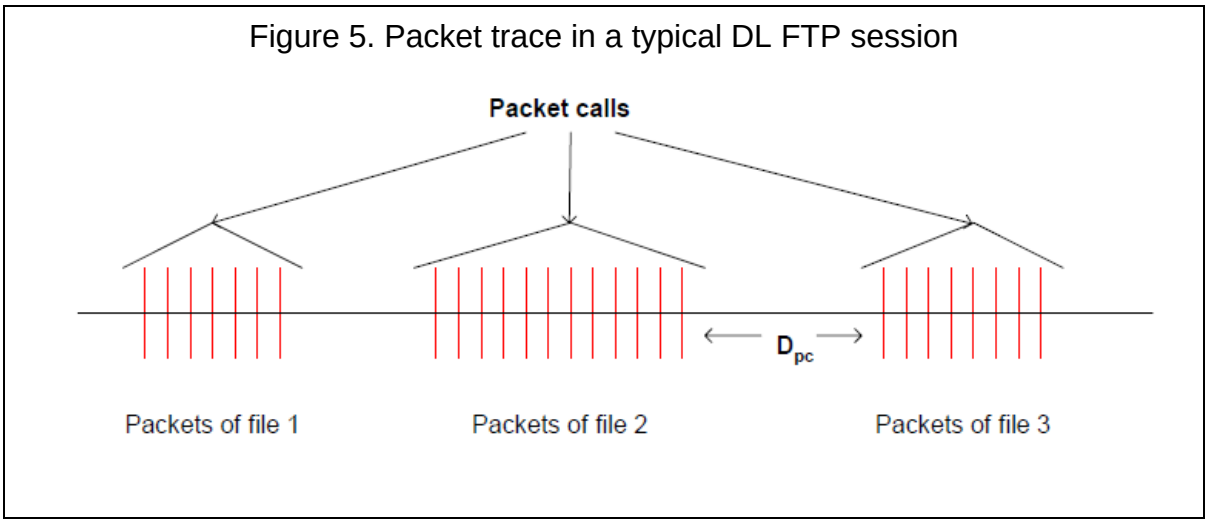
HTTP/1.1 is used for UL HTTP traffic. For details regarding the modeling of the interaction between HTTP traffic and the underling TCP connection, refer to 4.2.4.1, 4.2.4.2 of [3].

3.1.2 FTP model (UL and DL) [3][4]

3.1.2.1 DL FTP traffic model characteristics

For DL FTP, activities within a FTP session can be found in Figure . A typical FTP session consists of a sequence of file transfers separated by reading time. Each file transfer can be treated as a packet call. Reading time can be treated as the OFF period within a session. Within each packet call, only the file size is randomly generated.

Figure 5. Packet trace in a typical DL FTP session



3.1.2.2 DL FTP traffic model parameters

Hence, there are two main parameters for a DL FTP session:

S: size of file to be transferred;

D_{pc}: reading time. This is the time interval between end of download of the previous file and the user request for the next file.

The parameters distribution and values can be found in Table 3.

Table 3. DL FTP traffic model parameters

Component	Distribution	Parameters	PDF
File size (S)	Truncated Lognormal	Mean = 2Mbytes Std. Dev. = 0.722 Mbytes Maximum = 5 Mbytes	$f_X = \frac{1}{\sqrt{2\pi}\sigma x}\exp\left[\frac{-(\ln x-\mu)^2}{2\sigma^2}\right], x \geq 0$ $\sigma = 0.35, \mu = 14.45$
Reading time (D_{pc})	Exponential	Mean = 180 sec (TBD).	$f_X = \lambda e^{-\lambda x}, x \geq 0$ $\lambda = 0.006$

3.1.2.3 UL FTP traffic model characteristics

FTP traffic in the UL direction is generated mainly from file upload and email attachment upload. Each FTP upload user stays in the system until it finishes the transmission of its file. The FTP upload user leaves the system immediately after it finishes the transmission of its file.

Hence, for UL FTP traffic, each FTP session consists of 1 packet call. Within the packet call, only the file size is randomly generated.

3.1.2.4 UL FTP traffic model parameters

The only traffic model parameter is the upload file size and can be found in

Table 4.
For UL FTP traffic, users shall arrive according to a Poisson process with arrival rate λ.

Table 4. UL FTP traffic model parameter

Arrival of new users	Poisson with parameter λ
Upload file size	Truncated lognormal; lognormal pdf: $f_x = \frac{1}{\sqrt{2\pi}\sigma x}\exp\left[\frac{-(\ln x-\mu)^2}{2\sigma^2}\right], x \geq 0$ $\sigma = 2.0899, \quad \mu = 0.9385$ Min = 0.5 kbytes, Max = 500 kbytes If the value generated according to the lognormal pdf is larger than Max or smaller than Min, discard it and regenerate a new value. The resulting truncated lognormal distribution has a mean = 19.5 kbytes and standard deviation = 46.7 kbytes

3.1.2.5 FTP and TCP interactions

To model the FTP and TCP interactions, please refer to 4.1.4.2 of [3] for details.

3.1.3 Near real time video streaming (NRT video streaming) (UL and DL) [3][4]

A video streaming session is defined as the entire video streaming call time. It is equal to the simulation time for this model. Hence, a video streaming session occurs during the whole simulation period. No session inter-arrival time is needed. It is originally modeled for DL direction. However, the same model is proposed to be used for UL direction.

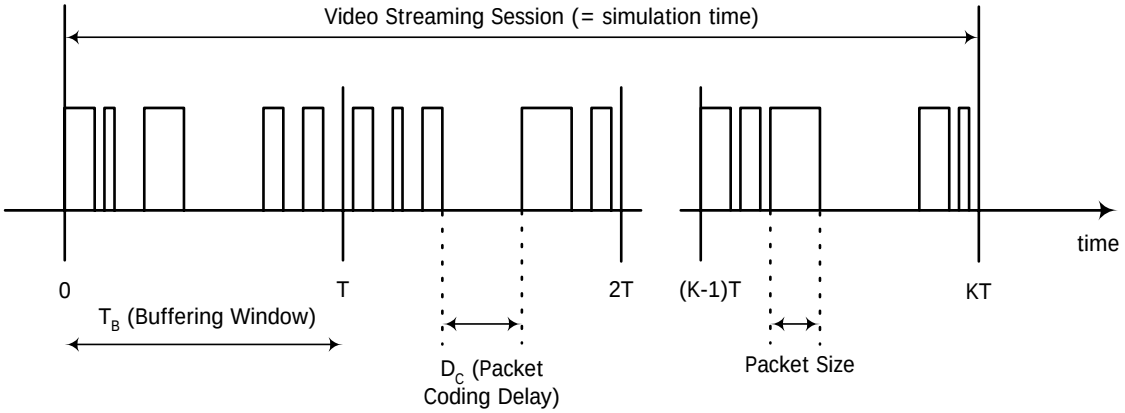
3.1.3.1 NRT video streaming traffic model characteristics

Figure describes a steady state of video streaming traffic from the network as observed by the base station. Call setup latency and overhead is not considered in this model.

Each frame of video data arrives at a regular interval T. Each frame can be treated as a packet call and there will be zero OFF duration within a session. Within each frame (packet call), packets (or datagrams) arrive randomly and the packet sizes are random as well.

To counter the jittering effect caused by the random packet arrival rate within a frame at the MS, the MS uses a de-jitter buffer window to guarantee a continuous display of video streaming data. The de-jitter buffer window for video streaming service is 5 seconds. At the beginning of simulation, the MS de-jitter buffer shall be full with video data. During simulation, data is leaked out of this buffer at the source video data rate and filled as DL traffic reaches the MS from the BS. As a performance criterion, the simulation shall record the length of time, if any, during which the dejitter buffer runs dry.

Figure 6. Video streaming traffic model



3.1.3.2 NRT video streaming traffic model parameters

The packet sizes and packet inter-arrival rate can be found in Table 5 when using a source rate of 64 kbps.

Table 5. Near Real-Time Video Traffic Model Parameters

Information types	Inter-arrival time between the beginning of each frame	Number of packets (slices) in a frame	Packet (slice) size	Inter-arrival time between packets (slices) in a frame
Distribution	Deterministic (Based on 10fps)	Deterministic	Truncated Pareto (Mean= 50bytes, Max= 125bytes)	Truncated Pareto (Mean= 6ms, Max= 12.5ms)
Distribution parameters	100ms	8	K=20bytes α = 1.2	K=2.5ms α = 1.2

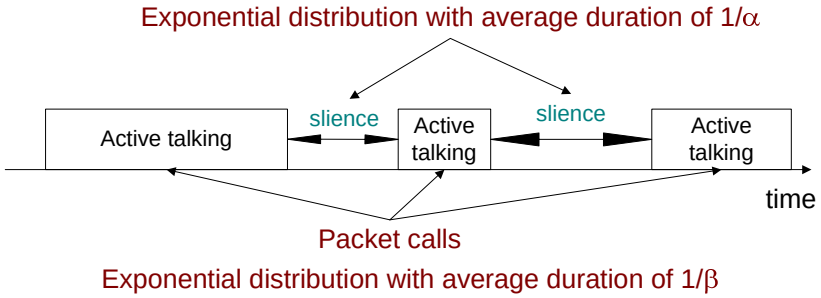
3.1.4 VoIP model [3][5][6][7]

VoIP refers to real-time delivery of packet voice across networks using the Internet protocols. A VoIP session is defined as the entire user call time and VoIP session occurs during the whole simulation period.

3.1.4.1 VoIP traffic model characteristics

A typical phone conversation is marked by periods of active talking interleaved by silence/listening period as shown in Figure .

Figure 7. Typical phone conversation profile



A two state Markov process (active-inactive) is used to model a VoIP source in Figure . The alternating periods of activity and silence are exponentially distributed with average durations of $1/\beta$ and $1/\alpha$ respectively. Hence, the fraction of time the voice source is active is $\alpha/(\alpha+\beta)$. For a voice activity factor of 40%, $1/\beta$ = 1s and $1/\alpha$ = 1.5s. Each active state period can be treated as a packet call and inactive period as the OFF period within a session.

During the active state, packets of fixed sizes are generated at a regular interval. During the inactive state, we have chosen to generate comfort noise with smaller packet sizes at a regular interval instead of no packet transmission. The size of packet and the rate at which the packets are sent depends on the corresponding voice codecs and compression schemes. Table 6 provides information on some common vocoders.

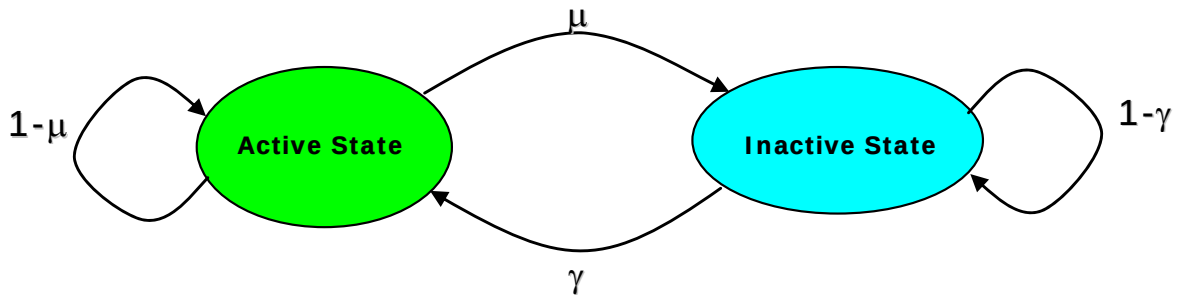
Table 6. Information on various vocoders

Vocoder	EVRC	AMR	G.711	G.723.1		G729A
Source Bit rate [Kb/s]	0.8/2/4/8.55	4.75-12.2	64	5.3	6.3	8
Frame duration [ms]	20	20	10	30	30	10
Information bits per frame	16/40/80/171	95-244	640	159	189	80

Among the various vocoders in Table 6, a simplified AMR (adaptive multi-rate) audio data compression can be used to simplify the VoIP modeling process. AMR is optimized for speech coding and was adopted as the standard speech codec by 3GPP and widely used in GSM. The original AMR uses link adaptation to select from one of eight different bit rates based on link conditions. If the radio condition is bad, source coding is reduced (less bits to represent speech) and channel coding (stronger FEC) is increased. This improves the quality and robustness of the network condition while sacrificing some voice clarity. In our simplified version,

we have chosen to disable the link adaptation and use the full rate of 12.2kbps in the active state. This will give us the worst case scenario.

Figure 8. Markov chain model of a VoIP source



Without header compression, AMR payload of 33 bytes are generated in the active state for every 20ms and AMR payload of 7 bytes are generated in the inactive state for every 160ms. Table 7 shows the VoIP packet size calculation for simplified AMR with or without header compression when using IPv4 or IPv6.

Table 7. VoIP packet size calculation for simplified AMR and G. 729

Description	AMR without Header Compression IPv4/IPv6	AMR with Header Compression IPv4/IPv6	G.729 without Header Compression IPv4/IPv6	G.729 with Header Compression IPv4/IPv6
Voice Payload	7bytes (inactive) 33 bytes (active)	7bytes (inactive) 33 bytes (active)	0 bytes (inactive) 20 bytes (active)	0 bytes (inactive) 20 bytes (active)
Protocol Headers	40 bytes / 60 bytes	2 bytes/ 4 bytes	40 bytes / 60 bytes	2 bytes/ 4 bytes
RTP	12 bytes		12 bytes	
UDP	8 bytes		8 bytes	
IPv4 / IPv6	20 bytes / 40 bytes		20 bytes / 40 bytes	
802.16 Generic MAC Header	6 bytes	6 bytes	6 bytes	6 bytes
CRC	4 bytes	4 bytes	4 bytes	4 bytes
Total VoIP packet size	57 bytes/ 77 bytes (inactive) 87 bytes / 103 bytes (active)	19 bytes/ 21 bytes (inactive) 45 bytes/ 47 bytes (active)	0 bytes (inactive) 70 bytes / 90 bytes (active)	0 bytes (inactive) 32 bytes/ 34 bytes (active)

3.1.4.2 VoIP traffic model parameters

During each call (each session), a VoIP user will be in the Active or Inactive state. The duration of each state is exponentially distributed. Within the Active/Inactive state, packets of fixed sizes will be generated at a fix interval. Hence, both the datagram size and datagram arrival intervals are fixed within a packet call. Parameters associated with the VoIP traffic model can be found in Table 8.

Table 8. VoIP traffic model parameters specification

Component	Distribution	Parameters	PDF
Active state duration	Exponential	Mean = 1 second	$f_X = \lambda e^{-\lambda x}, x \geq 0$ $\lambda = 1/Mean$
Inactive state duration	Exponential	Mean = 1.5 second.	$f_X = \lambda e^{-\lambda x}, x \geq 0$ $\lambda = 1/Mean$
Probability of transition from active to inactive state	N/A	μ (=0.6)	N/A
Probability of transition from inactive to active state	N/A	γ (=0.4)	N/A

3.1.5 Gaming model (UL and DL) [3][8]

Gaming traffic is generated by users engaged in interactive gaming of multiple users in different locations via the internet. A gaming session is defined as the time duration that a user plays a game and a gaming session occurs during the whole simulation period.

3.1.5.1 Gaming traffic model characteristics

The packet arrival time and the frame boundary are random and shall be simulated. Gaming packets are relatively small in size. Due to the interactive nature of gaming, packet delay must be short. Any packets that are generated and not transmitted at the PHY layer within 160ms shall be dropped.

3.1.5.2 Gaming traffic model parameters

Gaming traffic model parameters for DL and UL can be found in Table 9[8]. Largest Extreme Value distribution is used for random packet size generation. Since packet size has to be an integer, the largest integer less than or equal to X is used as the actual packet size.

Table 9. Gaming traffic model parameters

Component	Distribution		Parameters		PDF
	DL	UL	DL	UL	
Initial packet arrival	Uniform	Uniform	a=0, b=40ms	a=0, b=40ms	$f(x)=\frac{1}{b-a}, a \le x \le b$
Packet inter-arrival time	Extreme	Extreme	a=48ms, b=4.5ms	a=40ms, b=6ms	$f(x)=\frac{1}{b}e^{-\frac{x-a}{b}}e^{-e^{-\frac{x-a}{b}}}, b>0$ $X=\lfloor a-b\ln(-\ln Y)\rfloor, Y \in U(0,1)$
Packet size	Extreme	Extreme	a=330bytes, b=82bytes	a=45bytes, b=5.7	$f(x)=\frac{1}{b}e^{-\frac{x-a}{b}}e^{-e^{-\frac{x-a}{b}}}, b>0$ $X=\lfloor a-b\ln(-\ln Y)\rfloor+2, Y \in U(0,1)$ Addition of 2 in the equation is due to 2 bytes of UDP header size after header compression.

3.2 Traffic mix proposal

To test various aspect of the system, we propose the following traffic mixes:
1. Five cases of HTTP, FTP, NRT Video Streaming, Gaming, or Voice only.
2. Three cases of mixed traffic from Mix -1 to Mix -3 referenced in Table 10. The percentage of the traffic mix in these 3 cases is expressed in terms of data capacity (i.e., bps) of a given targeted cell.

Table 10. Proposed traffic mixes

	VoIP	FTP	HTTP	NRT video	Gaming
Voice Only	100% #users = Nv	0%	0%	0%	0%
FTP only	0%	100%	0%	0%	0%
HTTP only	0%	0%	100%	0%	0%
NRT Video only	0%	0%	0%	100%	0%
Gaming only	0%	0%	0%	0%	100%
Traffic Mix 1	0.5 Nv	Remaining Capacity for Data Users 100%	0%	0%	0%
Traffic Mix 2	0.5 Nv	Remaining Capacity for Data Users 30%	30%	30%	10%
Traffic Mix 3	0.75 Nv	Remaining Capacity for Data Users 30%	30%	30%	10%

Nv is the system voice capacity that satisfy outage criteria at system and user level.

4 Performance Metrics

The performance metrics are divided into two categories. They are:

- Single-user performance; and
- Multi-user performance.

Examples of single-user performance metrics are the link budget margins, C/I area coverage and data rate area coverage. These metrics are evaluated assuming that a single user is in a particular cell area utilizing all the resources in that cell while external interference may be evaluated assuming that at least a single active user is available in the external cell (for both forward and UL). These metrics are not end-to-end performance metrics and therefore, could be evaluated without modeling higher layer protocols and is independent of applications.

However, when multiple users are in the system the system resources have to be shared and a user's average data rate will be smaller than the single-user rate. Therefore, multi-user metrics are proposed which show how a system behaves under a multi-user environment.

In order to evaluate multi-user performance accurately, scheduling and higher layer traffic behaviors and protocols need to be modeled. However, simulation run times can be prohibitively large. Specially, in the case of multihop systems which may be part of 802.16 m Project proposals, each sector can have several relay stations and there are a large number of relay stations and relay to user and relay to base links need to be modeled sand simulated. Therefore, such simulations can be very CPU intensive. Therefore, we suggest that initial design validations be done using a simple but representative analysis using a full queue traffic without modeling higher layers. These are described under multi-user performance metrics.

4.1 Single-user performance Metrics

Note that the area coverage mentioned below is equivalent to the percentage of users meeting a given requirement when the users are uniformly distributed in the interested geographical area.

4.1.1 Link Budget and Coverage Range (Noise Limited) – single-cell consideration

Link budget evaluations is a well known method for initial system planning and this needs to be carried out for BS to MS links, and if relays are included, RS to BS and RS to MS links. Although a link budget can be calculated separately for each link, it is the combination of the links that determines the performance of the system as a whole. The parameters to be used needs to be agreed upon after obtaining consensus. Using the margins in the link budget, the expected signal to noise ratio can be evaluated at given distances. Using these results, the noise limited range can be evaluated for the system. The link budget template is TBD.

Since relays can be used to extend the range covered by a cell under noise limited environment (i.e. negligible interference from other cells but the limitation coming from the fact that the transmit power is not enough to provide a sufficient signal strength above thermal noise) coverage range is a metric of importance in such cases.

Coverage range is defined as the maximum radial distance to meet a certain percentage of area coverage (x%) with a signal to noise ratio above a certain threshold (target_snr) over y% of time, assuming no interference signals are present. It is proposed that x be 99 and y be 95.

4.1.2 C/I Coverage – interference limited multi-cell consideration

The C/I coverage is defined as the percentage area of a cell where the average C/I experienced by a stationary user is larger than a certain threshold (target_ci).

4.1.3 Data Rate Coverage – interference limited multi-cell consideration

The percentage area for which a user is able to transmit/receive successfully at a specified mean data rate using single-user analysis mentioned above. No delay requirement is considered here.

4.2 Multi-user Performance Metrics

Although a user may be covered for a certain percentage area (e.g. 99%) for a given service, when multiple users are in a sector/BS, the resources (time, frequency, power) are to be shared among the users. It can be expected that a user's average data rate may be reduced by a factor of N when there are N active users (assuming resources are equally shared and no multi-user diversity gain), compared to a single user rate.

For example, assume that there is a system, where a shared channel with a peak rate of 2 Mbps can serve 99% of the area. If a user wants to obtain a video streaming service at 2 Mbps, that particular user will be able to obtain the service, but no other user will be able to get any service during the whole video session (which may extend for more than an hour). Therefore, in this example although 99% area is covered for the video service, this service is not a viable service for the operator and performance of coverage need to be coupled with the capacity in order to reflect viable service solutions. Coverage performance assessment must be coupled with capacity (# of MSs), to obtain a viable metric.

The users having poor channel quality may be provided more resources so that they would get equal service from the cellular operator. This could adversely impact the total cell throughput. Thus, there is a trade-off between coverage and capacity. Any measure of capacity should be provided with the associated coverage. .

Since an operator should be able to provide the service to multiple users at the same time, an increase in the area coverage itself does not give an operator the ability to offer a given service

Therefore, the number of users that can be supported under a given coverage captures actual coverage performance for a given service from a viability point of view.

The suggested performance metric is the number of admissible users (capacity), parameterized by the service (R_{min}), and the coverage (allowable outage probability).

4.2.1 Combined Coverage and Capacity Index (*cc*)

The number N of simultaneous users per cell that can be supported achieving a target information throughput $R_{\min}$ with specified coverage reliability.

This performance metric can be approximated using either a simplified approximate evaluation methodology or a more detailed simulation as described below. Both methods are useful since the approximation methodology can be used to quickly compare two coverage enhancement techniques during the initial system concept development stage. The detailed simulations are useful to evaluate more carefully the most promising concepts. When results are presented the evaluation method used should be reported.

4.2.2 Method 1: Simplified Combined Coverage and Capacity Index Evaluation

This is a Simplified Methodology to evaluate Combined Coverage and Capacity Index (cc) using only the rate capability of each user. This can be evaluated without modeling higher layer protocols.

Assume that in a simulation N users are dropped uniformly in the service area. Let the required coverage for a given service be x% and the required information rate for that service be R_{min}. The first step in evaluating *cc* is to sort the MSs in descending order of achievable rate, assuming each utilizes the entire resources. Then, only the top x% of the MSs are considered.. Assume the number of users in the remaining group is *k*, and the data

rate capability of user i is r_i (i = 1 to N) by using a scheduler that provides equal throughput to all the serviced users.

Then,

if the min(r_i) < R_{min}, $cc = 0$ (i.e. indicating that the service cannot be provided with the required coverage, regardless of the number of users).

Else, $$cc = \frac{k}{\sum_{i=1}^{k} \frac{R_{\min}}{r_i}} \quad ,$$

Letting N become large, *cc* approaches the expected value of the number of users that can be supported by the system for that service with the given coverage (i.e. x%).

If a user communicates directly with BS, r is its effective rate to BS.

4.2.3 Method 2: Detailed Combined Coverage and Capacity Index Evaluation

The following is a more detailed methodology to evaluate the combined coverage and capacity metric.

Coverage reliability for a particular system (cell radius, shadow fading environment, relay station placement if present, and so on) with a particular number of users n each requiring information throughput $R_{\min}$ is calculated using a static system simulator. The static simulator shall model all other-user interference affects using appropriate path loss models and power control models (if any). The static simulator shall model a scheduler and resource manager that allocates resources to as many users as possible (and all relays supporting those users if present,) such that the target information throughput $R_{\min}$ is achieved. The static system simulator is run repeatedly with each run modeling a different instance of random drops of n MSs. Each simulator run results in $n_{s,i}$ MSs being served with the required information throughput and $n_{b,i}$ MSs being blocked due to insufficient carrier to interference plus noise ratio and/or insufficient time-frequency (or power) resources. $n = n_{b,i} + n_{s,i}$. In this equation, i is an index identifying a particular simulation run. Coverage reliability is a function of n and is:

$$\frac{1}{M \times n} \sum_{i=1}^{M} n_{s,i}$$

where M is the total number of simulation runs. The Combined Coverage and Capacity Index *cc* is the largest n for which

$$\frac{1}{M \times n} \sum_{i=1}^{M} n_{s,i} > x$$

4.3 Definitions of Performance Metric

4.3.1 System data throughput

The data throughput of a BS is defined as the number of information bits per second that a site can successfully deliver or receive using the scheduling algorithms.

4.3.2 Packet call throughput:

Packet call throughput which is the total bits per packet call divided by total packet call duration.

$$\text{Packet Call Throughput} = \frac{1}{K}\sum_{k=1}^{K}\frac{\text{bits in packet call k}}{(t_{end\_k} - t_{arrival\_k})}$$

4.3.3 Effective system spectral efficiency

Effective system spectral efficiency should be normalized by the downlink/uplink ratio of TDD system. For the DL case:

$$\text{DL Site Spectral Efficiency} = \frac{\text{DL System Data Throughput}}{\text{Total Site BW allocated to DL}}$$

Both physical layer spectral efficiency and MAC layer spectral efficiency should be evaluated. Physical layer spectral efficiency should represent the system throughput measured at the interface from the physical layer to the MAC layer, thus excluding physical layer overhead but including MAC and upper layer protocols overhead. MAC layer spectral efficiency should represent the system throughput measured at the interface from the MAC layer to the upper layers, thus including both physical layer and MAC protocols overhead.

The MAC efficiency of the system should be evaluated by dividing the MAC layer spectral efficiency by the physical layer spectral efficiency.

4.3.4 CDF of data throughput per user

The throughput of a user is defined as the ratio of the number of information bits that the user successfully received divided by the amount of time the user was actively involved in data packet transfer.

4.3.5 The CDF of packet delay per user

CDF of the packet delay per user provides a basis in which maximum latency, x%-tile, average latency as well as jitter can be derived.

4.3.5.1 Maximum Packet Latency

The maximum packet latency is defined as the maximum interval between packets originated at the source station (either MS or BS) and received at the destination station (either BS or MS) in an system for a given packet call duration.

4.3.5.2 X%-tile Packet Latency

The x%-tile packet latency is simply the packet latency number in which x% of packets have latency below this number.

4.3.5.3 Average Packet Latency

The average packet latency is defined as the average interval between packets originated at the source station (either MS or BS) and received at the destination station (either BS or MS) in a system for a given packet call duration.

4.3.5.4 The CDF of X%-tile Packet Latencies

The CDF of x%-tiles of packet latencies is used in determining the y%-tile latency of the x%-tile per user packet delays.

4.3.5.5 The Y%-tile of X%-tile Packet Latencies [3]

The y%-tile is the latency number in which y% of per user x%-tile packet latencies are below this number. This latency number can be used as a measure of latency performance for delay sensitive traffic.
A possible criteria for VoIP, for example, is that the 98th %-tile of the 98%-tile of packet latencies per user is 30ms.

4.3.5.6 Jitter

This parameter defines the maximum delay variation (jitter) for the packets of a given packet call duration in a system.

4.3.6 Packet Loss Ratio

The packet loss ratio per user is defined as:

$$Packet\ Loss\ Ratio = \frac{Total\ Number\ of\ Successfully\ \mathrm{Re}ceived\ Packets}{Total\ Number\ of\ Successfully\ Transmitted\ Packets}$$

Typically for a VoIP application, 2% packet loss ratio is tolerable. For gaming and video streaming applications, packet loss ratio is typically less than 1%. Both the single link packet latency and the packet loss ratio per user are important performance metrics for assessing different QoS schemes.

4.4 Fairness Criteria

It may be an objective to have uniform service coverage resulting in a fair service offering for best effort traffic. A measure of fairness under the best effort assumption is important in assessing how well the system solutions perform.

The fairness is evaluated by determining the normalized cumulative distribution function (CDF) of the per user throughput. The CDF is to be tested against a predetermined fairness criterion under several specified traffic conditions. The same scheduling algorithm shall be used for all simulation runs. That is, the scheduling algorithm is not to be optimized for runs with different traffic mixes. The owner(s) of any proposal are also to specify the scheduling algorithm.
Let $T_{put}[k]$ be the throughput for user k. The normalized throughput with respect to the average user throughput for user k, $\tilde{T}_{put}[k]$ is given by

$$\tilde{T}_{put}[k] = \frac{T_{put}[k]}{\underset{i}{avg}\, T_{put}[i]}.$$

4.4.1 Fairness Index

Since CDF does not provide a quantitative measure of fairness it is important to define a metric to measure fairness. Since fairness of a system can be increased by providing more resources to low rate users which result in a reduction of the system capacity, when performance is measured it is important to specify the associated fairness. Then, the performance of two systems can be compared under same fairness conditions. For this purpose, fairness index of a resulting throughput distribution is defined as,

Fairness Index (FI) $= e^{-\sigma}$

where σ is the standard deviation of the normalized per user throughput distribution.

Note that higher the *FI* higher is the fairness of a system and *FI* =1 corresponds to the case where all the users receive same throughput.

Depending on the service type and test case being simulated, different fairness requirements may be specified. Three such fairness criteria are specified in this document for this purpose. The evaluation methodology should specify what fairness criterion has to be met for a given test case.

4.4.2 Equal Throughput or Full Fair Criterion:

To satisfy equal throughput requirement, all the users who are admitted to the system should get equal per user throughout if they have same amount of traffic to send/receive. In a full queue scenario, where traffic is assumed to be always available for transmission, the equal throughput requirement can be achieved by allocating time slots to users, such that the time allocated during a certain period for that user is inversely proportional to the data rate capability of the user.

If the data rate capability of the ith user is *r(i)*, under the equal throughput criterion, time allocated to each user should be proportional to 1 / *r(i)* (assuming equal input traffic).

The resulting equal aggregate throughput is, $C = \frac{1}{\sum_{i=1}^{n} 1/(r(i)}$

For some systems usch as those involving relays, one of the primary objectives of is to provide uniform service offering across users, so the total aggregate throughput under equal throughput criterion, is a good metric to compare systems.

4.4.3 Moderately Fair Solution:

The CDF of the normalized throughputs with respect to the average user throughput for all users is determined. This CDF shall lie to the right of the curve given by the three points in Table 11.

Table 11. Criterion CDF

Normalized Throughput w.r.t average user throughput	CDF
0.1	0.1
0.2	0.2
0.5	0.5

4.4.4 Minimum Average Throughput Fairness Criterion

This fairness criterion ensures a level of fairness of average user throughputs, with a requirement of a specific minimum average user throughput, R_{min}. The minimum average user throughput that all systems must satisfy is not a normalized value, but rather a minimum performance for a given service. The CDF of the normalized throughputs and the minimum normalized average user throughput, k, with respect to the average user throughput for all users, are determined This CDF shall lie to the right of the curve given by the three points in Table 11, where the line passes through the point {0.5, 0.5}, as in the Moderately Fair Solution.

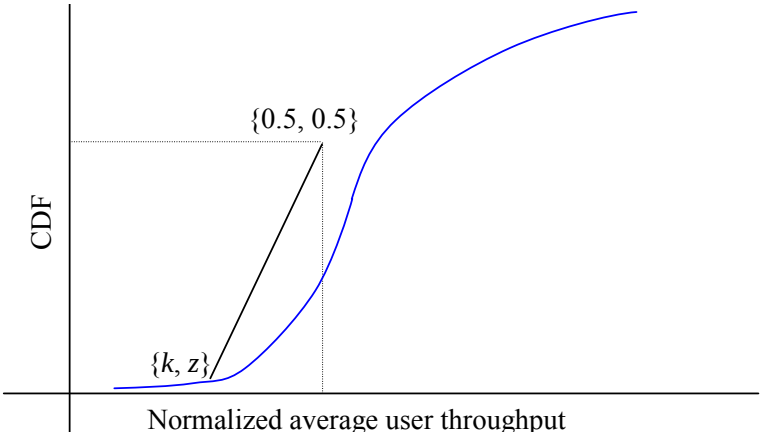
Table 12. Criterion CDF

Normalized Throughput w.r.t average user throughput (x)	CDF
k	z
$k \le x \le 0.5$	$\frac{(1-2z)x+z-k}{1-2k}$
0.5	0.5

As some systems may tolerate some number of users in outage for the entire simulation window, up to a fraction of z users do not have to satisfy the minimum system requirements. The minimum average user throughput R_{min} is set to the service minimum. The allowed fraction of users in outage is z, and is selected based on the service.

The criterion address both coverage for a minimum throughput through point {k, z}, and user throughput fairness through the requirement of being to the right of curve defined by Table 12. Note that this criteria defaults to the moderately fair solution if k = 0.1 and z = 0.1. The equal throughput criteria can achieved by setting k = 1.0 and z = 0.0. Note that for k >0.5 the criteria is effectively described by a single point at {k, z}, which a minimum normalized average user throughput k, given some allowable outage z.

Figure 9. Minimum average throughput fairness criterion



4.4.5 Fairness Criterion to meet a Specified Fairness Index

Under this fairness criterion, the fairness index of the normalized per user throughput should be higher than a target value. This target value is to be specified under each test case. i.e., the fairness requirement is,

Fairness Index of the resulting distribution > target_fairness_index.

5 References

[1] IEEE 802.16j-06/013r3, "Multi-hop System Methodology (Channel Model and Performance Metric), 2007-02-19

[2] IEEE 802.16m-XY/XYZ, "Comments on Channel Model Requirements", March 5, 2007.

[3] 3GPP2/TSG-C.R1002, "1xEV-DV Evaluation Methodology (V14)", June 2003.

[4] IEEE 802.16j-06/013, "Multi-hop System Methodology (Channel Model and Performance Metric), 2006-05-19.

[5] Chen-Nee Chuah, "A Scalable Framework for IP-Network Resource Provisioning Through Aggregation and Hierarchical Control", PhD dissertation, UC Berkeley, 2001.

[6] IEEE P 802.20™ PD-09 Version 1.0, "*802.20 Evaluation Criteria – Ver. 1.0*," September 23, 2005.

[7] IEEE 802.20 Working Group on Mobile Broadband Wireless Access, "*IEEE 802.20 Evaluation Criteria (EC) – Simulation of Common Traffic Types*," May-5-2005.

[8] Hua Xu, Pranav Joshi, Eren Gonen, Y.C. Chen, Xiao Xu, "First Person Shooter Gaming Traffic Model for 802.16jMMR", IEEE 802.16j-06/094.

Appendices

A.1 Multi-Cell Layout

In Figure , a network of cells is formed with 7 clusters and each cluster consists of 19 cells. Depending on the configuration being simulated and required output, the impact of the outer 7 clusters may be neglected. In those cases, only 19 cells (and associated relays if present) may be modeled. These cases are identified in the sections below.

For the cases where modeling outer-cells are necessary for accuracy of the results, the 7 cluster network can be used. However, the six of the seven clusters are just virtual clusters repeating the middle cluster in its surroundings as shown in the figure. Each cell with generic hexagonal grid is separated to 3 sectors, each is formed by a panel directional antennas.

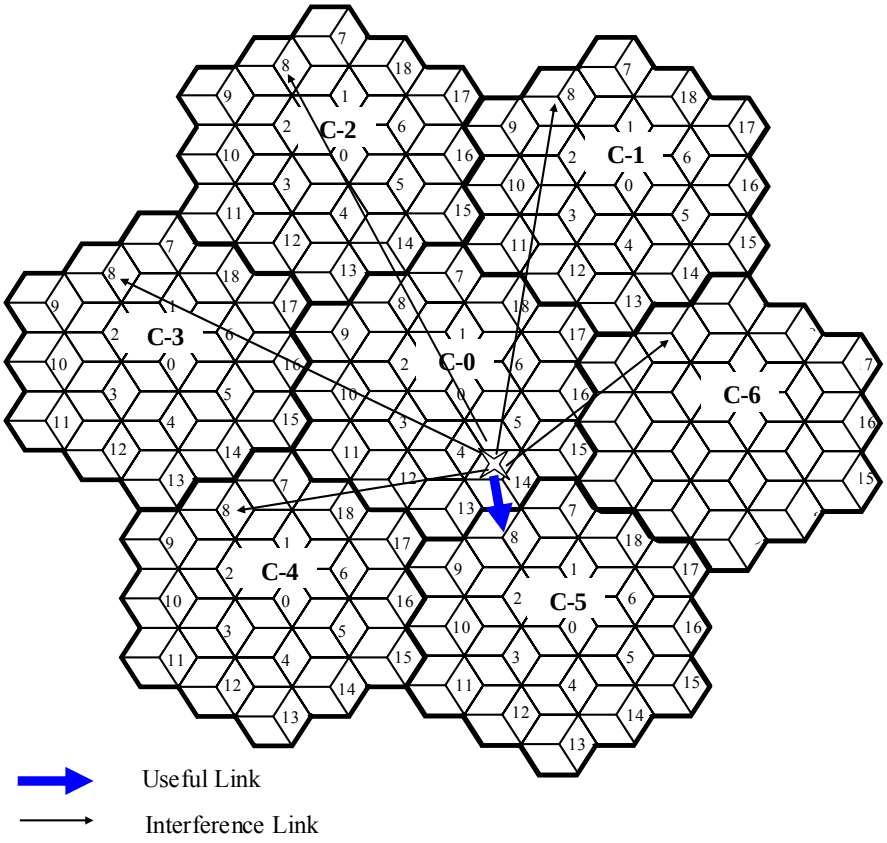


Figure A.1. Multi-cell Layout and Wrap-around Example

A1.1 Obtaining virtual MS locations

The number of MSs is predetermined for each sector, where each MS location is uniformly distributed. The MS assignment is only done in the cluster-0 from where the decided MSs are replicated in the other six clusters. The purpose to employ this wrap-around technique, as will be discussed in later section, is to easily model the interferences from other cells.

A1.2 Determination of severing cell for each MS in a wrap-around multi-cell network

The determination of serving cell for each MS is carried out by two steps due to the wrap-around cell layout; one is to determine the shortest distance cell for each MS from all seven logical cells, and the other is to determine the severing cell for each MS based on the strongest link among 19 cells related to the path-loss and shadowing.

To determine the shortest distance cell for each MS, the distances between the target MS and all logical cells should be evaluated and select the cell with a shortest distance in 7 clusters. Figure 2 illustrates an example for determination of the shortest distance cell for the link between MS and cell-8. It can be seen that the cell-8 located in cluster-5 generates the shortest distance link between MS and cell-8.

To determine the severing cell for each MS, we need to determine 19 links, whereby we may additionally determine the corresponding path-loss, shadowing and transmit/receive antenna gain in consideration of antenna pattern. The serving cell for each MS should offer a strongest link with a strongest received long-term power. It should be noted that the shadowing experienced on the link between MS and cells located in different clusters is the same.

B Link Budget

The link budget can be divided into two parts: The system gain reflects the performance of the transmitter and receiver, including aspects such as antenna gain and receiver sensitivity. The link budget template and values are TBD.

C Link-to-System Mapping

As outlined in Section 1, system level simulations use link level curves in order to simplify the determination of packet error or success on the system level simulations. The link level curves are generated in AWGN channels and determine packet error rate (PER) performance for each modulation and coding scheme used in the simulation. PER curves over a given signal to noise (SNR) ratio range are produced.

C.1 Equivalent SNR Method

In system level simulations, an encoder packet may be transmitted over a selective channel. For example, OFDM systems may experience frequnecy selective fading, and hence the channel gains of each subcarrier may not be equal. Additionally, the channel gains of subcarriers can change in time due to fading process and possible delay involved in HARQ re-transmissions. The result on a transmission of a large encoder packet is encoded symbols of unequal signal-to-noise ratios at the input of the decoder due to the selective channel response over the encoder packet transmission. As the link level curves are generated assuming a flat channel response at given SNR, an effective SNR, SNR_{eff} is required to accurately map the system level SNR onto the link level curves to determine the resulting probability of packet error.

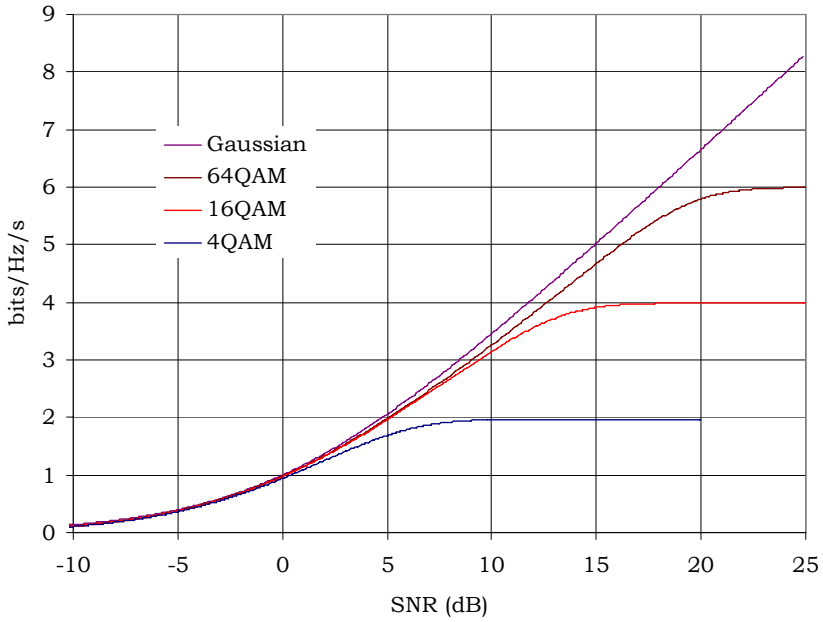
The Equivalent SNR Method, ESM, with the Convex Metric (ECM) [3] is a useful approach to map link-to-system level SNR's. This method maps the SNR of each segment (symbol or set of symbols) of the encoder packet to a capacity curve for a given modulation format. The mean of component capacities is then found, and mapped back to an effective SNR.

Consider the following procedure. Given a set of N received encoder symbol SNR from the system level simulations, $SNR_1, SNR_2, SNR_3, \ldots, SNR_N$, the channel capacity for each component SNR is found by mapping to the channel curve for given modulation scheme. The curves are shown in figure C.1. In this example, the

process will be demonstrated by mapping to the capacity curve for Gaussian signaling. This gives a capacity of $C_n = \log_2(1 + QSNR_n)$ for the *nth* received sample. The mean capacity of the of the symbols is then found, $\overline{C} = \sum_{n=1}^{N} C_n$, and then mapped back to the effective SNR by demapping from the capacity curve, $SNR_{eff} = \frac{2^{\overline{C}} - 1}{Q}$. The effective SNR can then be used to acquire accurate link level performance. The effective SNR takes into the account the variation in encoder symbol gains for each channel realization.

In order to calibrate link-to-system mapping properly, it may be necessary to use an additional factor of *Q*. Link level curves in each channel model for each MCS are required for full calibration of *Q* values. In some cases, it may also be desirable to accept possible link-to-system offsets and to use *Q* =1 for all cases to simplify the system simulation calibration.

Figure C.1. Approximate capacity curves for QPSK, 16 QAM, 64 QAM, and Gaussian signaling.



C.2 Modeling of System Imperfection

C.2.1 Back-off factors

System imperfections such as imperfect channel estimation must also be included in the system level simulations in order to properly determine system performance. In order to minimize complexity of system level simulations, it may be sufficient to use back-off factors applied to each SNR value to account for the performance degradation due to system imperfections. Back-off factors can be obtained from long-term performance curves of the system and link level simulations. For example, system level simulations with calibrated back-off factors for channel estimation imperfections should be able to reproduce long-term link level curves with the actual channel estimation implemented. System degradation due to other imperfections and other impairments may also be represented by back-off factors. Back-off factors for each channel model and MCS, and possibly SNR region may be required.

C.2.2 Other modeling methods

More involved methods of modeling degradation of system level performance due to imperfections are TBD.