Capacity of Packetized Voice Services over Time-Shared Wireless Packet Data Channels

Patrick Hosein
Ericsson Inc.
5012 Wateridge Vista Dr., San Diego, CA 92121.
Email: patrick.hosein@ericsson.com

Abstract—In traditional CDMA wireless networks, real-time services, such as circuit-switched voice, are transported over the air via synchronous channels because of their stringent delay requirements. In the Third Generation networks (3G), 1xEV-DV [1] and HSDPA [2], an additional time shared channel was introduced on the forward link to support data services for which delay requirements are less stringent. The 1xEV-DO [3] standard provides a single time-shared forward link channel and no synchronous channels since it was designed primarily for data services. It has recently been suggested that certain voice services (e.g., Voice over IP (VoIP) and Push-To-Talk (PTT)), can be efficiently transported over such time-shared channels since they have less stringent delay requirements compared to circuit-switched voice. In this paper we investigate the capacity of VoIP users over these time-shared channels and investigate the sensitivity to various base station (BS) and mobile station (MS) design parameters, algorithms and features. Note that detailed simulations of each standard is not provided but rather a comparative approach is used whereby we investigate specific features of each standard. Our focus is on the forward channel since code division multiplexing is performed in the reverse link and hence comparable user capacities are achieved.

I. INTRODUCTION

In traditional CDMA wireless networks, circuit switched voice frames are transported via power controlled synchronous channels over the air in both forward and reverse directions. This means that no queuing of voice frames are necessary at the BS and hence no queuing delays are introduced. In the wireline network a 64 kbps (DS0) channel is used for the voice connection. In order to more efficiently utilize the limited resources (BS transmission power and carrier bandwidth) available over the air, the 64 kbps signal is converted back to analog form and then a more efficient voice codec is used to convert the signal back to digital form. The number of bits per frame produced by the standard codec can take one of four values and hence four bit rates. The maximum bit rate (9.6 kbps) is called full-rate and correspondingly there are half rate, quarter rate and eighth-rate frames. The transmission power required for a physical layer frame will thus be dependent on its corresponding rate. For example, during a silence period (i.e, the MS is listening and not talking), eighth-rate frames can be used with a significant reduction in transmission power requirements. These eighth-rate frames maintain the background noise of the silent party and are also used to maintain the integrity of the channel (power control). Therefore at any point in time the transmission power needs of a MS varies with the activity of the speech as well as the interference being experienced by the MS. If the total power required to serve all users exceeds the available power (a limited resource) then the system is said to be in outage and a subset of transmission frames will be incorrectly received due to insufficient power. If this occurs too often then the frame error rates (FER) of one or more users become unacceptable. Therefore, there is some multiplexing gains due to the variation of the power needs of each user but the number of supportable users is limited. This limit determines the sector user capacity.

Time-shared channels were introduced in 3G networks for more efficient transport of data which tends to be bursty but has less stringent delay requirements. These time-shared channels are slotted in time and one or more (as for the case of 1xEV-DO Revision A [4]) users are served within a time slot using all the power that is available to the channel. With this approach, one can take advantage of the fluctuation of a MS’s channel conditions (due to multi-path fading) and serve it during time periods when it is in good radio conditions. When many users are in the system then, at any point in time, typically one or more of them will experience a positive fade (i.e, high channel gain) and can hence be served at a high rate. However, since users are no longer deterministically served, this now introduces variable queuing delays at the BS scheduler. Capacity is increased by allowing increased queuing delays since the scheduler has a larger time horizon for positive fade opportunities. On the other hand, the delay and the delay jitter introduced by this approach will affect voice quality. Also note that, in order to serve a user at a high rate during a time slot, the user must have sufficient data in its buffer. Since the average bit rate for a VoIP session is low (at most 9.6 kbps) this means that several voice frames must be accumulated before transmission at these high rates. If this is not done then users will be served at low rates even when they are in very good radio conditions with a resultant loss in efficiency. The issue then becomes, given that a certain additional delay at the BS can be tolerated, how many users can be supported such that they all achieve frame losses comparable to that experienced by users of the traditional CDM approach. Furthermore, we would like to understand which factors can be varied to improve capacity while maintaining acceptable voice quality and operator costs.

In this paper we first provide a model for the code division multiplexed case and determine the user capacity for this...
model. The parameters for this model are then used for a
generic time-shared channel for which we determine the user
capacity. We then vary several important parameters of this
model and investigate the trade-offs between capacity and user
performance (delay and FER). In particular we investigate user
capacity dependence on:

- **Scheduling Delay:** As the maximum allowed queuing
delay at the scheduler is increased, the user capacity
increases due to increased diversity gains but at the cost of
reduced voice quality.
- **Blanked Silence Period Frames:** Silence periods result
in very low bit rates (eight-rate frames). The transport of
such a low rate stream while maintaining the delay
constraints is very expensive for the time-shared approach
so we investigate the gains achieved by removing such
frames (which provide background noise). Naturally this
degrades the voice quality. In practical systems some
subset of these silence period frames will need to be
conveyed for acceptable voice quality.
- **Frame Loss Rate:** The voice Frame Loss Rate (FLR) is
the sum of the Frame Discard Rate (FDR) and the Frame
Error Rate (FER). The FDR is the rate at which voice
frames are discarded because they exceed the maximum
scheduling delay. The FER is the rate of errored over
the air frames. By increasing the target FLR one can
increase capacity but again at the cost of voice quality.
Also one can adjust the performance criterion for capacity
by decreasing the percentage of users that must achieve
the specified FLR (i.e., reduced outage requirements).
- **Multi-User Packets:** Channel efficiency can be increased
by serving multiple users in a single slot if all payloads
can be reliably received by all users to which the packet
is destined.
- **Multiple Packet Data Channels:** By supporting multiple
(code-shared) packet data channels, multiple users can
be served at a time. Since less power and Walsh codes are
available for each channel then the maximum possible
rate is reduced.
- **Cell Size:** As the cell size is decreased then user capacity
increases but at the cost of increased BS density (and
hence operator capital and maintenance costs).
- **Increased Receiver Efficiency:** One can increase user
capacity with the use of multiple receive antennae and/or
improved MS receivers. This will increase capacity but
again at the cost of increased complexity (and hence cost)
of the MS station.
- **User Mobility:** As a user’s speed increases, the potential
user diversity gain also increases. However, the accu-

cacy of the reported channel quality information also
decreases. In a practical system a more conservative
estimate of channel conditions must be made for faster
moving mobiles due to the delay between the time the
measurement of these channel conditions is made by the
MS and the time at which a frame is transmitted to the
MS.

- **Channel Bandwidth and Peak Rate:** Naturally as the
carrier bandwidth is increased the user capacity also
increases. Typically this is not a linear increase. The
increased bandwidth provides increased user diversity
gains (due to more simultaneous uses) and multiplexing
gains (i.e., the coefficient of variation of the sum of the
user channel rates decreases with the number of
simultaneous channels). With a larger bandwidth, higher
peak rates can also be supported.

Papers have been written on the voice capacity of CDMA
networks (e.g., [5], [6], [7]) for code-shared channels and on
the data throughput of time-shared channels (e.g., [8], [9]).
However, little has been published on the voice capacity of
time-shared channels.

Our paper is organized as follows. In the next section
we provide the various models we use for both the code-
shared and time-shared approaches. We determine the model
parameters for the code-shared case that provides a capacity
similar to that experienced in the field. Using these parameters
we then determine the capacity for a baseline time-shared sce-
nario. Next we vary the model to determine the sensitivity of
the capacity to the various parameters and features described
above. We end with some insights gained from the exercise
as well as conclusions and directions for future work.

II. CAPACITY MODEL FOR THE CDM APPROACH

A. Traffic Model

Voice traffic is typically modeled by a Markov chain with a
state for each frame rate. For simplicity we instead model
voice traffic by an ON/OFF data arrival process for each
user. The ON state corresponds to voice activity while the
OFF state corresponds to a silence period. During the ON
period we assume full-rate voice frames arrive every 20ms
with a resultant bit rate of 9.6kbps. We therefore do not
model the lower rate frames during speech activity. However
since our intent is to provide a comparative analysis then our
simplification is reasonable.

During the OFF period (the silence period) we will consider
two cases. First we consider the case in which eighth-rate
frames arrive every 20ms with a resulting bit rate of 1.2kbps.
These frames are used to provide background noise informa-
tion at the receiver. Note that the transport of these low rate
streams is inefficient for time-shared channels. Secondly we
consider the case in which these frames are suppressed and not
transmitted over the air. This will naturally affect the quality
of the voice connection. In practice one can instead transmit
a subset of the eighth-rate frames during the silence period to
maintain some degree of realistic background noise but we do
not investigate that alternative. We call the fraction of time a
user’s connection is in the ON state, the activity factor, and
assume a default value of 0.4.

B. MS Location Model

We assume 19 hexagonal cells configured as shown in
Figure 1. Each cell consists of three sectors and each sector
antenna has a 120 degree coverage. We focus on the center
cell and in particular a portion of a sector in that cell (the shaded portion). We generate MS positions as follows. Denote the maximum distance between antenna and MS by $d$. This is the distance between the center of the hexagonal and any of its vertexes and is determined from the specified site to site distance for which we use a default value of 2.5 km. We denote the minimum MS distance from the antenna by $\varepsilon d$ and this has a default value of 35 m.

First we randomly generate a point within the two concentric circles centered at the antenna and with radii $\varepsilon d$ and $d$. We denote the MS’s distance from the antenna by $dx$ where $\varepsilon \leq x \leq 1$ is a random variable. One can show that the CDF of $x$ is given by

$$F(x) = \frac{x^2 - \varepsilon^2}{1 - \varepsilon^2} \quad \text{for} \quad \varepsilon \leq x \leq 1. \quad (1)$$

We focus on the shaded triangle depicted in Figure 1 and generate an independent random variable $\theta$ that is uniformly distributed between 0 and 30°. If $x \cos(\theta) < \frac{\sqrt{3}}{2}$ then the sample point is used (i.e. it lies within the hexagonal) otherwise the process is repeated (see Figure 2).

### C. Path Loss and Fading Models

Given the randomly generated distance $dx$ between the MS and antenna, we use the Modified Hata Urban Propagation Model to determine a path loss of $28.6 + 35 \log_{10}(dx)$ dB. Shadow Fading is modeled by a Log-Normal distribution with a Standard Deviation of 8.9dB. This random variable is obtained by generating a zero mean, unit Standard Deviation Gaussian random variable $s$. The Shadow Fading is then given by $8.9s$ dB. Fast fading gains are assumed to be Exponentially distributed and is denoted by $\gamma(t)$ at time $t$. Therefore the total gain (in linear units) experienced by a MS at time $t$ is given by the product of these three terms.

### D. Inter-cell Interference Model

The interference experienced by a MS is taken as the sum of the interference from all other sectors. Since we assume that the power radiated by a sector antenna is confined to 120 degrees then only one sector from each cell contributes to the interference on the MS. For each of the 18 sectors we compute the path loss using our knowledge of the MS location. The shadow fading from each sector is then generated and we follow the approach in the 3GPP2 strawman [10] to maintain 0.5 correlation between the shadow fading of any two of these sectors. The path loss and shadow fading is generated once per drop and the resulting interference is determined for each sector and summed over all sectors. We ignore the fast fading in this computation but do consider it when determining the intra-cell interference on the MS.

If the transmission power is denoted by $P$ and the inter-cell interference is denoted by $I$ then the SINR of the signal received by the MS denoted by $\Gamma(t)$ is given by

$$\Gamma(t) = \frac{P10^{0.89s}\gamma(t)}{10^{28.6}d^{3.5}x^{3.5}I} = g10^{0.89s}\gamma(t) \cdot x^{3.5}I \quad (2)$$

where $g$ contains all of the MS independent constant terms and $x, s$ and $\gamma(t)$ are all random variables.

### E. Soft Handoff Model

In the case of code-shared channels, a MS receives signals from all BSs in its active set [11]. It then combines these to determine the corresponding frame. This provides additional macro-diversity gains that are not possible with the TDM case. We model this macro-diversity gain as follows. For each MS in the concerned sector we compute the strength of the pilot received from each of the 19 sectors that faces it. If this pilot strength exceeds a specified value then the corresponding sector is added to the active set. This process determines the active set of each MS in the concerned sector. The pilot strength threshold (commonly referred to as the T-ADD threshold) is chosen so that the average number of sectors in the active set is comparable to what is observed in the field.

### F. Link Power Budget

We assume that the forward link pilot uses 17.8% of the total power (20W). The total overhead power (for pilot, paging and
synchronization) is assumed to be 25%. We assume that the average traffic power utilization is 80% (due to the fluctuation of the power needs of each MS). Therefore the average power that causes interference to other sectors is 85% of the maximum power.

G. Performance Criteria

A closed-loop algorithm (not modeled) is used to maintain the average Frame Error Rate (FER) of each user at some target value (for circuit-switched voice channels this target is 1%). We define user capacity as the maximum number of users that can be supported such that at most 2% of users experience a FER greater than 1%.

H. User Capacity for the CDM Case

In our model we assume that in the ON state sufficient power must be provided to each active connection to achieve the desired throughput of 9.6 kbps with the specified constraint on FER. In the OFF state we assume that sufficient power is assigned to the channel to maintain a throughput of 1.2 kbps (i.e., eighth-rate frames). We also include a component for background noise. We include intra-cell interference (i.e. interference from transmissions to all other MSs within the sector). Using the power budgets and assumptions provided above, our simulations provide a user capacity of 30.

III. Capacity Model for the TDM Approach

We assume a time-shared forward channel. This channel is slotted with slot duration denoted by $\tau$. Within each time slot one or more MSs can be served. We assume that voice frames arrive every 20ms but that each frame can be delayed by up to some multiple of 20ms. As this delay is increased, the quality of the voice connection decreases. However, since the period of time over which the frames can be scheduled is increased, the gains possible with opportunistic scheduling also increases which in turn increases the number of users that can be supported. Therefore this allows one to trade voice quality for capacity. We assume the same traffic model, MS location model and path loss model as used for the CDM case.

Note that Hybrid-ARQ (H-ARQ) can still be used even if the allowed delay is 20ms. Assuming a slot size of 5/3 ms and 4 H-ARQ channels, a frame can be transmitted a total of 3 times within a 20 ms period and hence we account for H-ARQ gains. In the case of multi-user packets, the frame is re-transmitted until all MSs served within that slot correctly receive their frames or until the number of allowable re-transmissions (which we assume to be 2) is reached. Note that this leads to some inefficiency since those frames that may have been correctly received in a multi-user frame are re-transmitted if one or more of the other voice frames are incorrectly received.

Finally we need to take into account scheduling so that users are served during positive fades. However, unlike data, there is a constraint on the maximum delay allowed per voice frame. Any frames that have reached this maximum must be discarded.

A. The Rate Estimation Model

Let $\Gamma(t)$ denote the $C/I$ of the signal received by a MS during a slot. We assume that Channel Quality Information (CQI) is reported to the BS every time slot. In the case of 1xEV-DO (for which the F-PDCH transmission power is fixed) the rate that can be supported by the MS is reported in a Data Request Channel (DRC) message. The BS uses the CQI information to determine the appropriate modulation and coding scheme to be used in order to achieve some specified residual FER (i.e., the FER after all physical layer re-transmissions are taken into account). The mapping from the reported $C/I$ information to the modulation and coding scheme (MCS) is determined in advance and a simple table lookup can be used by the BS. We model this mapping with a Shannon capacity bound but instead use an “effective” channel bandwidth as opposed to the real bandwidth, to take into account overheads and the fact that because of finite block lengths the Turbo codes used do not achieve the Shannon bound. Therefore the rate for a given $C/I$, $\Gamma(t)$, is given by

$$R(t) = B \log_2(1 + \Gamma(t)),$$

$\Gamma(t)$ is determined as described in Section II. We limit the maximum value of $R$ to the maximum rate supported by the standard under consideration. Note also that only a finite set of rates are supported whereas we assume any rate, up to the maximum rate value, can be used.

In Figure 3 we plot the MCS mapping used for 1xEV-DV together with the Shannon approximation. In this case the effective bandwidth is $B = 900$kHz. Note that this approximation is reasonable for the range of $C/I$ values considered.

In the case of HSDPA, which has a carrier bandwidth of 5 MHz, we use the same model. However, in this case the efficiency increases with the $C/I$ value and so the effective bandwidth is larger for large $C/I$ values. For a $C/I$ value of $x$ dB, the bandwidth $B$ is multiplied by the factor $0.2x^2$. In Figure 4 we plot this approximation together with a typical MCS mapping. In this case we find that $B = 2800$ kHz.
for a $C/I$ value of zero. Note that the ratio of the effective bandwidths of HSDPA and CDMA2000 is approximately 3.1.

B. MS Channel Model

We use a single channel model for our baseline problem (pedestrians moving at 10 mph). However, we also simulate a mix of user channel models each resulting in different fast fading fluctuations. The mix is based on that defined in the 3GPP2 strawman document.

C. The Fast Fading Model

The user capacity is dependent on the mobility of the MS. In order to capture this dependency we independently simulate the channel quality of each MS (sampled every time slot) following the Jakes model. We assume a feedback delay of two time slots. Therefore, if a user is scheduled based on its channel quality at time slot $n$, the channel conditions experienced during transmission is that for time slot $n + 2$. The path loss and shadow fading for the MS is computed as was done for the CDM case.

If the average fade duration of a MS is larger than its maximum queuing delay then it is impossible to serve the user’s voice frames during a negative fade. Figure 5 contains a plot of the normalized fast fading gain (in dB) for a pedestrian moving at 3 kph. Over a 20 ms duration the fading gain fluctuates slowly making it difficult to have any opportunistic scheduling gains. On the other hand, the channel quality is more easily predicted.

Next, in Figure 6, we consider the other extreme of a 120 kph vehicular user. In this case the normalized fading gain varies rapidly (over a 15 dB range) during each 20 ms period. This implies increased opportunistic scheduling gains. However, the channel quality is more difficult to predict because of the increased variation. In this case a very conservative estimate of the channel quality must be made.

D. Hybrid-ARQ

We model Hybrid-ARQ as follows. For each re-transmission we combine the energy of the re-transmission with a fraction of the energies of earlier transmissions. The energy of the second transmission is combined with 58% of the first transmission. The energy of the third transmission is combined with 58% of the second transmission plus 34% of the first transmission. These values were chosen to correspond with observed soft recombining gains obtained from more realistic simulations.

A multi-user physical layer frame is re-transmitted (at most twice) if one or more of the user frames are incorrectly received. This implies that correctly received user frames will also be re-transmitted within the physical layer frame because with Chase combining an identical encoder packet must be used for the re-transmissions as for the first one.

E. Sector Selection

We also assume optimal sector selection. For each user we determine the BS with the best forward link channel conditions (taking into account path loss and Shadow fading) and assume that this BS serves the MS. Note that, in practice, the optimal sector can change rapidly for MSs near the cell edge. Rapid changes to the serving sector of a MS can require significant
signaling resources and may also cause other stability problems. Hence the assumption of always serving the MS from the optimal forward link sector is very optimistic. However, again we emphasize that the objective of this exercise is the relative capacity performance as design parameters are varied.

F. Link Power Budget

For the baseline case we assume a single F-PDCH and that each slot transmission is made with the maximum available power which is 75% of the maximum power. This implies that 100% of the sector power causes interference to adjacent sectors. Note that in the case of 1xEV-DV and HSDPA, the power and number of Walsh codes that are available to the Forward Packet Data Channel (F-PDCH) depends on what is left over from the synchronous channels which are given higher priority. In our analysis we assume that all resources are allocated to the F-PDCH and hence no circuit switched voice calls are in progress. 1xEV-DV can support two F-PDCCHs while HSDPA can support four. When multiple F-PDCCHs are supported another issue that arises is the optimal allocation of power and Walsh codes to each of them. This determination should be coordinated with the scheduler being used and will be discussed further in a later section.

IV. THE TDM SCHEDULER

The scheduler for the time-shared channel plays an important role in determining capacity. We therefore provide a detailed discussion of the simulated scheduler. We first consider the case of a single user assigned per slot with one F-PDCH. We then consider extensions to the scheduler to take into account multi-user packets and multiple F-PDCCHs.

Kelly [12] provides a formulation of this shared resource allocation problem but using utility functions dependent on user throughput. In the case of VoIP, in order to maintain acceptable performance, the frame loss rate must be kept small. This implies that the throughput achieved by the user equals the bit arrival rate. Since all voice conversations have similar characteristics, the average bit arrival rate is the same for all active conversations (i.e., those that are experiencing a speech burst). This implies that all users must achieve comparable long term average throughputs under the assumption of low FLR. However, during short periods of time the frame queue length will vary due to variation in the number of bits per voice frame and fluctuations in the forward link radio conditions due to fading. Therefore, instead of making the utility a function of the user throughput, we make it a function of the user queuing delay. By appropriate choice of this utility function we can achieve the objective of maintaining low frame losses due to discards even for users in poor radio conditions.

Let \( U_i(d_i) \) denote the utility function of user \( i \) when experiencing a queuing delay of \( d_i \). We assume that this function is strictly concave and differentiable in the region \( d_i \geq 0 \). Denote the number of bits that arrive for user \( i \) in period \( n \) by \( b_i(n) \), its achievable rate (determined by its reported \( C/I \)) by \( r_i(n) \), its average throughput by \( \mu_i(n) \) and its present queue size by \( q_i(n) \). Its maximum rate if served is then given by \( r_i(n) = \min\{r_i(n), q_i(n)/\tau\} \). User \( i \)'s expected delay evolves every slot period as follows:

\[
d_i(n+1) = \begin{cases} 
\max \left\{ 0, d_i(n) + \frac{b_i(n) - r_i(n)\tau}{\mu_i(n)} \right\} & \text{if } i \text{ is served,} \\
\frac{d_i(n) + \frac{b_i(n)}{\mu_i(n)}}{\mu_i(n)} & \text{otherwise.}
\end{cases}
\]

(3)

Note that the average throughput will in fact vary based on whether or not \( i \) is served but we assume that \( \mu_i \) is averaged over a sufficiently long period of time and so the difference is negligible. Given this dependence of \( d \) on the decision variable \( r \) we can formulate the optimization in terms of \( d \) as follows:

\[
\text{maximize} \quad F(d) = \sum_{i=1}^{k} U_i(d_i(n+1)) \\
\text{subject to} \quad \sum_{i=1}^{k} d_i(n) > C \\
\text{over } \quad d_i(n) \geq 0, \quad 1 \leq i \leq k.
\]

(4)

(5)

(6)

where,

\( k = \) number of active users competing for the shared channel,
\( d_i(n) = \) the queuing delay of user \( i \),
\( C = \) lower limit on delay.

Note that \( d_i(n) \) cannot be directly controlled but rather is controlled through the choice of the scheduled user. Consider a queuing system with a stochastic service rate (in this case the channel bandwidth) and assume an arrival rate of \( N\lambda \) for some constant \( \lambda \). Due to the variation in the capacity, the queuing delay will vary over time and, for example, it 98th percentile value will grow with \( N \). Therefore we can limit the 98th percentile value by limiting \( N \), and this determines the user capacity. For a given value of \( N \) there is a time varying lower limit, \( C \), on the delay.

Under the assumption that the utility function of each user is strictly concave and differentiable, then the same also holds for the objective function \( F \). Since the feasible region is compact then an optimal solution exists, is unique and can be found by Lagrangian Methods. However, the optimal assignment changes with time due to the dynamic nature of the available resources (which determines \( C \)). Furthermore, resources can only be allocated via the assignment of slots. This implies that, whenever the optimal assignment changes, it may take several slots before one can again achieve optimality and by that time the optimal solution would have changed again. Therefore the best that can be done is to continuously move in the direction of the optimal solution. We achieve this using a gradient ascent algorithm. At each decision point (i.e., whenever a new slot assignment must be made), we compute, for each user, the gradient of \( F \) along the direction of serving that user and then we serve the user with the largest gradient.

Let us parameterize the movement along the ray corresponding to serving user \( j \) by \( \alpha \). The objective as a function of \( \alpha \) is then given by
\[
F_j(\alpha) = \sum_{i=1}^{k} U_i(d_i(n) + \alpha(d_i(n+1) - d_i(n))) \tag{7}
\]

Taking the derivative with respect to \( \alpha \) and evaluating the derivative at \( \alpha = 0 \) (corresponding to the derivative at the present point) we get

\[
F_j' = U_j'(d_j(n)) \frac{-r_j(n)}{\mu_j(n)} + \sum_{i=1}^{k} U_i'(d_i(n)) \frac{b_i(n)}{\mu_i(n)} \tag{8}
\]

Since the second term on the right hand side is common to all users then the user, \( j^* \), with the largest gradient can be obtained from

\[
j^* = \arg \max_j \left\{ \frac{-r_j(n)}{\mu_j(n)} U_j'(d_j(n)) \right\}. \tag{9}
\]

Interestingly, for the utility function \( U(d) = -kd \), the user with the largest ratio of achievable rate and throughput is picked. This is simply the proportional fair scheduler. If \( k \) is made user dependent then the result is a weighted proportional fair scheduler.

In a traditional gradient ascent algorithm, the direction with the largest gradient is first determined. Next, the optimal point along this direction is determined by solving the corresponding one-dimensional optimization problem. However, in our particular case, the objective function increase for the chosen user depends on the number of bits that can be transmitted in the slot and hence movement to the optimal point along the direction may not be possible.

In the case of VoIP, frame delay and loss rate determines the call quality. An end-to-end delay of at most 200-250 ms can be tolerated with a maximum voice frame loss rate of at most 1-2\%. We can budget these metrics across the various components through which the voice frames must pass, the access link (which could be wireless), the core network and the egress link which we assume is wireless. For the forward wireless link, a typical budget can be 140 ms for the scheduling and transmission delays and 2\% for the frame loss rate. We can enforce the 140 ms budget by dropping any frames that will experience a larger delay. If we do this then there are two sources of frame losses, those due to discards in the scheduling buffer and those due to transmission errors. The scheduler can be used to trade these two types of losses and ideally they should be made roughly equal.

We define the user capacity as the maximum number of users that can be supported such that, for each user, the scheduling delay is at most \( d_{\text{max}} \) and the frame loss rate (due to both discards and transmission errors) is at most \( \delta \). The scheduler must therefore be designed with these objectives in mind. If we schedule independently of the delay constraint then, as the number of users increases, those at the cell edge will not be sufficiently served and will exceed the required frame loss rate while those users nearer the antenna will be served often and will have a frame loss rate far below the threshold. Naturally, user capacity can be increased by assigning more slots to the poor radio condition users and reducing the number assigned to the good radio condition users. Ideally this should be done in such a way so that as the number of users is increased, all users simultaneously reach the frame loss rate threshold. If this is achieved then the system will accommodate the maximum number of users.

The above objective (allocating resources so that all users achieve comparable frame loss rates) can be achieved by including a delay constraint in the optimization problem. Instead of introducing a hard constraint we instead introduce a soft constraint by using a barrier utility function (see [9] for an explanation of barrier functions) that incurs a high cost for large delays. In this way the influence of the delay constraint is more gradual. This allows the system to react earlier to the needs of users that are approaching their delay limit rather than being forced to do so only when the limit is reached. In addition, when all users are experiencing low delays then the scheduler should try to achieve as much user diversity gains as possible.

We introduce the barrier function as the utility function. In other words, we assume a utility function that provides a high penalty for high delay users. This in turn reduces the frame loss rate since delays are kept small. Consider, for example, the following utility function,

\[
U(d) = \begin{cases} (d_{\text{max}} - d)^{1-\alpha} & \text{for } \alpha > 0, \ \alpha \neq 1, \\ \log(d_{\text{max}} - d) & \text{for } \alpha = 1, \end{cases} \tag{10}
\]

where \( \alpha \) is a parameter that can be fine-tuned to achieve optimal user capacity. With this utility function the user chosen at each scheduling decision point is given by

\[
j^* = \arg \max_j \left\{ \frac{r_j(n)}{\mu_j(n)} (d_{\text{max}} - d_j(n))^{-\alpha} \right\}. \tag{11}
\]

In the case of VoIP we can make some simplifications to the algorithm. Note that since the frame loss rate is small then the throughput \( \mu_j \) equals the bit arrival rate. The frame rate required at any point in time depends on the activity of the voice call at that instant. In practice the average bit rate will be nearly the same for all voice calls. This implies that all users will experience the same long term throughput and hence we can assume \( \mu_j(n) = \mu \) for all users in the long term. Therefore the throughput component in 11 which requires processing resources to continuously update can be removed since it is the same for all users.

Note however, if we make this simplification then we need to address the following issue. Consider the last few frames at the end of a talk burst. Since no new frames arrive, the queue size does not increase and hence the user’s priority remains low and the frames are stuck in the queue and will be useless to the MS when they are finally transmitted. There are ways to monitor this situation and address it. Had we used the throughput \( \mu_i \) in the priority function, then if the MS is not served then the average throughput drops and even though the queue does not increase, the MS’s priority will
increase because of its low throughput. Therefore we suggest that, if maintaining the average throughput per MS is not very costly, then the more accurate priority function, which includes average throughput, should be used.

Ideally the queuing delay $d$ used in the scheduling algorithm should be that experienced by the frame at the tail of the queue. However, this cannot be known in advance. One can use the queuing delay that was experienced by the frame at the head of the queue. This can be obtained by placing a time stamp on new frame arrivals and checking the elapsed time when the frame is served. However, this delay information will be old and may not accurately reflect the delays of new frame arrivals. We can instead compute the expected delay of a newly arriving frame. The frame arrival rate is constant at one per 20 ms. Since the system is near lossless then the average frame departure rate must also be one per 20 ms (one can discount this rate based on the frame loss rate but for small frame loss rates of 1-2% the effect is negligible). Therefore, if there are $q$ frames in the queue then the expected queuing delay is $20q$ ms. We can thus make the scheduler dependent on queue length, measured in voice frames, rather than on delay. If we let $q_{max} = \frac{d_{max}}{20}$, and make the assumption that all user throughputs are equal, then we can instead determine $j^*$ from

$$j^* = \arg \max_j \left\{ \frac{r_j(n)}{q_{max} - q_j(n)} \right\}^{\alpha}. \quad (12)$$

Therefore, when all queues are small (light loading), then essentially the user with the highest requested rate (max C/I algorithm) is served. As the queue for a particular user increases (e.g., because it is in poor radio conditions) then its priority level increases and it is served before reaching its delay threshold. We can compute the queue dependent utility function for the priority function given in 12 as

$$U(q) = \begin{cases} \frac{(q_{max} - q)^{1-\alpha}}{1-\alpha} & \text{for } \alpha > 0, \ \alpha \neq 1, \\ \log(q_{max} - q) & \text{for } \alpha = 1. \end{cases} \quad (13)$$

In Figure 7 we plot this utility function for various values of $\alpha$ (shifted vertically so they all start from the origin for a more convenient comparison). Note that, as $\alpha$ increases, the penalty incurred as the maximum delay is approached is more severe. The result is that delays are forcibly kept at low values. However, this in turn reduces user diversity gains since users are scheduled based more on their queue length rather on their radio conditions. For small values of $\alpha$ scheduling is more dependent on radio conditions for most delay values and it is only when the delay almost reaches its maximum value that the scheduler reacts. However, this late reaction will lead to more frame losses and hence higher FLRs. Therefore by fine tuning $\alpha$ we can satisfy the desired frame loss rate requirement while attaining some user diversity gains. Note also that as $\alpha$ is increased, the probability of packet discards decreases but the probability of transmission losses increases since a MS is forced to transmit irregardless of radio conditions. We found $\alpha = 2$ is suitable for most scenarios.

A. Scheduling for Multi-User Packets

With 1xEV-DO Revision A, up to eight users can be served in a single slot. The frame carries sufficient information for each MS to determine if it contains any of its data. Acknowledgments and negative acknowledgments are performed by each user to which the packet is destined. The BS re-transmits the frame as long as one or more users have not received their data correctly. This feature adds a new dimension to scheduling. In addition to maintaining the delay requirements for each user while trying to achieve diversity gains, the scheduler must now also find the optimal combination of users to serve within a slot.

We can again formulate this as an optimization problem and use a gradient ascent approach as follows. At each decision point we find all feasible search directions. A search direction corresponds to a subset of users who can be simultaneously served within a time slot. These are users who have sufficiently good radio conditions to support the rate necessary for the larger payload made up of all individual user payloads. We can then determine the gradients for each of these search directions and pick the one with the largest gradient. Although the number of search directions can be significantly pruned using various properties of the optimal solution, it may still be computationally intensive to do such an exhaustive search.

We use a simpler but sub-optimal approach. At each decision point we order users based on the criterion given by 12. Starting from the top of the list we pack as many user data into the frame as possible. If more room is available we go to the next user on the list and see if any of its data can be feasibly packed and, if not, then we continue on to the next user and repeat until no more data can be packed or we have exhausted all users. Note that this is suboptimal since, for example, an assignment to one user may prevent a more beneficial assignment of two later users.
B. Scheduling over Multiple F-PDCHs

We next consider scheduling over multiple packet data channels. Here we assume only a single user is allocated per slot per channel. A new issue that now arises is the optimal assignment of power and Walsh codes to each of these channels. Furthermore, since each incoming MS has to be assigned to a F-PDCH then one also has to consider the load balancing problem.

One can formulate the combined power/Walsh code allocation and user scheduling problems as an optimization problem and solve as we have done previously. In order to simply get an idea of the benefits of multiple F-PDCHs we provide a lower bound on capacity by considering independent channels each with an equal share of power and Walsh code resources. Scheduling within a channel is done as previously described for the single user per slot, single F-PDCH, case.

V. CAPACITY AND SENSITIVITY FOR THE TDM CASE

A. The Baseline Case

We first determine the capacity for a baseline scenario and then investigate the effect on capacity as parameters for this baseline problem are varied. For this baseline we assume, (a) a maximum scheduling delay of 140ms, (b) blanked silence period frames, (c) a target FLR of 2%, (d) Multi-user packet transmissions, (e) a single F-PDCH, (f) a site to site distance of 2.5 km, (g) a single receive antenna, (h) a bandwidth of 1.25 MHz, (i) a single channel model (10kph pedestrian) for all users, and (j) a slot size of 5/3 ms. Capacity is defined as the maximum number of users that can be supported such that no more than 2% of users experience an FLR greater than 2%. Under these assumptions we obtain a capacity of 25 users.

B. Scheduling Delay

In this section we investigate the sensitivity of capacity on scheduling delay. In Figure 8 we plot the capacity for the baseline problem as a function of the maximum scheduling delay. The low capacity for low delays is due to the fact that for a 10 kph user, the scheduling delay is too small to wait for the user to get out of a deep negative fade. However, for larger delays the capacity increases steadily up to 44 users at 200 ms.

C. Blanked Silence Period Frames

In this section we investigate the loss in capacity if eighth-rate frames are transmitted during the silence period (as done for the CDM case). We simulated the situation where, during the silence period, eighth-rate frames arrive and are transmitted over the air interface. However, we provide relaxed delay requirements for the eighth-rate frames since they are not as essential as the full-rate frames. We assume that the delay threshold for the eighth-rate frames is eight times that of the full-rate frames (i.e., the same sized buffer is used whether the conversation is in active or silent mode). Therefore even in this case the voice quality is not as good as that obtained with circuit switched voice over a code-shared channel. Still, for this case we find that the capacity is significantly less than the capacity of the baseline case. In Figure 9 we plot the capacity as a function of delay for both cases.

D. Frame Loss Rate

The Mean Opinion Score (MOS) values for speech quality is highly dependent on the frame loss rate and hence there is very little room for varying this quantity. We investigated the dependence of capacity on both the FLR as well as the outage criterion used in determining capacity (i.e., the percentage of users that must have an FLR less than the target value). We find that user capacity is very sensitive to these values and that supporting the 1% FLR target used for the code-shared case is difficult for a time-shared channel.

The FLR is due to both packet discards at the scheduler (the FDR) and frame transmission errors over the air (the FER). In the baseline case the objective was that 98% of users must achieve a FLR no greater than 2%. In Figure 10 we provide, for the baseline case, the average (over all users and runs) of the packet discard rate and the frame transmission error rate. In this particular case, packet discards dominate over transmission errors. One can adjust this relationship by adjusting

![Fig. 8. Capacity versus Maximum Scheduling Delay](image)

![Fig. 9. Effect on Capacity of Silence Period Frames](image)
\begin{table}[h]
\centering
\caption{Capacity dependence on targeted FLR and Outage values}
\begin{tabular}{|c|c|c|}
\hline
FLR & Outage Percentage & Capacity \\
\hline
1% & 98% & 19 \\
1% & 99% & 17 \\
2% & 98% & 26 \\
2% & 99% & 17 \\
\hline
\end{tabular}
\end{table}

\begin{figure}[h]
\centering
\includegraphics[width=0.5\textwidth]{figure10.png}
\caption{Average Discard and Transmission Error Rates}
\end{figure}

$\alpha$ in the priority function or by adjusting how aggressively the radio conditions are predicted given the reported channel quality information.

Note that the average values are quite small which indicate that most users have low losses. The capacity is determined by those users at the cell edge that require significant resources to maintain their FLR. In the code-shared case these edge users were able to take advantage of the macro-diversity gains (i.e., receiving signals from multiple BSs) whereas for the time-shared case the gains come from sector selection.

\subsection*{E. Multi-User Packets}

The baseline problem assumes multi-user packet transmissions. However some 3G standards do not support this feature. In Figure 11 we plot the capacity versus delay for the baseline problem together with the case where we assume single user per slot transmissions. It shows that there is a significant benefit to serving multiple users simultaneously.

\subsection*{F. Multiple Packet Data Channels}

In this section we consider the benefit of supporting multiple F-PDCHs. Here we face the problem of the optimal allocation of resources between these channels on a slot by slot basis. Note that a special case is the assignment of all resources to a single channel which will provide similar performance to the single channel case. We investigated the other extreme, which is that all channels are used and resources are shared evenly among them. However, the performance under this assumption is poor for the following reason. Those users in poor radio conditions which can be supported if all power is used when transmitting frames to them, may not be supportable when that power is reduced by a factor of two or more. This implies that the flexibility to adjust the amount of resources (power and Walsh codes) assigned to each channel on a per slot basis must be included to experience the benefits of this approach. We can formulate and solve the corresponding joint power and Walsh allocation/slot scheduling optimization problem but have left this as a future research topic.

\subsection*{G. Cell Size}

Next we consider the effect of cell size on the user capacity. In Figure 12 we plot the user capacity as a function of delay for the baseline case cell separation of 2.5 km and also for a separation of 2 km. Note that there is a significant increase in capacity as the cell is shrunk. However the smaller cell size implies higher capital and maintenance costs.

\subsection*{H. Increased Receiver Efficiency}

By increasing the number of antennae on the MS and/or improving the receiver one can increase user coverage and/or
capacity. Simulation of multiple antennae systems is quite complicated. We therefore consider the simplified situation whereby the signal received by the MS is boosted by 3dB. In Figure 13 we plot the capacity as a function of delay for this case as well as the baseline case. We find significant capacity improvement even at low scheduling delays. Therefore dual receive antennae and enhanced mobile receivers should be included in any VoIP-centric network design especially if it replaces circuit switched services.

I. User Mobility

In this section investigate the effect of mobility on user capacity. We consider the cases of pedestrians moving at 10 kph which is the baseline case and pedestrians moving at 30 kph. We also simulated a mixed traffic model scenario as provided in Table II. In Figure 14 we plot the user capacity versus the maximum queuing delay. We find that low mobility users require sufficiently large queuing delays to achieve scheduling gains but the resulting capacity (at these high delays) continues to increase with delay because of good channel prediction. On the other hand, high mobility users achieve high scheduling gains with relatively small queuing delays but the resulting system capacity quickly saturates because of poorer channel prediction.

The capacity eventually becomes flat for the following reason. As the maximum delay is increased, service to a particular user can be delayed further until improved radio conditions are experienced. However as the delay before servicing a MS increases, the amount of data accumulated in its buffer also increases and so the rate that it must achieve to clear this data also increases. Therefore, as service to the MS is delayed, the rate that must be achieved also increases. At some point, having this additional room for delay is no longer beneficial because the corresponding rates needed, especially for users at the edge, cannot be achieved and the MS is forced to be served at relatively poor radio conditions (inefficient transmissions) or loses frames that exceed the maximum queuing delay.

J. Channel Bandwidth and Peak Rate

As the channel bandwidth available to the PDCH is increased, two things occur. Firstly, the diversity gain increases because of the larger number of users that can be supported. Secondly, there is an increase in the multiplexing gains. The latter is due to the following. The resources required for the forward link depends on the number of users as well as the voice activity of these users. The sum of the required rate is therefore a stochastic quantity. As the number of users increases, the ratio of the variance to the mean decreases thus allowing the system to operate at a higher level of utilization. In Figure 15 we plot the capacity as a function of delay for the baseline case as well as the case of a 5 MHz carrier bandwidth with a peak rate of 14 Mbps and using single user per slot transmissions. As expected, the increased bandwidth results in a significant increase in capacity even though single user transmissions are being used.

K. Capacity sensitivity to \( \alpha \)

Note that as \( \alpha \) is increased, the FDR decreases (because a user’s priority becomes highly dependent on queue size even for moderately sized queues) while the FER increases (since users are forced to be served even if not in relatively good radio conditions). There exists some optimal value of \( \alpha \) that provides the best trade-off between these two effects. In

---

**TABLE II**

<table>
<thead>
<tr>
<th>Channel Model(%)</th>
<th>Multi-path Model</th>
<th>No. of Fingers</th>
<th>Speed (kph)</th>
<th>Fading</th>
</tr>
</thead>
<tbody>
<tr>
<td>B (30%)</td>
<td>Ped B</td>
<td>3</td>
<td>10</td>
<td>Jakes</td>
</tr>
<tr>
<td>C (20%)</td>
<td>Ped A</td>
<td>2</td>
<td>30</td>
<td>Jakes</td>
</tr>
<tr>
<td>D (10%)</td>
<td>Veh A</td>
<td>1</td>
<td>120</td>
<td>Jakes</td>
</tr>
<tr>
<td>E (40%)</td>
<td>Single Path</td>
<td>1</td>
<td>0</td>
<td>Rician</td>
</tr>
</tbody>
</table>

**Fig. 13.** Capacity dependence on MS receiver improvements

**Fig. 14.** Capacity for Single-Speed and Mixed-Speed Users
Figure 16 we plot the capacity for different values of $\alpha$ for the baseline case but with a maximum queuing delay of 160 ms. Note $\alpha = 0$ results in the max $C/I$ scheduler and clearly this under-performs the queue based scheduler. For $\alpha = 3$ the capacity also begins to drop. A value of $\alpha = 1$ or 2 provides near-optimal capacity. We decided to use $\alpha = 2$ for our simulations but similar results hold for the case $\alpha = 1$. Also note that non-integer values of $\alpha$ can also be used.

VI. SUMMARY, CONCLUSIONS AND FUTURE WORK

In this paper we investigated the efficiency of using time-shared channels for real-time services such as VoIP. We found that, although these channels cannot support circuit switched quality voice services, they can support these lower quality voice applications. We presented a delay or queue based scheduler that can be used for these services and demonstrated its effectiveness. We also investigated the dependence of user capacity on various system and network design parameters. In general, capacity can be improved but typically only through increased cost and/or implementation complexity.

These results are interesting in that it shows that CDM performs well for circuit switched services since it can more easily divide available resources among users to provide exactly what is needed. However, the advantage of the time-shared approach is that a single mechanism can be used to provide all operator services and this reduces the complexity of the design and may even decrease operational costs. Therefore, both code-shared and time-shared approaches are necessary in order to offer varying service quality.

We emphasize that this study should not be used for understanding absolute capacities for the systems discussed since we made many simplifying assumptions and did not model important details. However, it can be used to understand the gains achievable by the various features and design parameters.

In order to determine standard specific capacities, detailed simulations that take into account other supporting channels must be performed. Further refinements of the scheduler can also be made, and in particular, the multiple users per slot scheduler. Admission and congestion control is also more intricate especially if best effort or other QoS classes are also being supported over the same time-shared channel. We found that by improving the received signal at the MS (e.g., through the use of dual antennae), the capacity can potentially be significantly increased but more detailed simulations are needed to quantify this gain.

ACKNOWLEDGMENT

The author would like to thank several people who provided feedback and suggestions to this work. In particular, Mahesh Makhijani, Magnus Persson, Anthony Soong, Rath Vannithamby and Shawn Tsai. Thanks to Jonas Wiorek for supporting this work.

REFERENCES