Compressed Transport of Baseband Signals in Radio Access Networks

Dragan Samardzija, John Pastalan, Michael MacDonald, Susan Walker, and Reinaldo Valenzuela

***Abstract*—In current wireless base station solutions it is becoming common to physically separate baseband units and radio subsystems. In many wireless technologies this architecture requires allocation of significant transport network resources. In this paper a low-latency baseband signal compression scheme is presented. The compression scheme significantly lowers the transport data rate while maintaining low levels of signal distortion, thus resulting in a lower-cost transport network. Considering the importance of packet-based networks, a number of additional novel compression schemes are proposed. They are optimized for transport networks that implement a quality-of-service (QoS) mechanism and/or multi-link transmission. The compression schemes are parameterized such that a smooth trade-off between the required signal quality and compression performance can be achieved through operator choice of the suitable parameter values. An attractive feature of these schemes is that they can be applied to different wireless technologies, with appropriate parameter settings, without disrupting the present architecture. The proposed solutions will lead to a cost-effective implementation of collocated and distributed network-centric baseband processing, coordinated multi-point (CoMP) and/or distributed antenna system (DAS) which are critical topics for the entire wireless telecommunications industry and infrastructure.**

***Index Terms*—Compression, transport network, RRH, distortion, dithering, EVM, CoMP, DAS.**

I. Introduction

NOVEL wireless base station solutions, where baseband units (BBUs) and radio subsystems are physically separated, represent an important change in radio access network architecture. Specifically, the antennas, radio-frequency front-end and analog-to-digital interface are a part of the remote-radio heads (RRHs). The RRHs are connected to the BBUs via digital transport network. Digitized baseband complex inphase (I) and quadrature (Q) samples are transported over the transport links between the RRHs and BBUs. This architecture enables novel network deployments and implementation of advanced transmission techniques. It offers a significant potential to cost-effectively increase data rates and improve user experience. The key technical and economic issue is that this architecture requires significant transport network resources and the corresponding infrastructure investment. In this paper we address this particular problem, proposing a number of solutions that provide effective usage of transport network.

The above architecture represents a key platform for implementing a number of the current and future radio access network solutions. Three solution examples are the collocated and distributed network-centric baseband processing [1], coordinated multi-point transmission and reception (CoMP) [2]–[5], and the distributed antenna system (DAS) [6]–[8]. Each implementation is based on a transport network that connects the RRHs to a multiplicity of collocated or distributed baseband processing resources.

As the industry standard, the Common Public Radio Interface (CPRI) transport technology has been widely applied to connect RRHs and BBUs [9]. It supports different network architectures, and transports uncompressed I/Q samples. In many wireless technologies (3G and 4G), such a transmission requires allocation of significant transport network resources. For example, to transport a 10 MHz LTE waveform to a single antenna, CPRI requires 460.8 MBPS (excluding protocol overhead). Consequently, CPRI will require 1.843 GBPS per four-antenna multiple-input multiple-output (MIMO) cell, i.e., sector.

In this study we propose baseband signal compression schemes (i.e., I/Q compression) that lower the required transport data rates. In Section II we describe a baseline solution, providing detailes on the trade-off between the compression rates, latency and signal quality. For example, in LTE, the solution results in three times lower data rates than in the case of uncompressed I/Q transmission. The proposed solution is general in nature and hence, can be applied to different wireless technologies, (e.g., LTE/LTE-Advanced and UMTS/HSPA), as well as, on the uplink and downlink. Furthermore, the proposed solution maintains the overall signal quality, i.e., error vector magnitude (EVM) and adjacent carrier leakage power ratio ($ACLR$) that are required by a particular wireless technology, i.e., standard.

The above solution applies a set of well-known signal processing techniques. It should be viewed as a baseline that we use to propose two novel compression schemes. The novel schemes are described in Section III. They are optimized for packet-based transport networks that implement a quality-of-service (QoS) mechanism and/or multi-link transmissions. Due to a broad adoption of those networks (e.g., Ethernet or mesh wireless), enabling efficient I/Q transport over those networks is considered particularly critical for the corresponding cost-effective implementation. Furthermore, the proposed schemes exploit multiplexing and diversity aspects of transport networks. To the best of our knowledge, this represents our original contribution.

In addition, the proposed techniques are parameterized such that a smooth trade-off between the required signal quality and compression performance can be achieved through operator

Manuscript received July 18, 2011; revised December 1, 2011 and February 15, 2012; accepted May 4, 2012. The associate editor coordinating the review of this paper and approving it for publication was Y. Li.

The authors are with Bell Laboratories, Alcatel-Lucent, Holmdel, NJ 07733, USA (e-mail: {dragan.samardzija, john.pastalan, mike.macdonald, susan.walker, reinaldo.valenzuela}@alcatel-lucent.com).

Digital Object Identifier 10.1109/TWC.2012.12.111359

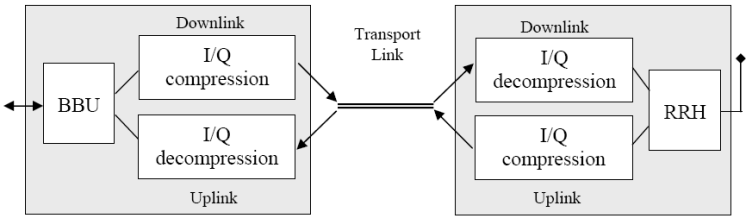

Fig. 1. System consisting of a BBU, RRH and I/Q compression and decompression.

choice of suitable parameter values. Furthermore, the processing delay is limited (i.e., compression and decompression latency), compared to the uncompressed version.

The paper is organized as follows. In Section II we present the baseline algorithm details. The corresponding subsections will describe removal of redundancies in spectral domain, block scaling, and optimized quantization. The two novel schemes for transmission over multiple links with QoS are proposed in Section III. Numerical and experimental results are presented in Section IV. We discuss benefits and applications of the proposed solutions in Section V.

II. Algorithm Details

Figure 1 depicts the basic functional blocks of a system that is the subject of this study. Namely, the system consists of an RRH and a BBU connected via a transport link. On the uplink, the RRH radio-frequency front-end and analog-to-digital converter (ADC) convert the received analog radio signal into the digital I/Q sample form. Typically, the ADC is a conventional high-resolution converter. After the analog-to-digital conversion, the proposed I/Q compression is applied, and its output is transported to the BBU. The decompression is applied at the BBU, followed by the receiver baseband processing (i.e., physical layer) of a particular wireless technology.

Conversely, on the downlink, the BBU transmitter generates a sequence of I/Q samples, which are compressed using the proposed scheme. The output of the compression is then transported to the RRH, where the decompression takes place. Following the decompression, the RRH digital-to-analog converter (DAC) and radio-frequency front-end convert the sequence of decompressed I/Q samples into the analog radio signal that is being transmitted. Typically the DAC is a conventional high-resolution converter.

Note that the RRH radio-frequency front-end, ADC, DAC as well as BBU processing are identical to the ones ordinarily applied to a given wireless technology. Namely, there is nothing in those subsystems that is specifically implemented to accommodate the proposed I/Q compression and decompression. The solution may be viewed as a 'black-box' with respect to other subsystems.

The functional block scheme on the proposed baseline I/Q compression is presented in Figure 2. It consists of basic functional blocks that are described in this section.

A. Removal of Redundancies in Spectral Domain

Based on the current practice the sampling rate of the ADC, DAC and BBU processing is higher than the minimum required according to the Nyquist sampling theorem. In UMTS/HSPA as well as cdma2000/EV-DO, 2-time and 4-time oversampling is customary. Similarly, in LTE the sampling rate exceeds the signal bandwidth. This results in redundancies in the spectral, i.e., frequency domain. Namely, in the uncompressed form, a spectrally broader signal is transmitted than what is necessary. For example, in 10 MHz LTE, the sampling rate is 15.36 MHz (both for the BBU processing and in the case of CPRI uncompressed transmission), where approximately one-third of the spectrum carries no information relevant to the LTE transmission. Consequently, this particular function is designed to remove those redundancies. It is implemented as a multi-rate filter. Starting from the original sampling rate f_s, this function downsamples the input signal to the lower sampling rate f_{ds}. The downsampling factor F is a rational number

$$F = \frac{f_s}{f_{ds}} = \frac{L}{K} \geq 1, \tag{1}$$

where L and K are positive integers.

The input signal $s(k)$ is sampled at the original sampling rate f_s. After K-time upsampling, where $K-1$ zeros are

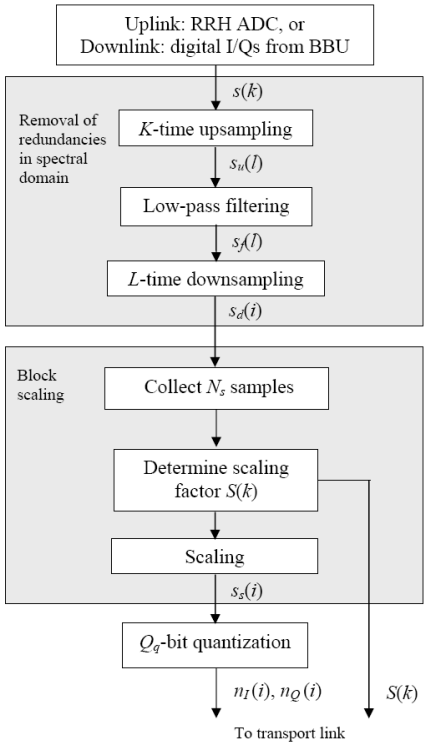

Fig. 2. I/Q compression block scheme.

inserted, the signal is

$$s_u(l) = \begin{cases} s(k) & l = K\,k \\ 0 & l \neq K\,k \end{cases} \quad (2)$$

where k and l are integers denoting time samples. The upsampled signal $s_u(l)$ is then low-pass filtered with the bandwidth limited to $[-f_{ds}/2, f_{ds}/2]$. The filter's finite impulse response is

$$g(l) = w(l)\frac{\sin(\pi l/L)}{\pi l/L} \quad (3)$$

where $w(l)$ is the window function. For example, we have implemented Hamming window that is defined as

$$w(l) = 0.54 - 0.46\cos\left(\frac{2\pi(l + N_w/2)}{N_w}\right) \quad (4)$$

where N_w represents the dimension of the window function, i.e., the filter length. Note that any other window function may be applied, depending on the desired filter length and impulse response. The low-pass filter output is

$$s_f(l) = \sum_{i=-N_w/2}^{N_w/2-1} s_u(l-i)g(i) \quad (5)$$

sampled at the frequency Kf_s. After the filtering, every L-th sample is selected such that

$$s_d(i) = s_f(Li). \quad (6)$$

The signal $s_d(i)$ is sampled at the frequency f_{ds}, thus completing the downsampling process.

Conceptually, in this function there is nothing exclusive to a particular wireless technology. Only the sampling rates f_s and f_{ds}, and filter length N_w should be specified. In general, the above parameters (L, K, N_w) should be selected to optimize complexity-versus-performance tradeoff. For example, in the 10 MHz LTE case, we have implemented the above multi-rate filter to lower the sampling rate from 15.36 MHz down to 10.24 MHz ($L = 3, K = 2, N_w = 64$) with no measurable signal distortion and low latency (2.08 usec).

In order to efficiently implement filtering in (5), other filter structures may be used. For example, polyphase or cascaded integrator-comb filters [10] should be considered, which is beyond the scope of this study.

B. Block Scaling

In mobile communications, a typical radio signal has a large dynamic range. For example, in the uplink with multiple simultaneous users, due to different large- and small-scale propagation effects and mobility, the received signal power may experience significant variations. In 3G and 4G, the signal variations are further exasperated by the downlink and uplink multiuser scheduling that dynamically activates and/or terminates transmissions. Individual transmissions may be short because the scheduling decisions are performed on a millisecond basis.

In order to address the above problem, in the case of conventional, i.e., uncompressed I/Q transport, high sample resolution is applied. Typically, LTE samples are transported using 15-bit resolution per each complex component (as in CPRI [9]).

Lowering the sample resolution would lead to correspondingly lower transport data rates. In order to achieve this, while maintaining the ability to transport a signal with high dynamic range, our proposal applies a fast digital automatic gain control (AGC). It is implemented as a block scaling function. Block scaling is also known as block floating-point where for a block of N_s I/Q samples, a scaling factor is determined such that the subsequent quantization error is minimized.

After the block scaling, I/Q samples are quantized using a quantizer with Q_q-bit resolution per each complex component. The scaling factor is sent once per block, adding to the transport data rates. The scaling factor and quantized I/Q samples may be organized and transported as given in Figure 3, where Q_s bits are used to represent the scaling factor. However, due to minimized quantization error, a lower sample resolution is applied than in the uncompressed case, resulting in overall lowering of the transport data rates.

In each block of N_s samples, a sample with the largest absolute value is determined as

$$A(k) = \max_{i=N_s k,\cdots,N_s(k+1)-1} \{|\Re(s_d(i))|, |\Im(s_d(i))|\} \quad (7)$$

where the integer k denotes the block index. The corresponding scaling factor is determined as

$$S(k) = \begin{cases} \lceil A(k) \rceil & \text{for } \lceil A(k) \rceil \leq 2^{Q_s} - 1 \\ 2^{Q_s} - 1 & \text{for } \lceil A(k) \rceil > 2^{Q_s} - 1 \end{cases} . \quad (8)$$

The above scaling factor is an integer and it does not exceed $2^{Q_s} - 1$. This is done so that the scaling factor is quantized to Q_s bits. This is required because the scaling factor is also transported and is only allocated Q_s bits per one block of samples (as depicted in Figure 3). Each sample in the block is then scaled as

$$s_s(i) = s_d(i)\frac{2^{Q_q-1} - 1}{S(k)} \quad (9)$$

for $i = N_s k, \cdots, N_s(k+1) - 1$.

For example, in the 10 MHz LTE case, we have implemented the block scaling that is performed on a 32-sample block. Due to the block size $N_s = 32$, the latency incurred by the block scaling is 3.125 usec, for $f_{ds} = 10.24$ MHz. The particular block size is selected to capture fast signal variations due to multiuser scheduling decisions and channel variations. In this example, $N_s = 32$ samples corresponds to a fraction of the LTE symbol duration. A different block size may be selected for a particular implementation platform and wireless technology.

C. Quantization

After the block scaling, I/Q samples are quantized using a quantizer with Q_q-bit resolution per each complex component. This function is performed sample-by-sample.

A simple linear (i.e., uniform) quantizer may be applied. However, the application of a quantizer with optimized distances between the quantization levels will result in lower quantization error, and improved signal quality. The distances between quantization levels are not necessarily equal, therefore it is denoted as non-linear (i.e., non-uniform) quantizer. The

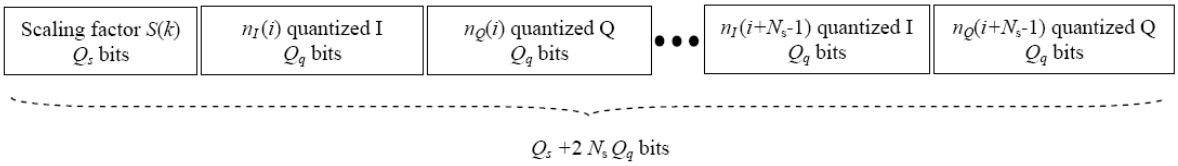

Fig. 3. Possible frame arrangement of quantized scaling factor and I/Q samples.

following off-line adaptive procedure is used to determine the quantization levels.

Initially, the quantization levels are uniformly distributed in the range $[-(2^{Q_q-1}-1),\cdots,2^{Q_q-1}-1]$, centered at zero, with total of 2^{Q_q} levels. The levels are denoted as $q_n(m)$, $n=1,\cdots,2^{Q_q}$ where m is the iteration index. A signal with zero-mean and unit-variance Gaussian distribution is a subject of the above block scaling. The output of the block scaling is denoted as $s_{sg}(m)$. The index of the closest quantization level with respect to $s_{sq}(m)$ is determine as

$$n_{min}(m) = \arg\min_n |q_n(m) - s_{sg}(m)| \tag{10}$$

where $q_n(m)$ is the n-th quantization level at iteration m. The closest quantization level is then adapted as

$$q_{n_{min}}(m+1) = q_{n_{min}}(m) - \mu_q(q_{n_{min}}(m) - s_{sg}(m)) \tag{11}$$

where μ_q is the adaptation coefficient. The above iterative procedure belongs to a broad set of gradient algorithms [11], [12]. It adaptively minimizes the mean square error between the quantization level selected in (10) and $s_{sq}(m)$.

Using the quantization levels obtained from the above off-line procedure, the output of the block scaling is quantized as

$$\begin{aligned} n_I(i) &= \arg\min_n |q_n - \Re(s_s(i))|, \\ n_Q(i) &= \arg\min_n |q_n - \Im(s_s(i))| \end{aligned} \tag{12}$$

and where $s_s(i)$ is given in (9). The quantization level indices $n_I(i)$ and $n_Q(i)$ may be transported as given in Figure 3.

Higher resolution will improve the signal quality (i.e., lower quantization error), while increasing the transport data rates. Therefore, the resolution Q_q is a design parameter derived from the trade-off analysis between the required signal quality and the desired data rate.

Considering both the block scaling and quantization, the average number of bits used to transport a complex I/Q sample is

$$Q = \frac{2N_sQ_q + Q_s}{N_s} = 2Q_q + \frac{Q_s}{N_s} \tag{13}$$

where the second term corresponds to the block scaling factor contribution.

At the decompression side the inverse operations are performed in the following order: (i) dequantization, (ii) block rescaling, and (iii) resampling to the original sampling rate. In order to quantify the signal quality after the dequantization and block rescaling, we measure the signal to quantization noise ratio ($SQNR$) defined as

$$SQNR = \frac{E|s_d(i)|^2}{E|\bar{s}_d(i) - s_d(i)|^2} \tag{14}$$

where $s_d(i)$ is the block scaling input in (6) and $\bar{s}_d(i)$ is the output of the block rescaling.

III. Transmission over Multiple Links

In certain transport networks there may be multiple physical and/or logic transport links between each RRH and BBU. Those links may be assigned different quality of service (QoS) attributes. For example, modern packet-based networks have the QoS mechanism such that each packet may be assigned a guaranteed maximum latency and minimum data rate according to a QoS class it is associated with [13]. Alternatively, in wireless mesh networks [14] multiple links may be used for I/Q transport, each link supporting specific data rates and latency. In order to exploit the above network architecture and improve I/Q transport we propose two solutions: (i) successive transmissions of quantization errors, and (ii) multiple transmissions of dithered signals.

A. Successive Transmissions of Quantization Errors

Let us assume that there are M possible links, each associated with a unique QoS class. For example, link 1 has the lowest guaranteed latency, link 2 the second lowest guaranteed latency and so on. The I/Q compression for link 1 is performed as previously described in Section II. Locally, at the compression side, decompression is performed. $\bar{s}_{d_1}(i)$ denotes the output of the local dequantization and block rescaling. It corresponds to the original signal $s_d(i) = s_{d_1}(i)$, as in (6), where the subscript 1 denotes link 1. The link 1 quantization error is

$$e_{q_1}(i) = s_{d_1}(i) - \bar{s}_{d_1}(i). \tag{15}$$

For link 2, the above error for link 1 is compressed and sent over link 2, i.e., $s_{d_2}(i) = e_{q_1}(i)$. In Figure 4 the proposed solution is depicted for two links. In general, the quantization error for link m is compressed and sent over link $m+1$, i.e.,

$$s_{d_{m+1}}(i) = e_{q_m}(i) \tag{16}$$

for $m = 1,\cdots,M-1$.

Note that for each link the block scaling and quantization are performed individually, while the removal of the redundancies in spectral domain is performed once (as in Section II), prior to per-link processing. Per link, the resolution Q_m is selected such that the transmission data rate matches the assigned QoS data rate for that particular link m.

At the decompression side the dequantization and block rescaling is performed for each link, and the outputs are denoted as $\bar{s}_{d_1}(i),\cdots,\bar{s}_{d_M}(i)$. Assuming successful reception for each link, the results are summed up as

$$\bar{s_d}^*(i) = \sum_{m=1}^{M} \bar{s}_{d_m}(i) \tag{17}$$

representing a composite output of the multi-link compressed I/Q transmission.

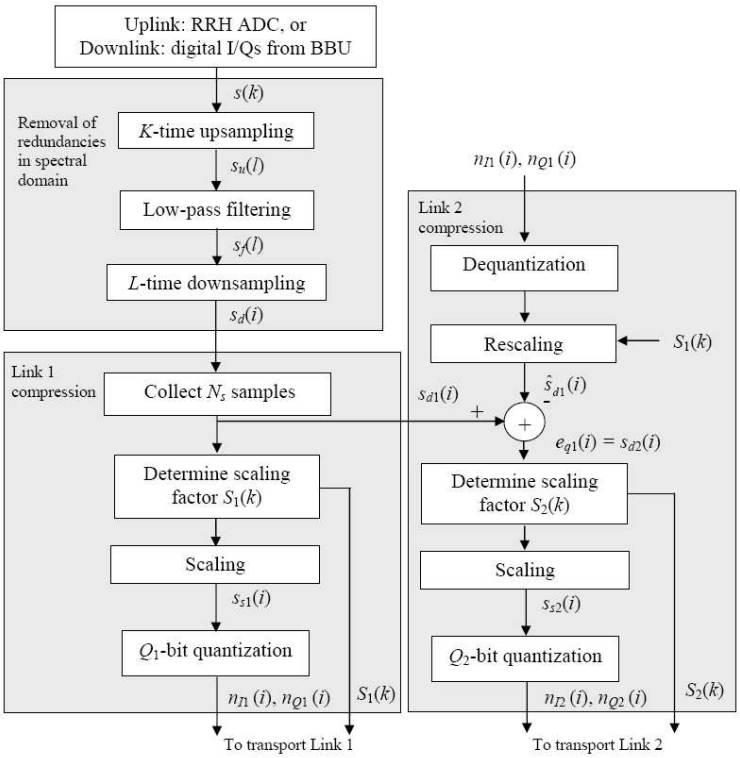

Fig. 4. Successive transmission of quantization errors over multiple transport links.

In general, for certain links the transmission may fail (e.g., packet transmission exceeding the maximum allowed latency). The links with less stringent QoS attributes (e.g, longer guaranteed maximum latency) will have a higher probability of failure. In the case when link L_M+1 has not been received correctly, the above summation is performed for the first L_M links as

$$\tilde{s}_d^*(i) = \sum_{m=1}^{L_M} \bar{s}_{d_m}(i). \tag{18}$$

In this case, only the first L_M links contribute to the decompressed signal.

The above solution exploits multiple links, taking into account their individual QoS attributes to improve the overall signal quality. Successful transmissions over each successive link incrementally contributes to improving the quality of the composite decompressed signal, which is expressed by the following proposition.

Proposition 1: Successful transmission over L_M successive links results in a signal quality that is equal to the signal quality of a single-link transmission with the resolution

$$Q^* = \sum_{m=1}^{L_M} Q_m \tag{19}$$

where Q_m is the number of bits per I/Q sample for link m. SQNR quantifies the signal quality, and it is assumed to increase exponentially with the resolution.

Based on this proposition, each successful successive link transmissions will exponentially improve the signal quality depending on its resolution. The corresponding proof is presented in Appendix, including discussions on exponential dependency between the resolution and signal quality.

B. Multiple Transmissions of Dithered Signals

Due to transmission of successive quantization errors the above scheme may be viewed as a differential transmission. Consequently, if a transmission over link $L_M + 1$ fails, the successful transmissions over subsequent links ($L_M+2, L_M+3, \cdots, M$) will not contribute to improving the decompressed signal quality (as expressed in (18)). As an alternative that addresses this particular problem, instead of applying successive transmissions of quantization errors, dithered versions of the same signal may be transmitted. Basic aspects of dithering are analyzed in [15].

Let us assume that there are M possible links. For each link, after the block scaling that is described in Subsection II-B, dithering is performed by adding a pseudo noise to the quantizer input as

$$s_{s_m}^d(i) = s_s(i) + p_m(i) \tag{20}$$

where $p_m(i)$ is a pseudo noise with the subscript m denoting the link index ($m = 1, \cdots, M$). The above signal is then quantized as described in Subsection II-C. Different instantiations of the pseudo noise are used for different links. In addition, the variance of the pseudo noise depends on the particular quantizer resolution Q_{q_m} ($m = 1, \cdots, M$). The goal of the above dithering is to produce independent quantization noise between links $m = 1, \cdots, M$.

Note that for each link the dithering and quantization are performed individually, while the block scaling and removal of redundancies in spectral domain are performed once (as in Section II), prior to per-link processing.

Per each link, the added pseudo noise is also known at the

decompression side. After the dequantization, it is removed as

$$\bar{s}_{s_m}(i) = \bar{s}^d_{s_m}(i) - p_m(i) \tag{21}$$

where $\bar{s}^d_{s_m}(i)$ is the output of the link m dequantizer. The composite dequantized signal is

$$\bar{s}^*_s(i) = \frac{\sum_m \bar{s}_{s_m}(i)}{L_D} \tag{22}$$

where the summation is performed only for the links with successful transmission, and L_D is the number of those links ($L_D \leq M$). A successful transmission over each link incrementally contributes to improving the quality of the composite decompressed signal, which is expressed by the following proposition.

Proposition 2: Assuming independent quantization noise between the links, successful transmission over L_D links results in the composite signal $SQNR$

$$SQNR^*_{L_D} \leq L_D \, SQNR_1 \tag{23}$$

where $SQNR_1$ corresponds to link 1, which is assumed to have the highest resolution. The equality holds in the case when each link has equal resolution, i.e., data rate.

The corresponding proof is presented in Appendix. Based on the above proposition, for the multiple transmissions of dithered signals the signal quality improves linearly with the number of *any* successful link transmissions (e.g., losing a transmission over one link will not affect usefulness of other links). In the case of successive transmissions of quantization errors, as described in the previous subsection, the signal quality improves exponentially. However, the links must be successfully received in the successive order (e.g., loss of a transmission over link $L_M + 1$, will render other successful links useless ($L_M + 2, \cdots, M$)).

IV. Numerical and Experimental Results

In this section we first numerically investigate performance of the block scaling and quantization that are described in Subsections II-B and II-C, respectively. In order to assess the performance and compare against the idealized quantizer, we consider independent identically distributed (*i.i.d.*) input samples $s_d(i)$ with complex Gaussian distribution $\mathcal{N}_{\mathcal{C}}(0, P_s)$[1]. For the given resolution Q in (13) the performance is quantified in terms of $SQNR$, and compared against its idealized upper bound. As discussed in Appendix, the $SQNR$ upper bound is $SQNR_{ub} = 2^Q$ [16]. In Figure 5 $SQNR$ is presented as a function of the resolution Q. Both the proposed non-linear and liner quantizer are considered. The non-linear quantizer has a clear advantage over the linear one for lower resolutions (e.g, for $Q_q = 3$, i.e., $Q = 6.5$ bits the difference between the two is 1.6 dB). The difference is diminishing with higher resolution. For example, for $Q_q > 7$, i.e., $Q > 14.5$ bits the performance between the two quantizers is practically indistinguishable. In further analysis we apply the non-linear quantizer. Furthermore, $SQNR$ increases exponentially with Q (i.e., linearly in the logarithmic domain). We note that the proposed practical block scaling and non-linear quantization

[1]The *i.i.d.* assumption corresponds to a case when there is no redundancy in the spectral domain, with a constant, i.e., flat power spectrum density.

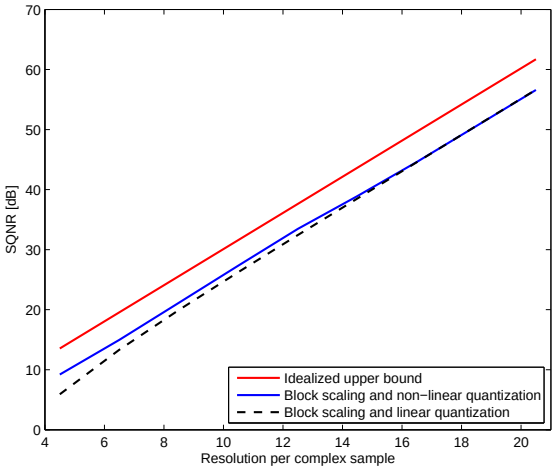

Fig. 5. $SQNR$ as a function of the resolution Q, for *i.i.d.* complex Gaussian distribution.

performance is approximately 4.5 dB lower than the upper bound $SQNR_{ub}$. However, in order to achieve $SQNR_{ub}$, the idealized vector quantization would have to be applied, incurring infinite processing delay [16].

We now consider I/Q transmission over multiple links, as described in Section III. Both solutions have been considered: (i) successive transmissions of quantization errors, and (ii) multiple transmissions of dithered signals. In Figure 6, $SQNR$ is presented as a function of the number of links M. In general, each link may support different data rates. However, in this particular example, each link has an identical data rate supporting resolution $Q_q = 3$ bits, $Q_s = 16$ bits, and $N_s = 32$, resulting in $Q = 6.5$ bits per complex I/Q sample, per link. Furthermore, transmission over each link is assumed to be error-free. The upper bound corresponds to the previously described idealized quantizer being applied to each link, resulting in the aggregate multi-link performance that is described in Proposition 1 and Appendix. In the idealized case and for successive transmissions of quantization errors, $SQNR$ increases exponentially with the number of links M (i.e., linearly in the logarithmic domain). However, in this error-free example, for multiple transmissions of dithered signals, the increase in $SQNR$ is significantly slower, resulting in lower $SQNR$.

To illustrate the multi-link performance in the presence of transmission errors, we expand the above example, setting the number of links to $M = 4$. In general, each link may have different transmission error probability. However, in this example it is assumed that each link has equal probability of transmission errors (i.e., $P_e = P_{e1} = P_{e2} = \cdots = P_{eM}$). The errors are uncorrelated in time and between the links. If there is a transmission error in link m ($m = 1, \cdots, M$) the data transmitted over that particular link will not be available during the decompression, thus adversely affecting the performance. In Figure 7, $SQNR$ is presented as a function of the transmission error probability P_e. From the results we note that successive transmissions of quantization errors are very sensitive to the transmission errors, i.e., $SQNR$ is decreasing rapidly with P_e. On the other hand, multiple transmissions of

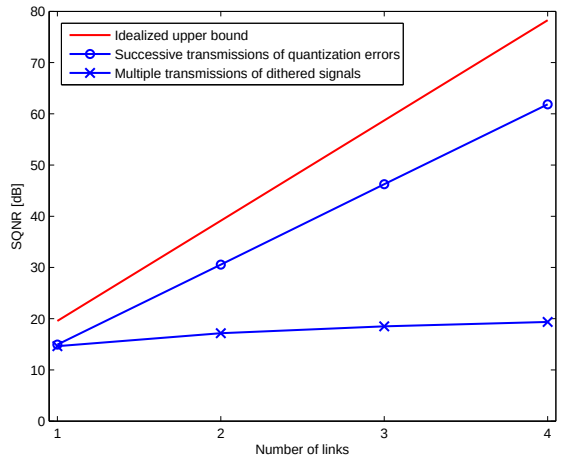

Fig. 6. $SQNR$ as a function of the number of transport links M, error-free transmission, for *i.i.d.* complex Gaussian distribution.

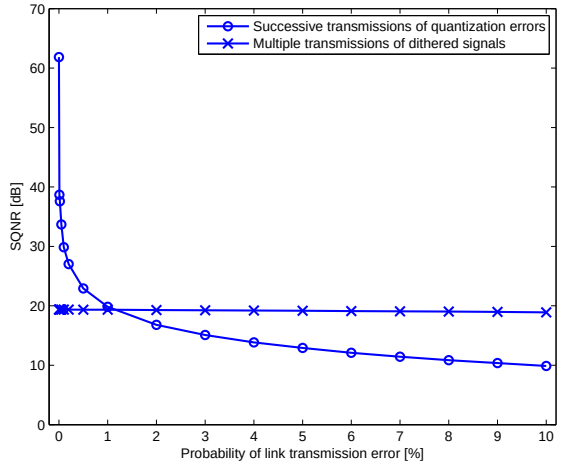

Fig. 7. $SQNR$ as a function of the probability of transmission error, for *i.i.d.* complex Gaussian distribution.

dithered signals are very robust, i.e., the decrease in $SQNR$ is imperceptible for the P_e range in Figure 7. In this particular example for $P_e \geq 1.2\%$, multiple transmissions of dithered signals outperform successive transmissions of quantization errors.

Superior performance by multiple transmissions of dithered signals in the presence of transmission errors may be explained as a consequence of its ability to exploit multi-link diversity. On the other hand, exceptional performance of successive transmissions of quantization errors in the error-free case may be explained as a consequence of its ability to exploit multi-link multiplexing. Additional analysis will be needed to establish a more rigorous relationship between the diversity and multiplexing aspects of the proposed multi-link solutions. For example, approach in [17] may be used as a basis for the future work.

When choosing between the two proposed multi-link transmissions schemes, both the particular link data rates and error probabilities need to be considered. The above numerical analysis may be used to decide on the application of the most suitable transmission scheme.

TABLE I
DOWNLINK LTE REQUIREMENTS [21].

Modulation Scheme	Maximum $EVM_{\%}$
QPSK	17.5
16-QAM	12.5
64-QAM	8

A. LTE Experimental Results

In this subsection we present experimental performance evaluations of the proposed single-link I/Q compression scheme applied to LTE. The performance is quantified in terms of error vector magnitude (EVM) and adjacent carrier leakage power ratio ($ACLR$). Both quantities are well-established figures of merit, widely applied in wireless industry [18]–[20]. In this study EVM is used to quantify in-band distortions introduced by the compression scheme, and is defined as

$$EVM_{\%} = \sqrt{\frac{E[|\bar{x} - x|^2]}{E[|x|^2]}}\, 100\ [\%] \tag{24}$$

where $\bar{x}$ is the output signal (after the compression and decompression have been performed), while x is its idealized noise-free version. Higher values of EVM correspond to higher levels of in-band distortions. EVM directly relates to the maximum achievable SNR, i.e., SNR ceiling as

$$SNR_{ceiling} = -20 \log EVM\ \ [\text{dB}]. \tag{25}$$

The maximum level of in-band distortion that a base station or a mobile terminal may introduce is typically specified by a particular standard. For example, according to the LTE 3GPP specification [21], the maximum downlink EVM requirements are given in Table I.

In this study $ACLR$ is used to quantify out-of-band distortions introduced by the compression scheme affecting adjacent frequency channels. It is defined as

$$ACLR = 10 \log \frac{P(f_c)}{P(f_c + \Delta f)}\ \ [\text{dB}] \tag{26}$$

where $P(f_c)$ and $P(f_c + \Delta f)$ are the signal power densities (after the compression and decompression have been performed) at the assigned and adjacent channel frequencies f_c and $f_c + \Delta f$, respectively. For example, in LTE the required minimum downlink $ACLR$ is 45 dB.

To generate test LTE signals and evaluate the performance we apply independent third-party software and test equipment. A wide variety of downlink and uplink LTE signal arrangements have been synthesized using [22]. After the compression and decompression, the output is generated using [23]. The signal quality is analyzed using [24]. The signal analyzer performs physical-layer LTE reception, including synchronization, channel estimation and estimation of in-band and out-of-band distortions. The in-band distortions are estimated for reference signals, resource blocks with QPSK, 16-QAM and 64-QAM, individually. Both the transmission and reception are performed in real-time, at f_c = 763 MHz.

In Figure 8 we present experimentally measured EVM for 10 MHz downlink LTE, per antenna[2]. In this example, we

[2]Each antenna is processed individually, incrementally increasing the transport data rates.

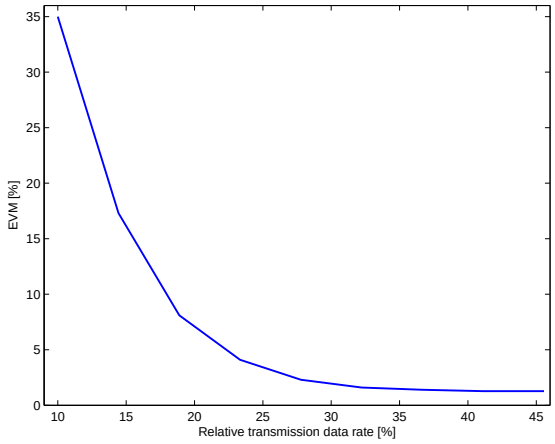

Fig. 8. EVM as a function of the relative transmission data rate with respect to the uncompressed rate of 460.8 MBPS, 10 MHz downlink LTE, per antenna.

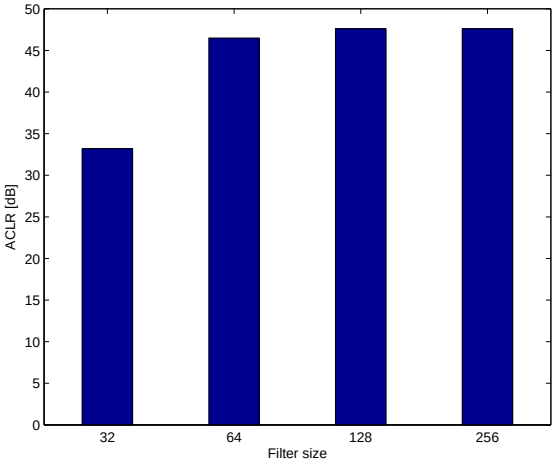

Fig. 9. $ACLR$ as a function of the multi-rate filter size, 10 MHz downlink LTE.

implemented the multi-rate filter with $L = 3$, $K = 2$ and $N_w = 64$. The scaling factor resolution is set to $Q_s = 16$ bits, while the I/Q sample resolution Q_q is changing according the the presented data rates (from $Q_q = 2$ to 10 bits, corresponding to the lowest and the highest data rate, respectively). The downlink scheduling decisions are performed once every 2 msec. The modulation schemes are randomly assigned to each resource block. The scheduler may also randomly decide not to use a particular resource block, lowering overall transmission power, and consequently increasing the signal dynamic range. We believe that this dynamic signal arrangement corresponds to a realistic case of downlink LTE transmission. We note that for the relative rate of 32.22% (i.e., 148.48 MBPS) and higher, the proposed compression scheme will introduce negligible in-band distortion ($EVM = 1.5\%$, which is significantly better than the required maximum $EVM_{64-QAM} = 8\%$). Furthermore, for the relative rate of 27.78% (i.e., 128 MBPS) and higher, $ACLR$ is 45 dB or better, meeting the LTE $ACLR$ requirements.

Out-of-band distortions are also affected by the multi-rate filter size. To investigate this effect in Figure 9 we present experimentally measured $ACLR$ as a function of the filter size N_w. In this example, the data rate is set to 148.48 MBPS ($N_s = 32$ samples, $Q_s = 16$ bits, $Q_q = 7$ bits). Note that the smallest filter size $N_w = 32$ will not meet the LTE $ACLR$ requirements, while for $N_w = 64$ and larger, the requirement is achieved and exceeded.

Based on the above measurement results, the transport data rates of approximately 150 MBPS and the filter size $N_w = 64$ will guarantee very low levels of in-band and out-band distortion comfortably meeting the LTE signal-quality requirement. In this case, the proposed I/Q compression scheme will provided better than threefold data rate reduction compared to the uncompressed CPRI transmission. Furthermore, for the given parameters the minimal algorithmic compression and decompression latency is 8.33 usec, introduced by the multi-rate filtering and block scaling. In our real-time FPGA implementation, additional processing latency is less than 3 usec. This overall latency is significantly lower than the available LTE latency budget. For example, the uplink latency budget is 4 msec, easily allowing for the application of the proposed compression scheme [3].

Equivalent uplink measurements have been performed with two test user terminals (UEs). At this stage, the results and conclusions are identical as in the above downlink case. However, more specific uplink measurement scenarios will have to be addressed. Namely, uplink channel variations due to mobility and large differences in the uplink signal power between multiple UEs will be considered.

V. Applicability and Conclusions

The application of the proposed I/Q compression leads to a number of benefits that are listed below.

- Significantly lower transport data rate, resulting in a lower-cost transport network.
- Low processing latency, which is critical for many advanced coordinated transmission and reception techniques.
- Ability to exploit multiple links with different QoS attributes, which may be particularly beneficial in packet-based and wireless mesh transport networks.
- May be applied to different wireless technologies, with appropriate parameter settings, while keeping the architecture identical, i.e., technology and implementation agnostic.

The above characteristics will lead to a cost-effective implementation of a number of novel radio access network solutions. Those solutions are briefly addressed in the following.

Collocated network-centric baseband is a wireless base station architecture which relies on multiple RRHs being connected to baseband processing resources at a centralized location via a low latency, high bandwidth transport network. In addition, in distributed network-centric baseband, processing is distributed across multiple physical locations [1]. Those architectures are expected to gain importance in light of the

[3]The LTE uplink latency is limited by the hybrid automatic repeat request (HARQ) mechanism.

general trend in the industry to move to network-based computing. Ability to perform cost-effective radio access network (RAN) processing in the 'network cloud' will greatly depend on the required transport rates as well as associated latencies. Therefore, the proposed compression scheme is viewed as a key solution enabling implementation of the above novel concepts and architectures.

Coordinated multi-point (CoMP) is a set of novel inter/intra base station coordination techniques, resulting in significant improvements in wireless data rates. For example, under idealized conditions coherent joint processing (JP) CoMP transmission is proven to provide significant mean and cell-edge data rate gains [2], [25]. Assuming perfect channel state information (CSI), the uplink mean and cell-edge data rate gains are approximately 100% and 200%, respectively. Different versions of CoMP are expected to be standardized under LTE-Advanced. However, stringent requirements are imposed on channel state information (CSI) availability, frequency stability, transport network throughput and latency [26]. The proposed I/Q compression scheme enables efficient implementation of coherent JP CoMP, realizing low latency, and efficient usage of transport network resources.

APPENDIX
PROOF OF PROPOSITION 1

In this appendix we provide a proof of Proposition 1. First we assume that the link m SQNR is an exponential function of the resolution Q_m as

$$SQNR_m = \frac{E|s_{d_m}(i)|^2}{E|e_{q_m}(i)|^2} = \Gamma^{Q_m} \tag{27}$$

where Γ is a positive constant and $m = 1, \cdots, M$. Consequently, the power of quantization noise for link 1 is

$$E|e_{q_1}(i)|^2 = \frac{E|s_d(i)|^2}{\Gamma^{Q_1}}. \tag{28}$$

The quantization noise power for link 2 is

$$E|e_{q_2}(i)|^2 = \frac{E|s_{d_2}(i)|^2}{\Gamma^{Q_2}} = \frac{E|e_{q_1}(i)|^2}{\Gamma^{Q_2}} = \frac{E|s_d(i)|^2}{\Gamma^{Q_1+Q_2}}. \tag{29}$$

Based on the above, for link L_M the quantization noise is

$$E|e_{q_{L_M}}(i)|^2 = \frac{E|s_d(i)|^2}{\Gamma^{Q_1+Q_2+\cdots+Q_{L_M}}}. \tag{30}$$

As given in (15), the link 1 dequantization and block rescaling output is

$$\bar{s}_{d_1}(i) = s_d(i) - e_{q_1}(i), \tag{31}$$

while the corresponding link 2 output is

$$\bar{s}_{d_2}(i) = e_{q_1}(i) - e_{q_2}(i). \tag{32}$$

Combining the link 1 and link 2 outputs yields

$$\bar{s}_{d_1}(i)+\bar{s}_{d_2}(i) = s_d(i)-e_{q_1}(i)+e_{q_1}(i)-e_{q_2}(i) = s_d(i)-e_{q_2}(i). \tag{33}$$

Consequently, the summation of the link 1 to link L_M outputs (as in (18)) results in

$$\tilde{s}^*_d(i) = \sum_{m=1}^{L_M} \bar{s}_{d_m}(i) = s_d(i) - e_{q_{L_M}}(i). \tag{34}$$

$SQNR$ for the above composite signal is

$$\tilde{s}^*_d(i) = \frac{E|s_d(i)|^2}{E|e_{q_{L_M}}(i)|^2} = \Gamma^{Q_1+Q_2+\cdots+Q_{L_M}}. \tag{35}$$

The above $SQNR$ also corresponds to a single-link quantizer with the resolution $Q^* = \sum_{m=1}^{L_M} Q_m$ in (19), which concludes the proof of Proposition 1.

In order to address the initial assumption that $SQNR$ increases exponentially with the resolution Q, we consider the following idealization. According to rate distortion theory, for an infinite sequence of independent and identically distributed samples, with complex Gaussian distribution $\mathcal{N}_{\mathcal{C}}(0, P_s)$ there is a lower bound on the quantizer resolution Q_{min} for the given distortion D [16], [27]. The bound is given as

$$Q_{min} = \log_2\left(\frac{P_s}{D}\right). \tag{36}$$

The idealized quantizer is achieving the above bound by performing vector quantization over infinitely-long sequence of samples. The distortion D is the mean square error between the original samples and the output of quantization and dequantization. It corresponds to the power of the quantization noise. Based on the above expression, there is a clear exponential dependency between $SQNR$ and resolution given as

$$SQNR = \frac{P_s}{D} = 2^{Q_{min}}. \tag{37}$$

Therefore, we conclude that Proposition 1 is valid if the above idealized quantization is used in conjunction with successive transmissions of quantization errors.

APPENDIX
PROOF OF PROPOSITION 2

In this appendix we provide a proof of Proposition 2.

After the removal of the dithering signal in (21), we assume that the quantization error, i.e., noise for each link is independent. Namely,

$$E|e_{q_m}(i)e_{q_n}(i)| = 0, \text{ for } m \neq n \tag{38}$$

where m and n are link indices. The averaging in (22) results in

$$\bar{s}^*_d(i) = s_d(i) - \frac{\sum_m e_{q_m}(i)}{L_D} \tag{39}$$

where the summation is performed for L_D links. Assuming that link 1 supports the highest data rate and consequently the highest resolution Q_1, the power of link 1 quantization noise is the lowest compared to other links. Therefore, the power of the composite noise is

$$\frac{\sum_m E|e_{q_m}(i)|^2}{L_D^2} \geq \frac{E|e_{q_1}(i)|^2}{L_D} \tag{40}$$

where the equality holds in the case when each link has equal resolution $Q_1 = Q_2 = \cdots = Q_M$, and thus equal power of the quantization noise. Using the above inequality, $SQNR$ for the composite signal for multiple transmissions of dithered signals is

$$L_D^2 \frac{E|s_d(i)|^2}{\sum_m E|e_{q_m}(i)|^2} \leq L_D \frac{E|s_d(i)|^2}{E|e_{q_1}(i)|^2}, \tag{41}$$

thus

$$SQNR^*_{L_D} \leq L_D \, SQNR_{L_1}, \tag{42}$$

which concludes the proof of Proposition 2.

REFERENCES

[1] "C-RAN: the road towards green RAN," China Mobile Research Institute, Apr. 2010, Technical White Paper.

[2] M. K. Karakayali, G. J. Foschini, and R. A. Valenzuela, "Network coordination for spectrally efficient communications in cellular systems," *IEEE Trans. Wireless Commun.*, vol. 13, pp. 56–61, Aug. 2006.

[3] "Feasibility Study for Further Advancements for E-UTRA (LTE-Advanced)," 3GPP TR 36.912, 2009.

[4] M. Sawahashi, Y. Kishiyama, A. Morimoto, and M. T. D. Nishikawa, "Coordinated multipoint transmission/reception techniques for LTE-advanced," *IEEE Trans. Wireless Commun.*, June 2010.

[5] A. Ghosh, R. Ratasuk, B. Mondal, N. Mangalvedhe, and T. Thomas, "LTE Advanced: next generation wireless broadband technology," *IEEE Trans. Wireless Commun.*, vol. 17, pp. 10–22, June 2010.

[6] A. A. M. Saleh, A. J. Rustako, and R. S. Roman, "Distributed antennas for indoor radio communications," *IEEE Trans. Commun.*, vol. 35, p. 12451251, Dec. 1987.

[7] H. Hu, Y. Zhang, and E. J. Luo, *Distributed Antenna Systems: Open Architecture for Future Wireless Communications*. CRC Press, May 2007.

[8] Z. Ma, M. Zierdt, J. Pastalan, A. Siegel, T. Sizer, A. J. Wijngaarden, P. R. Kasireddy, and D. Samardzija, "RADIOSTAR: providing wireless coverage over gigabit Ethernet," *Bell Labs Technical J.*, vol. 14, pp. 7–14, 2009.

[9] "CPRI Specification," V 4.2, Nov. 2010.

[10] E. B. Hogenauer, "An economical class of digital filters for decimation and interpolation," *IEEE Trans. Acoustics, Speech, and Signal Process.*, vol. 29, pp. 155–162, Apr. 1981.

[11] D. Bertsekas, *Constrained Optimization and Lagrange Multiplier Methods*. Academic Press, 1982.

[12] S. Haykin, *Adaptive Filter Theory*, 2nd edition. Prentice-Hall, 1991.

[13] Z. Wang, *Internet QoS: Architectures and Mechanisms for Quality of Service*. Morgan Kaufman Publishers, 2001.

[14] H. Viswanathan and S. Mukherjee, "Throughput-range tradeoff of wireless mesh backhaul networks," *IEEE J. Sel. Areas Commun.*, vol. 24, no. 3, pp. 593–602, 2006.

[15] L. Schuchman, "Dither signals and their effect on quantization noise," *IEEE Trans. Commun.*, vol. 12, pp. 162–165, Dec. 1964.

[16] T. Berger, *Rate Distortion Theory: A Mathematical Basis for Data Compression*. Prentice-Hall, 1971.

[17] D. Tse, P. Viswanath, and L. Zheng, "Diversity-multiplexing tradeoff in multiple access channels," *IEEE Trans. Inf. Theory*, vol. 50, pp. 1859–74, Sep. 2004.

[18] Agilent Product Note 89400-8: Using Vector Modulation Analysis in the Integration, Troubleshooting, and Design of Digital RF Communications Systems.

[19] R. Hassun, M. Flaherty, R. Matreci, and M. Taylor, "Effective evaluation of link quality using error vector magnitude technique," in *Proc. 1997 IEEE Wireless Communications Conference*, pp. 89–94.

[20] M. Heutmaker, "The error vector and power amplifier distortion," in *Proc. 1997 IEEE Wireless Communications Conference*, pp. 100–104.

[21] "Base Station (BS) radio transmission and reception," 3GPP TS 36.104, Apr. 2011.

[22] Agilent Signal Studio for 3GPP LTE FDD, No. version 7.9.3.0, Apr. 2010.

[23] Agilent MXG Vector Signal Generator (100 KHz - 3 GHz).

[24] Agilent N9030A PXA Signal Analyzer with LTE Software 89600 Vector Signal Analyzer, No. version 11.20 (3 Hz - 8.4 GHz).

[25] D. Samardzija and A. Domazetovic, "Coherent joint-processing CoMP in pico-cellular lamp-post street deployment," *2011 IEEE International Workshop on Signal Processing Advances in Wireless Communications*.

[26] H. Huang and D. Samardzija, "Determining backhaul bandwidth requirements for Network MIMO," *2009 European Signal Processing Conference*.

[27] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 1st edition. Wiley Publications, 1991.

Dragan Samardzija received his B.S. degree in electrical engineering and computer science from the University of Novi Sad, Serbia, in 1996 and M.S and Ph.D. degrees in electrical engineering from Wireless Information Network Laboratory (WINLAB), Rutgers University, USA, in 2000 and 2004, respectively. Since 2000 he has been with Bell Laboratories, Alcatel-Lucent, where he is involved in research in the field of the next generation wireless systems. He is also teaching at the University of Novi Sad. He authored over 50 journal and conference publications. As a member of the BLAST project team, he is a recipient of the 2002 Bell Labs President's Award. He received the Central Bell Labs Teamwork Award for the HSDPA Demonstration Team, 2003.

John Pastalan is a member of technical staff in Bell Labs' Wireless Communication Research Department in Murray Hill, New Jersey. He received a B.S. degree in physics and a M.S. degree in applied physics from the State University of New York at Binghamton, His research interests include digital design prototyping and wireless networking infrastructure architecture. He holds three patents and has contributed to numerous technical publications.

Michael MacDonald received his B.S.degree from the University of Notre Dame in physics in 1979, M.S. (1980), and Ph.D. (1984) from the University of Illinois at Urbana-Champaign in physics. He joined the research staff at Bell Laboratories in 1984. He is involved in research on distributed antenna systems and localization in wireless networks.

Susan Walker received a BS degree in engineering from Rutgers University, New Brunswick, NJ, USA in 1984 and a M.S degree in material science and engineering from Stevens Institute of Technology, Hoboken, NJ, USA in 1988. She began her career in 1984 when she joined Bell Laboratories Monolithic Optics Department, Murray Hill, NJ, USA as a Technical Assistant working on epitaxial growth of laser structures. After many years in optics research, she joined the Wireless Communication Research Department, where she now works as a Member of Technical Staff at Alcatel-Lucent's facility in Holmdel, NJ. USA. Mrs. Walker is the recipient of the 2003 Bell labs President's Gold Award, author of over 25 publications, and inventor of five patents.

Reinaldo A. Valenzuela received a B.Sc. degree from the University of Chile, and a Ph.D. degree from Imperial College, London. He is an IEEE Fellow. In 2010 he received the IEEE Eric E. Sumner Award. He is a director of the Wireless Communications Research Department, and a Distinguished Member of Technical Staff, Bell Laboratories. He has been engaged in MIMO and space-time systems achieving high capacities using transmit and receive antenna arrays. He has published over 130 papers and 12 patents. He has over 15000 Google Scholar citations and is a 'Highly Cited Author' in the Thomson ISI and a Fulbright Senior Specialist.