Voice traffic bicasting enhancements in mobile HSPA network

Voice Traffic Bicasting Enhancements in Mobile HSPA Network

Vesa Hytönen
Department of Mathematical Information Technology,
University of Jyväskylä, Finland
vesa.a.hytonen@jyu.fi

Abstract—This paper discusses methods for improving the effectiveness of delivering voice traffic over High-Speed Downlink Packet Access (HSDPA) network by employing transmission diversity with Single-Frequency Dual-Cell (SF-DC) Aggregation that is part of the Multiflow specification. SF-DC Aggregation allows the user to be served at the same time by two different cells. The enhancements discussed in the paper capitalize strongly on the availability of the composed Channel Quality Indication (CQI) feedback at both serving cells. According to the results obtained from the network simulations, the ability to select the better channel for each voice transmission significantly decreases the required transmission power of the cell, thus also improving the residual cell capacity available for best effort (BE) traffic. The required updates are completely software-based and are applied only in the access network, making them transparent to user terminal.

Keywords—HSPA, Multiflow, SF-DC Aggregation, Bicasting, CSoHS, VoHS.

I. INTRODUCTION

According to [1], the growth in the number of High-Speed Packet Access (HSPA) subscriptions will increase in the coming years along with the volume of real-time communication in mobile networks. In addition to growing amount of video content delivered in the network, there is an interest towards extending the usage of HSPA to deliver voice traffic originating from circuit switched network or IP Multimedia Subsystem (IMS), creating a demand for an efficient control of real-time traffic delivery over the wireless channel [2].

One of the recent entrants in the HSPA+ standard, developed by the 3rd Generation Partnership Project (3GPP), is a multipoint transmission concept called Multiflow that was introduced in the Release 11 for HSDPA [3]. Resembling Co-operative Multi-point (CoMP) concept in Long Term Evolution (LTE), Multiflow aims to improve the data plane performance particularly in cell edge areas by utilizing additional traffic flows towards user terminal (UE). In its simplest form, SF-DC Aggregation, a UE can be served concurrently by two cells, known as the primary and secondary serving cells. Multiflow is conventionally designed to elevate the user-specific data rate in Downlink direction. However, network operators have an option to enable duplication of user plane data over the primary and secondary data flows, which is useful especially with traffic, such as voice traffic, that demands a higher level of quality of service (QoS) [4]. Transmission diversity with multiple low- or uncorrelated channels provided by Multiflow enables protection against multipath attenuation as well as shadow fading, and transmitting the same data over these channels greatly improves the likelihood of a successful packet reception.

The current study aims at improving the effectiveness of voice calls delivered over HSPA network by utilizing SF-DC Aggregation related methods. Specifically, enhanced voice-traffic flow management and scheduling are examined and evaluated by network simulations. Some of the recent efficiency improvement studies related to voice over HSPA have concentrated on the Enhanced Serving Cell Change (E-SCC) ([5]) as well as duplication of voice data flow by utilizing connections with different network operators ([6]). Research on using Multiflow for improving voice service is, however, exiguous, mainly due to the novelty of the multipoint transmission concept. Many of the principles presented in this paper can also be applied for DL CoMP in LTE Advanced. However, the existing X2 interface between eNodeBs in LTE already enables an efficient dynamic cell selection, making the applicability of certain parts unnecessary. In a network with non-saturated voice capacity, the presented voice traffic control, in conjunction with an efficient power control operated by the NodeB, is capable of saving significant amount of transmission resources at the NodeB which also results in an extended residual cell throughput. One should notice that the methods proposed should be completely software-based that can be taken into use by a firmware update, and that the updates can be enabled in the access network devices, NodeB in particular, which makes the changes transparent to UEs.

The remainder of the paper is organized as follows. The proposed flow management mechanisms are discussed in Section II. The system model and scenarios for simulations are described in Section III, and the obtained results are presented in Section IV. Section V concludes the paper.

II. VOICE BICASTING ENHANCEMENTS

Two fundamental targets regarding voice traffic service are tackled in this paper. First, the data should be delivered to UE as reliably as possible while complying the temporal limitations in order to maintain a good call quality and to avoid outage condition. Second, the resource requirements of voice users ought to be minimized to achieve a better residual network capacity. Employment of SF-DC Aggregation allows the above to be performed by operations related to flow management and scheduling in NodeB.
A. Traffic Flow Utilization

This section concentrates on discussing the rules based on which the two traffic streams in SF-DC Aggregation should be used or deactivated in relation to voice bicasting. The underlying target is to serve the UE over the channel that is likely to be the better from the two available channels. The roles of Radio Network Controller (RNC) and NodeB and their suitability for participating the flow control process are considered.

1) The role of RNC in enhanced voice traffic management: RNC supporting Multiflow is responsible for either splitting or duplicating the user plane data between the flows in the Radio Link Control (RLC) protocol layer [7], [4]. With conventional unspecified bit rate (UBR) QoS traffic (e.g., FTP, web browsing), RNC splits the information for the Multiflow cells, and the cells forward all the data to UE according to their internal scheduling operation. Applying a slightly different approach for voice traffic, the RNC can duplicate each Protocol Data Unit (PDU) to both serving Multiflow cells [4].

It is possible to augment the responsibilities of RNC in bicasting process so that it may either duplicate or split the voice data based on certain conditions. The RNC would have to rely upon the channel measurement reports from UEs, for example, to deliver the user data to the serving cell with a stronger channel for downlink transmission. A disadvantage of this approach relates to the accuracy of how the "best" cell can be selected. First of all, this approach assumes that periodical measurement reports are available to the RNC, which can be enabled by the network service provider. Assuming periodical reporting is enabled, channel measurement reports are usually filtered over a longer period of time, which means that they represent the average quality level of the channel. This will rule out the possibility of fast adaptation to channel fluctuation, especially since the reporting interval may range from hundreds of milliseconds to several seconds. Since the enhanced traffic control by RNC would likely lead to very ineffective voice bicasting, the function of RNC in this study is restricted to simple duplication of each voice-traffic related RLC PDU to both cells, as depicted in Fig. 1, leaving further flow control to be performed by NodeBs.

2) NodeB controlled flow management: The primary control of the voice traffic flows is assigned to NodeB. As concluded in the previous section, RNC will duplicate all data to both Multiflow cells and the cell decides whether it should serve the UE or not in certain Transmission-Time-Interval (TTI). The actual advantage of this method originates from the fact that the UE will report a composite CQI and Hybrid Automatic Repeat Request (HARQ) acknowledgement of both serving cells on one High-Speed Dedicated Physical Control Channel (HS-DPCCH) that can be decoded by both cells (see Fig. 1), allowing a cell to estimate the second Multiflow link quality and compare it to the cell’s own link [8].

For achieving the best performance from the viewpoint of overall resource consumption, each RLC PDU should be transmitted only from one cell. Sending the same packet from both cells always dissipates the remaining capacity. Since the RNC duplicates all data over both Iub interfaces, it is the cell’s responsibility to ensure no duplicate transmission takes place, unless explicitly desired. Therefore, the cell needs to discard the PDUs that it should not transmit from its queue. The PDU extraction will be done immediately when the cell receives a new packet from the RNC. By applying the same rules on both Multiflow cells and assuming the primary serving cell knows which PDUs are deliverable also to the secondary serving cell, it can be guaranteed that each PDU will be transmitted exactly from one cell. This method is selected for the study, as it allows a controlled way of handling PDUs in both Multiflow cells.

The cell relies on the uplink CQI reports when deciding whether to discard a packet. When a cell receives a new PDU on the Iub and detects that the Multiflow operation is enabled, it inspects the composite HS-DPCCH including the CQI report for both flows. Upon the CQI lookup, the cell calculates a relative channel quality between both serving cells by

$$\Delta r_n(l) = r_{n,0}(l) - r_{n,1}(l).$$

Since the CQI report received in TTI $l$ from UE $n$ contains only the index of the supported Modulation and Coding Scheme (MCS), the cell may match the index to transport block size, which is here denoted by $r_{n,i}(l)$.

Subscripts 0 and 1 represent the current cell and opposite Multiflow cell, respectively. Both serving cells will also trace the slope of the channel by utilizing channel quality information from previous TTIs. As most of the physical layer errors take place during an adverse channel progression (assuming a constant interference level), i.e. channel quality decreased between the creation of CQI report and the data transmission, a favorable condition would be to switch the data flow if the opposite flow has an increasing trend in quality. A simple linear estimation of the slope for cell $i$ is performed as follows:

$$m_{n,i}(l) = r_{n,i}(l) - r_{n,i}(l - 1).$$

It can be assumed that the CQI already includes a filtered representation of the channel, therefore a direct subtraction of two subsequent data rate values based on supported CQIs will suffice. Another reason for not tracing the CQIs from a longer period is that the algorithm needs to follow the small-scale fading of the channel as well as possible, which would not be possible with averaging CQIs over a long duration.

After performing the steps above, cells that receive bicasted PDUs from the RNC follow simple heuristic rules for discarding the packet. In case the cell observes that the CQI related to its own channel is worse than the CQI of the second Multiflow channel, it discards the packet. When the CQIs are equal, which often can be the case as the CQI is a discrete variable,
the cell makes the decision based on the calculated channel slopes and discards the packets if the own slope is smaller than the opposite slope. Moreover, only the primary serving cell enqueues the new PDU if the slopes are also similar. In other form, set the discard flag as follows:

\[
\text{procedure Cell::DiscardPDU(psec)}
\]

\[\text{discard} \leftarrow \text{False}\]

\[\text{if } \Delta r_n(l) < 0 \text{ then}
\]

\[\text{discard} \leftarrow \text{True}\]

\[\text{else if } \Delta r_n(l) = 0 \text{ then}
\]

\[\text{discard} \leftarrow \text{False}\]

\[\text{else if } m_{n,0}(l) < m_{n,1}(l) \text{ then}
\]

\[\text{discard} \leftarrow \text{True}\]

\[\text{end if}\]

\[\text{end if}\]

\[\text{return discard}\]

\[\text{end procedure.}\]

The rules should guarantee that the UE will always be served over either of the flows with a sufficient channel quality, which affects not only on the successful packet reception probability but also the transmission power requirement. When the conditions are met in a cell, i.e. it is allowed to serve the UE, the cell’s internal scheduler will weigh the priorities of each of its active UEs. The actual scheduling is discussed in the next section.

### B. Scheduling and Power Control

After the schedulability of each UE is resolved, the cell should order its UEs based on their scheduling metrics. The scheduling and power control methods described here are applied to both baseline and Multiflow simulations in this study.

For best effort UEs, basic proportional fair scheduling is a commonly used method, which tries to provide some level of fairness, while at the same time promoting UEs whose relative channel capability is above those of other UEs [9]. The basic formula for proportional fair user selection is

\[n = \arg \max_{n \in N} \frac{\tau_n(l)}{\tau_n(l)},\]

where \(n\) denotes a user belonging to a group of active users \(N\) and \(\tau_n(l)\) is the instantaneous achievable throughput for user \(n\) in TTI \(l\) based on the CQI report. \(\tau_n(l)\) represents the moving average data rate obtained by

\[\tau_n(l+1) = (1 - \alpha)\tau_n(l) + \alpha r_n(l),\]

Here, \(r_n(l)\) denotes the actual data rate from previous transmission in TTI \(l\) and \(\alpha\) is the forgetting factor. This method is applied to the best-effort traffic flows in this study. According to the formulas above, if a UE is scheduled, its priority metric will decrease in the following TTI, whereas if an active UE was not selected by the scheduler, \(r_n(l)\) becomes zero and the priority is increased. This is to prevent the UE from being blocked for long periods of time.

With QoS-dependent voice data it is crucial to deliver the packets within a certain delay budget in order to maintain a good call quality. This is often done by applying a delay coefficient that modifies UE’s scheduling priority based on how long the voice packets have stayed in the transmission buffer. In this paper a similar extended approach is used, with a basis on proportional fair. For voice users, (3) becomes

\[n = \arg \max_{n \in N} \frac{r_n(l)}{D_n(l)},\]

where \(D_n\) is the delay factor.

\[D_n = \max\{w \cdot c_n, 2^{\alpha(x - (db - z))}\},\]

\(c_n\) denotes the number of packets in UE’s transmission buffer. This number is multiplied by a coefficient \(w\), that in this study is set to 0.25. The latter part in the maximization function is the exponentially increasing delay term, where \(r_n\) represents the Head-of-Line (HoL) delay in milliseconds, matched to delay budget \(db\) with an offset \(x\). In the simulations, \(db\) is given the value of 80 milliseconds and the offset \(x\) is set to 20 milliseconds. As a result, the delay coefficient will reduce the scheduling priority of UEs whose HoL delay is small. Once the delay starts to approach the maximum delay budget, the priority will increase exponentially to ensure the packet can be delivered in time. The optional offset \(x\) is used to cope with the maximum hard limit of buffering time, which in this case is 80 milliseconds. The offset ought to guarantee that the packet will be transmitted before the hard limit is reached and the packet is discarded.

Another small but important modification for the voice traffic scheduling is required in the data rate filtering. The separation of \(r_n\) and \(r'_n\) in (3) and (4) should be revised for voice as the \(r_n\) represents the theoretical upper bound of the channel capability, while \(r'_n\) is the previous real data rate. As mentioned earlier, BE user’s priority will be increased in the next TTI if it does not get scheduled (e.g. \(r' = 0\)), in order to prevent a long term starvation. This is actually dispensable for voice traffic, because a separate delay factor is utilized. Furthermore, data rate of the used Adaptive Multi-Rate (AMR) audio codec is relatively low and independent from the actual channel quality, which means \(r'_n\) does not provide useful information for measuring the relation between current channel quality and filtered, long term average of the channel. Due to these reasons, (4) for voice UEs should be replaced by

\[\tau_n(l+1) = (1 - \alpha)\tau_n(l) + \alpha r_n(l),\]

which now allows ordering the voice traffic flows correctly according to their current channel quality relative to the average quality level.

Proportional fair-based scheduling is chosen to reduce the power requirement by scheduling UEs with relatively good channel conditions. For low velocity users, the CQI report is able to provide a channel quality estimation with a fairly high accuracy, since the de-correlation time of fast fading may often be clearly longer than the reporting delay. In other words, the scheduling algorithm in NodeB is able to benefit from fast variation of channel caused by multipath propagation. Increasing the user’s velocity results in a more random fluctuation in adjacent CQI reports, due to fast fading becoming uncorrelated. However, slow fading and path loss create a slowly-changing trend in the channel envelope. By
nature, the moving average filtering given in (7) tends to avoid scheduling UEs during a negative slope in the channel quality, which actually is a very likely time for erroneous packet reception, as already discussed.

Selection of proportional fair scheduling for voice traffic is partially linked to power control implementation. When best effort traffic is transmitted, the served UE is allocated full transmission power of the cell so as to achieve a target BLER for the selected MCS. As discussed in [4], apart from bundling of packets in one transport block, usually the same MCS is used for voice, since the packet size remains the same. Conventional link adaptation is therefore of little use, and the BLER target can instead be reached by adaptive power control.

III. SIMULATION MODEL

A system simulator employing a hexagonal grid of a total of 21 cells was used for the evaluation of the voice enhancements. Voice users are deployed randomly within the simulation area, where they move in random directions during the simulation. When a user moves from the coverage of one cell to another, a hard handover is executed if the channel of the neighboring cell remains better during the whole time-to-trigger (TTT) duration than the channel of the primary serving cell plus the hysteresis value. The serving cell change (SCC) process includes a constant delay of 100 milliseconds that models the uplink signaling and RNC processing latencies, after which the handover takes place. This allows modeling the mobility procedures with a sufficient accuracy to assess the behavior of the algorithms in realistic scenarios where the channel quality of the primary serving cell may become worse than the channel of a neighboring cell, due to delays of the mobility procedures. Updating the secondary serving cell when Multiflow is enabled follows a similar RLC procedure as that for primary serving cell. Best effort UEs are also dropped in random positions in the network, but the mobility for them is disabled. The reason for including best effort users is to measure the remaining cell capacity that is left after scheduling the voice users.

A set of simulation parameters is provided in Table I. They are mostly based on the 3GPP simulation assumptions (e.g. [3]). The reporting range logically matching to 1A/1B measurement events, acts as the secondary serving cell activation threshold. When the UE enters this "handover region", it is possible to utilize the second Multiflow stream for transmission as procedure that as for primary serving cell. Best effort UEs are also dropped in random positions in the network, but the mobility for them is disabled. The reason for including best effort users is to measure the remaining cell capacity that is left after scheduling the voice users.

A set of simulation parameters is provided in Table I. They are mostly based on the 3GPP simulation assumptions (e.g. [3]). The reporting range logically matching to 1A/1B measurement events, acts as the secondary serving cell activation threshold. When the UE enters this "handover region", it is possible to utilize the second Multiflow stream for transmission as procedure that as for primary serving cell. Best effort UEs are also dropped in random positions in the network, but the mobility for them is disabled. The reason for including best effort users is to measure the remaining cell capacity that is left after scheduling the voice users.

A set of simulation parameters is provided in Table I. They are mostly based on the 3GPP simulation assumptions (e.g. [3]). The reporting range logically matching to 1A/1B measurement events, acts as the secondary serving cell activation threshold. When the UE enters this "handover region", it is possible to utilize the second Multiflow stream for transmission as procedure that as for primary serving cell. Best effort UEs are also dropped in random positions in the network, but the mobility for them is disabled. The reason for including best effort users is to measure the remaining cell capacity that is left after scheduling the voice users.

A set of simulation parameters is provided in Table I. They are mostly based on the 3GPP simulation assumptions (e.g. [3]). The reporting range logically matching to 1A/1B measurement events, acts as the secondary serving cell activation threshold. When the UE enters this "handover region", it is possible to utilize the second Multiflow stream for transmission as procedure that as for primary serving cell. Best effort UEs are also dropped in random positions in the network, but the mobility for them is disabled. The reason for including best effort users is to measure the remaining cell capacity that is left after scheduling the voice users.
from the generally higher power requirement in the VA30 than in the PA3 simulation, a high similarity is observed between the results. Extracted from the figure, with bicasting, on average approximately 55% of baseline’s transmission power in the PA3 case is required to serve the voice users in the handover region. One should notice that the major improvement is the reduced number of TTIs when very high transmission power is required.

On average, only 49% of the baseline power is required in the VA30 scenario after enabling Multiflow. Since the path loss and shadowing variation becomes faster than in PA3, the relative superiority of the average channel qualities of neighboring cells may change often, which allows good prospects for utilizing the second traffic channel. In the baseline VA30 simulations, around 10% of the samples reach the maximum High-Speed Downlink Shared Channel (HS-DSCH) transmission power of 16 watts. It is a result from bad channel quality where only the lowest MCS capable of transmitting one RLC layer PDU in single transport block is supported. To reach the 10% BLER target with such MCS the cell must allocate the maximum available power for one voice user, which naturally disables user multiplexing in one TTI. In most cases, this can be avoided by using bicasting. The small steps visible especially in VA30 curves also relate to the selection of MCS and are produced by the power control algorithm. The allocated power is based on the reported CQI, which is a discrete value. Thus there is only a finite number of possible powers to be designated for a user.

Data rates of the best effort users are gathered from the TTIs when the voice users are active in the handover zone. The average rates from the simulation scenarios are depicted in Fig. 3. The achieved BE data rate improvement is quite moderate. Depending on the scenario, the gain varies approximately from 2% to 12% with PA3 and from 4% to 22% with VA30. Although the power saving is significant, it cannot be translated linearly to higher throughput. Instead, the data rate gain depends on what is the effect of additional available cell power to the MCS selection. For example, being able to choose a more efficient MCS might often require at least around 1 dB (or approximately 26%) higher transmit power than the previous supported MCS. Nevertheless, when the voice load of the network approaches the saturation point, it is likely possible to benefit from the bicasting enhancements more frequently, resulting in an increased total cell throughput compared to conventional single-cell operation.

As mentioned earlier, packet drop rates are not significant in majority of the simulated scenarios. The only exception is the baseline VA30 simulation with a mean of 60 voice users per cell. There, the network starts to saturate, resulting in approximately 2% of the users being in outage condition if a 2% packet drop limit is set as an outage threshold. In the corresponding bicasting scenario, 0% of UEs are in outage, thus by applying the discussed methods, the voice capacity of the network can be extended. However, further investigation would be needed to determine the exact improvement.

V. CONCLUSIONS

This paper discussed enhanced voice-bicasting concept which allows the Multiflow-capable cell edge user to be served over either of the configured data streams. In the presented approach, the RNC duplicates each PDU related to voice traffic to both serving cells, and the cell decides whether it should discard the PDU or transmit it to the target UE. Utilization of SF-DC Aggregation for improving the real-time voice traffic delivery in downlink direction can be done without hardware updates and the software changes concentrate only on the HSPA access network. It was shown that delivering voice traffic in HSPA network may require a substantial amount of cell’s resources if the user resides in the border of the serving cell, whereby conditional voice bicasting saves a significant amount of the transmission power in such cases which grants higher residual cell capacity and more efficient user multiplexing. Further optimizations voice bicasting are possible for the channel selection conditions. The suitability of bicasting for other types of real-time traffic is also a sound target for research in the future, as well as the study on the impact of different CQI reporting delays on the accuracy of the discussed methods.

ACKNOWLEDGMENT

The author would like to thank Alexander Sayenko from Nokia Solutions and Networks for his guidance.

REFERENCES