Publication P8

Jani Lakkakorpi and Alexander Sayenko. 2009. Uplink VoIP delays in IEEE 802.16e using different ertPS resumption mechanisms. In: Jaime Lloret Mauri, Joseph A. Meloche, Sergey Balandin, Malohat Ibrohimova, and Junya Nakata (editors). Proceedings of the Third International Conference on Mobile Ubiquitous Computing, Systems, Services and Technologies (UBICOMM 2009). Sliema, Malta. 11-16 October 2009, pages 157-162.

2009 Third International Conference on Mobile Ubiquitous Computing, Systems, Services and Technologies

Uplink VoIP Delays in IEEE 802.16e Using Different ertPS Resumption Mechanisms

Jani Lakkakorpi
Department of Communications and Networking
Helsinki University of Technology
Espoo, Finland
jani.lakkakorpi@tkk.fi

Alexander Sayenko
Research, Technology & Platforms
Nokia Siemens Networks
Espoo, Finland
alexander.sayenko@nsn.com

***Abstract*—In this paper, we present different IEEE 802.16e uplink channel access mechanisms that can be used to activate extended real-time polling service (ertPS) voice over IP (VoIP) connections after a silence period. The performance, especially uplink delay, of different resumption mechanisms is compared with each other using simulations. In addition to uplink VoIP delay, we study the uplink resource usage with different mechanisms. In our studies, we have found that using the fast feedback channel or multicast polling are the most promising approaches for efficient ertPS VoIP resumption.**

***Keywords*-IEEE 802.16e, WiMAX, QoS, ns-2**

I. INTRODUCTION

IEEE 802.16e, often also referred to as Mobile WiMAX, is an IEEE standard for wireless broadband access network [1, 2]. The main advantages of IEEE 802.16e are long range and sophisticated support for quality of service (QoS) at the MAC level. The MAC layer supports convergence between several different types of applications and services. The standard defines two basic operational modes: mesh and point-to-multipoint (PMP). In the former mode, subscriber stations (SS) can communicate with each other and with the base station (BS). In the PMP mode, the SSs are only allowed to communicate through the BS. It is anticipated that providers will use the PMP mode to connect their customers to the Internet. In this case, the SSs do not send data to each other but rather communicate through the BS. Thus, the provider can control the environment to ensure the QoS requirements of its customers.

At the IEEE 802.16e BS, all downlink (DL) connections have dedicated buffers and resources are allocated per connection. There can be multiple connections per SS. In uplink (UL) direction, however, the BS grants slots per SS and not per connection. It is the SS that decides how these slots are used. The effective air interface bandwidth that a connection gets may vary substantially because there are no dedicated radio channels.

IEEE 802.16e has three QoS classes that can be used for real-time connections. In unsolicited grant service (UGS), the BS allocates fixed-size grants periodically; UGS connections do not send any bandwidth requests. In real-time polling service (rtPS), the BS periodically polls the SS by granting one slot for sending a bandwidth request, while the goal of extended real-time polling service (ertPS) is to combine the advantages of UGS and rtPS. In ertPS, the BS continues granting the same amount of bandwidth (by default, the size of this allocation corresponds to maximum sustained traffic rate of the connection) until the ertPS connection explicitly requests a change in polling size. Extended piggyback request field of the grant management subheader can be used for this purpose. If the bandwidth request size is zero, the BS may provide allocations for bandwidth request header only or nothing at all. In the latter case contention request opportunities or fast feedback channel, i.e., channel quality indicator channel (CQICH) may be used when there is a packet to send after a silence period.

There are also two classes for non-real time connections: non-real time polling service (nrtPS) is similar to rtPS except that connections are polled less frequently and they can also use contention request opportunities. Best effort (BE) connections are never polled and they can receive resources only through contention.

In this paper, we are interested in different mechanisms an ertPS VoIP connection can use to resume sending packets after a silence period (during which no packets are sent), i.e., we study different ertPS resumption mechanisms and their performance. Based on our simulation studies, it seems that using the fast feedback channel or multicast polling are the most promising approaches for efficient and low-delay ertPS VoIP resumption.

There is a good amount of recent research articles on uplink scheduling in IEEE 802.16e. For example, in [3] the authors propose a delay constrained uplink scheduling policy for rtPS/ertPS services and in [4] the performance of UGS, rtPS and ertPS is compared to each other. However, we found no articles that were dedicated to different ertPS resumption methods and their performance.

The rest of this paper is organized as follows: Section II presents the different uplink channel access and ertPS resumption mechanisms, Sections III and IV present our simulator and the simulation results, respectively, while Section V concludes the paper.

II. DIFFERENT UPLINK CHANNEL ACCESS MECHANISMS

The IEEE 802.16e standard supports several mechanisms that the SSs can use to request uplink

978-0-7695-3834-1/09 $26.00 © 2009 IEEE
DOI 10.1109/UBICOMM.2009.15

bandwidth. Depending on the QoS and traffic parameters associated with a service, one or more of these mechanisms may be used by the SS. Once the SS has an allocation for sending traffic, it is allowed to request more bandwidth by transmitting a stand-alone or a piggybacked bandwidth request [5].

A. Polling

The BS allocates dedicated or shared resources periodically to each SS. The SS can then use these resources to request bandwidth. This process is called polling. If an ertPS VoIP connection is polled regularly also during the silence period, we can send the first packet of the next talkspurt without additional delay. Some uplink resources are wasted, though. With rtPS class, this is our only alternative.

B. Contention Resolution

The contention resolution mechanism in IEEE 802.16e allows SSs to send their bandwidth requests to the BS without being polled. This kind of mechanism is necessary for scheduling service classes that are polled irregularly or not at all, i.e., ertPS, nrtPS and BE. Contention resolution parameters are the number of bandwidth request transmission opportunities per frame and backoff start/end values. The backoff start value determines the initial backoff window size, from which the SS randomly selects a number of transmission opportunities to defer before sending the bandwidth request. If there is a collision, the backoff window is increased and the contention resolution is repeated. The SS continues to retransmit the bandwidth request until the maximum number of retransmissions expires.

When orthogonal frequency-division multiple access (OFDMa) is used as physical layer, the uplink contention comprises several phases. First, the SS sends a code division multiple access (CDMA) request code. If the code is received correctly (no collisions), the BS grants an uplink CDMA allocation, which the SS can use for sending a bandwidth request.

Contention resolution mechanism is useful, e.g., with VoIP connections that support silence suppression. We assume here that ertPS is used for these connections. During the silence periods, we can either have periodic allocations just big enough for a stand-alone bandwidth request or no polling at all. Naturally, the latter option is more bandwidth-efficient. However, in this case we need to use the contention resolution mechanism when the connection becomes active again, i.e., when the connection has a packet to send. If there are lots of connections that participate in contention resolution, this could result in considerable UL packet delays. Moreover, the same set of backoff parameters is used for all connections, which may not suit well for real-time connections – assuming that the backoff parameters were selected based on BE connection requirements.

C. Multicast Polling

In multicast polling, the SS does not send its bandwidth request codes during the common bandwidth request contention slots but during slots that have been assigned for a particular group of SSs. Multicast polling and VoIP has already been studied in [6], where the use of separate request backoff parameters for different multicast polling groups is proposed in order to fulfill VoIP delay (and packet loss) requirements.

D. CQICH / Fast Feedback Channel

An alternative to contention resolution (and polling) is to use the fast feedback channel (CQICH, see Fig. 1) for informing the BS that the SS has a packet it wants to send after the silence period. The fast feedback channel is mainly used for periodical (e.g., every four frames) transmission of the signal-to-noise ratio (SNR), which can be used by the BS in link adaptation. As we cannot fit the SNR information and the ertPS resumption codeword into the same message (the length is only six bits), sending the latter may have some implications on the link adaptation. However, switching from silence period to talkspurt should be a rare event. Assuming that an average talkspurt and silence period have lengths of 1.2 seconds and 0.8 seconds, correspondingly, there should be only one ertPS resumption message per two seconds on average. When the ertPS resumption codeword arrives at the BS, we immediately grant enough slots for one VoIP packet and re-schedule the next grant, i.e., we reset the frame counter of the connection.

III. IEEE 802.16E NETWORK SIMULATION MODEL

The basic implementation of our IEEE 802.16e module is described in [8]. The module includes the following features: orthogonal frequency-division multiplexing (OFDM) and OFDMa physical layers, hybrid automatic repeat request (HARQ), transport and management connections, fragmentation, packing, ranging and bandwidth request contention periods, CDMA codes for ranging and bandwidth requests, support for the most important MAC level signaling messages and the ARQ mechanism that allows retransmitting dropped PDUs. Additionally, the module includes several different BS schedulers and has a simple, trace-based model for link adaptation. These features are described in more detail in the following sections.

A. MCS, Link Adaptation and Errors

Modulation and coding scheme (MCS) defines how many bits can be sent in a single slot. The BS can dynamically change both the DL and UL MCS of an SS. Link adaptation is based on reported SNR values and carefully tuned transition thresholds. Naturally, we have a different set of link adaptation thresholds for HARQ and non-HARQ connections. Our error model analyzes the PDU SNR, maps it to the forward error correction (FEC) block error rate (BLER) based on the channel performance curves, and decides whether the PDU is erroneous or not. Each SS has a randomly selected, trace file, where the SNR values are read from[1]. In the simulations of this paper, we model only

[1] 60% of our traces correspond to ITU PedB model, 40% to ITU VehA model.

one sector, while the trace files have been obtained from 19-cell system simulations.

B. Scheduler

The BS scheduler grants slots for the SSs according to the QoS parameters and bandwidth request sizes of the individual connections. Uplink virtual queue sizes are updated based on bandwidth requests and received UL packet sizes. For DL connections, we use the BS queue sizes and the QoS parameters. In our basic scheduler, slots are assigned in deficit round-robin (DRR) [9] fashion. Quantum size is a configuration parameter (default is 17 slots); a bigger quantum size decreases the MAP overhead as we then serve fewer connections per frame.

We have implemented support for three IEEE 802.16e data delivery services[2]: extended real-time variable rate service (ERT-VR), real-time variable rate service (RT-VR) and best effort (BE). ERT-VR and RT-VR connections are served before BE connections; they are assigned slots until all ERT-VR and RT-VR queues are empty or until there are no more slots left for real-time traffic. Connection admission control should take care of that there are always enough slots for real-time connections and that all slots are not used for real-time connections. Moreover, rate limiters are used at the BS to enforce the minimum reserved traffic rate (MRTR) of real-time connections; excess real-time traffic gets BE treatment.

In order to waste as little bandwidth as possible, silence suppression detection at the BS is done for the ertPS connections: whenever an UL PDU is received, a connection-specific timer is started. When this timer expires, silence state is started. If we let the ertPS connections participate in contention (or if CQICH based resumption is deployed), no polling is done during the silence state. Otherwise, periodical polling slots are granted for ertPS connections during silence periods.

If ARQ is enabled for a connection, the following connection-internal scheduling order is applied: 1) ARQ feedback messages, 2) retransmissions and 3) all other PDUs. In this paper, we study only cases with one UL/DL transport connection per SS.

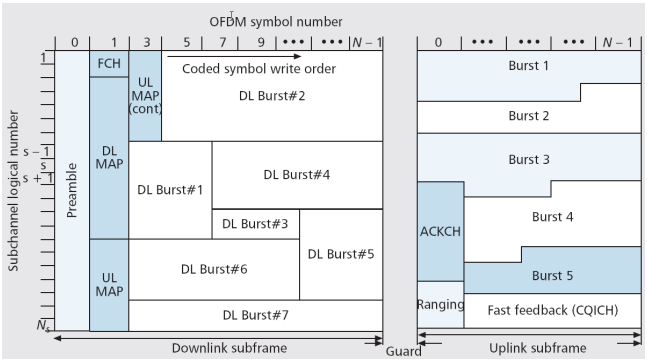

Fig. 1. Mobile WiMAX TDD frame structure [7].

[2] Data delivery services are defined for both UL and DL direction whereas scheduling service classes (ertPS and others) are defined for UL direction only.

IV. Performance Evaluation

We use a modified version of the ns-2 simulator [10]. The WiMAX related modifications have been described in the preceding section. Multiple simulations are run in each case in order to obtain small enough 95% confidence intervals. Simulation time is always 200 seconds. One-way core network delay between a server and the BS[3] is set to 31 ms. The only bottleneck in our system is the air interface (see Fig. 2). The most important IEEE 802.16e network parameters are listed in Table I[4]. We simulate the following traffic mix: 120–130 or 95–105 VoIP connections and 10 or 50 file downloading connections per BS. All connections are active during the whole simulation run.

TABLE I
IEEE 802.16e Related Simulation Parameters

Parameter	Value
PHY	OFDMa
Bandwidth	10 MHz
FFT size	1024
Cyclic prefix length	1/8
TTG (transmit-receive transition gap)	296 PS
RTG (receive-transmit transition gap)	168 PS
Duplexing mode	TDD
Frame length	5 ms
DL/UL ratio	35/12 OFDM symbols
DL/UL permutation zone	FUSC/PUSC
Channel report type and interval	CQICH, 20 ms
MAP MCS	QPSK-1/2, REP 2
Compressed MAP	Yes
Number of ranging opportunities	1
Ranging backoff start/end	0/15
Number of request opportunities	3[5]
Request backoff start/end	3/15
CDMA codes for ranging and bandwidth requests	64/192
HARQ (CC)	For VoIP connections only
Number of HARQ channels	16
HARQ buffer size	2048 B per channel
HARQ shared buffer	Yes
Max. number of HARQ retransmissions	4
HARQ ACK delay	1 frame
PDU SN	With HARQ (no ARQ)
Fragmentation/Packing	Yes/Yes
Maximum MAC PDU size	100 bytes
ARQ	For FTP connections only
ARQ feedback types	All
ARQ block size / window size	64 bytes / 1024
ARQ block rearrangement	No
ARQ feedback frequency	5 ms
ARQ retry timer	50 ms
ARQ block lifetime	1500 ms
ARQ rx purge timeout	2000 ms
MRTR for VoIP connections	11800 bps
Max. SS/BS queuing delay for VoIP SDUs	150 ms

[3] Since there is only one BS in our system, there are no handovers.
[4] Most parameters are taken from Mobile WiMAX system profile [11].
[5] In scenarios, where multicast polling is used, there are two request opportunities for the basic contention and one opportunity for the multicast polling group.

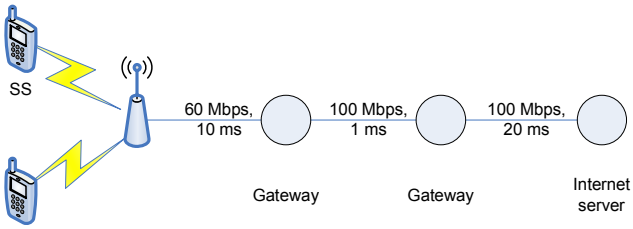

Fig. 2. Simulation topology.

Our VoIP traffic source is a simple G.723.1 model, where both on and off period lengths are exponentially distributed with mean lengths of 1.2 s and 0.8 s, correspondingly. 24 bytes of payload is sent every 30 ms during the active periods. Altogether, RTP, UDP and IP add 40 bytes of overhead, which results in a total packet size of 64 bytes. Packet header compression (from 40 bytes to 4 bytes) is applied at the BS and the SS. VoIP connections are given ertPS treatment with different resumption mechanisms.

Our file downloading traffic source is a simple FTP model, where a single 250 kB file is downloaded over and over again. Time between two downloads is uniformly distributed between 1 and 5 seconds. A single NewReno TCP connection is utilized. File downloading traffic is given BE treatment. Even though this traffic is downloading and not uploading, there are a lot of TCP acknowledgements that need to be sent upstream. This will cause a heavy load on the bandwidth request opportunities and CDMA codes that are shared with all SSs – including the SSs that host VoIP connections. To better illustrate this phenomenon, additional simulation scenarios have been run in addition to the basic scenario. In the second and third scenario, we have increased the number of file downloading SSs from 10 to 50 (and decreased the number of VoIP users by 25).

A. *Basic Scenario (Case 1)*

In our basic scenario, there are 120 or 130 VoIP users and 10 file downloading users. As one could easily guess, ertPS with polling consumes more uplink resources (see Fig. 3)[6] than the other three mechanisms and therefore the (resumption) delays (see Fig. 4–5) start to grow with this mechanism.

Contention resolution is not a bottleneck in this scenario. However, if there had been more connections using contention (or multicast polling), we could have had high delays due to CDMA code collisions and the uplink CDMA allocations.

CQICH based resumption uses a bit more resources than contention based resumption. This is likely due to non-optimal link adaptation. As we speculated earlier, link adaptation may not work optimally if the SNR value is replaced with the ertPS resumption codeword. However, we also ran simulations, where the SNR value and the ertPS resumption codeword were put into the same message but this had no major effect on the results. Things could be different with larger CQICH report interval, though.

We chose not to allocate additional resources for multicast polling but one request opportunity per frame for the multicast polling group was taken from the basic contention region. In the first case, the poor performance of multicast polling based resumption was due to non-optimal request backoff parameters (start/end: 3/15). In order to limit uplink VoIP delay with multicast polling based resumption, we also applied backoff parameters different (start/end: 1/15) from the basic contention resolution parameters, as proposed in [6]. The results were indeed better in the latter case. This, however, would require changes in the specification. Moreover, we did not want to apply request backoff parameters optimized for ertPS VoIP resumption for BE traffic as that would have led to a large number of collisions and thus lower TCP goodput.

B. *50 File Downloading Users (Case 2)*

In order to have meaningful results with 50 (instead of 10) file downloading users, we decreased the number of VoIP users by 25 in all cases. Thus, the amount of non-controllable UL resources (that are allocated to CQICH reports, HARQ acknowledgements etc.) stays more or less the same as in the previous scenario.

In this scenario, polling based ertPS resumption leads to excessive delays with a high number of VoIP users (see Fig. 6–8), while the other resumption mechanisms do not – except for the first multicast polling case, which suffers from non-optimal request backoff parameters. This makes us conclude that it is not the CDMA code collisions but the uplink CDMA allocations that are the reason for a bottleneck in the unicast allocations. With contention and multicast polling based ertPS resumption, the BS grants resources upon receiving the CDMA code. With polling based resumption, however, the polling slots are granted periodically and only after the CDMA codes (from BE connections) have been responded to. If there are no slots left in the present frame, we have to wait. The same phenomenon could happen with CQICH based resumption, only with higher traffic load, as CQICH based resumption consumes less resources.

C. *50 File Downloading Users, Limit for CDMA Allocations (Case 3)*

If we do not limit the number of uplink slots that can be granted as a response to CDMA codes, the BS allocates slots for sending bandwidth requests as a response to all CDMA codes it has received. Since adding CDMA allocation IEs to the UL-MAP happens before scheduling any user traffic, it is likely that the non-real time connections "steal bandwidth" from the real-time connections.

We limited the number of uplink resources that could be granted as a response to CDMA codes to ten slots, and the results changed dramatically (see Fig. 9–11). The reason for this is that now there are more slots available for ertPS connections. In this scenario, most of the SSs that participate in contention resolution have an ARQ feedback (or TCP acknowledgment) to send. Delaying ARQ feedbacks of BE

[6] This figure illustrates the averaged number of free uplink slots after the real-time (ertPS) connections have been served.

connections is a better alternative than letting ertPS VoIP connections suffer. Of course, it may every now and then happen that we delay the resumption of an ertPS VoIP connection when contention or multicast polling based resumption is used. (However, the latter should be a rare event. In our simulator, the contention region for multicast polling comes before the basic contention region in the UL subframe and thus multicast polling SSs get the first CDMA allocations.) This can be seen from Fig. 11: now polling and CQICH based resumption give the best delay performance. Multicast polling with optimized backoff parameters performs well, too.

In any case, however, it is possible that a high number of SSs that are hosting BE connections can cause problems to ertPS and other real-time connections. This is due to CQICH reports, HARQ acknowledgements, HARQ retransmissions and the aforementioned uplink CDMA allocations that are all granted slots before any user connection. A partial solution would be to have larger CQICH report intervals for the BE users. However, since an SS can host many connections (of different types) it might make more sense to introduce admission control for all active SSs – no matter what connections they might host.

V. Conclusions and Discussion

In this paper, we have presented different uplink channel access and ertPS resumption mechanisms for VoIP traffic in IEEE 802.16e systems. Our simulation studies indicate that we can have more ertPS VoIP connections (or better QoS), if we get rid of polling during the silence periods. However, this can result in high delays if the SS has to participate in contention after each silence period.

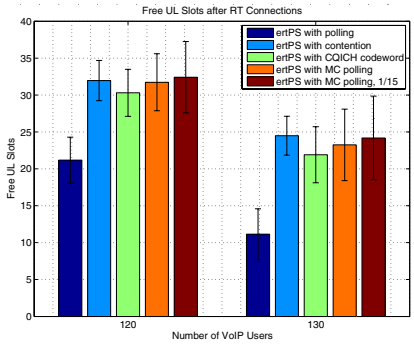

Fig. 3. Case 1: free UL slots.

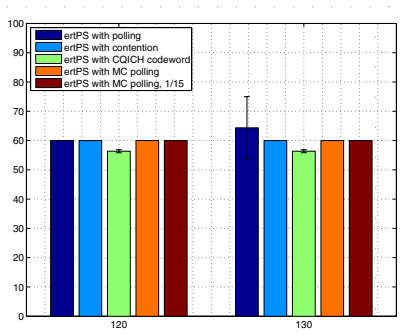

Fig. 4. 95th percentile UL VoIP delay.

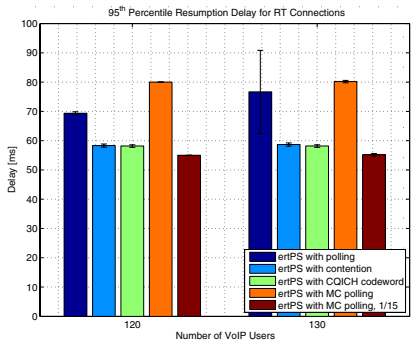

Fig. 5. 95th percentile resumption delay.

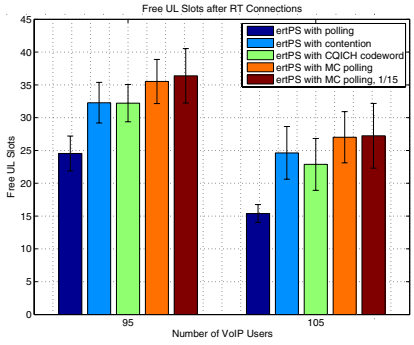

Fig. 6. Case 2: free UL slots.

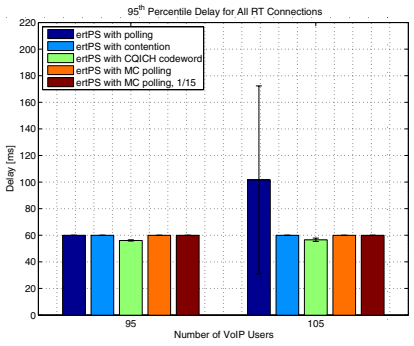

Fig. 7. 95th percentile UL VoIP delay.

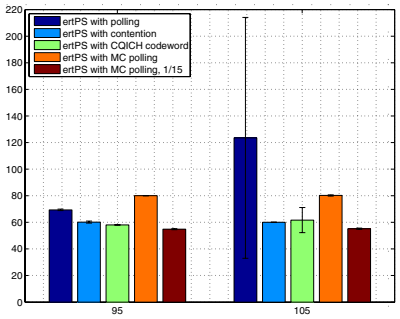

Fig. 8. 95th percentile resumption delay.

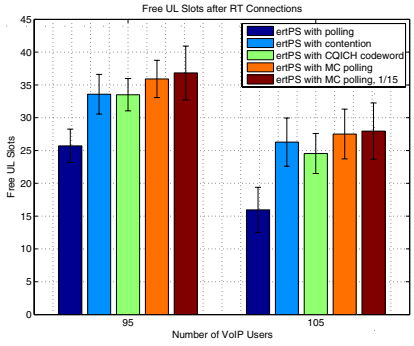

Fig. 9. Case 3: free UL slots.

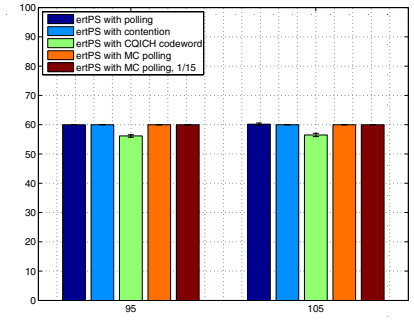

Fig. 10. 95th percentile UL VoIP delay.

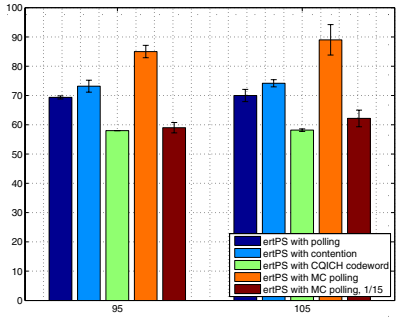

Fig. 11. 95th percentile resumption delay.

Multicast polling, with appropriate request backoff parameters, can lower the resumption delays. However, when there are lots of VoIP connections, we might need more multicast polling groups, which would take resources from the basic contention. Of course, multicast polling does not bring any gains if the basic contention is not a bottleneck.

Multicast polling group members (i.e., those SSs that host VoIP connections) send their bandwidth request CDMA codes in a dedicated multicast polling region, which prevents these codes from colliding with the codes sent by the BE connections. In our implementation, we have also prioritized uplink CDMA allocations based on the contention region and thus multicast polling group members always get the first uplink CDMA allocations.

CQICH can also be used for ertPS resumption. With CQICH based resumption, delays are lower when compared to contention based resumption – assuming that contention is a bottleneck. However, with large CQICH reporting intervals, this approach could cause some problems to link adaptation: if we use the CQICH message for resumption, we cannot update the SNR in the same CQICH message.

If the CQICH reporting interval is short enough, our recommendation is to use CQICH based ertPS VoIP resumption. If that is not the case, multicast polling with its own request backoff parameters should be used. With contention based resumption, we cannot guarantee low enough ertPS VoIP resumption delays unless the number of other connections participating in contention is somehow limited.

Moreover, we have observed some issues that make resource management and connection admission control in IEEE 802.16e quite challenging. A common approach is that CQICH reports, HARQ acknowledgements, HARQ retransmissions, and CDMA uplink allocations are always granted slots before any real-time connection. Therefore, it seems that admission control for real-time connections only is not sufficient but in addition to connection admission control we should have admission control for the SSs when they are entering the network.

ACKNOWLEDGMENT

This work was performed when Jani Lakkakorpi was still with Nokia Devices R&D.

The authors would like to thank everyone involved in simulator development at the Telecommunication laboratory of University of Jyväskylä and especially Olli Alanen for his help with the multicast polling issues.

REFERENCES

[1] Air interface for fixed broadband wireless access systems. IEEE Standard 802.16, Jun. 2004.

[2] Air interface for fixed broadband wireless access systems – amendment for physical and medium access control layers for combined fixed and mobile operation in licensed bands. IEEE Standard 802.16e, Dec. 2005.

[3] D.J. Deng, L.W. Chang, C.H. Ke, Y.M. Huang and J. Morris Chang, "Delay Constrained Uplink Scheduling Policy for rtPS/ertPS Service in IEEE 802.16e BWA Systems," International Journal of Communication Systems, vol. 22, pp. 119–133, Feb. 2009.

[4] J.W. Soo, "Performance Analysis of Uplink Scheduling Algorithms for VoIP Services in the IEEE 802.16e OFDMA System," Wireless Personal Communications, vol. 47, pp. 247–263, Oct. 2008.

[5] J. Andrews, A. Ghosh and R. Muhamed, "Fundamentals of WiMAX: Understanding Broadband Wireless Networking," Prentice Hall, Feb. 2007.

[6] Olli Alanen, "Multicast Polling and Efficient VoIP Connections in IEEE 802.16 Networks," Proceedings of MSWIM'07, Chania, Greece, Oct. 2007.

[7] F. Wang, A. Ghosh, C. Sankaran, P.J. Fleming, F. Hsieh and S.J. Benes, "Mobile WiMAX Systems: Performance and Evolution," IEEE Communications Magazine, pp. 41–49, Oct. 2008.

[8] A. Sayenko, O. Alanen, H. Martikainen, V. Tykhmyrov, A. Puchko and T. Hämäläinen, "WINSE: WiMAX NS-2 Extension," Proceedings of 2nd International Conference on Simulation Tools and Techniques, Rome, Italy, Mar. 2009.

[9] M. Shreedhar and G. Varghese, "Efficient fair queueing using Deficit Round-Robin," IEEE/ACM Transactions on Networking, vol. 4, pp. 375–385, Jun. 1996

[10] UCB/LBNL/VINT, "Network Simulator – ns (version 2)," Feb. 2008.

[11] WiMAX Forum, "Mobile System Profile Specification: Release 1.5 Common Part (Revision 0.2.1: 2009-02-02)," Feb. 2009.