IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received May 1, 2018, accepted May 25, 2018, date of publication May 31, 2018, date of current version July 6, 2018.

Digital Object Identifier 10.1109/ACCESS.2018.2842233

Enabling High Order SCMA Systems in Downlink Scenarios With a Serial Coding Scheme

YUXI HAN[1], WUYANG ZHOU[1], (Member, IEEE), MING ZHAO[1], AND SHENGLI ZHOU[2], (Fellow, IEEE)

[1]Key Laboratory of Wireless-Optical Communications, Chinese Academy of Sciences, University of Science and Technology of China, Hefei 230027, China
[2]Department of Electrical and Computer Engineering, University of Connecticut, Storrs, CT 06269, USA

Corresponding author: Wuyang Zhou (wyzhou@ustc.edu.cn)

This work was supported by the Natural Science Foundation of China under Grant 61461136002 and in part by the Key Program of National Natural Science Foundation of China under Grant 61631018.

ABSTRACT In the future fifth generation communication systems, sparse code multiple access (SCMA) shows strong competitiveness as a novel non-orthogonal multiple access (NOMA) technique. Attributing to multi-dimensional codebooks, the SCMA can obtain shaping gain and provide a better performance than some other NOMA schemes, such as pattern division multiple access, low-density signature (LDS), and so on. However, under higher user load, each orthogonal resource in high-order SCMA systems is occupied by more users and the constellations of these systems possess smaller Euclidean distances, both of which will degrade the link performance. The iterative algorithm employed in SCMA decoding, namely message passing algorithm, will also bring an exponential growth of computational complexity. In this paper, a scheme with serial codes for system order reduction in downlink scenarios, namely serial SCMA, is proposed to tackle these issues. Based on the idea of hierarchical coding, the processes of encoding and decoding in a high-order system are decomposed into several processes in low-order systems. Specific mapping modules are designed to recover the original user binary bits. Simulation results show that serial SCMA will substantially cut down the detection complexity and enjoy better link performance while maintaining sparser codebooks. Besides, it still inherits the advantages of original SCMA systems such as overloading and average aggregate energy efficiency over OMA and other NOMA schemes.

INDEX TERMS High order SCMA systems, serial SCMA, pattern matrix, Log-MPA, detection complexity.

I. INTRODUCTION

In the future Internet of Things (IoT) scenarios, both people and things will connect to the networks. The fifth generation (5G) communication systems should accommodate higher demands including massive connections, high spectrum utilization, lower latency, etc [1]. For example, 5G requires 1 million device connections per square kilometer which can not be met by conventional orthogonal multiple access (OMA) schemes due to their orthogonality of time/frequency resources. Therefore, some potential techniques are proposed to meet these demands in 5G such as millimeter wave communications, device to device (D2D) communications, all-spectrum access, ultra dense networks (UDN), non-orthogonal multiple access (NOMA) and so on [2]. Compared with conventional OMA, NOMA will achieve overloading while suffering from multiple interferences and increasing receiver complexity [3].

Many researchers have done ongoing studies toward NOMA. A competitive low density signature (LDS) scheme which limits the maximum degrees of each chip is proposed in [4]. Some other NOMA techniques such as multi-user shared access (MUSA) [5], bit division multiplexing (BDM) [6], pattern division multiple access (PDMA) [7] and sparse code multiple access (SCMA) [8] have also been proposed. Compared with other NOMA techniques, SCMA enjoys better link performance [9].

In SCMA, the encoders straightly convert incoming bit streams into complex codewords with multiple dimensions. All the codewords are chosen from a predesigned codebook set. Each user is assigned a special codebook. These codebooks have sparse structures such that the message passing algorithm (MPA) detector can decode the overlapped codewords with a reasonable computational complexity [10]. Both SCMA and LDS systems can achieve overloading. However, in SCMA information bits are carried over multi-dimensional complex constellations while in LDS they are carried over spread quadrature amplitude modulation (QAM) symbols. This brings SCMA potential shaping gain. Different from conventional code domain SCMA, the authors propose power domain SCMA (PSMA) in [11]. Both MPA and successive interference cancellation (SIC) are used in decoding. The spectral efficiency can be improved through codebook reuse while suffering from a higher computational complexity.

In general, SCMA has the advantages of overloading, sparse codewords and shaping gain contributing to multiple dimensions. On the other hand, the detection complexity of MPA iterations grows exponentially especially under high user load due to the brute force search.

A. LITERATURE

There have been extensive researches on SCMA codebook design and receiver algorithms.

Constellations with low projections [12], spherical codebooks [13], QAM based constellations [14], [15] and constellation rotation [16] are proposed to improve the coding gain of SCMA systems. Besides, an improved SCMA scheme in which each specific user may contribute only one symbol during each channel use is proposed in [17]. It can also support higher user load and enjoy better link performance at the expense of larger signaling overheads.

To provide guidelines for SCMA codebook design, some references focus on the system data rate. The balance between rate and energy for SCMA is studied in [18]. The sum rate of uplink SCMA is studied in [19] and a method of codebook design based on maximizing this index is also proposed. The cutoff rate of downlink SCMA is derived in [20] and the authors also aim to find more beneficial codebooks through the rate analysis.

For multi-user detection, based on the low-projection structure of SCMA codebooks, a low complexity detector is proposed in [21]. This scheme can save more than 60% computational complexity with negligible performance loss (0.5dB). To reduce the computational complexity of multiplications and exponent arithmetics, the authors expand the conventional MPA to the logarithm domain in [22]. Multiplications and exponent arithmetics are converted into simple additions and comparisons. Based on partial marginalization (PM), an improved MPA detector which will reduce the complexity by fixing some symbols in each iteration at the expense of some performance loss is proposed in [23]. Instead of the iterative process in MPA, the authors introduce weight factors to obtain the final probability distribution in [24]. Based on tree search, a low complexity detection scheme which exploits the lattice characteristic of SCMA codebooks is proposed in [25]. However, this search is only operated on each orthogonal resource and may bring a sub-optimal solution. Motivated by the Monte Carlo Markov Chain (MCMC), the authors design a low complexity SCMA detector in [26].

In high order SCMA systems with larger overloading factors, the codebook design is more complicated. The above detection methods will not reduce the degrees of each orthogonal resource and the codebook dimensions fundamentally. Therefore, they may merely cut down the exponential computational complexity to a limited degree while enduring performance loss. The specific methods of applying SCMA in real physical scenarios have not come to an agreement. A scheme to enable high order SCMA systems with moderate computational complexity while maintaining high performance is of great importance to accommodate the demands in 5G systems.

B. CONTRIBUTION

To further address the problems in high order SCMA systems, we apply the idea of hierarchical coding. For NOMA systems, discrete hierarchical modulations which consists of non-uniformly spaced constellation points, is designed to provide various levels of error protection for overlapped signals in [27]. Considering that non-uniformly constellations will improve link performance, in this paper, a coding scheme for SCMA system order reduction in downlink scenarios, namely serial SCMA, is proposed.

In serial SCMA, the information per user is transmitted through multiple serial low order subsystems. The overloading factor of the whole SCMA system is approached step by step by these subsystems. The signal between each group of adjacent subsystems can be regarded as a complex vector and the coding schemes of the adjacent subsystems should satisfy specific mapping rules. A low complexity receiver is also designed for the proposed serial SCMA system to recover the original binary bits.

Our main contributions are exhibited as follows:

- **Codebook Design Simplification**: The proposed serial SCMA system decreases the system order and maintain sparser codebooks. This also simplifies the codebook design.
- **Computational Complexity Reduction**: The total computational complexity of the proposed serial SCMA system mainly depends on the computational complexity of each low order subsystem and is reduced greatly especially under high user load.
- **Lower Block Error Rate**: In high order SCMA systems with larger codebook sizes, more dense constellations will also bring higher block error rate (BLER) especially under high user load. Based on the idea of system order reduction, the proposed SCMA system will enjoy lower BLER than the original SCMA system under the same user load.
- **Good Inheritance of Original SCMA**: The proposed SCMA system still inherits the advantages of the original SCMA system over OMA and other NOMA schemes.

C. STRUCTURE

The remainder of this paper is organized as follows. Section II illustrates original SCMA with channel coding. The key idea of this paper is also introduced. Section III proposes the serial SCMA system. We also discuss its transmitter design in this part. Section IV describes the receiver design of the proposed

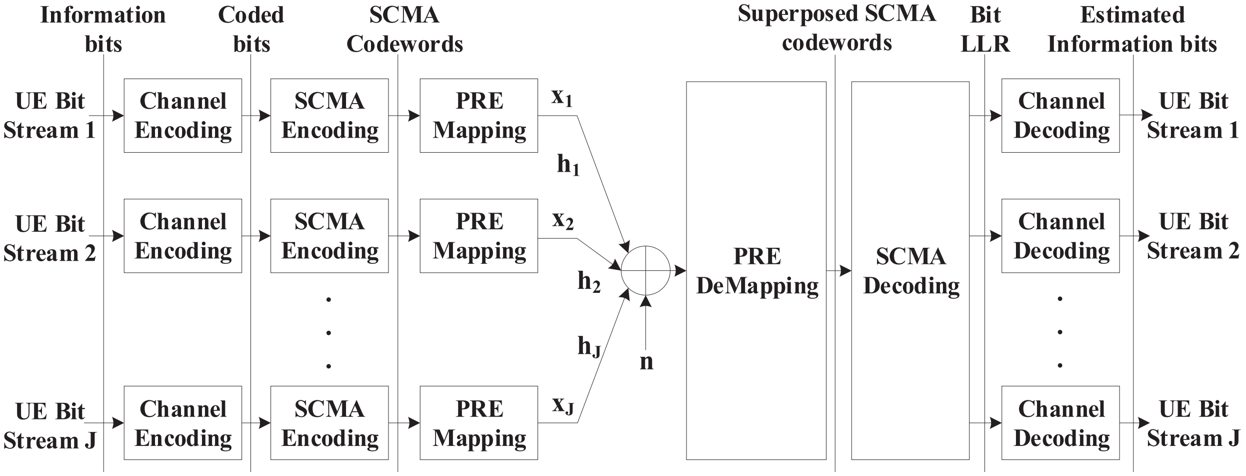

FIGURE 1. Original SCMA system model with channel coding.

serial SCMA and discusses corresponding low complexity detection algorithms. Section V derives some performance indices of different SCMA systems. Section VI presents the related simulation results where the proposed serial SCMA is compared with other schemes such as long term evolution (LTE), LDS, original SCMA and PSMA. Section VII summarizes the whole paper.

D. NOTATIONS

Throughout this paper, the symbols x, $\mathbf{x}$ and $\mathbf{X}$ stand for a scalar, a vector and a matrix, respectively. $\mathbf{X}^T$ and $\mathbf{X}^\dagger$ denote the transpose and conjugate transpose of matrix $\mathbf{X}$. diag($\mathbf{x}$) stands for a diagonal matrix in which all its diagonal elements consists of the vector $\mathbf{x}$. diag($\mathbf{X}$) denotes a vector whose elements are chosen from the diagonal elements of matrix $\mathbf{X}$. $\mathbb{B}$ $\mathbb{Z}$ and $\mathbb{C}$ represents the set of binary values, integer values and complex values, respectively. $\binom{K}{N}$ represents the number of N-combinations selecting from K elements. det($\cdot$) stands for the determinant of a square matrix. $e^{i(\cdot)}$ represents the complex exponential operation.

II. ORIGINAL SCMA AND KEY IDEA

This part mainly describes original SCMA and defines primary parameters.

Fig. 1 in the next page illustrates a coded SCMA system. J users transmit their signals through K shared orthogonal resources. SCMA decoders process the overlapped signals from each user through the near optimal MPA detector and exchange the extrinsic soft information of coded bits with channel decoders. The final estimated bits can be obtained from the channel decoders.

Denote the index of each user by $j \in \{1, 2, \ldots, J\}$. For user j, m_0 binary bits $\mathbf{b}_j$ are first encoded into n_0 binary bits $\mathbf{c}_j$ by channel encoding with code rate $R_0 = \frac{m_0}{n_0}$. Then, every $\log_2 M$ bits in $\mathbf{c}_j$ are mapped into an N-dimensional constellation point by an SCMA encoder, where M stands for the cardinality of multi-dimensional constellations and N represents the number of orthogonal resources occupied by each user. Furthermore, the N-dimensional constellation point is converted into a K-dimensional complex codeword by the mapping matrix $\mathbf{V}_j$, which generates the codeword set $\mathbf{X}_j$ and is given in [8]. The dimensions of $\mathbf{V}_j$ is $K \times N$ and we can regard SCMA encoding as a mapping from a binary vector $\mathbf{b}_j$ to a complex signal vector $\mathbf{x}_j$ following f : $\mathbb{B}^{\log_2 M} \rightarrow \mathcal{X}$ for user j, where $\mathcal{X} \in \mathbb{C}^K$.

In the base station, the received signals is given as follows.

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

where $\mathbf{h}_j = [h_{j,1}, h_{j,2}, \ldots, h_{j,K}]^T$ stands for the channel coefficient vector between the base station and user j. $\mathbf{n} = [n_1, n_2, \ldots, n_K]^T$ is a K-dimensional Gaussian noise vector. Its distribution follows $\mathcal{CN}(\mathbf{0}, \sigma^2\mathbf{I})$. σ^2 denotes the variance of each element of $\mathbf{n}$ and $\mathbf{I}$ is the identity matrix. $\mathbf{x}_j = [x_{j,1}, x_{j,2}, \ldots, x_{j,K}]^T$ stands for the K-dimensional sparse codeword of user j.

In downlink scenarios, all the users send their signals from the same transmit point and share the same channel conditions $\mathbf{h}$, i.e. $\mathbf{h}_j = \mathbf{h}, \forall j \in \{1, 2, \ldots, J\}$. Considering that each $\mathbf{x}_j$ has only N non zero elements, we can rewrite (1) as

$$\begin{aligned}\mathbf{y} &= \mathrm{diag}(\mathbf{h}) \sum_{j=1}^{J} \mathbf{V}_j\mathbf{s}_j + \mathbf{n} \\ &= \mathrm{diag}(\mathbf{h})\mathbf{V}\mathbf{s} + \mathbf{n},\end{aligned} \tag{2}$$

where $\mathbf{s}_j = [s_{j,1}, \ldots, s_{j,N}]^T$ is a N-dimensional complex codeword of user j which dose not include zero elements and $\mathbf{x}_j = \mathbf{V}_j\mathbf{s}_j$. $\mathbf{V} = [\mathbf{V}_1, \ldots, \mathbf{V}_J]$ is a $K \times NJ$ matrix and $\mathbf{s} = [\mathbf{s}_1^T, \mathbf{s}_2^T, \ldots, \mathbf{s}_J^T]^T$ is an $NJ \times 1$ column vector.

A sparse factor graph matrix $\mathbf{F}$ which has K rows and J columns can be used to represent an SCMA coding scheme. User node (UN) j occupies resource node (RN) k if and only if $f_{kj} = 1$. Each UN employs a K-dimensional complex codebook and occupies only N effective RNs while each RN is shared by d_f UNs. From [8], we can obtain the following relations: $K \leq J \leq \binom{K}{N}$, $d_f = \frac{JN}{K}$.

A factor graph with 6 UNs and 4 RNs is depicted in Fig. 2.

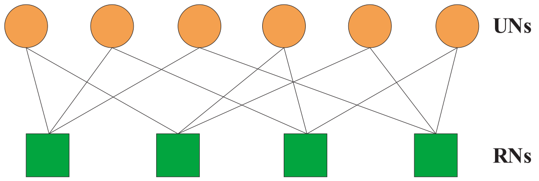

FIGURE 2. Factor graph of $J = 6$ UNs colliding over $K = 4$ RNs for MPA iterations.

In this paper, a single SCMA system is represented by SCMA(K, N, M). The key idea of system order reduction can be expressed as follows.

After N is determined, the overloading factor of an SCMA system mainly depends on the degrees of each RN d_f, which will incur high detection complexity and awful BLER. Reduction of d_f will substantially reduce the computational complexity and control the link performance within an applicable range. We propose a serial coding scheme which can decompose d_f of high order SCMA systems.

III. SERIAL SCMA SYSTEMS

In this part, we will introduce the system with serial codes to come through SCMA system order reduction.

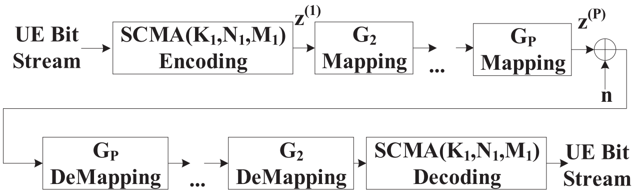

FIGURE 3. Serial SCMA system model.

Fig. 3 in the next page illustrates the block diagram of a serial SCMA system which consists of a single SCMA system and $P - 1$ pattern matrix $\mathbf{G}_p, p \in \{2, 3, \ldots, P\}$ modules. The dimension of each $\mathbf{G}_p$ corresponding to a single SCMA(K_p, N_p, M_p) system is set as $K_p \times J_p$, where $J_p = K_{p-1}, p \in \{2, 3, \ldots, P\}$. Moreover, each $\mathbf{G}_p$ has the same non-zero element positions as the factor graph matrix of SCMA(K_p, N_p, M_p). In a serial SCMA system, the binary bits per user are directly mapped into complex codewords of SCMA(K_1, N_1, M_1) at first and then processed by multiple serial subsystems $\mathbf{G}_p$.

The capacity gain of the high order SCMA system is approached step by step by these subsystems. Denote the overloading factor of the whole system by λ_{se}. We can obtain that

$$\lambda_{\text{se}} = \lambda_1 \prod_{p=2}^{P} \lambda_p, \tag{3}$$

where $\lambda_1 = \frac{J_1}{K_1}$ and $\lambda_p = \frac{J_p}{K_p}$ are the overloading factors of SCMA(K_p, N_p, M_p) and each $\mathbf{G}_p$, respectively. Given λ_{se}, we should select reasonable λ_1, λ_p and P. The overloading factor of each subsystem can not be set too high to guarantee the low order property.

The signals between adjacent subsystems can be regarded as a complex vector $\mathbf{z}^{(p)}$, which is given as follows.

$$\mathbf{z}^{(p)} = \mathbf{G}_p \mathbf{z}^{(p-1)}, \quad p \in \{2, 3 \ldots, P\}. \tag{4}$$

Define $\mathbf{z}^{(1)} = \mathbf{V}\mathbf{s}$, where $\mathbf{V}$ and $\mathbf{s}$ is defined as Section II. Their dimensions are $K_1 \times N_1 J_1$ and $N_1 J_1 \times 1$, respectively. The pattern matrices of the adjacent subsystems must satisfy specific mapping rules to make sure that the whole SCMA system can be decoded. In brief, the problem of enabling a high order SCMA(K, N, M) system can be broken down into several lower order systems which have sparser factor graphs. Furthermore, we can continue breaking down the lower order systems until they are practically tractable. From this perspective, serial codes can be expended to large-scale systems.

A. SERIAL SCMA CODEBOOK DESIGN

Regarding the transmitter design of serial SCMA, we can get the codebook set of SCMA(K_1, N_1, M_1) at first. Although there has been a lot of references about SCMA codebook design, the optimal design criterion is undetermined. In this paper, we mainly study the downlink scenarios where each UN shares the same channel coefficients over the d_f RNs. For the low order SCMA codebook design, we select a method which is widely used in most references. Hence, the minimum Euclidian distance $d_{\min}$ is a key performance indicator (KPI). Given $d_{\min}$ with a minimum energy E_s, we can always construct SCMA codewords from any complex constellation points. These constellations also have the minimized figure of merit $\eta = \frac{E_s}{d_{\min}}$. To simplify SCMA codebook design, the constellations of all the users can be built up through a mother constellation and user-specific operators.

The mother multi-dimensional constellation design can refer to some classic requirements in communication systems:

1) The minimum Euclidean distance between adjacent constellation points should be maximized while all the symbols can evenly cover the complex vector plane [28].

2) The binary bits of any adjacent constellation points should follow the Gray map rule and possess the minimum Hamming distance.

3) The symbols should have the maximum varieties of both in-phase and quadrature components [29]. Even if one or more symbols are lost, we can restore the original bit stream through different symbol quadrature interleavers.

4) When changing symbols, zero crossing should be avoided and the symbol modulus should be constant. Thus we can cut down the linear requirements of transmitters.

As an alternative method, the N_1-dimensional mother constellation can be built up through Cartesian product of N_1 independent QAM constellation sets. Lattice rotations can be operated on the constellations while maintaining its Euclidian distance fixed. Across the dimensions of codewords, power imbalance can bring near-far effect among overlapped users. It helps SCMA decoders to remove interferences among overlapped users more effectively.

A subset of lattice $\mathbb{Z}^2$ is given by [16]

$$\mathbf{S}_1 = \sqrt{2}e^{i\cdot\frac{\pi}{4}}[1-M_1, 3-M_1, \ldots, M_1-1]^T. \tag{5}$$

Set $\mathbf{S}_l = \mathbf{U}_l\mathbf{S}_1, \mathbf{U}_l = \text{diag}(\mathbf{1}e^{i\theta_{l-1}}) \in \mathbb{C}^{M_1\times M_1}$, where $\mathbf{1}$ is an $M_1 \times 1$ all one vector. θ_{l-1} can be given as follows.

$$\theta_{l-1} = \frac{(l-1)\pi}{M_1N_1}, \quad \forall l \in \{1, 2, \ldots, N_1\}. \tag{6}$$

Then we can build up a N_1-dimensional mother constellation as follows.

$$\mathbf{M}_c = [\mathbf{S}_1, \ldots, \mathbf{S}_{N_1}]^T = \begin{bmatrix} s_{11} & s_{12} & \cdots & s_{1M_1} \\ s_{21} & s_{22} & \cdots & s_{2M_1} \\ \vdots & & & \\ vdots & \cdots & \vdots & \\ s_{N_11} & s_{N_12} & \cdots & s_{N_1M_1} \end{bmatrix}. \tag{7}$$

After obtaining the mother constellation, we should determine the user-specific constellation operators. Three typical operators [30] are listed as follows:

1) Phase rotation

$$(\circ : \phi)z := e^{i\phi}z;$$

2) Complex conjugate

$$(* : r) := \begin{cases} z, & r = 0 \\ z^*, & r = 1; \end{cases}$$

3) Vector permutation

$$(\otimes : \pi)z := \pi z,$$

where π stands for the permutation operator. For user j, the constellation operators Δ_j is then given as follows.

$$\begin{aligned} \Delta_j &= (\otimes \circ * : \pi\phi r)z = (\otimes : \pi)(\circ : \phi)(* : r) \\ &= \pi\text{diag}(e^{i\phi})(* : r) : z. \end{aligned} \tag{8}$$

Then the codebook of user j can be expressed as,

$$\mathbf{X}_j = \mathbf{V}_j\Delta_j\mathbf{M}_c, \tag{9}$$

where $\mathbf{V}_j$ is a $K_1 \times N_1$ mapping matrix.

B. PATTERN MATRIX DESIGN

Compared with non-linear modulation in SCMA, the pattern matrices $\mathbf{G}_p, p \in \{2, 3, \ldots, p\}$ can be regarded as linear modulators. They are designed to provide more diversity between multiple users. A pattern matrix should be sparse, so that the iterative algorithm can be applied in decoding conveniently. It also plays a role of power distribution and phase rotation. The serial codes combine non-linear and linear modulations, i.e. hybrid modulation, which can improve the shaping gain contributing to the non-uniform distribution of equivalent constellations.

The pattern matrix design criteria can be concluded as follows:

1) $\mathbf{G}_p$ should have a sparse factor graph in accordance with its overloading factor λ_p. Although N_p with high values will bring better performance, the solvability of the system should also be considered.

2) After the factor graph of $\mathbf{G}_p$ is determined, we should find proper power scaling and phase rotation factors.

As a special case, if we only consider the phase rotation difference and each $\mathbf{G}_p$ has the same non zero element positions as the factor graph matrix of regular SCMA(K_p, N_p, M_p), $\mathbf{G}_p$ can be set as a Latin generator matrix referred in [16]. An example of $\mathbf{G}_p$ with dimensions of 4×6 which depends on M_p is given as follows.

$$\mathbf{G}_p = \begin{bmatrix} \varphi_0 & \varphi_1 & \varphi_2 & 0 & 0 & 0 \\ \varphi_1 & 0 & 0 & \varphi_0 & 0 & \varphi_2 \\ 0 & \varphi_2 & 0 & 0 & \varphi_1 & \varphi_0 \\ 0 & 0 & \varphi_0 & \varphi_1 & \varphi_2 & 0 \end{bmatrix}. \tag{10}$$

$\varphi_u = e^{i(\frac{2\pi u}{M_p d_{f_p}} + \frac{2\pi w}{M})}$, where $u = 0, 1, \ldots, d_{f_p} - 1, w \in \mathbb{Z}$ and $d_{f_p} = \frac{J_pN_p}{K_p}$.

Furthermore, if both power scaling and phase rotation are considered, we can select $\mathbf{G}_p$ by referring to the pattern matrix design of a particular PDMA system [31]. For this case, $\mathbf{G}_p$ may have an irregular factor graph. An example of $\mathbf{G}_p$ with dimensions of 4×6 which is independent of M_p is given as follows.

$$\mathbf{G}_p = \begin{bmatrix} 1 & 0 & 0 & 0.577 & -0.577 & 0 \\ 0 & 1 & 0 & 0.577 & 0.577 & 0 \\ 0 & 0 & 1 & 0.577 & 0.577 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}. \tag{11}$$

3) In the whole serial SCMA system, each user has an equivalent constellation which mainly depends on all the subsystems. An equivalent factor graph for the whole system also exists. We can assign all the users different diversity orders through the selection of $\mathbf{G}_p$. From this point of view, the matching of each $\mathbf{G}_p$ and the SCMA(K_1, N_1, M_1) codebook set should also be considered.

IV. SERIAL SCMA RECEIVER DESIGN

The factor graphs of adjacent subsystems $\mathbf{F}_{p-1}$ and $\mathbf{F}_p$ in serial SCMA are depicted in Fig. 4, where $p \in \{2, 3, \ldots, P\}$. $\mathbf{F}_{p-1}$ and $\mathbf{F}_p$ make up of a two-stage bipartite graph which includes UNs, connection nodes (CNs) and RNs. Suppose that the receiver knows the channel condition perfectly. The MPA detector with acceptable complexity transmits messages through the factor graph iteratively. In [32], the authors propose that channel and SCMA decoders exchange soft decision information to improve link performance.

Define the a posteriori log-likelihood ratio (LLR) as the soft decision messages between SCMA and

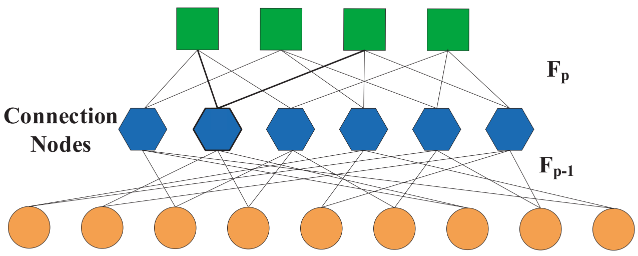

FIGURE 4. **Factor graph of adjacent subsystems.**

channel decoders. It can be given by

$$\Lambda(b_{j,m}) = \log \frac{p(b_{j,m}=1|\mathbf{y})}{p(b_{j,m}=0|\mathbf{y})}, \tag{12}$$

where $b_{j,m}$ is the m-th bit from user j and $\mathbf{y}$ stands for the received signal.

Based on Bayes rule, (12) can be converted into

$$\begin{aligned}\Lambda(b_{j,m}) &= \log \frac{p(\mathbf{y}|b_{j,m}=1)}{p(\mathbf{y}|b_{j,m}=0)} + \log \frac{p(b_{j,m}=1)}{p(b_{j,m}=0)} \\ &= \Lambda_e(b_{j,m}) + \Lambda_a(b_{j,m}),\end{aligned} \tag{13}$$

where $\Lambda_e(b_{j,m})$ and $\Lambda_a(b_{j,m})$ represent the extrinsic information transmitted by the SCMA decoder and the a priori LLR, respectively. Before the initial iteration process, we can set $\Lambda_a(b_{j,m}) = 0$. $\Lambda_e(b_{j,m})$ can be obtained by exhaustively searching all possible symbol combinations.

$\mathbf{G}_p$, $\mathbf{y}$ and $\mathbf{n}$ contain complex values. For a serial SCMA system, its received signals $\mathbf{y}$ is given by

$$\mathbf{y} = \sum_{j=1}^{J} \operatorname{diag}(\mathbf{h}'_j)\mathbf{G}_P\mathbf{G}_{P-1}\ldots\mathbf{G}_2\mathbf{V}_j\mathbf{s}_j + \mathbf{n}, \tag{14}$$

where $\mathbf{h}'_j = [h'_{j,1}, \ldots, h'_{j,K}]^T$ denotes the channel conditions between UN j and each RN. $\mathbf{V}_j$ and $\mathbf{s}_j$ are defined as Section II. The dimensions of $\mathbf{G}_p$ ($p \in \{2, 3, \ldots, P\}$), $\mathbf{V}_j$ and $\mathbf{s}_j$ are $K_p \times J_p$, $V \times N_1$ and $N_1 \times 1$, respectively. Each complex equation can be converted into two real equations.

For the downlink scenario in which $\mathbf{h}'_j = \mathbf{h}', \forall j \in \{1, 2, \ldots, J\}$, the receiver design can be simplified and (14) is converted into

$$\begin{aligned}\mathbf{y} &= \operatorname{diag}(\mathbf{h}')\mathbf{G}_P\mathbf{G}_{P-1}\ldots\mathbf{G}_2 \sum_{j=1}^{J}\mathbf{V}_j\mathbf{s}_j + \mathbf{n} \\ &= \mathbf{H}_2\mathbf{V}\mathbf{s} + \mathbf{n},\end{aligned} \tag{15}$$

where $\mathbf{H}_2 = \operatorname{diag}(\mathbf{h}')\mathbf{G}_P\mathbf{G}_{P-1}\ldots\mathbf{G}_2$ is a $K_P \times K_1$ matrix and $\mathbf{V}\mathbf{s} = \sum_{j=1}^{J}\mathbf{V}_j\mathbf{s}_j$. The dimensions of $\mathbf{V}$ and $\mathbf{s}$ are $V \times N_1\ J$ and $N_1\ J \times 1$, respectively.

In conventional SCMA systems, the maximum overloading can reach 300% due to the tradeoff between overloading and link performance [33] while in serial SCMA only two subsystems are needed to reach this requirement. In order to contrast with conventional SCMA, we firstly consider a two-stage serial SCMA system ($P = 2$) which includes J UNs, V CNs, and K RNs, i.e. $J_1 = J$, $J_2 = K_1 = V$, $K_2 = K$. The optimal ML detection can be represented by

$$\widehat{\mathbf{s}} = \operatorname*{argmin}_{\mathbf{s}} ||\mathbf{y} - \mathbf{H}_2\mathbf{V}\mathbf{s}||^2, \tag{16}$$

where $\widehat{\mathbf{s}}$ denotes the detected symbols and $\mathbf{H}_2 = \operatorname{diag}(\mathbf{h}')\mathbf{G}_2$ is a $K \times V$ matrix. Since the ML detection endures extremely high computational complexity, we propose a detection scheme which enjoys lower complexity and can still achieve the optimal ML performance.

Obviously, (16) stands for an undetermined system. Motivated by the analytical methods in [34], we can set $\mathbf{H}_2$ as $\mathbf{H}_2 = \operatorname{diag}(\mathbf{h}')\mathbf{G}_2 = [\mathbf{H}^A_{2,K\times K}, \mathbf{H}^B_{2,K\times(V-K)}]$. The modified pattern matrix is given as follows.

$$\widetilde{\mathbf{H}}_{2,V\times V} = \begin{bmatrix} \mathbf{H}^A_{2,K\times K} & \mathbf{H}^B_{2,K\times(V-K)} \\ \mathbf{0}_{(V-K)\times K} & \mathbf{I}_{(V-K)\times(V-K)} \end{bmatrix}. \tag{17}$$

In the same way, rewrite $\mathbf{z}^{(1)}$ as $\begin{bmatrix} \mathbf{z}^{(1),A}_{K\times 1} \\ \mathbf{z}^{(1),B}_{(V-K)\times 1} \end{bmatrix}$. The modified received signal is given as follows.

$$\widetilde{\mathbf{y}} = \widetilde{\mathbf{H}}_2\mathbf{z}^{(1)} + \widetilde{\mathbf{n}}, \tag{18}$$

$$\begin{bmatrix} \mathbf{y}_{K\times 1} \\ \mathbf{0}_{(V-K)\times 1} \end{bmatrix} = \widetilde{\mathbf{H}}_2 \begin{bmatrix} \mathbf{z}^{(1),A}_{K\times 1} \\ \mathbf{z}^{(1),B}_{(V-K)\times 1} \end{bmatrix} + \begin{bmatrix} \mathbf{n}_{K\times 1} \\ -\mathbf{z}^{(1),B}_{(V-K)\times 1} \end{bmatrix}. \tag{19}$$

$\mathbf{G}_2$ should satisfy $\det(\mathbf{G}^A_{2,K\times K}) \neq 0$. Then $\det(\mathbf{H}^A_{2,K\times K}) = \det(\operatorname{diag}(\mathbf{h}')) \cdot \det(\mathbf{G}^A_{2,K\times K}) \neq 0$ and $\det(\widetilde{\mathbf{H}}_2) = \det(\mathbf{H}^A_{2,K\times K}) \neq 0$. Now $\widetilde{\mathbf{H}}_2$ is a full rank matrix, the two sides of (19) can be multiplied by $\widetilde{\mathbf{H}}_2^{-1}$ on the left,

$$\begin{aligned}&\widetilde{\mathbf{H}}_2^{-1}\begin{bmatrix} \mathbf{y}_{K\times 1} \\ \mathbf{0}_{(V-K)\times 1} \end{bmatrix} \\ &\quad = \mathbf{z}^{(1)} + \widetilde{\mathbf{H}}_2^{-1}\begin{bmatrix} \mathbf{n}_{K\times 1} \\ -\mathbf{z}^{(1),B}_{(V-K)\times 1} \end{bmatrix}.\end{aligned} \tag{20}$$

$$\begin{aligned}\widetilde{\mathbf{H}}_2^{-1} &= \begin{bmatrix} (\mathbf{H}^A_{2,K\times K})^{-1} & -(\mathbf{H}^A_{2,K\times K})^{-1}\mathbf{H}^B_{2,K\times(V-K)} \\ \mathbf{0}_{(V-K)\times K} & \mathbf{I}_{(V-K)\times(V-K)} \end{bmatrix} \\ &= \begin{bmatrix} (\mathbf{G}^A_2)^{-1}\operatorname{diag}(\mathbf{h}')^{-1} & -(\mathbf{G}^A_2)^{-1}\mathbf{G}^B_2 \\ \mathbf{0}_{(V-K)\times K} & \mathbf{I}_{(V-K)\times(V-K)} \end{bmatrix},\end{aligned} \tag{21}$$

where $\mathbf{G}_2 = [\mathbf{G}^A_2, \mathbf{G}^B_2]$. The dimensions of $\mathbf{G}^A_2$ and $\mathbf{G}^B_2$ are $K \times K$ and $K \times (V-K)$, respectively.

Considering that $\mathbf{z}^{(1)} = \mathbf{V}\mathbf{s}$, the ML detection problem can be expressed as

$$\begin{aligned}\widehat{\mathbf{s}} = \operatorname*{argmin}_{\mathbf{s}}(&||\widetilde{\mathbf{H}}_2^{-1}\widetilde{\mathbf{y}} - \mathbf{V}\mathbf{s}||^2 \\ &- ||\widetilde{\mathbf{H}}_2^{-1}\begin{bmatrix} \mathbf{0}_{K\times 1} \\ (\mathbf{V}\mathbf{s})^B_{(V-K)\times 1} \end{bmatrix}||^2).\end{aligned} \tag{22}$$

(22) stands for a well-defined system of equations. $(\mathbf{V}\mathbf{s})^B_{(V-K)\times 1}$ represents the last $V-K$ rows of $\mathbf{z}^{(1)} = \sum_{j=1}^{J}\mathbf{x}_j$. When $p = 2$, conventional Log-MPA detection is described in Algorithm 1.

$f_{1,vj}$ denotes the (v,j)-th element of $\mathbf{F}_1$, where $\mathbf{F}_1$ is the factor graph matrix of SCMA(K_1, N_1, M_1). $N_U(v)$ stands for

Algorithm 1 Conventional Log-MPA Detection When $p = 2$

```
Input: $\mathbf{y}, \mathbf{h}', \mathbf{G}_2, \mathbf{V}, T_{\max}$.
Output: $\mathbf{b}_j, j = 1, 2, \ldots, J$.
Calculate and store $\widetilde{\mathbf{H}}_2$, $\widetilde{\mathbf{H}}_2^{-1}$ and $\widetilde{\mathbf{y}}$.
for all $j \in 1, 2, \ldots J$ and $v \in 1, 2, \ldots V$ $(f_{1,vj} \neq 0)$
Set $t = 0$, $L^0_{g_v \to u_j}(\mathbf{s}_j) = 0$, $L^0_{u_j \to g_v}(\mathbf{s}_j) = 0$.
end for
while $t < T_{\max}$ do
    for all $j \in N_U(v)$, $v \in N_C(j)$, and $\mathbf{s}_j$
    $L^t_{u_j \to g_v}(\mathbf{s}_j) = \sum_{v' \in N_C(j) \backslash v} L^{t-1}_{g_{v'} \to u_j}(\mathbf{s}_j)$,
    $L^t_{u_j \to g_v}(\mathbf{s}_j) = L^t_{u_j \to g_v}(\mathbf{s}_j) - \max^*_{\mathbf{s}_j}\{L^t_{u_j \to g_v}(\mathbf{s}_j)\}$.
    end for
    for all $j \in N_U(v)$, $v \in N_C(j)$, and $\mathbf{s}_j$
    $\mathbf{d}(\mathbf{s}) = \widetilde{\mathbf{H}}_2^{-1} \begin{bmatrix} \mathbf{y}_{K \times 1} \\ (\mathbf{Vs})^B_{(V-K) \times 1} \end{bmatrix} - \mathbf{Vs}$,
    $L^t_{g_v \to u_j}(\mathbf{s}_j) = \max^*_{\mathbf{s}_{N_U(v) \backslash j}}\{-\frac{1}{2\sigma^2}||d(\mathbf{s})_v||^2 + \sum_{j' \in N_U(v) \backslash j} L^t_{u_{j'} \to g_v}(\mathbf{s}_{j'})\}$,
    $L^t_{g_v \to u_j}(\mathbf{s}_j) = L^t_{g_v \to u_j}(\mathbf{s}_j) - \max^*_{\mathbf{s}_j}\{L^t_{g_v \to u_j}(\mathbf{s}_j)\}$,
    where $d(\mathbf{s})_v$ stands for the $v$-th component of $\mathbf{d}(\mathbf{s})$.
    end for
$t = t + 1$
end while
Compute all $I(\mathbf{s}_j)$,
$I(\mathbf{s}_j) = \log(p(\mathbf{s}_j)) + \sum_{v \in N_C(j)} L^{T_{\max}}_{g_v \to u_j}(\mathbf{s}_j)$,
where $p(\mathbf{s}_j)$ is the prior probability of $\mathbf{s}_j$ from channel decoders.
```

the index set of all the UNs corresponding to CN v and $N_C(j)$ stands for the index set of all CNs corresponding to the UN j. $L(\cdot)$ and $I(\cdot)$ are functions of LLR. The max* operation can be expressed as

$$\max^*(a, b) = \log(e^a + e^b). \tag{23}$$

The above analytical methods also apply for $P \geq 3$.

A. COMPLEXITY REDUCTION VIA GROUP MULTI-USER DETECTION

The conventional MPA detector can not utilize the coding structure of serial SCMA. Moreover, by adjusting the codebook set of SCMA(K_1, N_1, M_1) and $\mathbf{G}_2$, the gaps between power levels in each CN from different UNs' contribution can be enlarged, which can make interference cancellation more applicable. Based on the idea of hierarchical decoding, we can improve the detection algorithm.

Reducing the size of the factor graph is also an applicable method to come through complexity reduction. During the process of MPA iterations, if some UNs' symbols with higher belief values have the trend of convergence, their signals can be determined legitimately at first. Furthermore, through the cyclic redundancy check (CRC) module, we can observe the decoding conditions of each user after each iteration in good time. Motivated by SIC, we can cut off the correctly decoded UNs. Hence, the decoding process for the signals of the rest UNs can be simplified.

According to the factor graph of SCMA(K_1, N_1, M_1), we can divide the J UNs into three disjoint sets: the set $\{\text{UN}^1\}$ corresponding to the UNs which only occupy N_1 out of the first K CNs, the set $\{\text{UN}^2\}$ corresponding to the UNs which occupy N_1 CNs out of both the first K CNs and the last $V - K$ CNs and the set $\{\text{UN}^3\}$ corresponding to the UNs which only occupy N_1 out of the last $V - K$ CNs. Based above, we can modify the conventional Log-MPA detection as Algorithm 2.

Algorithm 2 Multi-User Detection in Groups With SIC When $p = 2$

```
Input: $\mathbf{y}, \mathbf{h}', \mathbf{G}_2, \mathbf{V}, d, T_{\max 1}, T_{\max 2}, \beta$.
Output: $I(\mathbf{s}_j), j = 1, 2, \ldots, J$.
for all $1 \leq j \leq J$ and $1 \leq v \leq V$ $(f_{1,vj} \neq 0)$
Set $t = 0$, $L^0_{g_v \to u_j}(\mathbf{s}_j) = 0$, $L^0_{u_j \to g_v}(\mathbf{s}_j) = 0$.
end for
for all $v \leq K$ and $j \in \{\text{UN}^1\}$ $(f_{1,vj} \neq 0)$
    while $t < T_{\max 1}$, $L^t_{g_v \to u_j}(\mathbf{s}_j) \leq \beta$ do
    Compute $L^t_{g_v \to u_j}(\mathbf{s}_j)$ and $L^t_{u_j \to g_v}(\mathbf{s}_j)$.
    $t = t + 1$
    end while
Compute $I(\mathbf{s}_j)$.
end for
Run CRC check for all the UNs in $\{\text{UN}^1\}$ and cut off the UNs in $\{\text{UN}^1\}$ with high belief values.
Set $t = 0$.
for all $v \leq K$ and $j \in \{\text{UN}^2\} \cup \{\text{UN}^3\}$ $(F_{1,vj} \neq 0)$
    while $t < T_{\max 2}$ do
    Compute $L^t_{g_v \to u_j}(\mathbf{s}_j)$ and $L^t_{u_j \to g_v}(\mathbf{s}_j)$.
    $t = t + 1$.
    end while
Compute $I(\mathbf{s}_j)$.
end for
```

The core thought of Algorithm 2 is to divide the factor graph into several subgraphs. The UNs in $\{\text{UN}^1\}$ and their corresponding edges make up of the first subgraph. The messages transmitted by other edges which belong to UNs in $\{\text{UN}^2\} \cup \{\text{UN}^3\}$ can be considered as interference signals. We can apply Gaussian approximation to these messages. During each iteration, both the expectation and the variance are updated as the feedback information.

The compensate vector in (22) is given by

$$\begin{aligned} \mathbf{f}_1(\mathbf{s}) &= -\widetilde{\mathbf{H}}_2^{-1} \begin{bmatrix} \mathbf{0}_{K \times 1} \\ (\mathbf{Vs})^B_{(V-K) \times 1} \end{bmatrix} \\ &= -\widetilde{\mathbf{H}}_2^{-1} \mathbf{I}_0 \mathbf{Vs} \end{aligned} \tag{24}$$

where $\mathbf{I}_0 = \begin{bmatrix} \mathbf{0}_{K \times K} & \mathbf{0}_{K \times (V-K)} \\ \mathbf{0}_{(V-K) \times K} & \mathbf{I}_{(V-K) \times (V-K)} \end{bmatrix}$ and $\mathbf{f}_1(\mathbf{s})$ is independent of UNs in $\{\text{UN}^1\}$. Thus the v-th element $(1 \leq v \leq K)$ of $\mathbf{f}_1(\mathbf{s})$ is given by $f_1(\mathbf{s})_v = -(\widetilde{\mathbf{H}}_2^{-1} \mathbf{I}_0 \sum_{j \notin \{\text{UN}^1\}} \mathbf{x}_j)_v$.

For the v-th CN ($1 \leq v \leq K$),

$$\begin{aligned} Y &= \sum_{j\in\{\mathrm{UN}^1\}} x_{j,v} + \sum_{j\in\{\mathrm{UN}^2\}} x_{j,v} + f_1(\mathbf{s})_v + (\widetilde{\mathbf{H}}_2^{-1}\widetilde{\mathbf{n}})_v \\ &= \sum_{j\in\{\mathrm{UN}^1\}} x_{j,v} + n'_v, \end{aligned} \tag{25}$$

where $Y = (\widetilde{\mathbf{H}}_2^{-1}\widetilde{\mathbf{y}})_v$ stands for the v-th component of vector $\widetilde{\mathbf{H}}_2^{-1}\widetilde{\mathbf{y}}$. $n'_v = \sum_{j\notin\{\mathrm{UN}^1\}} x_{j,v} + f_1(\mathbf{s})_v + (\widetilde{\mathbf{H}}_2^{-1}\widetilde{\mathbf{n}})_v$ and $n'_v \sim \mathcal{CN}(u_{n'_v}, \sigma^2_{n'_v})$.

During the t-th iteration, the mean $u_{n'_v}$ is given by

$$\begin{aligned} u^t_{n'_v} = ((\mathbf{I} - \widetilde{\mathbf{H}}_2^{-1}\mathbf{I}_0) &\sum_{j\in\{\mathrm{UN}^2\}} \sum_{\mathbf{x}_j\in\mathbb{X}_j} P^t(\mathbf{x}_j)\mathbf{x}_j \\ &+ (-\widetilde{\mathbf{H}}_2^{-1}\mathbf{I}_0) \sum_{j\in\{\mathrm{UN}^3\}} \sum_{\mathbf{x}_j\in\mathbb{X}_j} P^t(\mathbf{x}_j)\mathbf{x}_j)_v, \end{aligned} \tag{26}$$

where $\mathbf{I}$ is a $V \times V$ identify matrix and $\mathbb{X}_j$ stands for the set which includes all the possible values of $\mathbf{x}_j$. $P^t(\mathbf{x}_j) = p(\mathbf{s}_j)e^{L^t_{u_j\rightarrow g_v}(\mathbf{x}_j)}$ is the probability of $\mathbf{x}_j$ during the t-th iteration.

Denote $\mathbf{x}^1_j = (\mathbf{I} - \widetilde{\mathbf{H}}_2^{-1}\mathbf{I}_0)\mathbf{x}_j$ and $\mathbf{x}^2_j = -\widetilde{\mathbf{H}}_2^{-1}\mathbf{I}_0\mathbf{x}_j$. The variance $\sigma^{2,t}_{n'_v}$ is given by

$$\begin{aligned} \sigma^{2,t}_{n'_v} = &\sum_{j\in\{\mathrm{UN}^2\}} \mathrm{Var}(x^1_{j,v}) \\ &+ \sum_{j\in\{\mathrm{UN}^3\}} \mathrm{Var}(x^2_{j,v}) + ||\widetilde{\mathbf{h}}_{2,v}||^2\sigma^2, \end{aligned} \tag{27}$$

where $x^1_{j,v}$ and $x^2_{j,v}$ represents the v-th component of $\mathbf{x}^1_j$ and $\mathbf{x}^2_j$, respectively. $\mathrm{Var}(\cdot)$ stands for the variance function. $\widetilde{\mathbf{h}}_{2,v}$ represents the v-th column of $\widetilde{\mathbf{H}}_2^{-1}$.

The CN-to-UN update can be rewritten as

$$\begin{aligned} L^t_{g_v\rightarrow u_j}(s_j) = \max_{\substack{\mathbf{s}_{j'}:j'\neq j \\ j'\in\{\mathrm{UN}^1\}}}{}^{*} \{&-\frac{1}{2\sigma^{2,t}_{n'_v}}||Y - \sum_{j\in\{\mathrm{UN}^1\}} (\mathbf{V}_j\mathbf{s}_j)_v - u^{t-1}_{n'_v}||^2 \\ &+ \sum_{j'\in\{\mathrm{UN}^1\}\backslash j} L^t_{u_{j'}\rightarrow g_v}(\mathbf{s}_{j'})\}, \end{aligned} \tag{28}$$

while the UN-to-CN update remain unchanged.

After the UNs in $\{\mathrm{UN}^1\}$ are decoded, their decoded information can be regarded as the prior knowledge of the other subgraphs. Based on SIC-MPA, we can cut off their information from the received signals and mitigate their associated edges. Then the information of $\{\mathrm{UN}^2\}$ and $\{\mathrm{UN}^3\}$ can be decoded in the same way.

B. FURTHER COMPLEXITY REDUCTION

In reality, even if we apply the methods in Section IV-A, the computational complexity is still high especially in high user load.

To reduce the complexity further, we can only consider the detected symbols which fall within a hypersphere of radius d:

$$0 \leq ||\mathbf{f}_2(\mathbf{s})||^2 - ||\mathbf{f}_1(\mathbf{s})||^2 \leq d^2, \tag{29}$$

where $\mathbf{f}_2(\mathbf{s}) = \widetilde{\mathbf{H}}_2^{-1}\widetilde{\mathbf{y}} - \mathbf{V}\mathbf{s}$ and $d^2_{\min} \leq ||\mathbf{f}_1(\mathbf{s})||^2 = ||\widetilde{\mathbf{H}}_2^{-1}\begin{bmatrix} \mathbf{0} \\ (\mathbf{V}\mathbf{s})^B_{(V-K)\times 1} \end{bmatrix}||^2 \leq d^2_{\max}$.

From (22), the optimal $\widehat{\mathbf{s}}$ should satisfy that

$$||\mathbf{f}_2(\mathbf{s})||^2 \leq d^2 + d^2_{\max} = \sum_{v=1}^{V} d^2_v, \tag{30}$$

where d_v denotes the maximum norm of each dimension of $f_1(\mathbf{s})$.

Although $\mathbf{V}$ is a rank-deficient matrix, the methods of formulating a new full-rank detection problem are proposed in [25] and [34]. Through these methods, list sphere decoding (LSD) can still be employed for each CN and regarded as a search over a tree with a depth of L, where L mainly depends on the degrees of each CN. Hence, we can apply the analytical methods in [25] to our proposed system while some transforms are needed.

Similar as [25], the detector on CN v can be expressed as

$$\widehat{\mathbf{z}}_v = \underset{\mathbf{z}_v}{\mathrm{argmin}}\, ||Y - \mathbf{g}_{1,v}\mathbf{z}_v||^2, \tag{31}$$

where $\mathbf{g}_{1,v}$ is a $1 \times L$ row vector which mainly depends on the codebook set of SCMA(K_1, N_1, M_1), and $\mathbf{z}_v$ is an $L \times 1$ column vector whose components consist of $z_i = \pm 1, \pm 3, \ldots (i = 1, 2, \ldots, L)$. $\widehat{\mathbf{z}}_v$ is the detected vector with dimensions $L \times 1$. Denote $\widetilde{\mathbf{G}}_1 = \begin{bmatrix} \mathbf{g}_{1,v} \\ \gamma\mathbf{I}_{L\times L} \end{bmatrix} = [\mathbf{Q}_1, \mathbf{Q}_2]\begin{bmatrix} \mathbf{R} \\ \mathbf{0}_{1\times L} \end{bmatrix}$ and $\widetilde{\mathbf{Y}} = \mathbf{Q}_1^{\dagger}\begin{bmatrix} Y \\ \mathbf{0}_{L\times 1} \end{bmatrix} = [\widetilde{Y}_1, \widetilde{Y}_2, \ldots, \widetilde{Y}_L]^T$, where $\gamma > 0$ and $\mathbf{R}$ represents an $L \times L$ upper triangular matrix. $\mathbf{Q}_1$ is an $(L + 1) \times L$ matrix while $\mathbf{Q}_2$ is an $(L + 1) \times 1$ matrix. Then $\widetilde{\mathbf{G}}_1$ and $\widetilde{\mathbf{Y}}$ stand for an $(L+1) \times L$ matrix and an $L \times 1$ vector, respectively. The LSD iteration for the v-th CN can be represented by Algorithm 3. After LSD iteration, the size of the solution set decreases and it will substantially reduce the detection complexity of the MPA iteration.

V. PERFORMANCE ANALYSIS

Various performance indices of different SCMA systems are studied through closed-form expressions as follows.

A. CAPACITY GAIN

In this subsection, we will discuss the capacity gain of different SCMA systems. At first, let us derivate the overloading factor of different SCMA systems. For a single SCMA(K, N, M) system with J users, the overloading factor λ is given by

$$\lambda = \frac{J}{K} = \frac{d_f}{N}, \tag{32}$$

where $K < J \leq \binom{K}{N}$ and the degrees of each RN d_f satisfies $Kd_f = JN$.

Algorithm 3 LSD for the v-th CN

```
Input: $\mathbf{y}, \mathbf{h}', \mathbf{G}_2, \mathbf{V}, d$.
Output: $\{\mathbf{z}_v\}$ and $\{d(\mathbf{z}_v)\}$.
1: Calculate and store $\widetilde{\mathbf{Y}}$ and $\mathbf{R}$.
2: Initialization: $l = L, \widehat{Y}_L = \widetilde{Y}_L, T_L = 0$.
3: Set $z_l = \text{sign}(|\widehat{Y}_l|\cos(\theta_l))$ and search step $\Delta_l = -2\text{sign}(|\widehat{Y}_l|\cos(\theta_l))$, where $\theta_i$ is the angle of $\widehat{Y}_L$.
4: (Node Pruning) If $T_l + ||\widehat{Y}_l - r_{l,l}z_l||^2 > d_v^2$ or $z_l \notin \{-1, 1\}$, go to 4. else go to 5, where $r_{l,l}$ is the $(l, l)$-th element of $\mathbf{R}$.
5: If $l = L$, terminate and output $\phi_v$, where $\phi_v$ denotes the null set; else $l = l + 1, z_l = z_l + \Delta_l$, go to 3.
6: If $l = 1$, go to 6; else set $T_{l-1} = T_l + ||\widehat{Y}_l - r_{l,l}z_l||^2$, $\widehat{Y}_{l-1} = \widetilde{Y}_{l-1} - \sum_{j=l}^{L} r_{l-1,j}z_j$ and $l = l - 1$, go to 2.
7: If the lattice point $\mathbf{z}_v$ falls within the hypersphere, set $d(\mathbf{z}_v) = T_1 + ||\widehat{Y}_1 - r_{1,1}z_1||^2$.
8: Add $\mathbf{z}_v$ and its Euclidean distance $d(\mathbf{z}_v)$ to the solution set.
9: If LSD iteration reaches its maximum size, find $\mathbf{z}_v$ with maximum Euclidean distance, set $d = d(\mathbf{z}_v)$. $z_l = z_l + \Delta_l$, go to 3.
```

For a serial SCMA system, as mentioned in Section III, the overloading factor λ_{se} is given as follows.

$$\lambda_{\text{se}} = \frac{J_1}{K_1}\prod_{p=2}^{P}\frac{J_p}{K_p} = \frac{d_{f_1}}{N_1}\prod_{p=2}^{P}\frac{d_{f_p}}{N_p}. \tag{33}$$

Then, let us define the SCMA capacity region. For a single SCMA(K, N, M) system, $\mathbf{H}$ includes the channel knowledge. The SCMA capacity region $C(U)$ is given by [35]

$$\begin{aligned} C(U) = \{&(C^1, \ldots, C^J) : \\ &C(S) \leq \log(1 + \sum_{j\in S}\sum_{k=1}^{K}|h_{j,k}|^2 P_j^k), \forall S \subseteq U\}, \end{aligned} \tag{34}$$

where U denotes the user set. In downlink scenarios, $h_{j,k} = h_k, \forall j \in \{1, 2, \ldots, J\}$. $P_j^k = E[x_{j,k}x_{j,k}^*]$ stands for the average transmission power. The definition of capacity region mentioned above also applies for the serial SCMA system.

B. AGGREGATE ENERGY EFFICIENCY

Define the throughput per unit power consumption of an SCMA system as the aggregate energy efficiency EE($\mathbf{H}$). It is given as follows.

$$\text{EE}(\mathbf{H}) = \frac{\sum_{j\in U}(R^j(1 - I^j))}{\sum_{j\in U}(\sum_{k=1}^{K} P_k^j/\beta^j + P_{st}^j) + P_{BS}}, \tag{35}$$

where $(R^1, \cdots, R^J)$ stands for the rate region. For user j, β^j stands for the power amplifier efficiency. P_{st}^j represents its static power consumption. P_{BS} stands for the base station power consumption and I^j is the normalized channel condition of user j.

The average aggregate energy efficiency $\overline{\text{EE}}$ over all the possible fading blocks is given by

$$\begin{aligned} \overline{\text{EE}} &= \mathbb{E}_{\mathbf{H}}[\text{EE}(\mathbf{H})] \\ &= \frac{\mathbb{E}_{\mathbf{H}}[\sum_{j\in U}(R^j(1 - I^j))]}{\sum_{j\in U}(\sum_{k=1}^{K} P_k^j/\beta^j + P_{st}^j) + P_{\text{BS}}} \\ &= \frac{\sum_{j\in U}(R^j(1 - \rho_{out}^j))}{\sum_{j\in U}(\sum_{k=1}^{K} P_k^j/\beta^j + P_{st}^j) + P_{\text{BS}}}, \end{aligned} \tag{36}$$

where ρ_{out}^j is outage probability of user j. Its upper bound can be expressed as [22]

$$\begin{aligned} \rho_{out}^j \leq P_r(\epsilon^j < 2^{R^j} - 1) + &\left[\prod_{\overline{S^{j'}}} P_r(\sum_{j_0'} \epsilon^{j_0'} \geq 2^{\sum_{j_0'} R^{j_0'}} - 1)\right] \\ &P_r(\epsilon^j \geq 2^{R^j} - 1)P_r(\sum_{j_0\in U} \epsilon^{j_0} \leq 2^{\sum_{j_0\in U} R^{j_0}}), \end{aligned} \tag{37}$$

where $\epsilon^j = \sum_{k=1}^{K}|h_k|^2 \; P_k^j$ for all $j \in U$, $P_r(\cdot)$ is the probability of the inner event. For simplicity, we can suppose that $\mathbb{E}_{\mathbf{H}}[|h_k|^2] = \sigma_H^2$ for all the orthogonal resources and users. Besides, different users share the same transmit power, and rate region; i.e. $\beta^j = \beta, P_{st}^j = P_{st}, \sum_{k=1}^{K} P_k^j = P, R^k = R, \forall j \in \{1, 2, \ldots, J\}$.

In a serial SCMA system in which $P = 2$, P_k^j mainly depends on the codebook set of SCMA(K_1, N_1, M_1) and $\mathbf{G}_2$.

C. BLER

For a single SCMA system, define $\{\mathbf{s}\}$ as the set of all the possible M^J SCMA signals of J users. $\mathbf{s}^a \in \{\mathbf{s}\}$ is a specific element and $\mathbf{s}_j^a$ is the transmitted symbol in user j. Define $\mathbf{s}^b \in \{\mathbf{s}\}$ as a symbol which is different from $\mathbf{s}^a$ in user j. $\mathbf{s}_j^b$ stands for a detected symbol from user j related to $\mathbf{s}^b$ and $\mathbf{s}_j^a \neq \mathbf{s}_j^b$, then there are $(M - 1)M^{J-1}$ possible symbols of $\mathbf{s}^b$.

The average pairwise error probability (PEP) between $\mathbf{s}^a$ and $\mathbf{s}^b$ is given as follows,

$$P(\mathbf{s}^a \rightarrow \mathbf{s}^b) = \mathbb{E}_{\mathbf{H}}[Q(\sqrt{\frac{d^2(\mathbf{s}^a, \mathbf{s}^b)}{2N_0}})], \tag{38}$$

where $d^2(\mathbf{s}^a, \mathbf{s}^b)$ stands for the Euclidean distance between $\mathbf{s}^a$ and $\mathbf{s}^b$. N_0 represents the spectral density of noise power and $Q(x)$ is given by

$$Q(x) = \frac{1}{2\pi}\int_x^{\infty} \exp(-\frac{t^2}{2})dt. \tag{39}$$

Based on [36], in identical channels where $\mathbf{h}_j = \mathbf{h}$, $j \in \{1, 2, \ldots, J\}$, the union bound of a coded single SCMA system with code rate $R_0 = \frac{m_0}{n_0}$ can be expressed as

$$P_{\text{BL}} \leq \frac{L_0}{m_0}\sum_{d=d_{\min}}^{+\infty} W_d P(\mathbf{s}^a \rightarrow \mathbf{s}^b), \tag{40}$$

where $d_{\min}$ represents the minimum Hamming distance, L_0 is the date block length and W_d represents the sum weight of all error decisions for each d.

TABLE 1. Computational complexities of different SCMA systems.

		Single SCMA	Serial SCMA ($P=2$) with SIC and LSD
$\mathbf{G}_2$ Demapping and LSD	FLOP	0	$18K+6L+V\cdot\sum_{i=1}^{L}(2i+7)N_{\text{vs}_i}$
Log-MPA	ADD	$3M^{d_f}Kd_f^2T_{\max}+2MKd_f$	$3M_1^{d_c}Kd_cd_{f_1}T_{\max 1}+3\widehat{N}Kd_{c'}^2T_{\max 2}+2M_1K(d_c+d_{c'})$
	MUL	$3M^{d_f}Kd_fT_{\max}+4MKd_f$	$3M_1^{d_c}Kd_{f_1}T_{\max 1}+3\widehat{N}Kd_{c'}T_{\max 2}+4M_1Kd_c+4M_1Kd_{c'}$
	COM	$M^{d_f}Kd_fT_{\max}$ $+(N-2)MKd_fT_{\max}$	$M_1^{d_c}Kd_{f_1}T_{\max 1}+\widehat{N}Kd_{c'}T_{\max 1}$ $+(N_1-2)M_1K(d_cT_{\max 1}+d_{c'}T_{\max 2})$

For a serial SCMA system in the downlink scenario,

$$d^2(\mathbf{s}^a,\mathbf{s}^b)=||\text{diag}(\mathbf{h}')\mathbf{G}_P\mathbf{G}_{P-1}\ldots\mathbf{G}_2\mathbf{V}(\mathbf{s}^a-\mathbf{s}^b)||^2. \quad (41)$$

Combining (38) and (40) with (41), we can also obtain the error probability bounds of the turbo coded serial SCMA system.

D. DETECTION COMPLEXITY

In identical channel conditions, the detailed detection complexities of different SCMA systems are illustrated in Table 1. We use flop to denote the floating-point operation. Each comparison operation can be represented by a lookup table. Thus we suppose that it only costs one flop. Each addition of two complex values costs two flops. Each non-conjugated complex multiplication costs six flops.

In a serial SCMA system when $P=2$, we should conduct LU factorization of matrix $\widetilde{\mathbf{H}}_2$ at first and then calculate its inverse matrix. From (21), $(\mathbf{G}_2^A)^{-1}$ and $(\mathbf{G}_2^A)^{-1}\mathbf{G}_2^B$ can be calculated in advance. We need only calculate $(\mathbf{G}_2^A)^{-1}\text{diag}(\mathbf{h}')^{-1}$ and this procedure may require $12K$ flops. The calculation of $\widetilde{\mathbf{y}}$ costs $6K$ flops.

For Log-MPA iteration with SIC, The UNs in in $\{\text{UN}^1\}$ contribute more to the first K CNs and can be decoded first. This procedure requires $3M_1^{d_c}Kd_cd_{f_1}T_{\max 1}+2M_1Kd_c$ additions, $3M_1^{d_c}Kd_{f_1}T_{\max 1}+4M_1Kd_c$ multiplications and $M_1^{d_c}Kd_{f_1}T_{\max 1}+(N_1-2)MKd_cT_{\max 1}$ comparisons where d_c denote the number of UNs in UN_1 corresponding to the CN and $d_{f_1}=\frac{JN_1}{V}$. Then the other UNs are decoded and this procedure requires $3M_1^{d_{c'}}Kd_{c'}^2T_{\max 2}+2M_1Kd_{c'}$ additions, $3M_1^{d_{c'}}Kd_{c'}T_{\max 2}+4M_1Kd_{c'}$ multiplications and $M_1^{d_{c'}}Kd_{c'}T_{\max 2}+(N_1-2)MKd_{c'}T_{\max 2}$ comparisons, where $d_{c'}=\frac{JN_1}{V}-d_c$.

On the other hand, after we cut off the UNs in $\{\text{UN}^1\}$, we can conduct LSD iteration before Log-MPA for the UNs in $\{\text{UN}^2\}\cup\{\text{UN}^3\}$. Then QR factorization should be conducted on $\widetilde{\mathbf{G}}_1$ at first and it may require $2(L+1)L^2$ flops, where $L=d_{c'}\log_2 M_1$ and $\widetilde{\mathbf{G}}_1$ depends on the factor graph excluding the UNs in $\{\text{UN}^1\}$. This procedure can be finished in advance. And we only need $6L$ flops to get $\widetilde{\mathbf{Y}}$ as mentioned in IV-B. For LSD iteration, expected complexity [25] is given by

$$E(i)=\sum_{i=1}^{L}(2i+1)N_{\text{vs}_i}, \quad (42)$$

where N_{vs_i} is the average number of visited nodes over elementary arithmetic operations in level i.

Then the whole detection procedure of Log-MPA with SIC and LSD requires $3M_1^{d_c}d_cd_{f_1}T_{\max 1}+3\widehat{N}Kd_{c'}^2T_{\max 2}+2M_1K(d_c+d_{c'})$ additions, $3M_1^{d_c}Kd_{f_1}T_{\max 1}+3\widehat{N}Kd_{c'}T_{\max 2}+4M_1K(d_c+d_{c'})$ multiplications and $M_1^{d_c}Kd_{f_1}T_{\max 1}+\widehat{N}Kd_{c'}T_{\max 1}+(N_1-2)M_1K(d_cT_{\max 1}+d_{c'}T_{\max 2})$ comparisons, where $\widehat{N}<M_1^{d_{c'}}$ represents the total number of constellation points corresponding to UNs in $\{\text{UN}^2\}\cup\{\text{UN}^3\}$ after LSD iteration.

The above methods also apply for $P\geq 3$. We should balance the computational complexity and the overloading factor.

VI. EVALUATION

Considering downlink scenarios, we evaluate BLER of the proposed serial SCMA system in MATLAB over AWGN and Rayleigh fading channels. 10000 data blocks are simulated and each data block includes 200 bits. Other performance indices are evaluated over Rayleigh channels. The Jakes' Pedestrian-B model is used to generate the Rayleigh frequency selective channels and the Dropper frequency is normalized. Assume that each UN observes the same channel coefficients over d_f RNs, i.e. $\mathbf{h}_j'=\mathbf{h}', \forall j=1,2,\ldots,J$. We consider the normalized noise, i.e. $\mathbf{n}$ follows $\mathcal{CN}(\mathbf{0},\mathbf{I})$. SCMA codebooks are generated through QAM constellations. The pattern matrices in serial SCMA are designed based on the Latin generator matrices. Turbo code is used in channel coding with coding rate $R=\frac{1}{2}$, feedforward polynomial $F_1=1+D+D_3$ and feedback polynomial $F_2=1+D_2+D_3$. The interleaver size of the turbo coder is set as 2048 bits. Both SIC and LSD are used in decoding for serial SCMA. The maximum number of iterations in each Log-MPA detector is set as 6 times and the system bandwidth is 10MHz.

A. CAPACITY GAIN

We consider a single SCMA(8, 2, 4) system as a comparative benchmark. In the proposed serial SCMA system, set $p=2$ and $J=J_1=32, V=K_1=J_2=16, K=K_2=8, N_1=N_2=2, M_1=M_2=4$.

Fig. 5 illustrates the overloading abilities under fixed number of RNs for different systems. 4-QAM is employed in LTE. Compared with LTE, both the serial and the single SCMA systems can achieve overloading. In reality, the overloading

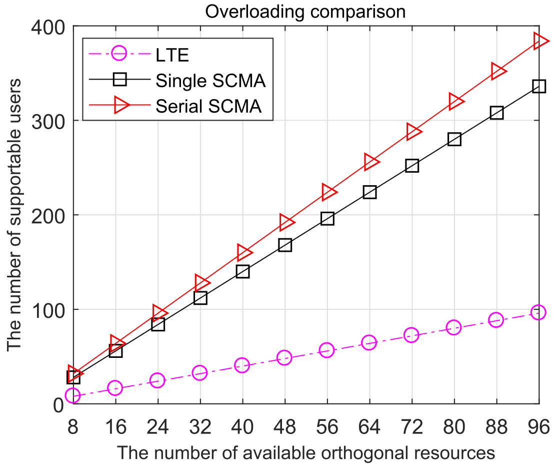

FIGURE 5. Overloading of different SCMA and LTE.

factor of a serial SCMA system equals the product of overloading factors of all the low order subsystems. It can even achieve higher user load than the single SCMA system while the balance between the overloading factor and the decoding complexity should also be stroke.

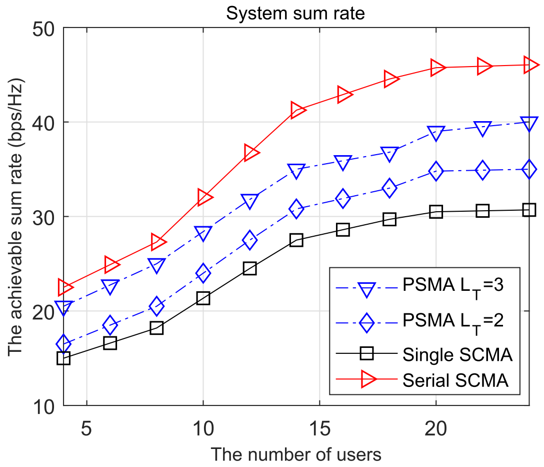

FIGURE 6. Achieveable system rate of different SCMA.

Fig. 6 exhibits the achievable data rate among different SCMA systems under the same user load, where L_T denotes the number of users which share the same codebook in PSMA. The parameters of PSMA are the same as [11]. Compared with PSMA, serial SCMA further improve the system sum rate. The serial SCMA system also enjoys the highest spectral efficiency under the same user load attributing to its high diversity gain.

B. AGGREGATE ENERGY EFFICIENCY

We consider the same power model as referred in [22] where $P_{\text{st}} = 44W, P_{\text{BS}} = 300W$. Fig. 7 demonstrates the average aggregate energy efficiencies of different systems. In single SCMA, $K = 8, J = 24, N = 2, M = 4$ while in the proposed serial SCMA, $p = 2, K = K_2 = 8, V = K_1 = J_2 = 12, J = J_1 = 24, N_1 = N_2 = 2, M_1 = M_2 = 4$.

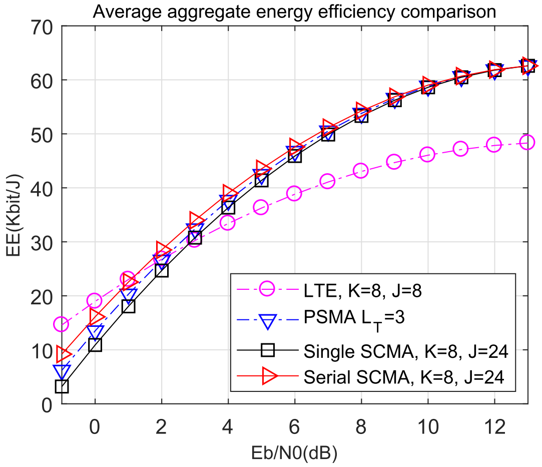

FIGURE 7. Average aggregate energy efficiency comparison.

In PSMA, $K = 8, J = 24, M = 4, L_T = 3$. In LTE, $K = 8, J = 8, M = 4$. $\frac{E_b}{N_0}$ stands for the SNR per bit. From Fig. 7, we can conclude that under low $\frac{E_b}{N_0}$, the outage probability of each user in each SCMA system is so high that LTE enjoys the highest average aggregate energy efficiency due to its orthogonality. As $\frac{E_b}{N_0}$ grows, the user outage probability of each SCMA systems decreases. All the three SCMA systems can reach higher average aggregate energy efficiency than LTE due to the non-orthogonality which will also bring them higher spectral efficiency. Moreover, the common BS power consumption P_{BS} is shared by all the users in these SCMA systems. This is also a factor to improve energy efficiency. The serial SCMA system can approach higher average aggregate energy efficiency than the other two SCMA systems for a low $\frac{E_b}{N_0}$ range. This is mainly because serial SCMA employs an equivalent factor graph in which each UN occupies more RNs than single SCMA and PSMA. In comparison, serial SCMA can utilize spectral resources and further decrease the user outage probability.

C. BLER

Both SCMA and LDS can achieve overloading while in SCMA the power variation of the incoming signals over each corresponding subcarrier helps the MPA detector to mitigate inter-layer interferences more efficiently. Fig. 8, Fig. 9, Fig. 10 and Fig. 11 exhibit the BLER of different SCMA systems and LDS with different M over different channels, respectively. Each user in different systems has the same data rate. The single SCMA and LDS share the same factor graph matrix which is given in (43), as shown at the bottom of the next page.

In the proposed serial SCMA system, $p = 2, K = K_2 = 8, V = K_1 = J_2 = 12, J = J_1 = 24, N_1 = N_2 = 2$. Under the same user load, the proposed serial SCMA enjoys lower BLER than LDS as single SCMA [12]. It inherits the nice structure of the single SCMA system. In reality, the link performance will improve when each user occupies a bit more

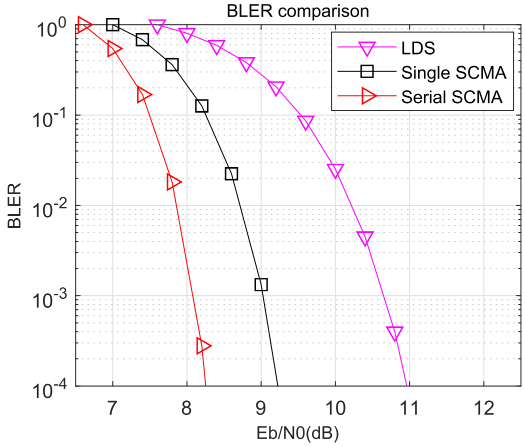

FIGURE 8. BLER comparison with $M = 4, \lambda = 300\%$ over AWGN.

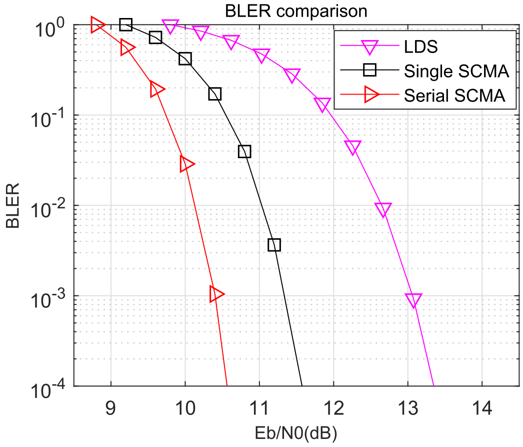

FIGURE 10. BLER comparison with $M = 4, \lambda = 300\%$ over Rayleigh block fading.

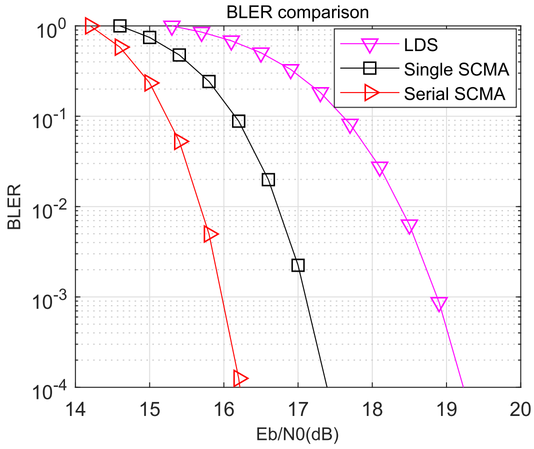

FIGURE 9. BLER comparison with $M = 16, \lambda = 300\%$ over AWGN.

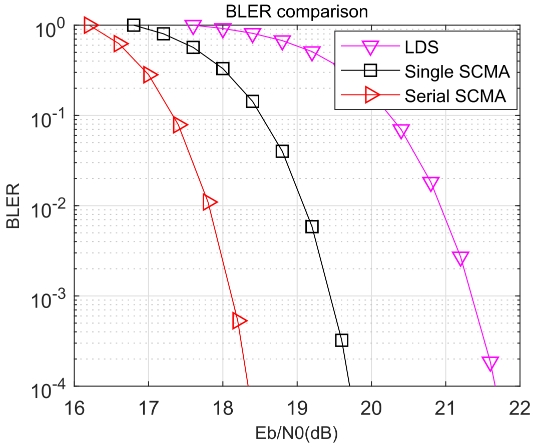

FIGURE 11. BLER comparison with $M = 16, \lambda = 300\%$ over Rayleigh block fading.

subcarriers under fixed user load due to the diversity gain. The serial SCMA system employs an equivalent factor graph in which each user occupies more subcarriers than the single SCMA system under the same capacity gain. This will make serial SCMA enjoy lower BLER than single SCMA.

D. DETECTION COMPLEXITY

In this subsection, we still suppose that each user in different systems has the same data rate. The detection complexities of different SCMA systems are compared under the same overloading factor and total system data rate. Consider a single SCMA system in which $K = 8, J = 24, N = 2$ and $M = 4$. In the proposed serial SCMA system, set $p = 2$ and $J = J_1 = 24, V = K_1 = J_2 = 12, K = K_2 = 8$, $N_1 = N_2 = 2$ and $M_1 = M_2 = 4$.

Fig. 12 demonstrates the computational complexity of additions, multiplications and comparisons with 300% load

$$\mathbf{F}_{8\times 24} = \begin{bmatrix} 0&0&0&1&0&0&0&1&0&0&0&1&0&0&0&1&0&0&0&1&0&0&0&1 \\ 0&1&0&0&0&1&0&0&0&0&0&1&0&0&1&0&0&0&1&0&0&1&0&0 \\ 1&0&0&0&0&0&1&0&0&1&0&0&0&1&0&0&0&0&1&0&0&0&0&1 \\ 0&0&1&0&1&0&0&0&0&1&0&0&1&0&0&0&0&0&0&1&0&1&0&0 \\ 0&0&1&0&0&0&1&0&0&0&1&0&0&0&0&1&0&1&0&0&0&0&1&0 \\ 1&0&0&0&1&0&0&0&0&0&1&0&0&0&1&0&1&0&0&0&1&0&0&0 \\ 0&1&0&0&0&0&0&1&1&0&0&0&0&1&0&0&1&0&0&0&0&0&1&0 \\ 0&0&0&1&0&1&0&0&1&0&0&0&1&0&0&0&0&1&0&0&1&0&0&0 \end{bmatrix}. \tag{43}$$

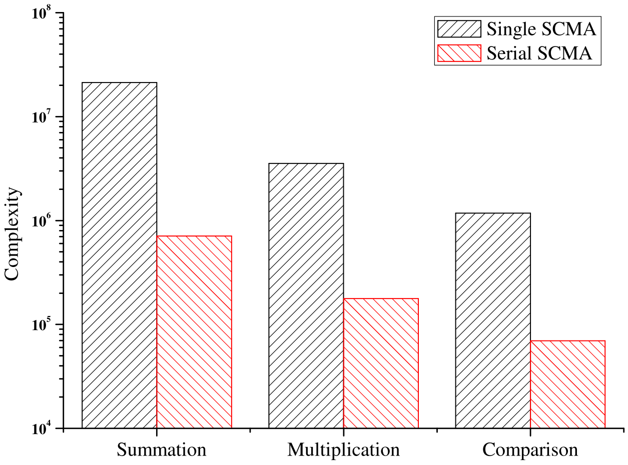

FIGURE 12. Computational complexity of additions, multiplications and comparisons in different SCMA systems with 300% load when $M = 4$.

for different SCMA systems when $\frac{E_b}{N_0} = 7dB$. We can observe that the serial SCMA system enjoys lower complexity due to the reduction of d_f.

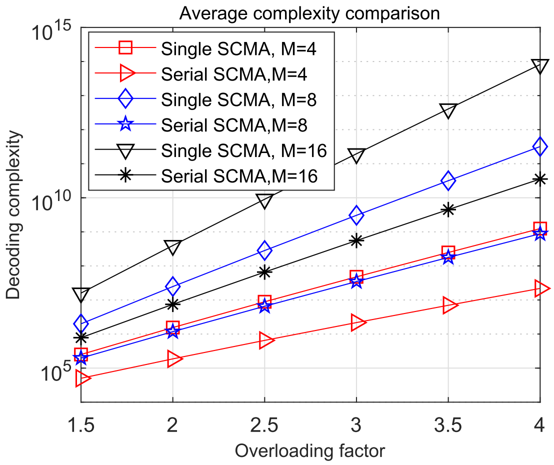

FIGURE 13. Average complexity versus overloading factor.

Fig. 13 illustrates the relationship between average decoding complexity and overloading factor for different schemes with different M. The average decoding complexity mainly denotes the total number of all the flops. $\frac{E_b}{N_0}$ is set as 7dB when $M = 4$ and 17dB when $M = 16$. Each complex addition costs two flops and each multiplication of two non-conjugated complex values costs six flops while each comparison costs only one flop. From the simulation results, we can conclude that the serial SCMA system will greatly save the detection complexity as the overloading factor and codebook size grow compared with the single SCMA system under the same user load.

VII. CONCLUSION

Motivated by hierarchical modulation, we propose a system with serial codes which is called serial SCMA to enable high order SCMA systems in downlink scenarios in this paper. The serial SCMA system achieves overloading step by step by multiple low order subsystems and can enjoy higher spectral efficiency compared with original SCMA systems. The signal between each group of adjacent subsystems is a complex vector and the coding schemes of the adjacent subsystems should satisfy specific mapping rules. Simulation results illustrates that serial SCMA inherits the advantages of original SCMA such as overloading and power variation over OMA and other NOMA techniques. Besides, serial SCMA maintains sparser codebooks and can substantially reduce the decoding complexity especially under high user load. Furthermore, sparser codebooks will also bring better link performance and make hardware implementation feasible.

REFERENCES

[1] J. G. Andrews *et al.*, "What will 5G be?" *IEEE J. Sel. Areas Commun.*, vol. 32, no. 6, pp. 1065–1082, Jun. 2014.

[2] F. Boccardi, R. W. Heath, A. Lozano, T. L. Marzetta, and P. Popovski, "Five disruptive technology directions for 5G," *IEEE Commun. Mag.*, vol. 52, no. 2, pp. 74–80, Feb. 2014.

[3] L. Dai, B. Wang, Y. Yuan, S. Han, C.-L. I, and Z. Wang, "Non-orthogonal multiple access for 5G: Solutions, challenges, opportunities, and future research trends," *IEEE Commun. Mag.*, vol. 53, no. 9, pp. 74–81, Sep. 2015.

[4] R. Hoshyar, F. P. Wathan, and R. Tafazolli, "Novel low-density signature for synchronous CDMA systems over AWGN channel," *IEEE Trans. Signal Process.*, vol. 56, no. 4, pp. 1616–1626, Apr. 2008.

[5] Z. Yuan, G. Yu, and W. Li, "Multi-user shared access for 5G," *Telecommun. Netw. Technol.*, vol. 5, no. 5, pp. 28–30, 2015.

[6] J. Huang, K. Peng, C. Pan, F. Yang, and H. Jin, "Scalable video broadcasting using bit division multiplexing," *IEEE Trans. Broadcast.*, vol. 60, no. 4, pp. 701–706, Dec. 2014.

[7] J. Zeng, B. Li, X. Su, L. Rong, and R. Xing, "Pattern division multiple access (PDMA) for cellular future radio access," in *Proc. Wireless Commun. Signal Process. (WCSP)*, Oct. 2015, pp. 1–5.

[8] H. Nikopour and H. Baligh, "Sparse code multiple access," in *Proc. IEEE 24th Int. Symp. Pers. Indoor Mobile Radio Commun. (PIMRC)*, Sep. 2013, pp. 332–336.

[9] M. Moltafet, N. M. Yamchi, M. R. Javan, and P. Azmi, "Comparison study between PD-NOMA and SCMA," *IEEE Trans. Veh. Technol.*, vol. 67, no. 2, pp. 1830–1834, Feb. 2018.

[10] F. R. Kschischang, B. J. Frey, and H.-A. Loeliger, "Factor graphs and the sum-product algorithm," *IEEE Trans. Inf. Theory*, vol. 47, no. 2, pp. 498–519, Feb. 2001.

[11] M. Moltafet, N. Mokari, M. R. Javan, H. Saeedi, and H. Pishro-Nik, "A new multiple access technique for 5G: Power domain sparse code multiple access (PSMA)," *IEEE Access*, vol. 6, pp. 747–759, 2018.

[12] M. Taherzadeh, H. Nikopour, A. Bayesteh, and H. Baligh, "SCMA codebook design," in *Proc. IEEE Veh. Technol. Conf. (VTC Fall)*, Sep. 2014, pp. 1–5.

[13] J. Bao, Z. Ma, M. A. Mahamadu, Z. Zhu, and D. Chen, "Spherical codes for SCMA codebook," in *Proc. IEEE Veh. Technol. Conf. (VTC Spring)*, May 2016, pp. 1–5.

[14] L. Yu, X. Lei, P. Fan, and D. Chen, "An optimized design of SCMA codebook based on star-QAM signaling constellations," in *Proc. Wireless Commun. Signal Process. (WCSP)*, Oct. 2015, pp. 1–5.

[15] J. Boutros and E. Viterbo, "Signal space diversity: A power- and bandwidth-efficient diversity technique for the Rayleigh fading channel," *IEEE Trans. Inf. Theory*, vol. 44, no. 4, pp. 1453–1467, Jul. 1998.

[16] D. Cai, P. Fan, X. Lei, Y. Liu, and D. Chen, "Multi-dimensional SCMA codebook design based on constellation rotation and interleaving," in *Proc. IEEE Veh. Technol. Conf. (VTC Spring)*, May 2016, pp. 1–5.

[17] M. Zhao, S. Zhou, W. Zhou, and J. Zhu, "An improved uplink sparse coded multiple access," *IEEE Commun. Lett.*, vol. 21, no. 1, pp. 176–179, Jan. 2017.

[18] D. Zhai, M. Sheng, X. Wang, Y. Li, J. Song, and J. Li, "Rate and energy maximization in SCMA networks with wireless information and power transfer," *IEEE Commun. Lett.*, vol. 20, no. 2, pp. 360–363, Feb. 2016.

[19] Z. Li, W. Chen, F. Wei, F. Wang, X. Xu, and Y. Chen, "Joint codebook assignment and power allocation for SCMA based on capacity with Gaussian input," in *Proc. IEEE Int. Conf. Commun. China (ICCC)*, Chengdu, China, Jul. 2016, pp. 1–6.

[20] L. Li, Z. Ma, L. Wang, P. Z. Fan, and L. Hanzo, "Cutoff rate of sparse code multiple access in downlink broadcast channels," *IEEE Trans. Commun.*, vol. 65, no. 8, pp. 3328–3342, Aug. 2017.

[21] A. Bayesteh, H. Nikopour, M. Taherzadeh, H. Baligh, and J. Ma, "Low complexity techniques for SCMA detection," in *Proc. IEEE Globecom Workshops (GC Wkshps)*, Dec. 2015, pp. 1–6.

[22] 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.*, Dec. 2014, pp. 4782–4787.

[23] 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.

[24] D. Wei, Y. Han, S. Zhang, and L. Liu, "Weighted message passing algorithm for SCMA," in *Proc. Wireless Commun. Signal Process. (WCSP)*, Oct. 2015, pp. 1–5.

[25] F. Wei and W. Chen, "Low complexity iterative receiver design for sparse code multiple access," *IEEE Trans. Commun.*, vol. 65, no. 2, pp. 621–634, Feb. 2017.

[26] J. Chen, Z. Zhang, S. He, J. Hu, and G. E. Sobelman, "Sparse code multiple access decoding based on a Monte Carlo Markov chain method," *IEEE Signal Process. Lett.*, vol. 23, no. 5, pp. 639–643, May 2016.

[27] M. Kaneko, H. Yamaura, Y. Kajita, K. Hayashi, and H. Sakai, "Fairness-aware non-orthogonal multi-user access with discrete hierarchical modulation for 5G cellular relay networks," *IEEE Access*, vol. 3, pp. 2922–2938, Dec. 2015.

[28] G. D. Forney and L.-F. Wei, "Multidimensional constellations. I. Introduction, figures of merit, and generalized cross constellations," *IEEE J. Sel. Areas Commun.*, vol. 7, no. 6, pp. 877–892, Aug. 1989.

[29] G. K. Karagiannidis and A. S. Lioumpas, "An improved approximation for the Gaussian Q-function," *IEEE Commun. Lett.*, vol. 11, no. 8, pp. 644–646, Aug. 2007.

[30] H. Nikopour and M. Baligh, "Systems and methods for sparse code multiple access," U.S. Patent 0 072 660 A1, Nov. 2, 2015

[31] B. Ren, Y. Wang, X. Dai, K. Niu, and W. Tang, "Pattern matrix design of PDMA for 5G UL applications," *China Commun.*, vol. 13, no. 2, pp. 159–173, Jan. 2017.

[32] Y. Wu, S. Zhang, and Y. Chen, "Iterative multiuser receiver in sparse code multiple access systems," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Jun. 2015, pp. 2918–2923.

[33] L. Lei, C. Yan, G. Wenting, Y. Huilian, W. Yiqun, and X. Shuangshuang, "Prototype for 5G new air interface technology SCMA and performance evaluation," *China commun.*, vol. 12, no. 9, pp. 38–48, Sep. 2015.

[34] M. Vameghestahbanati, E. Bedeer, I. Marsland, R. H. Gohary, and H. Yanikomeroglu, "Enabling sphere decoding for SCMA," *IEEE Commun. Lett.*, vol. 21, no. 12, pp. 2750–2753, Dec. 2017.

[35] R. Razavi, R. Hoshyar, M. Imran, and Y. Wang, "Information theoretic analysis of LDS scheme," *IEEE Commun. Lett.*, vol. 15, no. 8, pp. 798–800, Aug. 2011.

[36] G. Caire, G. Taricco, and E. Biglieri, "Bit-interleaved coded modulation," *IEEE Trans. Inf. Theory*, vol. 44, no. 3, pp. 927–946, May 1998.

YUXI HAN received the B.S. degree from the University of Science and Technology of China, Hefei, China, in 2011, where he is currently pursuing the Ph.D. degree. His research interests include non-orthogonal multiple access, constellation design, multi-user detection, and signal processing.

WUYANG ZHOU received the B.S. and M.S. degrees from Xidian University, Xian, China, in 1993 and 1996, respectively, and the Ph.D. degree from the University of Science and Technology of China (USTC), Hefei, China, in 2000. He is currently a Professor of wireless communication networks with the Department of Electronic Engineering and Information Science, USTC. He participated in the National 863 Research Project Beyond Third Generation of Mobile System in China (FUTURE Plan) and has been the Task Director for many projects, including Innovative Wireless Campus Experimental Networks Research on High Frequency Networking Technologies and Research on Transmission and Networking Technologies in Satellite Mobile Communications. His current research interests include green technologies for communication systems, satellite mobile communications, and underwater acoustic communications.

MING ZHAO received the B.E. and M.E. degrees from the University of Science and Technology of China (USTC), Hefei, China, in 1999 and 2002, respectively. He is currently a Lecturer with the Department of Electronic Engineering and Information Science, USTC. His research interests include non-orthogonal multiple access, green communications, and heterogeneous networks.

SHENGLI ZHOU (S'99–M'03–SM'11–F'14) received the B.S. and M.Sc. degrees in electrical engineering and information science from the University of Science and Technology of China, Hefei, China, in 1995 and 1998, respectively, and the Ph.D. degree in electrical engineering from the University of Minnesota, Minneapolis, MN, USA, in 2002. He is currently a Full Professor with the Department of Electrical and Computer Engineering, University of Connecticut, Storrs, CT, USA.

His general research interests lie in the areas of wireless communications and signal processing. He received the 2007 ONR Young Investigator Award and the 2007 Presidential Early Career Award for Scientists and Engineers. He was an Associate Editor for the IEEE Transactions on Wireless Communications (2005–2007), the IEEE Transactions on Signal Processing (2008–2010), and the IEEE Journal of Oceanic Engineering (2010–2016).