Performance Evaluation of VoIP over EV-DO Rev A

Sae-Young Chung
Dept. of EECS, KAIST, Yuseong-gu Guseong-dong 305-701, South Korea
(sychung@ee.kaist.ac.kr)
Mehdi Alasti, Pierre A. Humblet, M. Vedat Eyuboglu
Airvana, Inc.
19 Alpha Rd., Airvana, Inc., Chelmsford, MA 01824, USA
(malasti@airvana.com), (p-humblet@airvana.com), (v-eyuboglu@airvana.com)

ABSTRACT

1xEV-DO, a part of the cdma2000® family, has been enhanced to Revision A (Rev A) [1] to improve data throughput and to enable rich multi-media services such as voice over internet protocol (VoIP). EV-DO is designed to be IP friendly and therefore does not support circuit channels. This design approach leads to an obvious question: how well does EV-DO Rev A support VoIP? In this article, we present simulation results for VoIP over EV-DO Rev A. We show that EV-DO can support 40 Erlangs of VoIP traffic per sector using a 1.25 MHz carrier while maintaining the mouth-to-ear voice latency below 230 msec (with receive diversity) and still have bandwidth left for supporting high-speed data simultaneously with voice. It suggests that voice capacity and latency performance of actual EV-DO Rev A systems may perform comparably to that of other 3G standards optimized for circuit voice. We explain some of the advanced features in the EV-DO Rev A airlink, scheduler design, and quality of service (QoS) implemented to achieve this.

1. INTRODUCTION

VoIP has been growing steadily, both in the Internet and in wireless LANs such as 802.11, but has not yet been implemented in cellular networks. Cellular operators are motivated to transition to VoIP largely due to the increasing overhead in maintaining expensive circuit voice equipment such as mobile switching centers (MSC). There are many challenges to enable VoIP over cellular, but there have been enough activities in the standardization of VoIP over cellular to expect commercial cellular VoIP systems in the near future.

Together with IP multimedia subsystem (IMS) [2] and session initiation protocol (SIP) [3], multi-media services such as VoIP can be supported in a standardized way in various third generation cellular standards such as 3rd Generation Partnership Project (3GPP) in Europe and 3rd Generation Partnership Project 2 (3GPP2) in North America. However, most airlink standards in 3GPP and 3GPP2 typically support both circuit voice and packet data channels in the same RF carrier. VoIP over such systems will naturally use the same circuit channels as circuit voice. Therefore, VoIP traffic will use the same bandwidth as a circuit voice in such systems. Moreover, even with the help of link-layer assisted header compression, 3GPP VoIP capacity will be the same as circuit voice capacity. At the first glance, it may seem that running VoIP using the same circuit voice channels optimized for voice makes sense. However, as we will show in this article, carrying VoIP traffic using packet data channels can show additional benefits.

The EV-DO Rev A standard supports high speed packet data applications in a cellular environment at a speed of up to 3.1 Mbps. More importantly, Rev A provides QoS and a number of other features specifically designed to support delay-sensitive traffic such as VoIP. For example, to ensure the mobility of VoIP-based access terminals (ATs) moving at speeds up to about 200 km/h, Rev A supports soft handoffs in the uplink and fast selection handoffs in the downlink. Fast selection provides VoIP handoffs in the downlink with interruption time in the order of a few tens of msec. Overall, EV-DO is free from legacy circuit voice channels, which enables operators build cheaper and more efficient network.

In this article, we show our simulation results for VoIP over EV-DO Rev A. We show that EV-DO can support up to 40 Erlangs of VoIP traffic per sector per 1.25 MHz carrier while maintaining end-to-end latency below 230 msec – and still have room to support simultaneous high-speed data. Such high capacity for voice is achieved by having an efficient airlink that is free from circuit voice channels, by using advanced techniques such as opportunistic scheduling, and by providing a flow-based multiple-access control (MAC) that flexibly and efficiently allocates airlink resources.

The sections are organized as follows. First we present an overview of EV-DO Rev A and a description of QoS features for enabling VoIP over EV-DO Rev A. We present our simulation assumptions and then show forward and reverse link simulation results separately. We then show end-to-end latency and capacity performance by combining the forward and reverse link results.

2. OVERVIEW OF EV-DO REV A AIRLINK

EV-DO Rev A supports peak rates of 3.1 Mbps in the forward link and 1.8 Mbps in the reverse links using a pair of 1.25 MHz bands. Rev A can also support typical average sector throughputs of 1.2 Mbps and 600 Kbps in the forward and reverse links, respectively.

A wideband extension of EV-DO called NxEV-DO, or Revision B (Rev B), is standardized to improve throughput performance of EV-DO further, where up to 20 MHz consisting of 15 1.25 MHz carriers (i.e., N = 15 in NxEV-DO) can be assigned to a user to provide up to 46 Mbps and 27 Mbps in the forward and reverse links, respectively. The number of carriers assigned to a user can vary instantly and flexibly between 1 and 15 based on needs.

In EV-DO Rev A, QoS is tightly integrated into the airlink to support multimedia applications such as video streaming and VoIP. The minimum packet latencies are 1.67 msec and 6.67 msec in the forward and reverse links, respectively, which are short enough for delay-sensitive applications such as gaming and real time streaming.

Adaptive modulation and coding (AMC) is used in the forward link and power control is used in the reverse link to provide adaptation for rapidly varying channel conditions. Hybrid ARQ (HARQ) in the physical layer is used for both the forward and reverse links to provide additional level of adaptation.

Soft handoff is used in the reverse link for seamless handoff. Correspondingly, fast selection handoff provides nearly seamless handoff in the forward link – with interruption time of order of a few tens of msec. A new channel called the data source control (DSC) channel is added to provide an early warning signal before actual cell switching happens. This allows the radio access network (RAN) to reduce the time required to redirect an IP stream to the new serving cell. Typical interruption time is expected to be less than 20 msec. Thus the degradation of VoIP performance is expected to be minor.

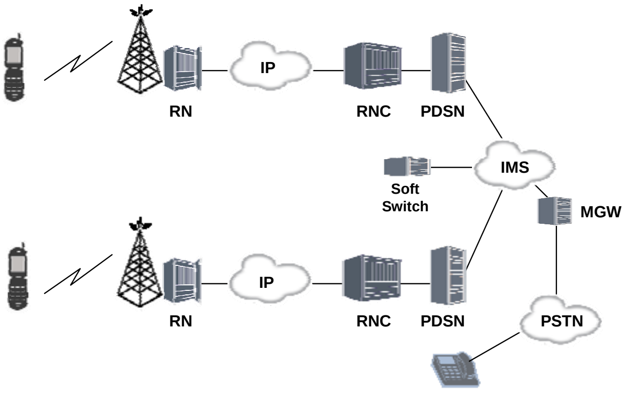


Figure 1: EV-DO Network with VoIP Support

Figure 1 shows a typical configuration of a network based on EV-DO supporting VoIP. The radio nodes (RNs) are base stations equipped with antennas, power amplifiers, and modems. Radio network controllers (RNCs) control multiple RNs and implement many control functions for the airlink such as power control, soft handoff, admission control, and the radio link protocol (RLP). Packet schedulers in the forward link run in the RNs rather than in the RNCs to minimize the scheduling delay. Media gateways (MGWs) are used to bridge between the cellular network running VoIP and Public Switched Telephone Networks (PSTNs) running circuit voice.

3. QOS IN EV-DO

One of the key components needed for VoIP over broadband wireless is the new QoS mechanism provided by EV-DO Rev A. QoS is needed in multiple places to enable VoIP, i.e., QoS in the airlink between the AT and the RAN, QoS between the RAN and the packet data service node (PDSN), QoS in the backhaul between RN and RNC, and QoS in the core network as shown in Figure 1. In this article, we focus on the airlink QoS as shown in Figure 2a and 2b since it is most relevant to the airlink of EV-DO.

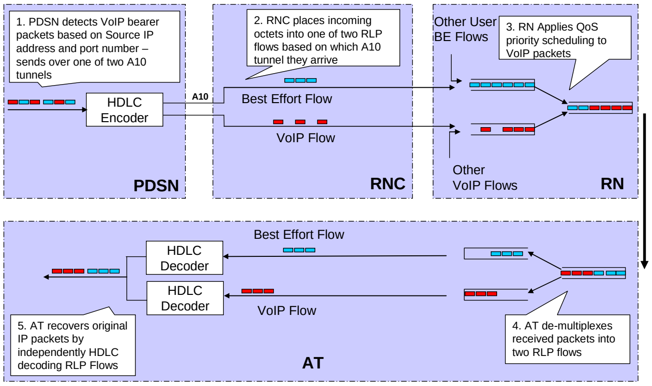


Figure 2a: Voice and data flows in the forward link

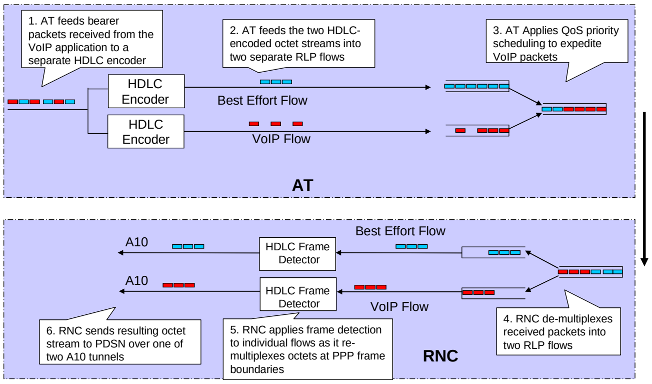


Figure 2b: Voice and data flows in the reverse link

An AT can have multiple IP flows with different QoS requirements. Commonly used traffic types are delay-sensitive traffic, constant-rate streaming, and best effort data. Depending on the requirements, the RAN needs to apply different QoS treatment.

When an IP packet arrives at the PDSN, a packet filter first identifies the packet to determine which IP flow it belongs to. Such information can then be conveyed from the PDSN to the RNC to enable different QoS treatment. Rev A defines multiple RLP flows for this purpose, i.e., each IP flow can be mapped to an RLP flow and an RLP flow identifier can be used to identify the flow at the AT. The scheduler can also use the flow information to apply appropriate scheduling priorities for each RLP flow. For example, the scheduler can be configured to prioritize VoIP packets over other packets.

Conversely, when an IP packet arrives for transmission at the AT, a packet filter can identify the IP flow and put it in the appropriate RLP flow. Rev A's multi-flow MAC is defined to provide different prioritization for each RLP flow in the reverse link. For example, Rev A can boost the transmit power for delay-sensitive packets, thereby reducing latency by facilitating earlier termination of the reverse link HARQ.

4. SIMULATION ASSUMPTIONS

In this section, we explain some of the assumptions in our VoIP simulation. First, we assume the following for VoIP.

- Voice CODEC: Enhanced Variable Rate Coder (EVRC) with full rate (9.6 Kbps), half rate (4.8 Kbps), quarter rate (2.4 Kbps), and eighth rate (1.2 Kbps).
- Markov model as defined in IS-871 to model random voice traffic generation with 29% full rate, 4% half rate, 7% quarter rate, and 60% eighth rate.
- No bundling, i.e., one voice frame per IP packet.
- Eighth-rate frames are blanked at the source, i.e., no IP packets are generated for eighth-rate packets (since eighth-rate frames are mainly used for transmitting silence, we assume it can be replaced by comfort noise at the receiver).

We assume Robust Header Compression (ROHC) header compression [4]. After compression, the header overhead is 1 byte most of the time and is 2 bytes occasionally. If an IPv6 UDP check sum of 2 bytes is added, then the overhead is 3 bytes most of the time and 4 bytes occasionally. In our simulations, we assume a fixed 2-byte header overhead after compression. Note that since the 256-bit EV-DO physical layer packets allow up to 4 bytes for overhead – assuming full rate voice frame without having to use the next bigger packet size – our results are not sensitive to the assumption on the extra bytes of overhead.

We mainly follow 3GPP2 evaluation methodology [5] for many of our simulation assumptions. Specifically we assume the following:

- 19 3-sectored cells
- 2 GHz carrier frequency
- 2km cell-to-cell distance
- 138 dB maximum path loss
- 17 dB maximum C/I
- Base station transmit power: 20W
- Mobile's maximum transmit power: 200 mW
- Base station noise figure: 5dB
- Mobile's noise figure: 9 dB
- Base station antenna gain: 17 dB
- Cable loss: 2 dB
- Body loss: 0 dB
- Mobile antenna gain: 0 dB
- Shadow fading with standard deviation of 8.9 dB and 50% base station correlation
- 5 channel models [5]
- 3km/h 1-path Rayleigh, Pedestrian A, 30% of ATs
- 10 km/h 3-path Rayleigh, Pedestrian B, 30% of ATs
- 30 km/h 2-path Rayleigh, Vehicular A, 20% of ATs
- 100 km/h 1-path Rayleigh, Pedestrian A, 10% of ATs
- 1.5 Hz Rician with K = 10 dB, 10% of ATs

Mobiles are distributed uniformly in the 19 hexagonal coverage area. Mobiles are re-dropped if they are within 35 meters from a base station, or if the path loss from the serving sector to the AT is greater than 138 dB. The latter is to remove ATs that are power limited and thus getting unacceptable voice quality. In an actual system, such removal can be done by admission and overload control. Coverage holes caused by this can be reduced by installing more base stations and/or by installing repeaters.

We do not simulate random arrivals and departures directly in a single simulation, but we do take them into account in our simulation. Our simulation can be considered as taking multiple snapshots of a system with random arrival and departures. We do simulate a random number of users resulting from random arrival and departure when we take each snapshot. Therefore, as we take more snapshots, our results will resemble that of the system with random arrival and departure closely. Simulation duration is 100 seconds per snapshot, which should be long enough to give sufficient averaging.

Forward and reverse links are simulated separately. All 57 sectors are simulated for the reverse link, where the total number of ATs dropped per simulation is equal to the number of target Erlangs per sector times the total number of sectors, which is 57. The number of ATs in each serving sector follows Poisson distribution closely since there are many sectors even though the total number of ATs is fixed. Note that the number of ATs can be different from sector to sector – closely matching the random number of users resulting from random arrival and departure. In the forward link, only the center sector is simulated. This is equivalent to simulating all 57 sectors simultaneously since we assume 100% inter-sector interference, which effectively decouples sectors. Forward link simulations are run many times with different numbers of users, following Erlang B model, to simulate random arrival and departure. Note that our assumption of 100% inter-sector interference is pessimistic for voice simulations, since not all time slots are utilized.

In the forward link simulation, we also take into account effects of overhead channels such as pilot, MAC, preamble, control channels. We assume the control channel uses 2 MAC packets at 76.8 Kbps in every control channel cycle of 256 slots, which corresponds to an overhead of about 6.3%. We simulate hybrid ARQ with misses and errors in the feedback channels such as the data rate control (DRC) and acknowledgment (ACK) channels. We assume single or dual receive diversity terminals. We simulate multi-user packets for voice users, which can contain up to 8 packets belonging to different users. For voice flows, we assume fast retransmission at the physical layer for packets lost in the physical layer. This is possible by using the ACK/NAK information used by the hybrid ARQ, i.e., when the last transmission of a physical layer packet results in a NAK, which means the packet is lost in the airlink, we let the scheduler immediately retransmit the packet in the next available slot. This is faster than the retransmission at the radio link protocol (RLP) layer, since it requires message exchanges between the AT and the RNC.

The forward link uses time-division multiplexing (TDM) and a scheduler serves packets in each sector as shown in Figure 3. We always prioritize voice packets over best effort data packets. Among VoIP packets, we apply the first-come first-serve scheduling with a modification to take into account channel variations. For example, when the channel condition degrades temporarily, the scheduler can wait until the channel condition improves. This may increase voice latency, but can provide capacity gain. Impact on the end-to-end voice latency can be minimized if the delay jitter caused by this is carefully controlled. Among data packets, we use the proportional fairness scheduler. The retransmission packets for voice will have the highest priority in any case.

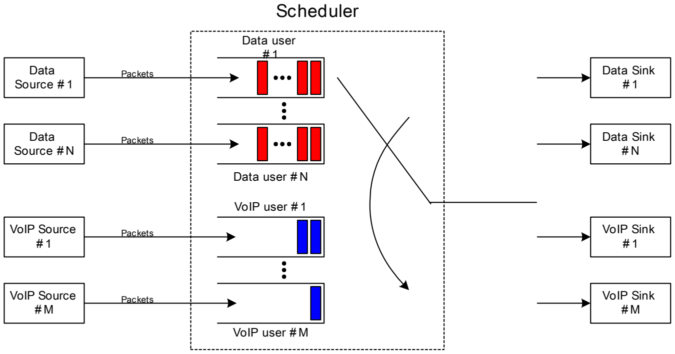


Figure 3: Scheduler

In the reverse link, we assume dual receive diversity at the base station. Four-way diversity, such as dual cross-polarization diversity, should improve the voice and data capacity further [6]. We simulate hybrid ARQ in the reverse link, which is a new feature added in Rev A to improve the reverse link performance significantly. In the multi-flow MAC, voice flows are configured to use low-latency mode [1]. This lowers the average voice latency by boosting the transmission power of VoIP packets. Data flows are configured to use high-capacity mode [1] where there is no power boosting, thus maximizing the HARQ gain at an expense of increased latency.

5. FORWARD LINK SIMULATION RESULTS

In this section, we present our forward link simulation results. We first show a VoIP-only case followed by a mixed voice and data case to see the impact of VoIP users on data users and vice versa.

Figures 4(a) and 4(b) show the distribution of latencies measured between the RN and the AT for each VoIP packet. Latency is defined as the sum of the queuing and transmission delay over the air and the effect of dejittering, as explained below. Two scenarios are considered, i.e., with and without dual receive diversity at the ATs.

Figure 4 (a) shows the distribution of latencies as a function of loading (VoIP Erlangs per sector) for ATs without receive diversity. It shows more than 90% of users get latencies less than about 130 msec when there are 40 Erlangs of loading. More than half of the users get latencies less than about 90 msec. Individual ATs see different latencies because they experience different channel conditions, including path loss and fading. Note that time-varying delay jitter will cause a packet-to-packet delay variation, which is assumed to be absorbed by a de-jitter buffer equalizing VoIP packet latencies for each user. We allow up to 2% packet loss rate in the airlink and in the de-jitter buffer. Since the packet loss rate in the forward airlink is very small due to retransmission at the physical layer – about 0.01% after retransmission and about 1% without – most of the packet losses are caused by the de-jitter buffer. This buffer drops packets that exceed a certain delay bound. In our forward link simulation results, de-jitter buffers are not simulated, but rather the 98th percentile delays are calculated for each AT and plotted as in Figures 4a, b, c, and d.

Similarly, Figure 4 (b) shows the distribution of latencies as a function of loading for ATs with dual receive diversity. Latencies are less when compared to the case without receive diversity. This is due to the diversity provided by the second receive antenna, which reduces average VoIP packet latencies as well as the delay jitter.

When there are only data users, the sector throughputs are 1 Mbps and 1.2 Mbps for 1 antenna and 2 antenna ATs, respectively, assuming 10 data users per sector.

Figure 4 (c) shows how voice latency distribution is affected by data users, where there are 40 Erlangs of voice per sector. In this case, dual receive diversity is assumed in all ATs. Latency is increased when data users are added – 10 data users per sector. But the additional latency for voice packets due to data users is minimal – about 10 msec. This is because voice traffic is prioritized over data packets by the scheduler. In some cases, however, voice packets may experience extra scheduling delay while the transmission of data packets is in progress.

Figure 4 (d) shows how data sector throughput is affected by voice users with 10 data users per sector. Two cases are considered – with and without dual receive diversity at ATs. For example, with dual receive diversity the data sector throughput is about 1.2 Mbps when there are no voice users. When voice users are introduced, data throughput starts to degrade at a rate of about 2% per voice user. Even so, with 40 voice users per sector, 400 Kbps of data throughput per sector is still available.

6. REVERSE LINK SIMULATION RESULTS

In this section, we present our reverse link simulation results. As in the previous section, we first show VoIP-only results followed by mixed voice and data results to see the impact of VoIP users on data users and vice versa.

Figure 5 (a) shows the distribution of latencies as a function of loading (VoIP Erlangs per sector), where the latency is defined as the queuing and transmission delay of VoIP packets from the AT to the RN. Unlike in the forward link, the user-to-user latency variation is small, i.e., about 13 msec. This is because the reverse link is power-controlled and users get consistent performance as long as the power control works. Power control also ensures that latency does not vary when the loading is changed – as long as the system remains stable. Of the 13 msec variation, 6.67 msec is due to uniform distribution of waiting time for the next boundary of the reverse link physical layer packets, and the remaining 6 msec is due to different channel models. As in the forward link, we plot the 98th percentile latency. This assumes that the de-jitter buffer equalizes latency except for 2% of packets, 0.14% of which are dropped in the airlink and the rest dropped by the de-jitter buffer.

Rise-over-thermal (ROT) is an important measure of reverse link loading. Reverse link loading needs to be limited to avoid instability in the reverse link power control mechanism. In our simulations, we maintain the 99th percentile ROT below 7 dB, which should be enough to guarantee the stability. Figure 5 (b) shows ROT distribution of all 57 sectors. It can be seen that the maximum capacity is limited to 40 Erlangs due to the requirements on the ROT.

Sector throughput is about 600 Kbps when there are 10 data users per sector and no voice users. Figure 5 (c) shows how voice latency distribution is affected by data users, where there are 30 or 40 Erlangs of voice per sector. The number of data users is varied between 0 and 10 per sector to see the effect on voice latency. Again, the latency remains the same as long as the power control works, i.e., when the loading is up to 30 Erlangs per sector regardless of whether there are data users or not. But when the loading is increased to 40 Erlangs per sector – and when there are 10 data users per sector as well – some mobiles become power limited. This is due to the increased transmission power caused by the increased reverse link loading from data users. Note that the 99th percentile ROT exceeds 7 dB when there are more than 15 Erlangs of voice and we add 10 data users per sector. This is because each data user is configured to receive a minimum throughput of 10 Kbps. Compared to a voice throughput of about 3 Kbps (after factoring in the voice activity of 40%), this is significantly higher. and can explain why the ROT reaches the maximum limit when there are only 15 Erlangs of voice.

Figure 5 (d) shows how data sector throughput is affected by voice users with 10 data users per sector. The data sector throughput is about 600 Kbps when there are no voice users and starts to degrade at a rate of about 3% per voice user. As voice users increase, the data sector throughput reaches a plateau of about 100 Kbps, due to the minimum guaranteed throughput of 10 Kbps per data user.

7. END-TO-END LATENCY

In this section, we show end-to-end mouth-to-ear latency and overall capacity for two cases, i.e., between two mobiles and between a mobile and a landline phone. To obtain the end-to-end latency, we need some assumptions on the amount delay in components other than the airlink. For this, we assumed the following fixed delays:

- Vocoder delay at encoder at mobile or at MGW: 35 msec
- Vocoder delay at decoder at mobile: 20 msec
- Vocoder delay at decoder at MGW: 15 msec
- Various processing delay at a mobile: 5 msec in each direction
- RAN and PDSN delay: 20 msec
- IMS delay: 15 msec
- PSTN delay: 20 msec

Combining these yields mobile-to-mobile end-to-end latency of 120 msec + forward link queuing and transmission delay + reverse link queuing and transmission delay + de-jitter buffer delay. Note that there are two RANs and PDSNs involved in a mobile-to-mobile call. But the sum of three time-varying delays components, i.e., forward link queuing and transmission delay + reverse link queuing and transmission delay + de-jitter buffer delay becomes constant after equalizing delay by the de-jitter buffer.

Note, however, that there are two different directions of voice flows in a mobile-to-mobile call, i.e., mobile #1-to-mobile #2 and mobile #2-to-mobile #1. In general, the latencies are different in the two directions. Since users cannot perceive the difference in the latencies in the two directions but can only perceive the round-trip delay, we define the end-to-end latency as the round trip delay divided by two – both in the mobile-to-mobile and mobile-to-landline cases.

Since the forward and reverse links are simulated separately, the end-to-end latency numbers need to be derived from the two separate results. One way to achieve this is to run forward and reverse link simulations separately, assume the same number of Erlangs per sector, randomly pick two users – one from the forward link and another from the reverse link – and treat them as a single user having both forward and reverse links. But such a model cannot take into account the correlation between the forward and reverse links of the AT. Therefore, we intentionally give 100% correlation in the 98th percentile VoIP packet delay between forward and reverse links when we pick the two users. As a result, we expect our end-to-end latency results to be pessimistic, since such a correlation will not be 100% in reality.

We also assume that IMS provides QoS for voice packets and thus voice packets do not experience significant queuing delays. Therefore, the IMS delay is assumed to be mainly due to propagation delay. Our assumption of 15 msec corresponds to the propagation delay of a fiber link of 3,000 km.

Figures 6 (a), (b), (c), and (d) show end-to-end latency distribution for four cases, i.e., mobile-to-mobile or landline cases with and without dual receive diversity at ATs. The worst case is the mobile-to-mobile case without diversity. In this case, the latency is less than 300 msec for about 98% of mobiles when there are 40 Erlangs of voice per sector.

Diversity reduces the latency to 230 ms. Overall capacity is 40 Erlangs and is limited by the reverse link due to the ROT requirements. If 4-way diversity [6] is used at the RN, then the capacity is expected to be higher.

8. CONCLUSIONS

We have shown our simulation results for VoIP over EV-DO Rev A. It shows we can support up to 40 Erlangs of voice traffic per sector using a single CDMA carrier, while still having some room for supporting data users simultaneously. We also showed throughput performance of data only case. Finally, we showed how voice users affect data users and vice versa when there are both types of users. Our results suggest that VoIP over EV-DO can perform at least comparably to other cellular technologies optimized for voice.

References

[1] 3GPP2, C.S0024-A, “cdma2000 High Rate Packet Data Airlink Specification,” April 2004

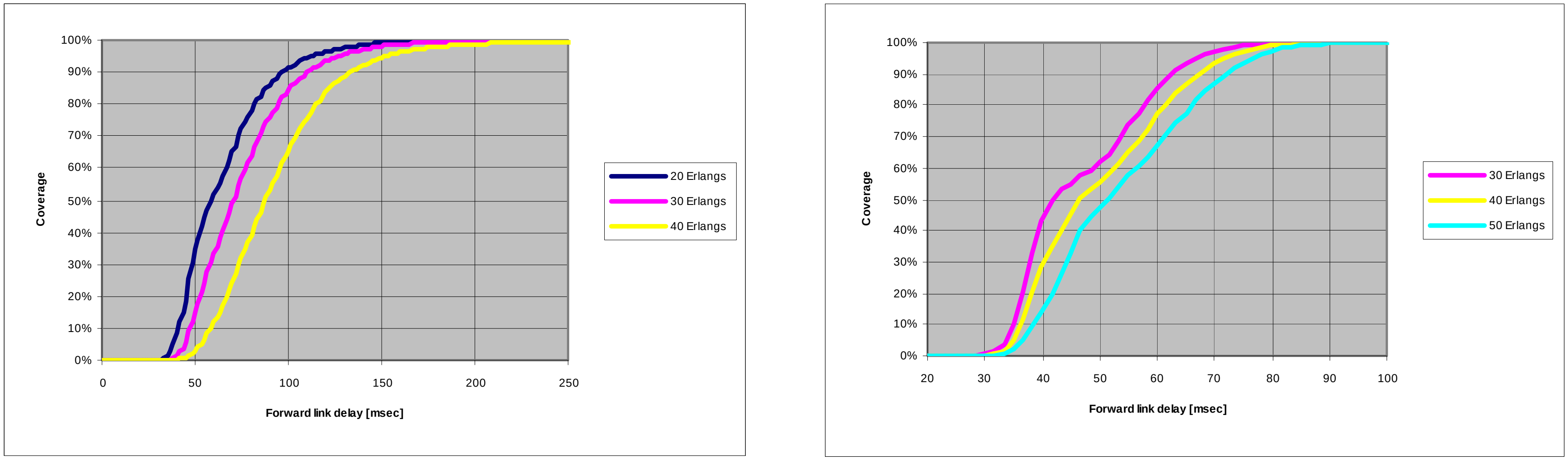
[2] 3GPP2, TS X.S0013-000, “All-IP Core Network Multimedia Domain: Overview”

[3] Session Initiation Protocol, IETF RFC 3261

[4] Robust Header Compression, IETF RFC 3095

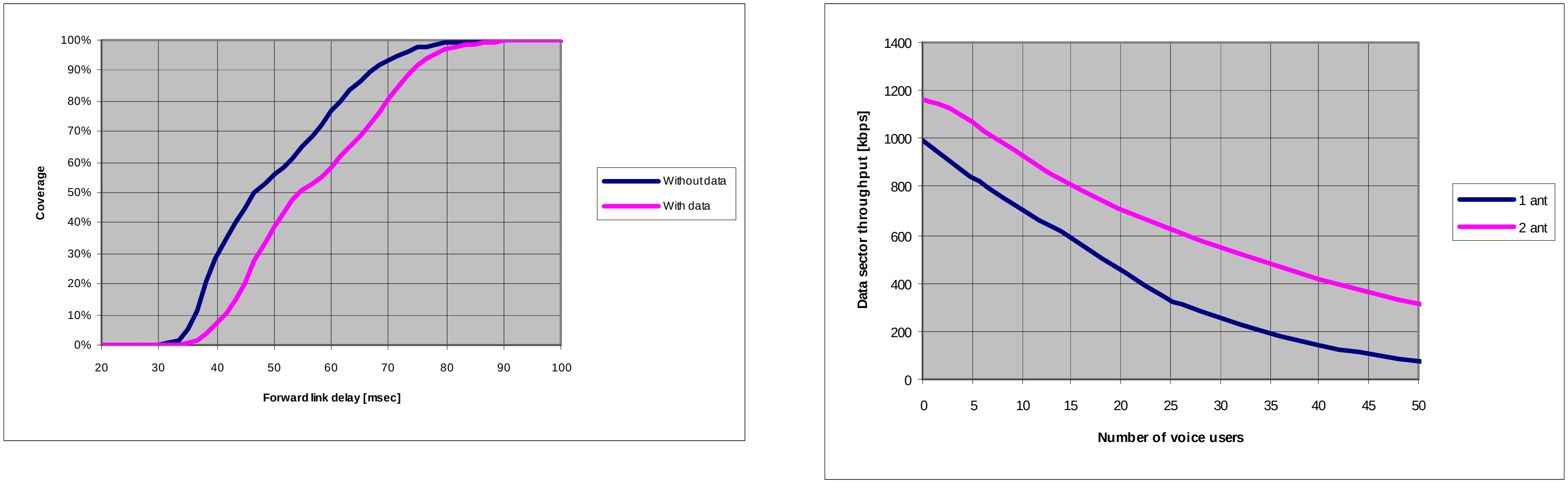
[5] 3GPP2, C.P1002-C-0 Version 0.4, “cdma2000 Evaluation Methodology Revision 0,” Sept. 2004

[6] L. Aydin, E. Esteves, and R. Padovani, “Reverse link capacity and coverage improvement for CDMA cellular systems using polarization and spatial diversity,” ICC 2002



(a) Latency distribution without receive diversity at mobiles

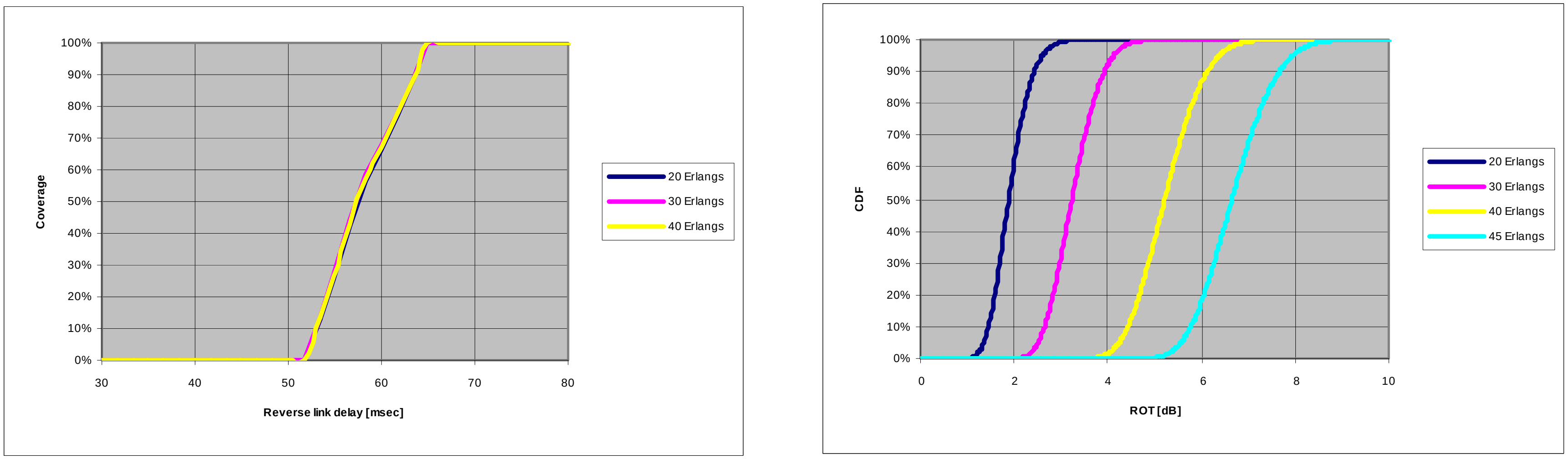
(b) Latency distribution with dual receive diversity at mobiles



(c) Voice latency with and without data users

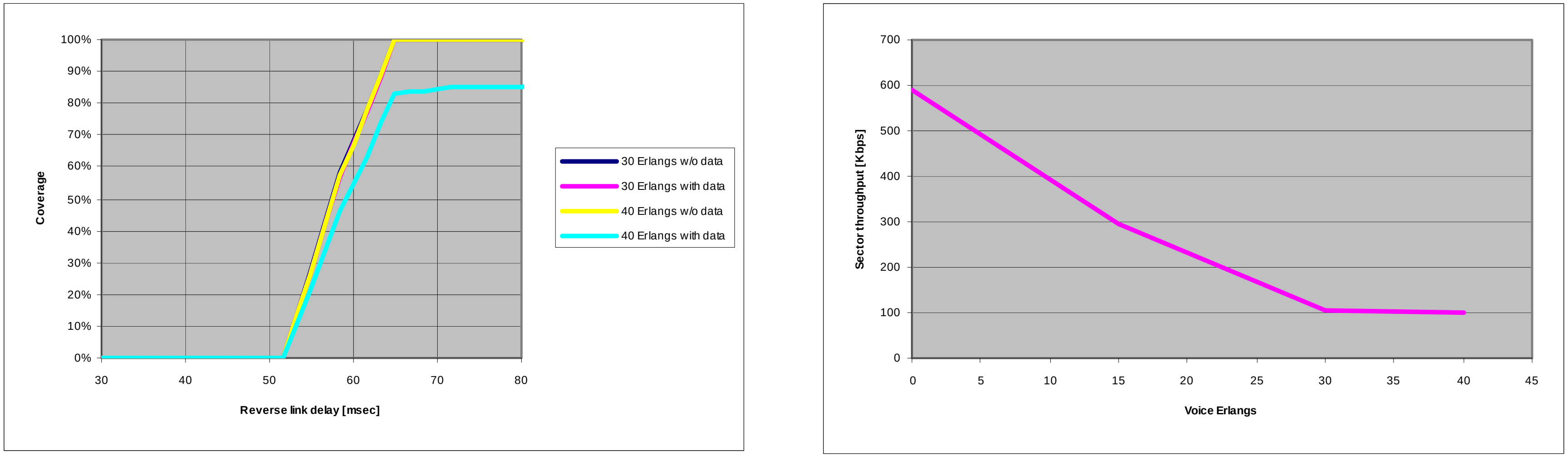
(d) Data throughput vs. number of voice users

Figure 4: Forward link performance



(a) Latency vs. loading

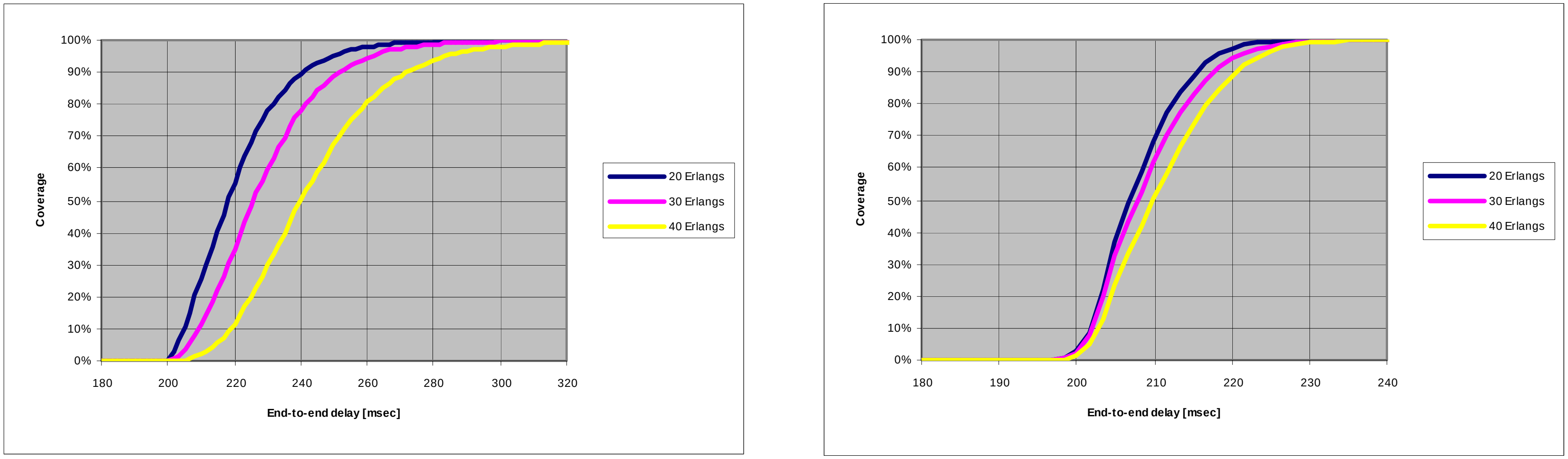
(b) ROT



(c) VoIP latency with and without data users

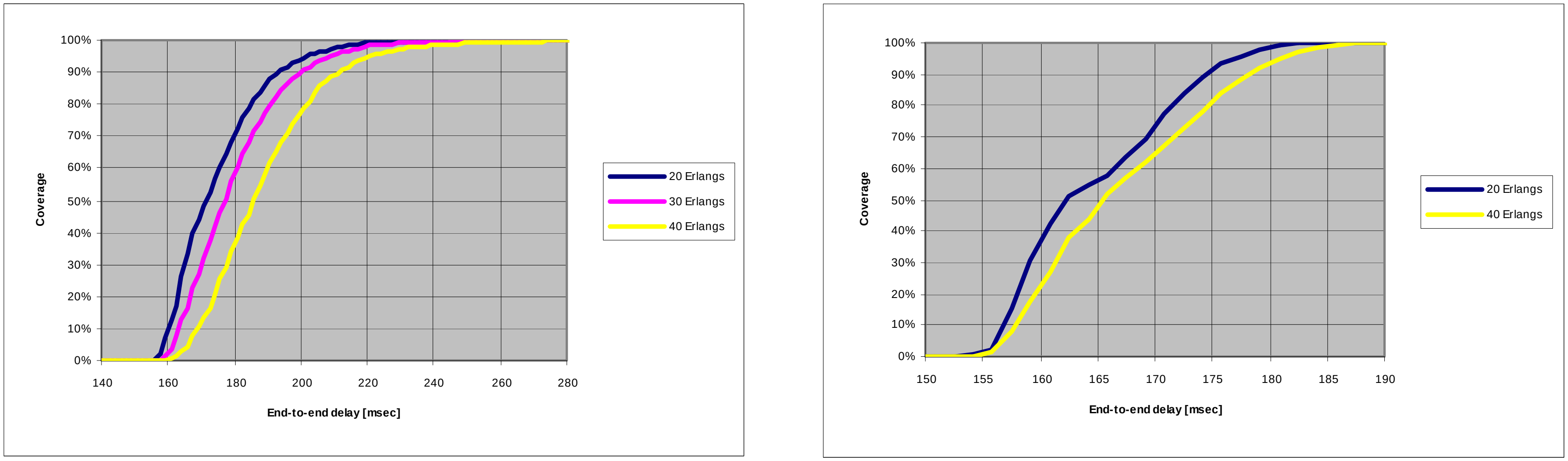
(d) Data sector throughput vs. voice loading (Erlangs)

Figure 5: Reverse link performance



(a) Latency distribution for mobile-to-mobile case without receive diversity at mobiles

(b) Latency distribution for mobile-to-mobile case with dual receive diversity at mobiles



(c) Latency distribution for landline case without receive diversity at mobiles

(d) Latency distribution for landline case with dual receive diversity at mobiles

Figure 6. Latency vs. Capacity curves for mobile-to-mobile and landline cases