IEEE Open Journal of
Circuits and Systems

Received 21 April 2023; revised 16 June 2023; accepted 18 July 2023. Date of publication 26 July 2023; date of current version 11 August 2023.

Digital Object Identifier 10.1109/OJCAS.2023.3299052

Grant-Free Sparse Code Multiple Access for Uplink Massive Machine-Type Communications and Its Real-Time Receiver Design

TI-YU CHEN [1] (Member, IEEE), ZHI-JING LIN [1,2], AND TZI-DAR CHIUEH [1,3] (Fellow, IEEE)

[1] Graduate Institute of Electronics Engineering, National Taiwan University, Taipei City 10617, Taiwan

[2] Department of Silicon Product Development (SPD), Mediatek Inc., Hsinchu 300, Taiwan

[3] Department of Electrical Engineering, National Taiwan University, Taipei City 10617, Taiwan

This article was recommended by Associate Editor C. Studer.

CORRESPONDING AUTHOR: T.-D. CHIUEH (e-mail: chiueh@ntu.edu.tw)

This work was supported in part by the Ministry of Science and Technology, Taiwan, under Grant MOST 109-2221-E-022-154.

ABSTRACT Massive Machine-type Communications (mMTCs) is a major use case for the 5G standard. The grant-free (GF) sparse-coded multiple access (SCMA) transmission is particularly spectrum efficient in the sporadic uplink traffic, which is characteristic of the mMTC networks. In this paper, an uplink GF-SCMA receiver with a user activity detection (UAD) function was designed and implemented. In particular, several new techniques were proposed to enhance the SCMA decoder performance; they include delayed serial update, early stopping, message passing algorithm equation reformulation, distance approximation, and sum circuit sharing. To meet the real-time operation requirements, we implemented key inner receiver function circuits, such as carrier frequency synchronization, user signature detection, channel estimation and compensation, soft SCMA detection, etc. on a Xilinx KCU1500 FPGA chip. Finally, an over-the-air (OTA) prototype has been constructed, demonstrating the efficient and reliable multi-user GF-SCMA uplink transmission of the proposed system.

INDEX TERMS 5G, massive machine type communications (mMTC), user activity detection (UAD), sparse code multiple access (SCMA), inner receiver, real-time, over-the-air (OTA), field programmable gate array (FPGA).

I. INTRODUCTION

MASSIVE Machine Type Communication (mMTC), one of the three major application scenarios in the 5G communication, is characterized by sparse and small data bursts and many connected devices. In this scenario, with decreasing transmitted data and an increasing number of potential users, the traditional grant-based access mechanism results in complicated and inefficient scheduling, more radio resource overhead, and severe network efficiency reduction. Therefore, the Grant-Free (GF) access, which skips the requesting procedure and directly commences data transmission, has been proposed. The advantages of GF include reduced transmission delay, less collision-induced network congestion, and lower power consumption of the connected devices. One of the problems with GF access is that without scheduling, the system cannot allocate resources and thus cannot apply traditional Orthogonal Frequency Division Multiple Access (OFDMA). Toward this end, Non-Orthogonal Multiple Access (NOMA) that allows multiple users to access the same time-frequency resource can be adopted, and GF-NOMA thus becomes a popular transmission technology in mMTC networks.

In an mMTC network with GF access, there is no scheduling mechanism. Moreover, due to transmission sparsity, only a few user equipments (UEs) transmit in each interval, so the base station (BS) must identify active users via user activity detection (UAD). Each user is assigned a pilot sequence known by the BS, and this sequence is usually carried by signals placed on several subcarriers in the frequency domain. UAD tries to detect the existing pilot sequences

corresponding to active UEs from the signals received by the BS. UAD algorithms include FO-CUSS [1], [2], [3], DGOMP [4], [5], SBL [6], [7], [8], and ALEM [9]. ALEM is based on the Expectation-Maximization (EM) technique and is shown to be robust to channel mismatch and variation.

Sparse Code Multiple Access (SCMA), proposed in [10], is one of the most popular NOMA technologies. In an SCMA system, instead of using traditional modulation techniques such as Quadrature Amplitude Modulation (QAM), UEs use their designated codebooks. In SCMA modulation, each UE picks one of the possible codewords in their codebooks and assigns signals, according to the codeword, to spread over a set of Resource Elements (REs). Proper codebook design can increase constellation shaping gain and improve spectral efficiency.

Taking advantage of the codeword sparsity, SCMA decoders can be implemented using the Message Passing Algorithm (MPA), whose performance can approach that of the optimal Maximum-A-Posteriori (MAP) decoding. MPA needs to compute the probabilities of all possible Codeword Combinations (CC) in an SCMA block, resulting in a complexity that scales exponentially with the number of active UEs. Therefore, low-complexity MPA is essential for SCMA decoding. The works in [11], [12], [13] proposed Partial Marginalization MPA (PM-MPA) and its improvements. Other low-complexity MPA algorithms include Dynamic Factor Graph MPA (DFG-MPA) [14] and Codebook Cardinality Reduction-based MPA (CCR-MPA) [15]. The above decoders reduce the MPA complexity by removing some CCs after a specific number of iterations. However, these algorithms may remove the correct CC prematurely, resulting in significant decoding performance degradation. The Max-log-MPA based on Serial and Threshold method [16] removes CCs based on bit log-likelihood ratio (LLR). The low-complexity algorithm has a decoding performance competitive with traditional MPA. The Serial Scheduling MPA (S-MPA) method [17] adjusts the message updating schedule to increase the convergence rate but ends up suffering from high data dependency.

This paper proposes a Delayed Serial updating MPA (DS-MPA) decoder with a performance and convergence rate competitive with the S-MPA-based SCMA decoders. We designed an efficient SCMA decoder architecture based on this algorithm and constructed a real-time SCMA transmission system to verify the efficiency of the proposed SCMA decoder. Practically, an mMTC network should involve a large number of users with sporadic traffic. However, supporting massive UE connections will require lots of hardware resources. So for verification and demonstration purposes, we densified the traffic and implemented only six UEs, and maintain a similar amount of average UE traffic in each time interval. To achieve real-time mMTC transmission, we have implemented all computational intensive signal processing, such as OFDM receiver processing, UAD block, and SCMA decoding, in hardware. Moreover, a companion software program manages the data input and output of the hardware receiver.

In order to validate the feasibility and accuracy of the implemented transmission system, six UEs transmit data streams through Over-the-Air (OTA) channels. The active transmission duration of each device is pre-scheduled subject to the transmission sparsity of the mMTC network. Note that the schedule is only for evaluating the UAD error rate and is unknown to the inner receiver. The received signal is processed and user activity information and LLR of the transmitted bits are generated. While there are several SDR implementations for NOMA [18], [19], to the best of our knowledge, the proposed prototype is the only prototype that demonstrates GF-SCMA. In the experiment, the data processing time is shorter than the signal transmission time, demonstrating the real-time reception capability of the implementation. The SCMA bit error rate is low even when no Forward Error Correction (FEC) code is adopted. Finally, The contributions of this article are summarized as follows:

- A low-data-dependency SCMA decoding algorithm with performance and convergence rate competitive with S-MPA method.
- Development of a high-efficiency SCMA decoder hardware architecture.
- Implementation of a GF-SCMA real-time transmission system.
- OTA validation of the system efficiency and transmission reliability.

The rest of this paper is organized as follows. Section II introduces the signal model and existing MPA algorithms. Section III presents the Delayed Serial updating MPA along with the high-efficiency SCMA decoder architecture. Then, Section IV describes how the real-time GF-SCMA transmission system is implemented. In Section V, we verify the proposed GF-SCMA transmission design by an FPGA-based software-defined radio prototype using an OTA channel. Finally, conclusions are given in Section VI.

Notations: Lowercase boldface and italic letters denote vectors and scalars, respectively. $()^T$ denote transpose. Uppercase boldface and uppercase italic letters denote sets with elements being vectors and scalars, respectively. $diag(\mathbf{x})$ is a diagonal matrix with diagonal terms being the elements of vector $\mathbf{x}$. $\|a\|$ stands for the modulus of complex number a. $()^{Re}$ and $()^{Im}$ denote the real part and the imaginary part, respectively.

II. SIGNAL MODEL AND MPA

A. SIGNAL MODEL

In the SCMA-based communication system, J users spread their codewords over K physical-layer resources, also known as resource element (RE). The overloading factor (OF), λ, is defined as $\lambda = J/K$. Note that the RE can be any orthogonal resource in the communication system. For the GF-SCMA communication system discussed in this paper, the RE refers to the time-frequency-domain resource grid in the OFDM signals. The users send their messages by

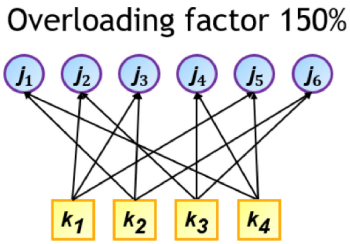


FIGURE 1. Factor graph of an SCMA system with 150% overloading.

selecting codewords from their respective codebooks. Each codebook contains M codewords, and each codeword carries $\log_2 M$ bits. These codewords are sparse K-dimensional vectors with only d_v non-zero entries.

In this paper, the codebook of the jth user is expressed as $\mathbf{C}_j = \{\mathbf{c}_{1,j}, \ldots, \mathbf{c}_{M,j}\}$, where $\mathbf{c}_{m,j} = [c_{1,m,j}, \ldots, c_{K,m,j}]^T$ denotes the mth codeword. The codeword sent by the jth user is represented by $\mathbf{x}_j = [x_{1,j}, \ldots, x_{K,j}]^T$. $\mathbf{a} = [a_1, \ldots, a_J]^T \in \{0, 1\}^J$ denotes the user activity vector. Then, the K-dimensional signal received at the BS, $\mathbf{y}_j = [y_1, \ldots, y_K]^T$, is given by

$$\mathbf{y} = \sum_{j=1}^{J} a_j diag(\mathbf{h}_j)\mathbf{x}_j + \mathbf{n}, \tag{1}$$

where $\mathbf{h}_j$ is a Rayleigh-distributed channel gain vector corresponding to user j on the K REs, and $\mathbf{n}$ is the Gaussian noise vector whose elements are independent zero-mean Gaussian random variables, each with variance σ^2. We assume no power control in the UE transmitters.

B. MESSAGE PASSING ALGORITHM

MPA is an iterative optimization algorithm based on factor graphs. Take the SCMA system with the factor graph shown in Fig. 1 as an example. The factor graph contains four resource nodes (RNs), six variable nodes (VNs), and edges connecting the RNs and VNs. Each VN corresponds to a user, and the probability of the user sending a particular codeword is called the VN message. On the other hand, each RN corresponds to one RE, and an RN message indicates the probability of a user transmitting a codeword based on the signal received by the RE. A pair of VN and RN connected by an edge denotes that the corresponding user's codewords have a non-zero element on the RE. In this paper, each VN connects to d_v RNs, and each RN connects to d_r VNs. Finally, note that there are six VNs ($J = 6$) and four RNs ($K = 4$), resulting in an OF of $\lambda = 150\%$.

At the beginning of MPA decoding, all M codewords are assumed to be transmitted with equal probability and

$$P_{k,j}^{(0)}(\mathbf{c}_{m,j}) = \frac{1}{M} \quad k \in V(j), j = 1, \ldots, J; m = 1, \ldots, M. \tag{2}$$

where $P_{k,j}^{(i)}(\mathbf{c}_{m,j})$ is the VN message for codeword $\mathbf{c}_{m,j}$ from the jth VN to the kth RN in the i-th iteration, and $V(j)$ is the set of RNs connected to the jth VN. Note that the VN messages corresponding to all M codewords are sent.

During an iteration step, the RN and VN messages are updated alternately. The conditional probabilities $p(y_k|\mathbf{x}^{[k]})$ (called initial probability) of all CCs are calculated first based on the received signal with (3). The CC $\mathbf{x}^{[k]}$ is a vector composed of $x_{k,j}$ for all $j \in R(k)$, where $R(k)$ is the set of VNs connected to the k-th RN. Equations (4) and (5) delineate the RN message $Q_{k,j}^{(i)}(\mathbf{c}_{m,j})$ and VN message $P_{k,j}^{(i)}(\mathbf{c}_{m,j})$ updating, respectively.

$$p\left(y_k|\mathbf{x}^{[k]}\right) = \frac{1}{\sigma\sqrt{2\pi}} exp\left(\frac{-1}{\sigma^2}\left\|y_k - \sum_{j\in R(k)} h_{k,j}x_{k,j}\right\|^2\right) \tag{3}$$

$$Q_{k,j}^{(i)}(\mathbf{c}_{m,j}) = \sum_{\mathbf{x}^{[k]}\in\mathbf{X}^{[k]}, \mathbf{x}_{k,j}=\mathbf{c}_{k,m,j}} \left\{p\left(y_k|\mathbf{x}^{[k]}\right) \prod_{j'\in R(k)\backslash j} P_{k,j'}^{(i-1)}(\mathbf{x}_{j'})\right\} \tag{4}$$

$$P_{k,j}^{(i)}(\mathbf{c}_{m,j}) = norm\left\{\prod_{k'\in V(j)\backslash k} Q_{k',j}^{(i)}(\mathbf{c}_{m,j})\right\}, \tag{5}$$

where the "norm" operator normalizes all probabilities so that they sum to unity, and the set $\mathbf{X}^{[k]}$ contains all possible 4^3 CCs. The RN updating in (4) is the reason for the high complexity due to the need to calculate the probability of 4^2 possible CCs.

After I iterations, the VN and RN messages will approach the actual probability distributions. The decoded soft outputs can be obtained from (6).

$$p(\mathbf{x}_j = \mathbf{c}_{m,j}) = \prod_{k\in V(j)} Q_{k,j}^{(I)}(\mathbf{c}_{m,j}) \tag{6}$$

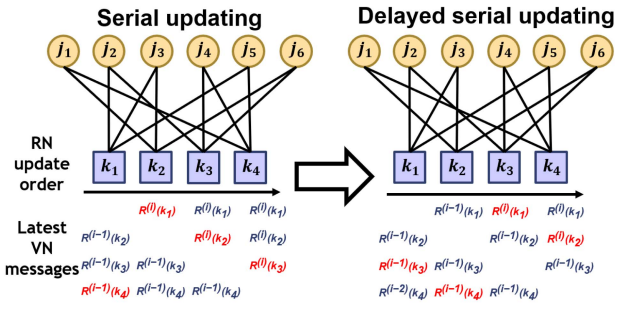
C. MAX-LOG MPA

RN updating requires massive multiplications and additions. Moreover, the exponential function used to calculate the initial probability is hard to implement in hardware. Logarithm approximation converts all computations to log-domain. It replaces multiplications with additions and approximates additions by taking maximum values [20]. This approximation significantly reduces the computational complexity with a minor performance penalty. Therefore, it is widely used in the SCMA decoder design. Through the approximation, the initial probability calculation, RN updating, and VN updating can be rewritten as follows:

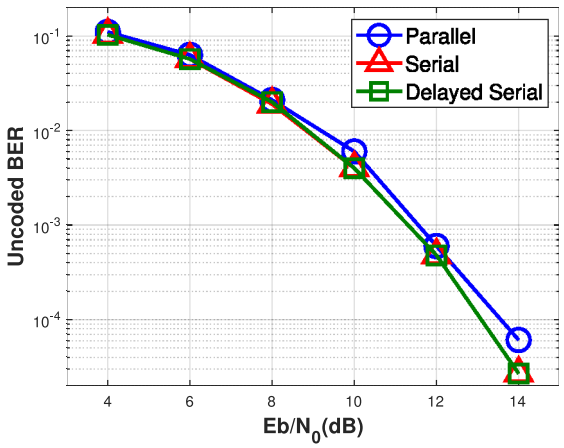
$$g_k\left(\mathbf{x}^{[k]}\right) = -\frac{1}{\sigma^2}\left\|y_k - \sum_{j\in R(k)} h_{k,j}x_{k,j}\right\|^2 \tag{7}$$

$$LQ_{k,j}^{(i)}(\mathbf{c}_{m,j}) = \max_{\mathbf{x}^{[k]}\in\mathbf{X}^{[k]}, \mathbf{x}_{k,j}=\mathbf{c}_{k,m,j}} \left\{g_k\left(\mathbf{x}^{[k]}\right) + \sum_{j'\in R(k)\backslash j} LP_{k,j'}^{(i-1)}(\mathbf{x}_{j'})\right\} \tag{8}$$

$$LP_{k,j}^{(i)}(\mathbf{c}_{m,j}) = \sum_{k'\in V(j)\backslash k} LQ_{k',j}^{(i)}(\mathbf{c}_{m,j}) \tag{9}$$



(a) Updating schedule



(b) Decoding performance

FIGURE 2. Updating schedule, and decoding performance of the proposed delayed serial updating MPA method.

III. SCMA DECODER DESIGN

A. PROPOSED DELAYED-SERIAL UPDATING MPA

In conventional MPA, the VNs are updated after all RNs have been updated in parallel. In contrast, S-MPA updates the VNs every time an RN is updated [17]. Such a modification increases the number of VN updates in an iteration, making the algorithm converge faster. S-MPA can achieve the same or even better decoding performance than conventional MPA with fewer iterations. However, it suffers from high data dependency, significantly reducing the hardware efficiency and the decoding speed in hardware implementation.

We proposed delayed serial MPA (DS-MPA) to achieve higher hardware efficiency and lower latency. The updating schedule and error rate performance of DS-MPA are shown in Fig. 2. Unlike S-MPA, DS-MPA allows RN (e.g., k_2) updating to occur without waiting for the update of the previous RN (e.g., k_1) and its associated VNs (e.g., j_2, j_3, j_5). This is because waiting for the previous RN update would result in idle cycles in the pipelined RN update circuit, leading to up to 50% loss in throughput. Instead, DS-MPA uses the VNs (j_2, j_3, j_5) computed from the previous iteration (step $i-1$) to update the RN of k_2. Moreover, in Fig. 2(b), DS-MPA achieves a competitive decoding error rate performance with S-MPA, and they both beat the parallel MPA method. The codebook design used in the simulation is referred to [10, Fig. 4]. Detail circuit pipeline descriptions according to the proposed scheduling will be presented later.

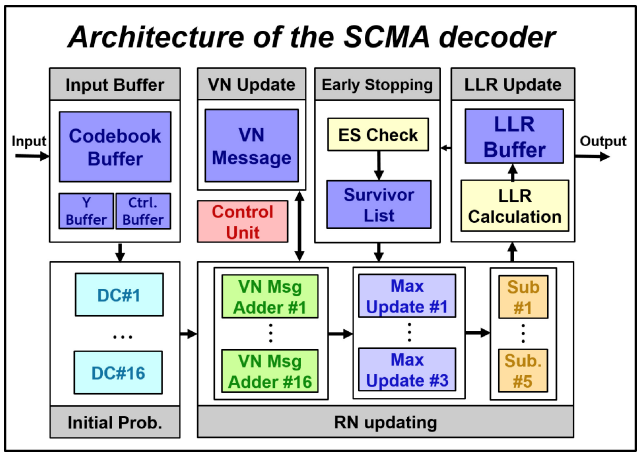


FIGURE 3. Hardware architecture of the proposed SCMA decoder.

B. DECODER ARCHITECTURE

Based on DS-MPA, we designed an SCMA decoder circuit to support conventional Max-log MPA and ST-Max-log-MPA [16]. Fig. 3 depicts the circuit architecture of the proposed SCMA decoder for $K = 4, J = 6, M = 4$, which includes major processing blocks such as distance-based initial log probability calculation, RN updating, VN updating, early stopping, and LLR updating. The first three blocks perform the Max-log MPA iterations given in (7), (8), (9). The early stopping block implements bit freezing in ST-Max-log-MPA, while the LLR updating block calculates the LLR. The LLR values serve as the final outputs. Also, the early stopping algorithm uses them to determine which bits should be frozen, and the corresponding candidate CCs be eliminated. The decision is made by the early stopping (ES) check block, and the survivor list block records the bits that have not been decided.

The initial probability calculation block will calculate the distance between all CCs and the frequency domain signal y as the initial probability of each RN. Each set of hardware is called a distance calculator (DC). To reduce circuit complexity, we have used an approximate distance that requires only additions/subtractions and no square or multiplication operations. Next, the RN updating block computes the new RN messages according to (8). During RN updating, the VN messages of all VNs connected to the RN except one will be summed with the initial probability. For each candidate CC, there will be a corresponding sum. The maximum among these sums becomes the new RN message. RN updating completes the current iteration when the RN messages of all codewords of all users have been updated. The number of CCs to consider during RN updating is 64 (4^3) since three users share one RN, and each has four possible codewords. To meet the throughput requirement, we have designed 16 sets of DC blocks and VN message adders to

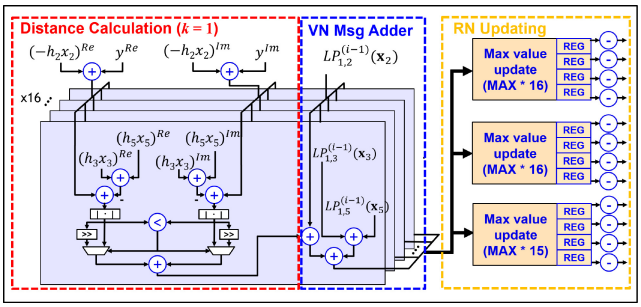


FIGURE 4. Architecture of the initial probability calculation & RN updating Circuit.

process 16 CCs in one round. Details of these circuits and the maximum finding circuit will be presented. The last part of the iteration is VN updating. However, since only two RNs are connected to each VN in Fig. 1, a summation is unnecessary, according to (9).

C. INITIAL PROBABILITY CALCULATION AND RN UPDATING CIRCUIT

The initial probability calculation and RN updating make up most of the computation complexity in MPA iterations. This results from the need to compute probabilities and RN messages corresponding to many CCs. As such, we carefully designed the architecture and scheduling for these two circuits. Fig. 4 illustrates the circuit architecture of the initial probability calculation and the RN updating blocks for the $k = 1$ case in Fig. 1. Sixteen sets of DC circuits compute the initial probabilities corresponding to 16 different CCs. The 16 CCs run through four-codeword combinations of two users (x_3, x_5) with the other user's codeword fixed (x_2). The frequency-domain received signal and the channel-faded codeword of the user with its codeword fixed in this round will be processed first $(y - h_2x_2$ in this example). This term will be distributed to 16 DC circuits; each handles one CC (16 CCs correspond to all combinations of x_3 and x_5). After the real and imaginary parts are calculated, the absolute value will be taken. The approximate distance is obtained by a weighted sum of the absolute real part (I) and absolute imaginary part (Q) by $\max(I, Q) + 0.25\min(I, Q)$. Note the VN message from the user with a fixed codeword $(j = 2)$ is distributed to all VN message adders.

For the two users that explore four codewords in the current round (x_3 and x_5), the maximum value for each of their codewords must be found. Four sums are compared to the stored current maximum for a particular codeword. As such, each of these two users requires a total of $4 \times 4 = 16$ sets of 2-input maximum circuits. For the user with a fixed codeword in this round (x_2), all 16 sums obtained within each cycle belong to a specific codeword. Therefore, the maximum value of the 16 sums is found by 15 2-input maximum circuits. The final subtraction circuit is activated when one iteration is complete (four rounds corresponding to four codewords of x_2). This final subtraction is necessary since the summation term in (8) is reformulated as below for lower complexity and reduced register access.

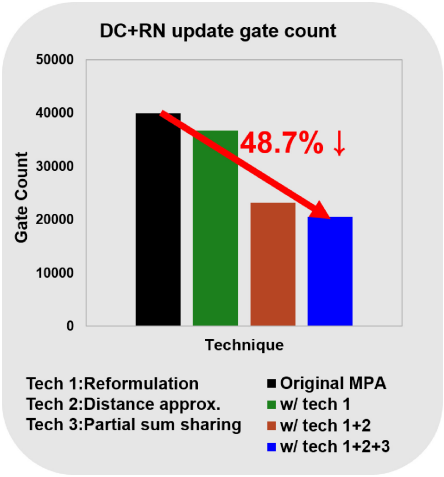


FIGURE 5. Gate count reduction of the DC & RN updating circuit.

$$\left\{\sum_{j' \in R(k)} LP_{k,j'}^{(i-1)}\left(\mathbf{x}_{\mathbf{j}'}\right)\right\} - LP_{k,j}^{(i-1)}\left(\mathbf{x}_{\mathbf{j}}\right) \tag{10}$$

To summarize, we applied three techniques to enhance the efficiency of the designed decoder. First, the updating formula (8) is reformulated, as shown in (10), decreasing the complexity of RN updating. Second, the distance calculation required by (7) is approximated to avoid the multiplication operations. Finally, sixteen copies of DC circuits that process the initial probabilities of sixteen different CCs reuse the partial sum $(y - h_2x_2$ in the example shown in Fig. 4). The gate count reduction of the DC and RN updating circuits when applying the three techniques is illustrated in Fig. 5. Compared to the direct implementation of the original Max-Log MPA, the proposed decoder can save up to 48.7% of the area.

Thanks to the DS-MPA mentioned in this section, the pipeline schedule of the RN updating, shown in Fig. 6, has no bubble. Without the proposed delayed update scheme, S-MPA causes the circuit to suffer five idle cycles between two adjacent RN updating iterations. On the other hand, with DS-MPA scheduling, the ensuing RN updating round (e.g., the first round of RN2) can start as soon as the last round of the previous RN updating (e.g., RN1) is completed. The proposed pipeline schedule thus significantly increases the hardware utilization and reduces the latency by up to 50%.

IV. GF-SCMA TRANSMISSION SYSTEM DESIGN

This section will first present the frame structure of the proposed GF-SCMA uplink transmission system for mMTC networks. Moreover, we also explain the baseband processing and FPGA circuit design of primary GF-SCMA receiver functions, including synchronization, channel estimation, user activity detection, and SCMA decoding.

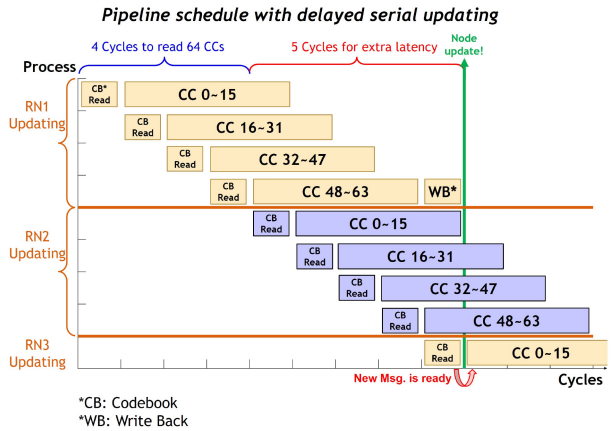


FIGURE 6. Pipeline schedule based on the delayed serial updating.

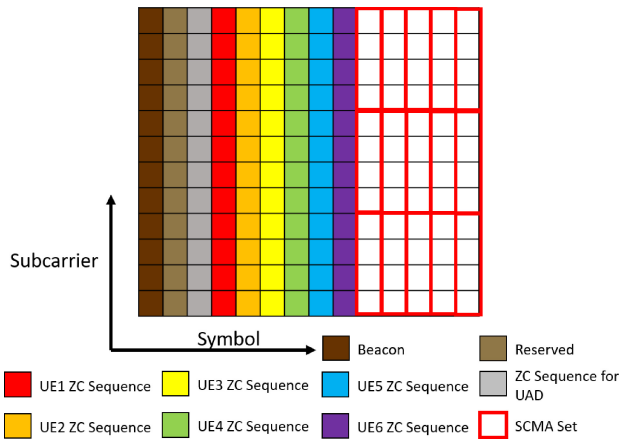


FIGURE 7. Uplink signal frame structure in the OTA experiments.

A. UPLINK FRAME STRUCTURE

The proposed uplink frame structure of the GF-SCMA transmission system is shown in Fig. 7. In practical mMTC networks with a massive number of UEs, the reserved second symbol can transmit the ACK signals to UEs. When two or more UEs with the same codebook and sequence assignment try to transmit simultaneously in one frame, ACK will not be sent due to collision. Without receiving ACK, the UEs would transmit again after some interval of waiting. Furthermore, the symbols carrying SCMA sets can be divided into multiple blocks. Thanks to the characteristic of small data bursts of mMTC, users can choose to transmit on only one of the blocks. With various combinations of codebooks and blocks, the probability of collision can be reduced. For mMTC demonstration purposes, the prototype that we built implements only the uplink traffic, and there is no downlink consideration. As such, all user TXs are assumed synchronized, and an uplink beacon symbol at the start of each frame helps achieve frame synchronization. Following the beacon, each active user in this frame will transmit a symbol made up of a unique Zadoff-Chu (ZC) sequence (gray) [21]. The sequence serves

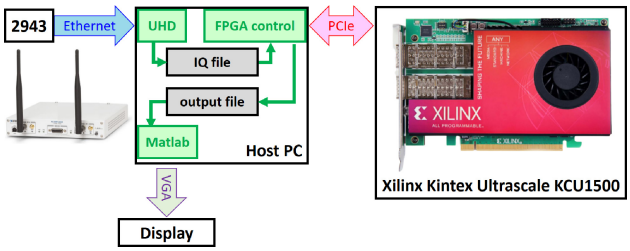


FIGURE 8. GF-SCMA receiver hardware/software configuration.

the purpose of the user signature for user activity detection at the BS.

Traditionally, channel estimation in OFDM receivers is usually performed on pilot symbols/subcarriers that are sparsely distributed in the time-frequency resource grid. Considering the demonstration of six mMTC users with light traffic loads, we decided to reserve six pilot symbols to carry the reference signals (the same ZC sequences used in UAD) for channel estimation in a 14-symbol frame. This will significantly reduce channel estimation complexity in the demonstration prototype's receiver. Naturally, in an actual mMTC network with tens of simultaneously active users, the reference signals for channel estimation should be sparsely deployed in the resource grid. Such resource allocation needs careful consideration to limit the spectrum overhead. The BS receiver's channel estimation task becomes more complicated with more convoluted reference signal allocation. Finally, the last five symbols carry sets of four subcarriers, marked by red boxes. As shown in Fig. 1, each set corresponds to four resource nodes that carry SCMA codewords of six users.

B. OVERALL GF-SCMA RECEIVER DESIGN

The GF-SCMA transmission prototype contains both hardware and software parts. The overall receiver structure is shown in Fig. 8. The inner receiver, including baseband signal processing, UAD, and SCMA decoding, is implemented in FPGA circuits. On the other hand, the software controls the hardware operation and the data transfer between the host PC, the FPGA board, and National Instruments (NI) 2943R that works as the RF frontend and digital down-converter. Finally, the GUI displays the decoding results and various OTA experiment outcomes.

The host computer connects to the receiving instrument 2943R via Ethernet and controls it using the USRP Hardware Driver (UHD). The captured waveform is digitized and down-converted to in-phase/quadrature-phase (I/Q) baseband signals and then stored in the IQ file on the host PC. The FPGA control program will send the baseband samples to the URAM on the FPGA through the PCIe bus and trigger the FPGA circuit for further baseband signal processing, UAD, and SCMA decoding. After the FPGA circuits complete all the processing, the control program then stores the SCMA decoding results in the output file on the host PC. Finally, the GF-SCMA transmission results are displayed on the GUI written in MATLAB.

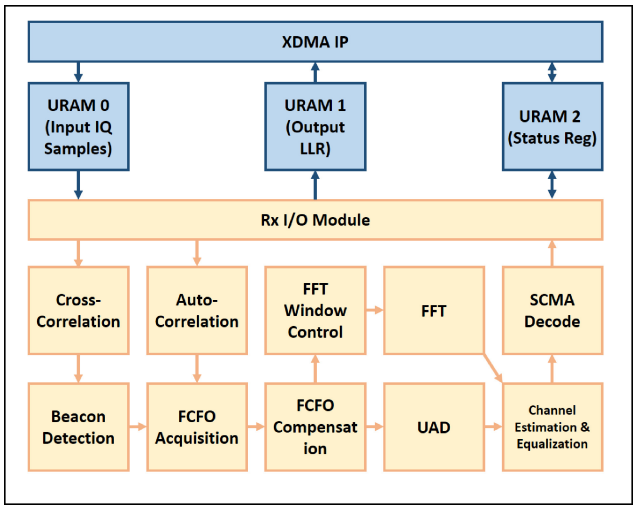


FIGURE 9. Block diagram of the FPGA-based GF-SCMA inner receiver.

C. INNER RECEIVER CIRCUITS

Hardware blocks and dataflow of the inner receiver and data buffer control are illustrated in Fig. 9. The blue blocks are in charge of data exchange with the host computer. The XDMA IP is a built-in module in Xilinx Vivado design environment for transmitting data through PCI-express. URAM 0 stores sample points of the captured signals and URAM 1 buffers the resulting LLR outcomes from SCMA decoding. URAM 2 is called *status register*, and it is used to indicate the current status of the receiver circuits, e.g., FPGA running, URAM full, URAM empty, etc. The status register can be read and written by the host computer and the inner receiver hardware, enabling effective communication between them.

The yellow blocks in Fig. 9 perform the baseband processing, UAD, and SCMA decoding. The RX I/O block controls the read/write operations of the input and output buffers. It also monitors and manages the status register to transfer data between the FPGA and the host computer. The I/Q baseband samples first pass through the Cross-Correlation and the Beacon Detection blocks to detect the symbol/frame boundary. The following Auto-Correlation, fractional carrier frequency offset (FCFO) Acquisition, and FCFO Compensation blocks are in charge of the FCFO detection and compensation. The subsequent dataflow is split into two: (a) the FFT Window Control and FFT blocks transform the signal into the frequency domain, and (b) the UAD block performs user activity detection in the time domain. Combining the active user list and the frequency-domain signals, the channel estimation & equalization block estimates the channel gains and then computes the channel-faded codewords. Finally, the faded codewords and the frequency-domain received signals are sent to the SCMA Decode block to find the LLR values of the transmitted bits. To achieve real-time decoding, the SCMA Decode block implements eight SCMA decoder cores presented in the previous section.

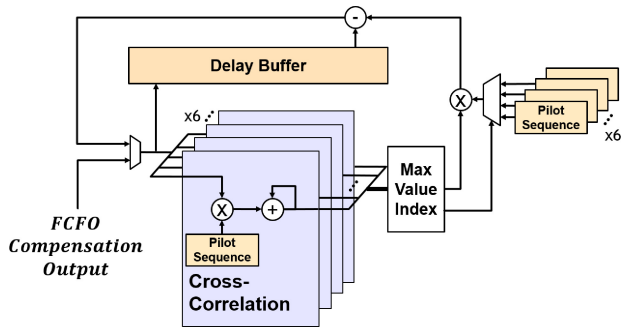


FIGURE 10. Circuit diagram of the SIC-based user activity detection (UAD) block.

D. UAD

Several algorithms have been proposed to address the UAD problem that arises with a large number of potential users and non-orthogonal pilots. However, for the demonstration and verification of the six-user mMTC network, we opted to use a simplified UAD algorithm for FPGA. This will significantly reduce the UAD complexity and latency in the inner receiver of the proposed prototype. For more complex UAD problems, advanced algorithms like ALEM [9] can be utilized. In light of the severe frequency-selective fading in the frequency domain, we applied the successive interference cancellation (SIC) algorithm in the time domain for user activity detection. This method selects the strongest user sequence based on the received signal's cross-correlations with the time-domain user sequences. Then the SIC UAD method deducts the strongest user's pilot sequence from the signal. This process is repeated to find other users until the maximum cross-correlation is below a preset threshold. The proposed UAD method is inherently regular and iterative and thus is hardware-friendly.

The hardware architecture for the SIC-based UAD block is shown in Fig. 10. Six multiply-and-accumulate units compute the cross-correlations of the interference-canceled signal and up to six user sequences. In the first iteration, the input signal comes from the previous CFO compensation block, and it is also buffered in preparation for the cancellation of the strongest user sequence once it is found. Note that the signal to be subtracted is the strongest user sequence multiplied by the detected channel gain found by the cross-correlation computation unit. The strongest detected user's signal component in the current iteration is removed from the residual signal using the subtractor on the top. Finally, the resulting residual signal goes through the multiplexer on the left, and the next iteration commences.

V. OTA EXPERIMENTS

We built a prototype and experimented with an over-the-air (OTA) channel to validate the proposed GF-SCMA transmission system design. This section describes the experimental environment and parameter settings. Also, the real-time decoding results of the GF-SCMA receiver are presented.

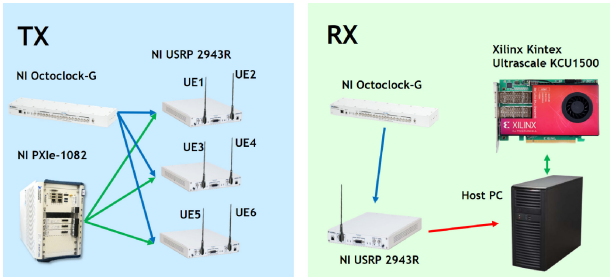


FIGURE 11. Overall instrument diagram of the OTA verification platform.

TABLE 1. Parameter settings in the OTA verification experiments.

System Parameters	Carrier Frequency	2.55GHz
	Baseband Sample Rate	15.36MHz
	Transmission Bandwidth	10MHz
	FEC	Non-Coded
Numerology	CP Type	Normal
	Subcarrier Spacing	15kHz
	FFT Size	1024
	Occupied Subcarriers	596
SCMA	Number of Users	6
	Overloading Factor	150%

A. EXPERIMENT SETTING

The GF-SCMA transmission system includes six single-antenna users and one single-antenna receiver. On the transmitter (TX) side, three NI 2943Rs are used to transmit signals, and each 2943R connects to two users' TX antennas. An NI PXIe-1082 based controller controls these three 2943Rs via a PXIe bus. A stable reference clock from an NI Octoclock-G device synchronizes all 2943Rs. On the receiver (RX) side, an NI 2943R captures the signal and sends it, via a 10 Gigabit Ethernet cable, to a host PC equipped with a Xilinx KCU1500 FPGA board that implements the baseband processing circuits. The overall architecture of the experimental prototype for the proposed GF-SCMA transmission system is depicted in Fig. 11.

Table 1 lists the parameter settings used in this demonstration prototype. When designing and verifying the SCMA decoder and the demonstration prototype, we focused on implementing the inner receiver of the transmission system. As a result, the transmitted data was not protected by forward error correction (FEC) codes. However, given the low uncoded error rate observed in the experiment, this approach was deemed sufficient. The data rate per user is approximately 596 (occupied subcarriers) $\times$ 0.5 (average bit per subcarrier) $\times$ 5 (data symbol per frame) $\times$ 1000 (frame per second) $\sim$ 1.42Mbps. Such uplink data rate is sufficient for the devices in mMTC networks, which inherently require only low data rate transmission. According to the 5G numerology, there are 600 subcarriers. However, due to severe interference (offest) around DC, we did not use the four subcarriers near DC, resulting in 596 used subcarriers. Finally, the six pilot sequences assigned to the UEs for UAD and channel estimation are the ZC sequences with a length of 599 and roots (u) equal to 1-6 [21].

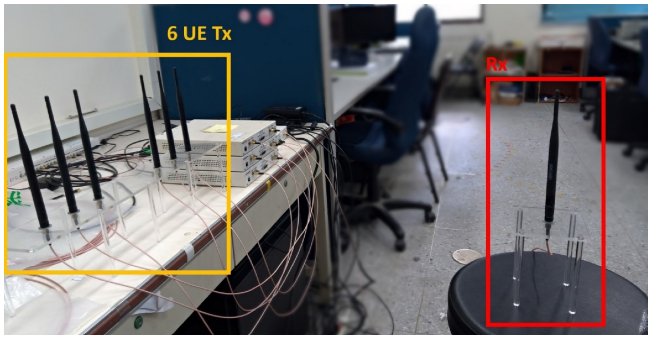


FIGURE 12. OTA verification environment.

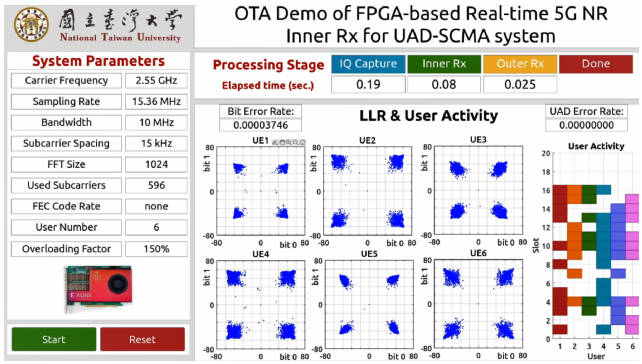


FIGURE 13. A screenshot for the GUI during SCMA-UAD decoding.

Our OTA verification experiment environment is illustrated in Fig. 12. The yellow rectangle on the left encloses the TX antennas of the six UEs, and the red rectangle on the right indicates the RX antenna of the BS. The distance between the TX and RX antennas is about 0.75 meters. Such limited TX-RX separation is because the OTA experiment was conducted in a crowded indoor environment. The experiment's transmission schedule and data for the six UEs are randomized and stored in the control host. The UEs transmit continuously according to the stored schedule during the experiment, while the BS RX keeps capturing the signals and conducting UAD and SCMA decoding. A colorful GUI then displays the receiver outcomes.

B. OTA RECEIVER RESULTS

Fig. 13 depicts the GUI showing the status of the GF-SCMA receiver, including the system parameter table on the left panel and the receiver status and outcomes on the right panel. The three receiver stages' execution time, the UAD error rate, the SCMA detection bit error rate (BER), detected active users, and SCMA detected LLR distribution are all illustrated in this panel. Our goal is to achieve real-time decoding by customized circuits, which means that the execution time of the inner receiver (the Inner Rx stage in the GUI) must be shorter than the IQ capture time. Otherwise, the captured IQ signals overflow the RX memory, causing receiver data loss.

We recorded a session of the GF-NOMA demonstration prototype executing UAD and SCMA detection smoothly and uploaded it for viewing in [23]. In this session, we manually turned off some UE TXs to demonstrate the validity of the prototype responding to real-time UE activities and successfully detecting the actual activities. The UAD error rate is always 0 except when we turned off some UEs manually. In those intervals, the pre-recorded UE activity patterns for UAD error rate calculation were no longer valid as some UEs were forced off, and the displayed UAD error rate did not reflect the true detection performance. Note that if there were more than 6 UEs, the interference among ZC sequences can be severe, and such zero error cannot be achieved by the SIC UAD method that we used in the prototype. Moreover, the detected SCMA BER was never higher than 10^{-4}, and this demonstrates that reliable SCMA OFDM transmission can be established with proper FEC protection.

Finally, the signal duration (IQ capture stage) takes about 0.19s, while the elapsed times of the inner and the outer receivers to process one segment of IQ signals are about 0.08s and 0.03s, respectively. We thus are confident that the hardware receiver outpaces the IQ capture stage, and the whole GF-SCMA BS RX can operate in real time.

VI. DISCUSSIONS AND CONCLUSION

While the constructed transmission system can achieve a low UAD error rate and BER with real-time operation, we have made some simplifications for demonstration purposes. In a practical grant-free NOMA network for mMTC, the enormous number of potential users introduces the challenges of codebook collisions. The collision reduction and handling techniques were addressed in Section IV-A. However, when the collision happens, how will the users be affected during multi-user detection (MUD), and how should the MUD algorithm be improved to avoid retransmission are topics worthy of further investigation.

In this paper, we developed a new SCMA decoding algorithm that supports high-throughput MPA-based decoder architecture. In addition, we also designed hardware circuits for this SCMA decoder and user-activity detection circuits needed for grant-free mMTC operation. Furthermore, to build a GF-SCMA transmission prototype, we designed and integrated key inner receiver function blocks, such as CFO synchronization, user signature detection, channel estimation and compensation, soft SCMA detection, etc. The inner receiver circuits were implemented on a Xilinx KCU1500 FPGA board. The whole prototype was demonstrated successfully for real-time SCMA decoding with a low uncoded error rate. Based on this feasibility demonstration, we can implement a UAD algorithm that supports more UEs and increase the SCMA overloading factor to accommodate tens or hundreds of UEs with sporadic traffic. In conclusion, the proposed GF-SCMA receiver hardware design and the real-time demonstration prototype present an important first step toward effective and efficient GF-SCMA transmission solutions to 5G massive MTC networks.

REFERENCES

[1] A. Bayesteh, E. Yi, H. Nikopour, and H. Baligh, "Blind detection of SCMA for uplink grant-free multiple-access," in *Proc. Int. Symp. Wireless Commun. Syst. (ISWCS)*, Barcelona, Spain, 2014, pp. 853–857.

[2] I. F. Gorodnitsky and B. D. Rao, "Sparse signal reconstruction from limited data using FOCUSS: a re-weighted minimum norm algorithm," *IEEE Trans. Signal Process.*, vol. 45, no. 3, pp. 600–616, Mar. 1997.

[3] Y. Chen. "Sparse signal reconstruction using FOCUSS." Nov. 9, 2014. [Online]. Available: https://www.slideserve.com/idona-cobb/sparse-signal-reconstruction-using-focuss

[4] J. Liu, G. Wu, S. Li, and O. Tirkkonen, "Blind detection of uplink grant-free SCMA with unknown user sparsity," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Paris, France, 2017, pp. 1–6.

[5] W. Xiong, J. Cao, and S. Li, "Sparse signal recovery with unknown signal sparsity," *Eurasip J. Adv. Signal Process*, vol. 1, no. 178, pp. 1–8, 2014.

[6] Y. Wang, S. Zhou, L. Xiao, X. Zhang, and J. Lian, "Sparse Bayesian learning based user detection and channel estimation for SCMA uplink systems," in *Proc. Int. Conf. Wireless Commun. Signal Process. (WCSP)*, Nanjing, China, 2015, pp. 1–5.

[7] K. Struminsky, S. Kruglik, D. Vetrov, and I. Oseledets, "A new approach for sparse Bayesian channel estimation in SCMA uplink systems," in *Proc. Int. Conf. Wireless Commun. Signal Process. (WCSP)*, Yangzhou, China, 2016, pp. 1–5.

[8] D. P. Wipf and B. D. Rao, "Sparse Bayesian learning for basis selection," *IEEE Trans. Signal Process.*, vol. 52, no. 8, pp. 2153–2164, Aug. 2004.

[9] N.-H. Huang and T.-D. Chiueh, "Sequence design and user activity detection for uplink grant-free NOMA in mMTC nNetworks," *IEEE Open J. Commun. Soc.*, vol. 2, pp. 384–395, 2021.

[10] H. Nikopour and H. Baligh, "Sparse code multiple access," in *Proc. IEEE PIMRC*, 2013, pp. 332–336.

[11] H. Mu, Z. Ma, M. Alhaji, P. Fan, and D. Chen, "A fixed low complexity message pass algorithm detector for up-link SCMA system," *IEEE Wireless Commun. Lett.*, vol. 4, no. 6, pp. 585–588, Dec. 2015.

[12] M. Jia, L. Wang, Q. Guo, X. Gu, and W. Xiang, "A low complexity detection algorithm for fixed up-link SCMA system in mission critical scenario," *IEEE Internet Things J.*, vol. 5, no. 5, pp. 3289–3297, Oct. 2018.

[13] L. Wei, B. Huang, and J. Zheng, "Low-complexity detectors for uplink SCMA: symbol flipping and dynamic partial marginalization-based MPA," in *Proc. IEEE VTC-Spring)*, Porto, Portugal, 2018, pp. 1–5.

[14] X. Ma, L. Yang, Z. Chen, and Y. Siu, "Low complexity detection based on dynamic factor graph for SCMA systems," *IEEE Commun. Lett.*, vol. 21, no. 12, pp. 2666–2669, Dec. 2017.

[15] H. Y. Lan, "Low complexity multi-user detector design for sparse code multiple access in 5G networks," M.S. thesis, Graduate Inst. Electron. Eng., Nat. Taiwan Univ., Taipei, Taiwan, Nov. 2017.

[16] G. Zhang, Z. Gu, J. Zhang, S. Li, J. Ren, and W. Lu, "A Max-log-MPA algorithm based on serial and threshold in SCMA system," in *Proc. Int. Wireless Commun. Mobile Comput. (IWCMC)*, 2020, pp. 1101–1105.

[17] Y. Du, B. Dong, Z. Chen, X. Wang, and P. Gao, "Improved serial scheduling-based detection for sparse code multiple access systems," *IEEE Wireless Commun. Lett.*, vol. 6, no. 5, pp. 570–573, Oct. 2017.

[18] J.-R. Garnier, A. Fabre, H. Farès, and R. Bonnefoi, "On the performance of QPSK modulation over downlink NOMA: From error probability derivation to SDR-based validation," *IEEE Access*, vol. 8, pp. 66495–66507, 2020, doi: 10.1109/ACCESS.2020.2983299.

[19] Y. Qi, X. Zhang, and M. Vaezi, "Over-the-air implementation of NOMA: New experiments and future directions," *IEEE Access*, vol. 9, pp. 135828–135844, 2021, doi: 10.1109/ACCESS.2021.3116613.

[20] S. Zhang, X. Xu, L. Lu, Y. Wu, G. He, and Y. Chen, "Sparse code multiple access: An energy efficient uplink approach for 5G wireless systems," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Austin, TX, USA, 2014, pp. 4782–4787, doi: 10.1109/GLOCOM.2014.7037563.

[21] "Wikipedia." Accessed: Apr. 19, 2023. [Online]. Available: https://en.wikipedia.org/wiki/Zadoff-Chu_sequence

[22] T. D. Chiueh, P. Y. Tsai, and I. W. Lai, *Baseband Receiver Design for Wireless MIMO-OFDM Communications*. Singapore: Wiley, Apr. 2012.

[23] T. Y. Chen. "6UE activity detection and SCMA receiver demo V3." Nov. 22, 2021. [Online]. Available: https://youtu.be/cy2kRoPy_5I

TI-YU CHEN (Member, IEEE) was born in Taiwan. He received the B.S. degree in electrical engineering from National Taiwan University, Taipei City, Taiwan, in 2020, where he is currently pursuing the Ph.D. degree in electronics engineering. His research interests include communication systems and digital circuit design.

ZHI-JING LIN was born in Taiwan. He received the B.S. degree in electrical and computer engineering from National Chiao Tung University and the M.S. degree in electronics engineering from National Taiwan University in 2021. He is currently with Mediatek, Inc. His interest is in wireless communication algorithms and VLSI design for digital signal processing.

TZI-DAR CHIUEH (Fellow, IEEE) was born in Taipei City, Taiwan, in 1960. He received the B.S.E.E. degree from the National Taiwan University, Taipei City, in 1983, and the M.S. and Ph.D. degrees in electrical engineering from the California Institute of Technology, Pasadena, CA, USA, in 1986 and 1989, respectively.

Since 1989, he has been with the Department of Electrical Engineering, National Taiwan University, where he is currently a Distinguished Professor and the Dean of the Graduate School of Advanced Technology. From 2004 to 2007, he served as the Director of the Graduate Institute of Electronics Engineering, National Taiwan University. He has held visiting positions with ETH Zurich, Switzerland, from 2000 to 2001 and the State University of New York, Stony Brook, from 2003 to 2004. From November 2010 to January 2014, he served as the Director General of National Chip Implementation Center Hsinchu, Taiwan. He also served as the Vice President of National Applied Research Laboratories from 2015 to 2017. His research interests include IC design for digital communication systems, neural network, and signal processing for bio-medical systems.

Prof. Chiueh's teaching efforts were recognized eleven times by the Teaching Excellence Award from NTU. He was the recipient of the Outstanding Research Award from National Science Council, Taiwan, from 2004 to 2007. In 2005, he received the Outstanding Electrical Engineering Professor from the Chinese Institute of Electrical Engineers, Taiwan, and was awarded the Himax Chair Professorship at NTU in 2006. In 2009, he received the Outstanding Industry Contribution Award from the Ministry of Economic Affairs, Taiwan. He received the Outstanding Technology Transfer Contribution Award from the Ministry of Science and Technology, Taiwan, in 2016. He is the NTU Macronix Chair Professor in 2021.