SCIENCE CHINA
Information Sciences

• **Supplementary File** •

A Layered Grouping Random Access Scheme Based on Dynamic Preamble Selection for Massive Machine Type Communications

Gaofeng CHENG[1], Huan CHEN[1], Pingzhi FAN[1], Li LI[1*] & Li HAO[1]

[1]*Information Coding & Transmission Key Lab of Sichuan Province, CSNMT Int. Coop. Res. Centre (MoST), Southwest Jiaotong University, Chengdu* 610031, *China*

Appendix A Layered Grouping Cellular Framework

We consider an mMTC cell having a radius R, where the BS equipped with a single antenna locates in the cell center and a total number of N coexisted users randomly distribute in the cell. It is assumed that all the mMTC users are mainly static, a typical scenario in mMTC applications [22], e.g. the interactions among machines in the industrial automation, the monitoring in smart agriculture, the environment monitoring for public safety, etc.

According to predefined system configurations including cell size, QoS requirement, maximum number of coexisted cellular users, etc. , BS will divide all the cellular users into a number of K clusters. Typically, K could be a small value, e.g. $K = 2, 4, 8$, or 16. The concerns behind this configuration suggestion are addressed in Appendix D. In order to facilitate the practical synchronization, it is better that the users in the same cluster experience similar transmission delays, normally located in a geographical area having roughly the same distance to the BS. This could be achieved by estimating the average of received signal strength (RSS) over a sufficient long duration at the BS [29], [30]. If it is necessary, the geolocation database (GDB) aided construction method proposed in [31] could be further invoked in. Ideally, the entire cellular is divided into a number of K rings and all the users located in the same ring will be assigned to the same cluster. This effect is visualized by the dashed ellipses in Fig.A1. In practice, owing to the limited localization accuracy, some edge users may be assigned to their adjacent cluster. But this potential mismatch will not impact the proposed two-phase RA scheme, as the proposed RA scheme is capable of adapting to unbalanced cluster loads.

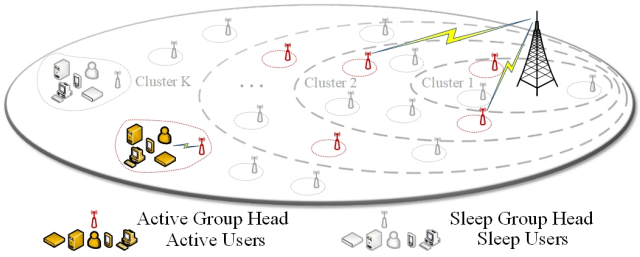

Figure A1 Network topology of the layered grouping in an mMTC cell.

Then, the users assigned to the k^{th} cluster c_k, $k = 1, 2, \cdots, K$ will further participate in a number of $M^{(k)}$ groups. These groups are initialized and updated in a self-organizing manner. The m^{th} group in k^{th} cluster is denoted by $g_{k,m}$. The users pertaining to a group is termed as group members and the set of their user ID is denoted by $\mathcal{G}_{k,m}$. The number of group members is termed as the group size and denoted by $|g_{k.m}|$. Since the D2D communication technique [31] will be employed to realize the internal message exchange among group members, we would like to constrain the group size by a small value. Because a large group size normally results in longer distances between GH and its group members, which implies less reliable device-to-device (D2D) links. Actually, it was shown in Fig.3 of [31] that packet error rate over D2D links becomes non-negligible after the group size exceeds 20. Although, this simulation result only reveals the situation in Bluetooth Low Energy communication service, it indeed verifies the fact that the group size should be constrained to a small number for guaranteeing reliable D2D links. In accordance with this spirit and without loss of generality, we will constraint the group size to a relatively small number when simulating the system performance in Appendix H. Then, since group size is small, we could further reasonably assume that all the group members are close to each other. As a benefit, the internal message exchange among group members can be reliably realized by the D2D communication technique [31], which could be interference-free to other groups and hence spectral efficient. Moreover, a particular group member, namely $\dot{u}_{k,m}$ will be selected as the GH of group $g_{k,m}$. If a normal group member u_n wants to communicate with the BS, it firstly sends the message to its GH $\dot{u}_{k,m}$. Then the GH $\dot{u}_{k,m}$ relays this message to BS, and vice versa. It implies that throughout the RA procedure, the GH $\dot{u}_{k,m}$ will communicate with the BS on behalf of all the members in the same group. The above-mentioned mechanism is visualized by the dotted ellipses in Fig.A1, where the GH is equivalently denoted as a relay node.

* Corresponding author (email: ll5e08@home.swjtu.edu.cn)

Finally, a GH $\dot{u}_{k,m}$ possesses two kinds of access preambles. The first one, namely $\mathbf{s}_k^I$ could be regarded as a signature of the cluster c_k. All cluster signatures are orthogonal to each other, i.e. $\langle \mathbf{s}_i^I, \mathbf{s}_j^I \rangle = \begin{cases} 0, & if\ j \neq i \\ 1, & if\ j = i \end{cases}$, where $i, j = 1, 2, \cdots, K$. Since K is relatively small, this orthogonality can be easily satisfied, even for a short preamble length. The second one, namely $\mathbf{s}_{k,m}^{II}$ could be regarded as a signature of the group $g_{k,m}$. Since $M^{(k)}$ is normally a very large number in mMTC scenarios, $\mathbf{s}_{k,m}^{II}$ employed by different groups in the same cluster have to be non-orthogonal sequences for reducing overhead. More details of preamble assignment of $\mathbf{s}_{k,m}^{II}$ could be found in Appendix E and Appendix G.

Appendix B Construction of Layered Grouping

The construction of clusters can be controlled by BS, where only an approximated distance from a user to BS is required. On the other hand, the formation and update of groups in each cluster may also be implemented in a centralized manner [31] [32], where BS controls the selection of GHs and assigns their group members. However, this centralized management requires a range of global information including users' accurate positions, propagations, data rates, battery levels, etc. Aggregating these information from millions devices in the mMTC scenario may become prohibitive. Hence distributed self-organized formation has a better applicability. Hence, the construction procedure of the layered network framework is designed, which is elaborated herein:

1. While a user (device) $u_n, n = 1, 2, \ldots, N$, firstly powers on in the cell and hears the system broadcast information (including the power level of control channel) from BS, its registration process will then start by sending a registration message containing user ID, device type, and a couple of reference signals, to BS in a contention free manner[1)].

2. Based on the reference signals contained in the registration message, BS is capable of approximating the distance between a user and itself by utilizing RSS aided positioning techniques [29], [30] and further assigning u_n to an appropriate cluster c_k. Then, BS assigns the generation method of a pair of preambles $\mathbf{s}_k^I$ and $\mathbf{s}_{k,m}^{II}$, the initial preamble lengths, as well as the group size $|g_{k,m}|$ to u_n. These information and a couple of reference signals are contained in the registration response message (RRM).

3. Based on the reference signals in RRM, user u_n is capable of estimating the channel from BS, and the associated channel state information (CSI) is denoted by $h_{n,b}$. It is assumed that all the channels are reciprocal and have a relative long coherent time based on the fact that the mMTC devices are mainly static in our application scenarios. According to RRM, user u_n becomes aware of its cluster index k. Then, user u_n will further autonomously select itself as a GH in a probability of $\frac{1}{|g_{k,m}|}$.

4. BS will periodically broadcast a group update opportunity message (GUOM) to all cellular users. Bearing the quasi-static property of our application scenario in mind, the group update period could be generally long, say daily or even weekly for reducing system overhead.

5. Once the cellular users hear GUOM, they will implement the group initialization[2)] or update[3)] procedure in a self-organizing manner and via D2D links. An Opt-EC based K-means grouping algorithm is designed to iteratively improve the grouping relationship and select the energy efficient GHs, which will be elaborated in Appendix C. With the aid of this K-means grouping algorithm, the groups in a cluster are constructed and updated.

6. If the role of a user changes (i.e. switches from a normal group member to a GH and vice versa), it will inform BS of its new state. The BS will add or remove the associated group ID from its group list.

The above-mentioned construction procedure is illustrated in Figure B1. For more information, the construction and maintenance of the layered grouping network framework consist of three major steps: a) the users register themselves at the BS and BS indicates some of them to act as initial group heads; b) the coexisted users autonomously constitute groups and periodically update the group members relying on Opt-EC based K-means algorithm; c) the new group heads inform BS of themselves and BS updates its group head list. This abstracted construction and maintenance procedures are shown in Figure B1 (a). Simultaneously, more detailed construction and maintenance procedures are depicted in Figure B1 (b).

Appendix C Maintenance of Layered Grouping: Opt-EC based K-means algorithm

Actually, the maintenance procedure of layered grouping has been briefly mentioned on the steps 4 and 5 in Appendix B. The critical technology employed is the proposed Opt-EC based K-means algorithm. To be concrete, after receiving the group update opportunity message from BS, the existing GHs $\dot{u}_{k,m}, 1 \leqslant m \leqslant M^{(k)}$ will broadcast grouping messages. Group ID is contained in these grouping messages[4)]. A registered neighboring non-GH user u_n will select the one from which it hears the strongest grouping message as its intended GH. Then, u_n broadcasts a subscribing message (SM), which carries its user ID, intended group ID. After receiving all the subscribing messages, a GH $\dot{u}_{k,m}$ is capable of determining all of its group members, and then feedbacks this temporary decision of $\mathcal{G}_{k,m}$ to all the group members. Herein, $\mathcal{G}_{k,m}$ denotes the set of user ID that regarded as group members of $g_{k,m}$. The transmission of above-mentioned intra-group signaling messages will rely on dedicated spectrum such as 5905-5925 MHz in 5G NR V2X (PC5 interface) [33], or unlicensed spectrum technologies, namely D2D outband communications [31]. Hence, the impact of these extra overheads on the cellular radio access network (RAN) can be negligible. These intra-group signaling are summarized as lines 2-9 in Algorithm C1.

---

1) Since the number of users simultaneously switching on is normally extremely low, contention-free transmission of registration message could be realized by predefining a small set of specific channel resources.

2) In the case a group has not been created before.

3) In the case a group has existed.

4) In order to adapt to different mMTC applications, a signature, which indicates the specific kind of device the group head seeks for grouping could be further added to the grouping message as a particular segment. Consequently, when the neighboring non-GH user selects its group head, it will further verify whether it is the right kind of device that the group head looks for.

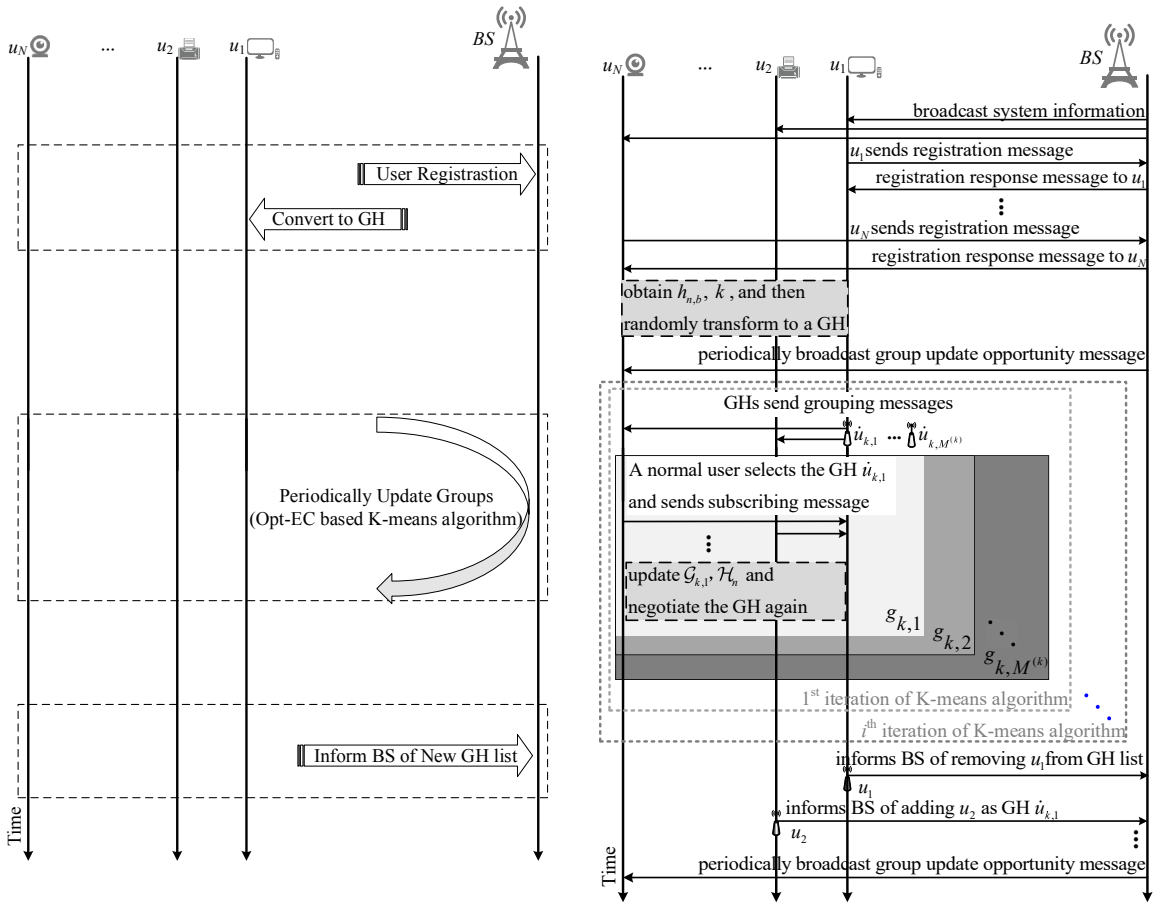

Figure B1 (a) Main steps involved in construction and maintenance of the layered grouping network framework. (b) More details of these construction and maintenance.

Algorithm C1 Opt-EC based K-means grouping algorithm

Require: cluster index k;
1: **while** (Predefined max iterations has not been reached.) **do**
2: **for** ($m = 1$ **to** $M^{(k)}$) **do**
3: $\dot{u}_{k,m}$ selected in last iteration broadcasts grouping message;
4: **end for**
5: **for** (($n = 1$ **to** N) **do**
6: **for** (($m = 1$ **to** $M^{(k)}$) **do**
7: u_n attempts to hear from $\dot{u}_{k,m}$ and estimate $h_{\dot{u}_{k,m},n}$;
8: **end for**
9: $u_n \xrightarrow{subscribe\ group} \underset{\dot{u}_{k,m}\ heard\ by\ u_n}{\arg\max} h_{\dot{u}_{k,m},n}$;
10: **end for**
11: **for** (($m = 1$ **to** $M^{(k)}$) **do**
12: Update $\mathcal{G}_{k,m}$ and every $\mathcal{H}_n$;
13: $\dot{u}_{k,m} = \underset{u_n \in \mathcal{G}_{k,m}}{\arg\min} \gamma_{k,m}(u_n)$;
14: **end for**
15: **end while**

The proposed Opt-EC based K-means grouping algorithm aims at selecting the best GH, which simultaneously minimizes the energy required by intra-group communications[5)] and that required by external cellular communications. In order to realize this optimization, every group member should be aware of all the channel conditions from other group members to itself, namely $\mathcal{H}_n = \{h_{i,n} \ : \ n \in \mathcal{G}_{k,m},\ i = 1, 2, \cdots, |g_{k,m}|,\ i \neq n\}$. Again, this requirement could be satisfied by exploiting D2D outband communications and the associated overhead is negligible for cellular RAN. The update of $\mathcal{H}_n$ for every group member is summarized as line 12 in Algorithm C1.

Herein, we further assume that the transmit power, packet size and bandwidth of the reference signals contained in the intra-group signaling messages are fixed to P, D, and B, respectively. Assume that u_n, u_i are group members of $g_{k,m}$, i.e. $n, i \in \mathcal{G}_{k,m}$. The achievable error-free data-rate from u_i to u_n could be characterized by $R_{i,n} = B \log_2 \left[1 + \frac{P|h_{i,n}|^2}{N_o}\right]$, where N_o denotes the power density of additive Gaussian noise. Accordingly, if we select u_n as the GH $\dot{u}_{k,m}$, the energy required by intra-group communications in ensuing payload transmission slot could be characterized by $\epsilon^{(n)}_{\text{inner}} = \sum_{\substack{i \in \mathcal{G}_{k,m} \\ i \neq n}} P \cdot \frac{D}{R_{i,n}}$.

Similarly, the achievable error-free data-rate from u_n to BS could be characterized by $R_{n,b} = B \log_2 \left[1 + \frac{P|h_{n,b}|^2}{N_o}\right]$. Again, if we select u_n as GH $\dot{u}_{k,m}$, the energy required by external communications between u_n and BS could be characterized by $\epsilon^{(n)}_{\text{outer}} = P \cdot \frac{D}{R_{n,b}}$.

Finally, the energy efficiency of selecting u_n as the GH $\dot{u}_{k,m}$ could be characterized by $\gamma_{k,m}(u_n) = \epsilon^{(n)}_{\text{inner}} + \epsilon^{(n)}_{\text{outer}}$, where a smaller value of $\gamma_{k,m}(u_n)$ implies a better energy efficiency. $\gamma_{k,m}(u_n),\ n \in \mathcal{G}_{k,m}$ will be calculated at the group member u_n and then be forwarded to current GH. Hence, by running the Opt-EC based K-means algorithm, the GH of $g_{k,m}$ could be selected according to $\dot{u}_{k,m} = \underset{u_n \in \mathcal{G}_{k,m}}{\arg\min}\ \gamma_{k,m}(u_n)$, which is summarized as line 13 in Algorithm C1.

The above-mentioned operations can be repeated again among the cellular users for further adjusting the grouping relationships and optimizing GH selections. But, in practice, owing to the limited energy budget, the iterations of Algorithm C1 has to be terminated within a predefined maximum number.

Appendix D Two-Phase Random Access

We summarize the major steps of the proposed two-phase random access procedure in Fig.D1 (a). In more details, as depicted in Fig.D1 (b), we divide the proposed RA procedure into phase-I and phase-II. During phase-I, after receiving the RAO message from the BS, the GHs of all the active groups in the cell will first transmit their cluster preambles $\mathbf{s}^I_k$. It is worth to mention that if the number of clusters is relatively small, then every cluster can be distinguished by a very short orthogonal preamble. It means if K is set to a relatively small value, then the length of $\mathbf{s}^I_k, k = 1, \ldots, K$ could be saved and the overhead is reduced. This is the reason we suggest to set $K = 2, 4, 8, 16$ in Appendix A. Moreover, please bear in mind that $\mathbf{s}^I_k$ of a GH $\dot{u}_{k,m}$ has been specified during the registration process introduced in Appendix A. Finally, a group $g_{k,m}$ is regarded as an active group if one or more members in this group want to transmit payload data to the BS. Accordingly, the signal received at the BS during phase-I can be written as

$$\mathbf{y}^I = \sum_{k=1}^{K} \sum_{m=1}^{M^{(k)}} a_{k,m} \sqrt{P_{k,m}} \mathbf{s}^I_k h_{k,m \to b} + \omega, \tag{D1}$$

where the activity state $a_{k,m} \in \{0, 1\}$ indicates the activity of a group $g_{k,m}$. If $g_{k,m}$ is active, $a_{k,m} = 1$, and so forth. $h_{k,m \to b}$ is the CSI of the channel from the GH $\dot{u}_{k,m}$ to the BS, which simultaneously encapsulates both the large-scale fading and small-scale fading. ω is an AWGN vector, whose elements obey i.i.d complex Gaussian distribution having zero mean and variance σ^2. $P_{k,m}$

5) Because, during the ensuing payload transmission slot, the group members will first send their data to the GH, then the GH relays all the data to the BS.

is the actual transmit power of the GH $\dot{u}_{k,m}$, which is given by

$$P_{k,m} = P \cdot \frac{\beta_{min}}{\beta_{k,m}}, \tag{D2}$$

where P is a common transmit power that can be afforded by all the GHs. The value of P could be assigned to the users during their registration process. $\beta_{k,m}$ is the average large-scale fading coefficient of the channel from the GH $\dot{u}_{k,m}$ to the BS. It could be continuously updated by testing the reference signals transmitted by the BS, e.g. the reference signals included in the registration response message, the GUO message and the RAO message. β_{min} is the minimum value among $\beta_{k,m}, k \in \{1, 2, \cdots, K\}, m \in \{1, 2, \cdots, M^{(k)}\}$, which could be carried in the RAO message. By substituting (D2) to (D1), it is apparent that specifying the actual transmit power $P_{k,m}$ according to (D2) is equivalent to employing an adaptive power control mechanism. Hence, at the BS, the average power of the signal received from $\dot{u}_{k,m}$ approximately equals to $P \cdot \beta_{min}$ [18].

During a RAO, the number of active groups in a cluster c_k is termed as its cluster load. Hence, based on the received signal $\mathbf{y}^I$, the BS is capable of estimating the cluster load of c_k, which is given by

$$\hat{M}_{ac}^{(k)} = \frac{\langle \mathbf{y}^I, \mathbf{s}_k^I \rangle}{\sqrt{P\beta_{min}}}. \tag{D3}$$

Correspondingly, the sparsity of cluster c_k is approximated by

$$\hat{\lambda}_k = \frac{\hat{M}_{ac}^{(k)}}{M^{(k)}}. \tag{D4}$$

After the cluster load estimation formulated by (D3), the BS could rank the access priority of different clusters according to their cluster loads $\hat{M}_{ac}^{(k)}$. A higher access priority is assigned to the cluster having a larger cluster load. As indicated by the largest streams in the middle and bottom of Fig.D1 (b), the GHs in the highest loaded cluster will firstly send their group access preambles and their payload data, respectively. Based on the cluster load estimation, as well as the compressive sensing theorem, the BS will adaptively select the group preamble length of $L_k = |\mathbf{s}_{k,m}^{II}|$ for different clusters. After ranking the access priority and selecting the group preamble length, the BS could arrange the access slots of every cluster. Then, the cluster-specific access slots and preamble lengths that will be used in phase-II are broadcasted by the BS. This message is called as the "Phase-II solution message" (PSM) in Fig.D1 (b). More details of DPS strategy is provided in Appendix E.

After receiving the phase-II solution messages, the GHs of all the active groups will generate their unique preambles of $\mathbf{s}_{k,m}^{II}$. The associated generation method has been determined in the user registration process[6)] and the preamble length is indicated by the phase-II solution messages. At this moment, the proposed two-phase RA procedure starts its phase-II operations. Firstly, the GHs of all the active groups in the same cluster will simultaneously send their group preambles during a specific access slot that has been indicated by the phase-II solution messages, which are illustrated by the shadowed arrows having dashed, solid, and dotted borders in the middle of Fig.D1 (b). During the access slot of cluster c_k, the signal received at the BS is given by

$$\mathbf{y}_k^{II} = \sum_{m=1}^{M^{(k)}} a_{k,m}\sqrt{P_k}\mathbf{s}_{k,m}^{II} h_{k,m\to b} + \omega_k = \sqrt{P_k}\mathbf{S}_k^{II}\mathbf{x}_k + \omega_k, \tag{D5}$$

where $a_{k,m}$ and $h_{k,m\to b}$ have been defined in (D1). ω_k is the AWGN vector in the access slot of c_k. P_k is the standard transmission power of all the GHs in phase-II. Let L_k denote the length of $\mathbf{s}_{k,m}^{II}$, then we have $\mathbf{S}_k^{II} = [\mathbf{s}_{k,1}^{II}, \mathbf{s}_{k,2}^{II}, \cdots, \mathbf{s}_{k,M^{(k)}}^{II}] \in \mathbb{R}^{L_k \times M^{(k)}}$ and $\mathbf{x}_k = [x_{k,1}, x_{k,2}, \cdots, x_{k,M^{(k)}}]^T$, where $x_{k,m} = a_{k,m}h_{k,m\to b}$.

As illustrated in the middle of Fig.D1 (b), after the active group detection of all the clusters are completed, the BS will broadcast payload data transmission solution (PDTS) message to every cluster. The active group ID detected by the BS and its payload data transmission time slot are carried in the PDTS message. Finally, as observed at the bottom of Fig.D1 (b) that the GH of every active group will relay the payload data of their group members to the BS at different time slots as indicated by PDTS messages.

Appendix E Dynamic Preamble Selection algorithm

In the context of our active group detection and according to the CS theorem [19, 36], to satisfy a certain detection accuracy, the minimum preamble length is related to the number of active groups in a cluster, namely $M_{ac}^{(k)}$ and to the total number of groups in a cluster, namely $M^{(k)}$. Apparently, in practice, the value of $M_{ac}^{(k)}$ and $M^{(k)}$ shall be salient different in different clusters. It implies employing a preamble with inappropriate length will result in either a serious detection inaccuracy or an excessive overhead. Accordingly, a DPS algorithm shown in algorithm E1 is designed and employed in phase-II of the random access. The technical challenge of algorithm E1 occurs at its line 3 that evaluates the MPL. Owing to analyzing complexity and importance, we specifically discuss it in Appendix G.

Appendix F MMSE Denoiser Based AMP Algorithm

The classical system model used in compressive sensing is represented as

$$\mathbf{y} = \mathbf{A}\mathbf{x} + \omega, \tag{F1}$$

where $\mathbf{x}$ is the original signal vector having a number of M elements. $\mathbf{A} \in \mathbb{R}^{L\times M}$ is the measurement matrix. ω is an AWGN vector. Since $M \gg L$, $\mathbf{y}$ is actually a compressed and corrupted observation of $\mathbf{x}$. As an efficient solution of recovering $\mathbf{x}$ from $\mathbf{y}$, the approximated message passing (AMP) algorithm is first proposed in [34]. Its theoretical derivation could be found in [35]. In more detail, the AMP algorithm could be formulated by the following iterative procedures

$$\mathbf{x}^{t+1} = \eta_t(\mathbf{A}^*\mathbf{z}^t + \mathbf{x}^t, \tau_t), \tag{F2}$$

6) In our work, complex gaussian sequences having zero mean and variance $\frac{1}{|\mathbf{s}_{k,m}^{II}|}$ are employed as $\mathbf{s}_{k,m}^{II}$.

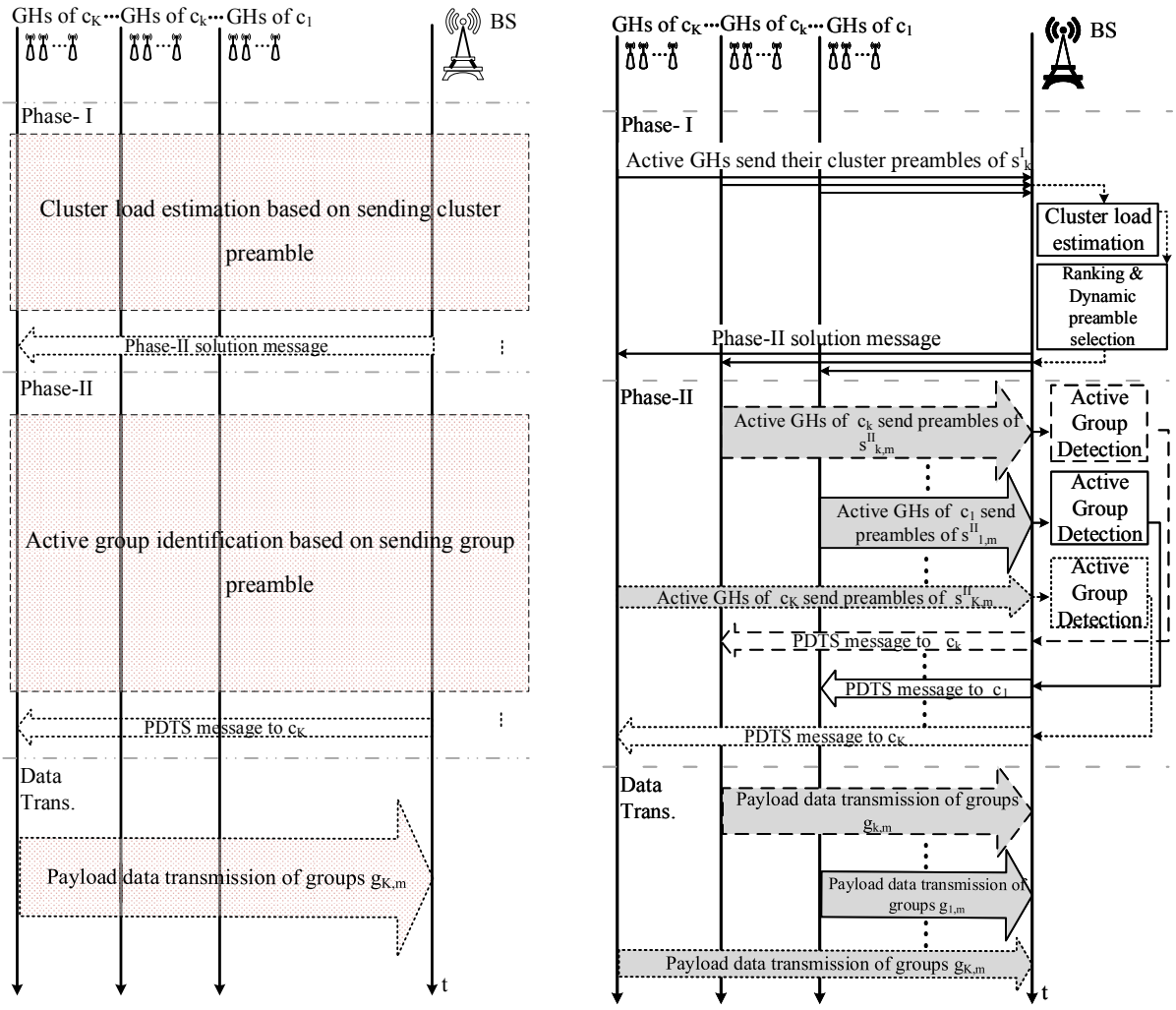

Figure D1 (a) Main steps of the two-phase random access procedure. (b) More details of the two-phase random access.

Algorithm E1 Dynamic preamble selection algorithm.

Require: $\boldsymbol{y}^I$;

Ensure: $\boldsymbol{V}$, $\boldsymbol{L}=[L_1,L_2,\cdots,L_K]$; // $\boldsymbol{V}$ indicates the access slot indices of every cluster, $\boldsymbol{L}$ indicates the preamble lengths selected for every cluster.

Initialize: K, target pF, target pM, $\left\{M^{(k)},k=1,2,\cdots,K\right\}$, $\left\{[R_1^{(k)},R_2^{(k)}],k=1,2,\cdots,K\right\}$; //$R_1^{(k)},R_2^{(k)}$ are the inner and outer radius of the k^{th} cluster, respectively.

1: **for** ($k=1$ **to** K) **do**
2: $\hat{M}_{ac}^{(k)}=$ Estimate_Cluster_Load($\boldsymbol{y}^I,k$); // according to (D3).
3: $L_k=$ Lower_Bound_on_MPL($\hat{M}_{ac}^{(k)},M^{(k)},R_1^{(k)},R_2^{(k)},pF,pM$); // according to (G8), (G10) and (G13).
4: $\boldsymbol{L}[k]=1.1*L_k$; // slightly enlarge L_k, see Appendix G.
5: $\hat{\mathbf{M}}_{ac}[k]=\hat{M}_{ac}^{(k)}$;
6: **end for**
7: $\boldsymbol{V}=$ Allocate_Access_Slot($\hat{\mathbf{M}}_{ac},\boldsymbol{L}$); //arrange the access priority of every cluster in the descending order of $\hat{M}_{ac}^{(k)}$, then allocate the access slot indices of every cluster according to $\boldsymbol{L}$.
8: **return** $\boldsymbol{V}$, $\boldsymbol{L}$.

$$\mathbf{z}^{t+1}=\mathbf{y}-\mathbf{A}\mathbf{x}^{t+1}+\frac{1}{\mu}\mathbf{z}^t\langle\eta_t^{'}(\mathbf{A}^*\mathbf{z}^t+\mathbf{x}^t;\tau_t)\rangle, \tag{F3}$$

$$\tau_t\approx\frac{1}{\sqrt{L}}\|\mathbf{z}^t\|_2, \tag{F4}$$

where the variables have the same definitions as that in [35, (1)∼(2)]. Furthermore, in [16], the soft thresholding denoiser $\eta_t(\cdot)$ is developed to an MMSE denoiser as follows

$$\eta_t(\hat{x}_m^t,g_m)=\mathbb{E}[X|\hat{X}^t=\hat{x}_m^t,G=g_m], \tag{F5}$$

where X, $\hat{X}^t$, $\hat{x}_m^t$, and g_m have the same definitions as that in [16].

This MMSE denoiser can employ the large-scale fading coefficient G known at the BS as a priori information of AMP algorithm. Hence it results in a better recovery accuracy. The above-stated MMSE based AMP algorithm is employed to solve the active group detection problem by replacing the classical compressive sensing model given in (F1) with the group access model given in (D5). Accordingly, the variables $\mathbf{y}$, $\mathbf{A}$, $\mathbf{x}$, ω involved in (F2)∼(F5) are replaced by the variables $\mathbf{y}_k^{II}$, $\mathbf{S}_k^{II}$, $\mathbf{x}_k$, ω_k given in (D5), respectively. The under-sampling ratio of μ in (F3) is calculated by $\frac{L_k}{M^{(k)}}$. The number of total elements M and nonzero elements M_{ac} in $\mathbf{x}$ is replaced by that of total groups $M^{(k)}$ and active groups $M_{ac}^{(k)}$ in a cluster c_k, respectively.

Consequently, in the application of active group identification, the formulation of MMSE based AMP algorithm could be rewritten as

$$\mathbf{x}_k^{t+1}=\eta_t(\mathbf{S}_k^{II*}\mathbf{z}^t+\mathbf{x}_k^t,\tau_t), \tag{F6}$$

$$\mathbf{z}^{t+1}=\mathbf{y}_k^{II}-\mathbf{S}_k^{II}\mathbf{x}_k^{t+1}+\frac{M^{(k)}}{L_k}\mathbf{z}^t\langle\eta_t^{'}(\mathbf{S}_k^{II*}\mathbf{z}^t+\mathbf{x}_k^t;\tau_t)\rangle, \tag{F7}$$

$$\tau_t\approx\frac{1}{\sqrt{L_k}}\|\mathbf{z}^t\|_2. \tag{F8}$$

Appendix G Theoretical Analysis on the Minimum Preamble Length

According to Appendix E, finding the MPL required by MMSE based AMP algorithm for achieving a certain data recovery accuracy makes great sense. Similar works have been attempted in [19] and [36]. However, two deficiencies of the MPL calculation method given in [36] prevent us from applying it in our DPS algorithm: (1) two constant parameters, namely C_1 and C_2 are involved, i.e., instead of an exact value, it only provides an asymptotical order; (2) it does not relate to a specific data reconstruction method. Furthermore, the authors of [37] and [38] attempted to answer this fundamental question from the perspective of classical asymptotic information theoretic analysis. In [38], seeking for the MPL is termed as the "minimum user-identification cost" problem. In their Gaussian many-access channel (MnAC) model, the minimum number of channel uses for guaranteeing an error-free random user identification is given by

$$L=\frac{N\cdot H_2(\frac{N_{ac}}{N})}{\frac{1}{2}\log(1+N_{ac}\gamma)}, \tag{G1}$$

where N denotes the total number of cellular users. In contrast, N_{ac} denotes the average number of active cellular users. γ denotes the signal-to-noise ratio (SNR) and it is assumed in [38] that every active user is subject to the same power constraint of γ. Besides, the entropy function is defined as $H_2(p)=-p\log(p)-(1-p)\log(1-p)$. However, the theorems provided in [38] are still not suitable in our scenarios owing to two reasons: (1) the result shown in (G1) does not relate to any specific active user detection algorithm, hence a significant gap may exist between the MPL required by AMP algorithm and that predicted by (G1), as illustrated in Fig.G1; (2) only Gaussian noisy channels are considered. However, both large-scale and small-scale fading effects are taken into account in our system for modelling a more practical random access scenario.

In the following, we will provide a tight lower bound on MPL for MMSE-AMP algorithm. The state evolution method proposed in [16] is employed, where the mean square error (MSE) of data reconstruction is regarded as a state variable. In more detail, at

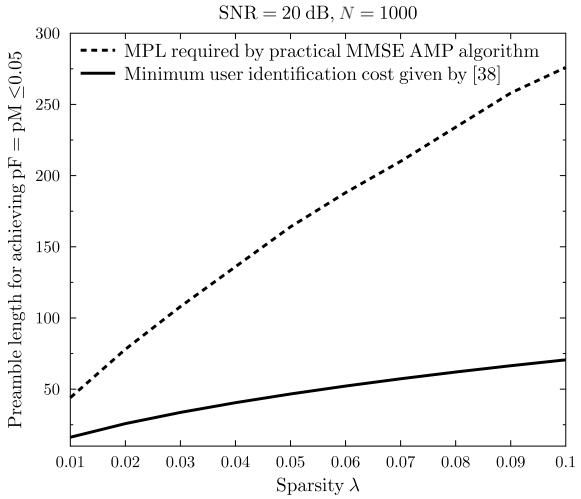

Figure G1 MPL versus sparsity. The MPL required by MMSE-AMP algorithm and that predicted by [38] are compared. AWGN channels are assumed.

every iteration of MMSE based AMP algorithm, $\hat{X}^t$ in (F5) is modeled as a noise corrupted signal. Hence, $\hat{X}^t$ could be formulated as

$$\hat{X}^t = X + \tau_t \cdot \omega_0, \tag{G2}$$

where the random variable ω_0 follows complex Gaussian distribution with zero mean and unit variance. Then, τ_t is referred to as the variance of the t^{th} estimation $\hat{X}^t$. Particularly, according to [16], τ_t is given by

$$\tau_{t+1}^2 = \frac{\sigma^2}{P_k L_k} + \frac{M^{(k)}}{L_k} MSE(\tau_t), \tag{G3}$$

where σ^2 is the variance of background noise ω involved in (D1). The function $MSE(\cdot)$ evaluates the MSE of its input variable and is specified in [16].

Based on the state evolution method [16], in order to achieve a high data reconstruction accuracy, the recursive reconstruction progress formulated by (F2)-(F4) has to converge. It means the variance of the AMP estimation, namely τ_t should constantly decrease to $\frac{\sigma^2}{P_k L_k}$. Hence, the following inequality holds

$$\tau_{t+1}^2 \leqslant \tau_t^2, \forall t. \tag{G4}$$

By substituting (G3) into the inequality (G4), a lower bound on preamble length for satisfying the convergence of AMP algorithm is given by

$$L_k \geqslant \frac{\frac{\sigma^2}{P_k} + M^{(k)} MSE(\tau_t)}{\tau_t^2}. \tag{G5}$$

In order to evaluate the performance of AMP algorithm, two metrics are invoked: (1) the probability of missed detection (pM) in cluster c_k; (2) the probability of false alarm (pF) in cluster c_k. They are defined as

$$pM^{(k)} = \frac{\sum_{m=1}^{M^{(k)}} \{\hat{x}_{k,m} < \theta \ \& \ x_{k,m} \neq 0\}}{\sum_{m=1}^{M^{(k)}} \{x_{k,m} \neq 0\}}, \tag{G6}$$

$$pF^{(k)} = \frac{\sum_{m=1}^{M^{(k)}} \{\hat{x}_{k,m} \geqslant \theta \ \& \ x_{k,m} = 0\}}{\sum_{m=1}^{M^{(k)}} \{x_{k,m} = 0\}}, \tag{G7}$$

where $x_{k,m}$ is defined in (D5), $\hat{x}_{k,m}$ is the estimation of $x_{k,m}$ given by the AMP based active group detection, and θ denotes the decision threshold employed by AMP algorithm. While $\hat{x}_{k,m} \geqslant \theta$, AMP algorithm will regard $g_{k,m}$ as an active group. Then, according to the state evolution method, the $\mathrm{pM}^{(k)}$ and $\mathrm{pF}^{(k)}$ that can be achieved in the t^{th} AMP iteration could be characterized by

$$\begin{cases} pF^{(k)} = e^{-\frac{\theta^2}{\tau_t^2}}, \\ pM^{(k)} = \frac{1}{M} \sum\limits_{m=1}^{M} (1 - e^{-\frac{\theta^2}{\tau_t^2 + g_m^2}}) = \int (1 - e^{-\frac{\theta^2}{\tau_t^2 + g^2}}) \cdot P_G^{(k)}(g)\, dg, \end{cases} \tag{G8}$$

where $P_G^{(k)}(g)$ is the probability density function of the large-scale fading coefficient g. The random variable g takes both the path-loss effect and the shadowing effect into account. In more detail, the path-loss effect is modeled as $\alpha + \beta log_{10}(d)$, where d is the distance between a group head and the BS. The shadowing effect follows log-normal distribution with a variance of σ_s^2.

In practical applications, we aim at a target performance of pM and pF, namely pM_{obj} and pF_{obj}, respectively. Then, by substituting the target pM_{obj} and pF_{obj} into (G8), the appropriate decision threshold θ and the required variance of t^{th} AMP

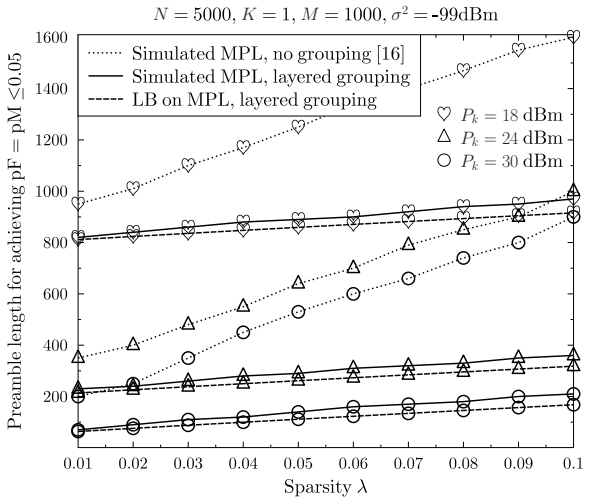

Figure G2 Comparison between simulated MPL and its lower bound with respect to different transmit powers. The practical preamble length required by conventional "no grouping" RA scheme [16] is provided for comparison.

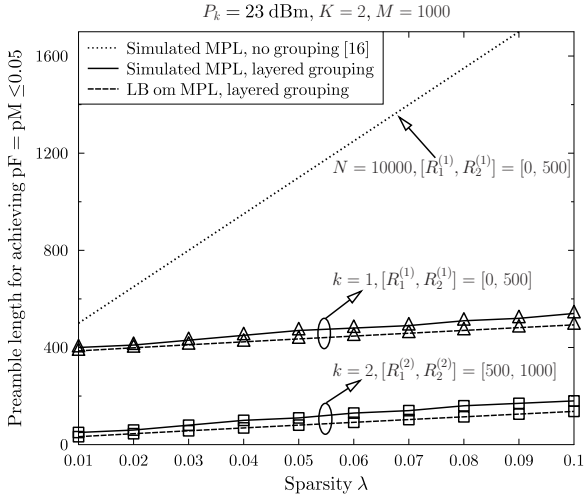

Figure G3 Comparison between simulated MPL and its lower bound with respect to different cluster coverages. The practical preamble length required by conventional "no grouping" RA scheme [16] is provided for comparison.

estimation τ_t can be determined while given the large-scale fading model. The associated solutions of θ and τ_t could be denoted as θ_{obj} and τ_{obj}, respectively.

Bear the above statements in mind, in order to obtain θ_{obj} and τ_{obj}, we shall specify the large-scale fading model. According to the proposed layered grouping network framework shown in Fig.A1, the users of a cluster uniformly locates in the same ring, whose inner and outer radius are represented by R_1 and R_2, respectively. Hence the distance between a user and the BS obeys $d \sim [R_1, R_2]$. Accordingly, the probability density function of large-scale fading coefficient could be formulated as

$$P_G(g) \triangleq a_1 g^{-\gamma_1} Q_1(g) - a_2 g^{-\gamma_2} Q_2(g), \tag{G9}$$

where

$$\begin{cases} a_1 = \frac{40}{(R_2-R_1)^2\beta\sqrt{\pi}} exp(\frac{2(\ln 10)^2\sigma_s^2}{\beta^2} - \frac{2\ln(10)\alpha}{\beta}), \\ a_2 = \frac{40R1}{(R_2-R_1)^2\beta\sqrt{\pi}} exp(\frac{(\ln 10)^2\sigma_s^2}{2\beta^2} - \frac{\ln(10)\alpha}{\beta}), \\ \gamma_1 \triangleq \frac{40}{\beta} + 1, \gamma_2 \triangleq \frac{20}{\beta} + 1, \\ Q_i(g) = \int_{b\ln g + c_{i1}}^{b\ln g + c_{i2}} exp(-s^2) ds, i \in \{1, 2\} \\ c_{i2} = \frac{-\alpha - \beta log_{10}(R_1)}{\sqrt{2}\sigma_s} - \frac{20}{i\beta b}, i \in \{1, 2\} \\ c_{i1} = \frac{-\alpha - \beta log_{10}(R_2)}{\sqrt{2}\sigma_s} - \frac{20}{i\beta b}, i \in \{1, 2\} \\ b = -\frac{10\sqrt{2}}{\ln(10)\sigma_s}. \end{cases} \tag{G10}$$

According to the state evolution method, if the AMP algorithm always achieves the target performance of pM_{obj} and pF_{obj}, then the following inequality has to be true as long as a sufficient large iteration number t is chosen

$$\tau_{t+1} \leqslant \tau_{obj} \leqslant \tau_t. \tag{G11}$$

By substituting (G3) into the above inequality, it results in

$$L_k \geqslant \frac{\frac{\sigma^2}{P_k} + M^{(k)} MSE(\tau_t)}{\tau_{obj}^2}. \tag{G12}$$

Then, it is provable that MSE $(\cdot)$ is a monotonically increasing function in the region of $g \in [0, 100]$. Hence we have $MSE(\tau_{obj}) \leqslant MSE(\tau_t)$ in practical scenarios. It implies replacing $MSE(\tau_t)$ by $MSE(\tau_{obj})$ in (G12) will yield a relaxed lower bound (LB) on MPL, which is given by

$$L_k \geqslant \frac{\frac{\sigma^2}{P_k} + M^{(k)} MSE(\tau_{obj})}{\tau_{obj}^2}. \tag{G13}$$

For example, in practice, simultaneously achieving $pM_{obj} = 0.05$ and $pF_{obj} = 0.05$ may be an acceptable active group detection performance. While considering the large-scale fading model given in (G9)-(G10), the associated solution of (G8) is $\theta_{obj} \approx 8.65 \times 10^{-8}, \tau_{obj} \approx 5 \times 10^{-8}$. Assuming that network configurations including the SNR, the cluster size $M^{(k)}$, as well as the sparsity λ are known. Then, by substituting $\tau_{obj} \approx 5 \times 10^{-8}$ into (G13), we could calculate the exact lower bound on MPL that enables the AMP algorithm to achieve the target active user detection performance.

It is evidenced in Fig.G2 that the lower bounds on MPL (G13) get quite close to the actual MPLs estimated by the Monte Carlo simulations for different transmit powers, although their discrepancy will be slightly enlarged while increasing the sparsity λ. The comparison between simulated MPL and its lower bound with respect to different coverage areas is illustrated in Fig.G3. Fig.G3

Table H1 System Configuration.

Parameter	Value
Radius of the cell	$1000m$
Path-loss model	$15.3 + 37.6 \log_{10}(d(m))$
Variance of shadowing σ_s^2	8
Background noise N_o	-99dBm
Total cellular devices N	$10000, 20000$
Number of clusters K	$2, 4, 8$
Group size $\lvert g_{k,m} \rvert$	5
Number of groups per cluster $M^{(k)}$	$M = 1000, 2000$
Preamble type and length of $\mathbf{s}_k^I$	Walsh Seq., $\lvert \mathbf{s}_k^I \rvert = 32$
TX power P defined in (D2)	23dBm
Preamble type of $\mathbf{s}_{k,m}^{II}$	Gaussian Random Seq.
TX power P_k defined in (D5)	23dBm

demonstrates that the discrepancy between the simulated MPL and its lower bound will be impacted by different cluster coverages. This phenomenon is due to the fact that the large-scale fading effect will be impacted by the cluster coverage as formulated in (G9) and (G10). On the condition of having a low sparsity of $\lambda \leqslant 0.05$, this discrepancy would not exceed 10% of the theoretical lower bound. Therefore, in algorithm E1, we first calculate the lower bound on MPL for a specific cluster c_k, then the L_k employed by the active groups in c_k will be 10% higher than the lower bound.

Furthermore, the conventional no grouping GF RA scheme proposed in [16], which also employs the MMSE based AMP algorithm, is shown in Fig.G2 and Fig.G3 as well. It is demonstrated in Fig.G2 and Fig.G3 that the MPL required by the proposed layered grouping based RA scheme is always significantly less than that required by its counterpart in [16] regardless of different SNR values, different sparsities, different coverages. For the sake of fair comparison, the total number of cellular users remains the same in Fig.G2 and Fig.G3.

Appendix H Simulation Results and Discussions

In this appendix, the active user detection performance of the proposed two-phase DPS aided RA scheme is simulated. The obtained results are compared with the classical CS aided RA that does not exploit any grouping strategy [16] and with the conventional group paging aided RA that does not exploit CS technologies [39]. Without loss of generality, the group size $\lvert g_{k,m} \rvert$ and the number of groups in different clusters $M^{(k)}$ are prefixed to constants regardless of the indices of k and m. Other system parameters are listed in Table.H1.

In Fig.H1, the pF, pM versus transmit power in the two-phase DPS aided RA is compared with that of CS aided RA [16], which also employs the MMSE based AMP algorithm. It is a general consensus that $pF = pM$ implies a good performance balance of active user detection can be achieved. Hence the decision threshold employed by MMSE based AMP algorithm is adjusted for achieving $pF = pM$.

Comparing the solid line labelled by diamonds with the dashed line labelled by crosses in Fig.H1, the two-phase DPS aided RA achieves a dramatic power gain with respect to the conventional CS based RA [16] while using the same preamble length of 400. Comparing the solid line labelled by triangles with the solid line labelled by squares in Fig.H1, the grouping strategy of $K = 4,\ M = 1000$ achieves a better performance than that of $K = 2,\ M = 2000$, even they employ the same preamble length of 800. It implies a RA power gain is available by adjusting the number of clusters and the number of groups.

Then, we define the probability of successful detection (pS) in cluster c_k as

$$pS^{(k)} = \frac{\sum_{m=1}^{M^{(k)}} \{\hat{x}_{k,m} \geqslant \theta \ \&\ x_{k,m} \neq 0\}}{\sum_{m=1}^{M^{(k)}} \{x_{k,m} \neq 0\}}. \tag{H1}$$

According to this definition and equation (G8), pF, pM and pS have following relationships

$$\begin{cases} pS^{(k)} = 1 - pM^{(k)} = 1 - \int (1 - e^{-\frac{\theta^2}{\tau_t^2 + g^2}}) \cdot P_G^{(k)}(g) dg, \\ pS^{(k)} = 1 - \frac{1-\lambda_k}{\lambda_k} pF^{(k)} = 1 - \frac{1-\lambda_k}{\lambda_k} e^{-\frac{\theta^2}{\tau_t^2}}. \end{cases} \tag{H2}$$

The parameter τ_t involved in (H2) can be determined according to (G3). Hence, (H2) enable us to theoretically analyze the active group detection performance of the proposed RA strategy.

As a result, the pS achieved by preamble length fixed strategy is compared with that achieved by DPS strategy in Fig.H2, where both of them employ two-phase RA framework and the grouping strategy is fixed to $K = 4, M = 1000$. Observe at Fig.H2 that while employing a predefined preamble length of $\lvert \mathbf{s}_{k,m}^{II} \rvert = 64,\ 128$, the pS of grouped RA still rapidly drops along with the growth of sparsity. In contrast, benefiting from the DPS aided RA scheme, the system is capable of achieving a high pS probability throughout the entire sparsity region. Actually, the DPS strategy approaches a similar active group detection performance to a preamble length fixed counterpart having $\lvert \mathbf{s}_{k,m}^{II} \rvert = 256$. However, the average preamble length required by the DPS strategy is always less than 210 throughout the entire sparsity region. It means the DPS strategy will further save considerable overhead.

We would like to further evaluate the energy consumption of two-phase DPS aided RA. According to interpretation in Appendix B, the transmission energy required for constructing the layered grouping network framework is negligible. Because the group

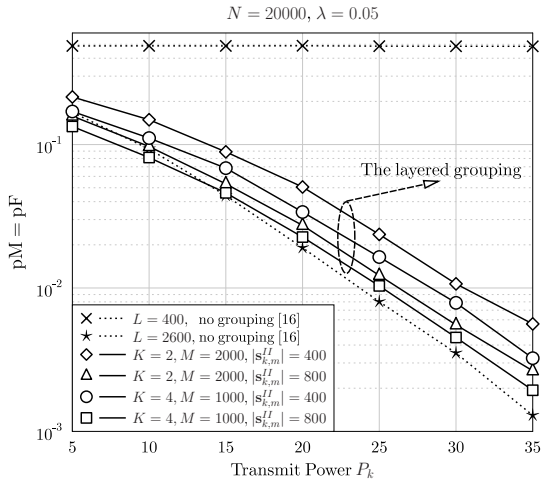

Figure H1 pF, pM versus transmit power of the two-phase DPS aided RA. It is clear that the layered grouping scheme employing a preamble length of only 800 can approach the performance of no grouping scheme [16] which employs a very long preamble length of 2600.

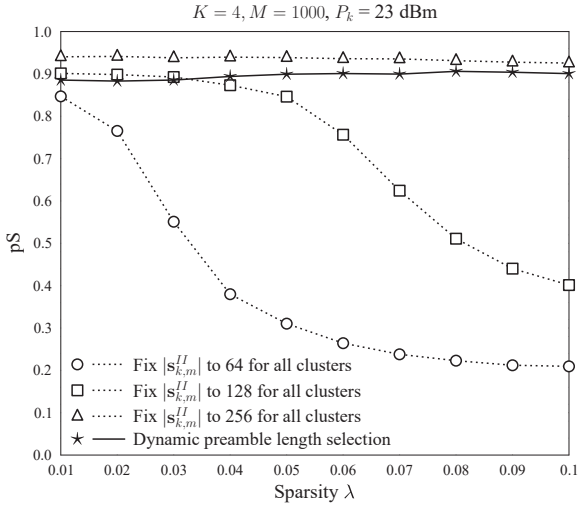

Figure H2 pS versus sparsity performance of two-phase DPS aided RA, where preamble length fixed strategy is compared with. The preamble length required by dynamic scheme is always less than 210 after averaged over all clusters.

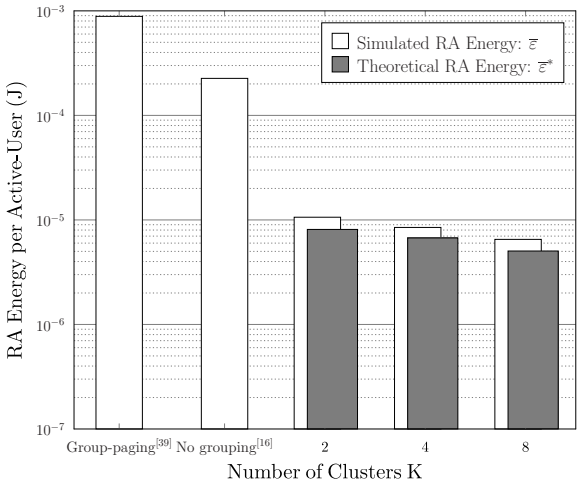

Figure H3 RA energy comparison between the proposed scheme, group-paging RA scheme [39] and no grouping RA scheme [16], where $N = 20000$, $\lambda = 0.05$, K increases from 2 to 8.

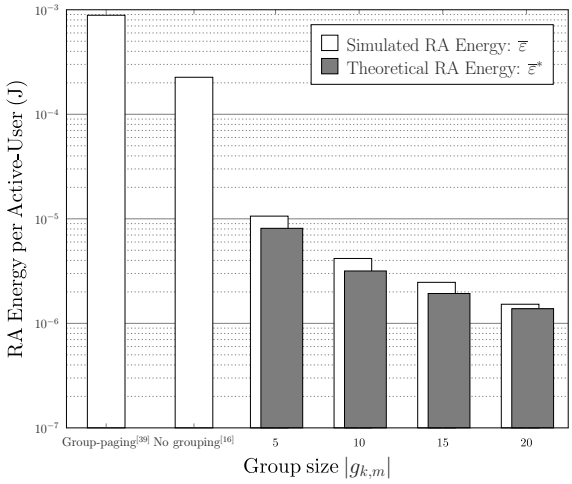

Figure H4 RA energy comparison between the proposed scheme, group-paging RA scheme [39] and no grouping RA scheme [16], where $N = 20000$, $\lambda = 0.05$, $|g_{k,m}|$ increases from 5 to 20.

construction only happens once after a GH user is registered. Besides, the group update only happens once during a very long period. Hence we only focus on the random access energy desired on the active user side in the entire "Phase-I" plus "Phase-II" durations. As shown in Fig.D1 (b), it consists of six parts: (1) energy required by transmitting cluster preambles for the sake of cluster load estimation, i.e. $\epsilon_{\text{pmb-I}} = \frac{1}{N\cdot\lambda}\sum_{k=1}^{K}\sum_{m=1}^{M^{(k)}} a_{k,m}P_{k,m}|\mathbf{s}_k^I|$, (2) energy consumed by waiting for the Phase-II solution message, i.e. $\epsilon_{\text{wait-I}} = P_{\text{wait}} \cdot \overline{T}_{\text{wait-I}}$, (3) energy required by processing the received Phase-II solution message, i.e. $\epsilon_{\text{pcs-I}} = P_{\text{pcs}} \cdot \overline{T}_{\text{pcs-I}}$, (4) energy required by transmitting access preambles for the sake of active group identification, i.e. $\epsilon_{\text{pmb-II}} = \frac{1}{N\cdot\lambda}\sum_{k=1}^{K}\sum_{m=1}^{M^{(k)}} a_{k,m}P_{k,m}|\mathbf{s}_k^{II}|$, (5) energy consumed by the waiting model (as defined in [40]) in the entire "Phase-II" duration, i.e. $\epsilon_{\text{wait-II}} = P_{\text{wait}} \cdot \overline{T}_{\text{wait-II}}$, (6) energy required by processing the PDTS message, i.e. $\epsilon_{\text{pcs-II}} = P_{\text{pcs-II}} \cdot \overline{T}_{\text{pcs-II}}$. Hence, the average random access energy per active user could be calculated as:

$$\overline{\epsilon} = \epsilon_{\text{pmb-I}} + \epsilon_{\text{wait-I}} + \epsilon_{\text{pcs-I}} + \epsilon_{\text{pmb-II}} + \epsilon_{\text{wait-II}} + \epsilon_{\text{pcs-II}} \tag{H3}$$

In the above energy parts, $P_{k,m}$ and P_k are defined in (D2) and (D5), respectively. Thus, the values of P_{wait}, P_{pcs}, $\overline{T}_{\text{wait-I}}$ and $\overline{T}_{\text{wait-II}}$ are specified according to similar parameters given in [39]- [40]. Particularly, $\overline{T}_{\text{wait-II}}$ is the average waiting time of an active group required in the entire "Phase-II" duration, which equals to subtracting length of $|\mathbf{s}_k^{II}|$ and $\overline{T}_{\text{pcs-II}}$ from duration of

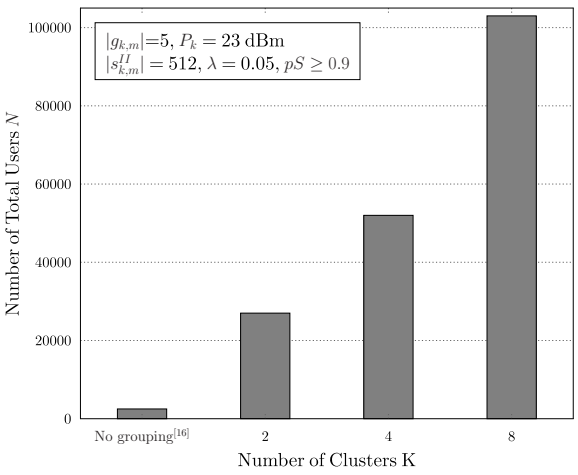

Figure H5 Comparison on affordable number of coexisted users in a cell between the proposed RA and no grouping CS aided RA.

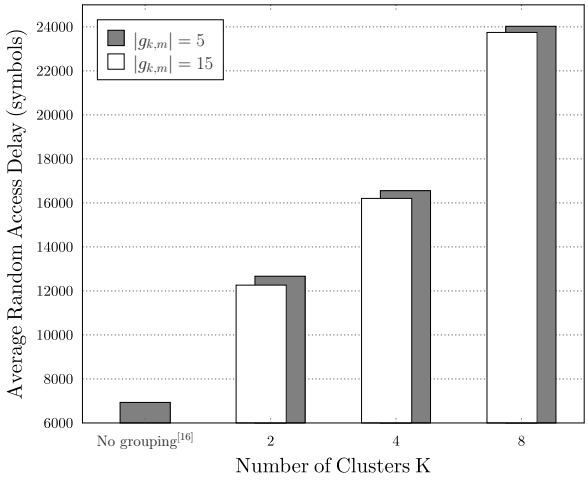

Figure H6 Access delay versus number of clusters in the proposed two-phase RA scheme.

"Phase-II" [7]. More specifically, according to LTE standard, up to 839 symbols can be transmitted within a single time slot (i.e. 0.5 ms). Hence, we equivalently employ the number of symbols as our time metric.

Furthermore, if we replace practical preamble length $|\mathbf{s}_k^{II}|$ used in (H3) by its theoretical lower bound (LB) specified in (G13), the LB of $\epsilon_{\mathrm{pmb-II}}$ is given by

$$\epsilon_{\mathrm{pmb-II}}^{*} = \frac{1}{N \cdot \lambda} \sum_{k=1}^{K} \sum_{m=1}^{M^{(k)}} a_{k,m} P_k \frac{\frac{\sigma^2}{P_k} + M^{(k)} MSE(\tau_{obj})}{\tau_{obj}^2}, \tag{H4}$$

Simultaneously, the length of $\overline{T}_{\mathrm{wait-II}}$ is also minimized, which results in the minimization of $\epsilon_{\mathrm{wait-II}}$, namely $\epsilon_{\mathrm{wait-II}}^{*}$. Substitute $\epsilon_{\mathrm{pmb-II}}^{*}$ and $\epsilon_{\mathrm{wait-II}}^{*}$ into (H3), we refer to the resultant $\bar{\epsilon}$ as theoretical RA energy of our two-phase DPS aided RA, namely $\bar{\epsilon}^{*}$.

Observe at Fig.H3 and Fig.H4 that on the condition of approaching the same $pS \geqslant 90\%$, the two-phase DPS aided RA scheme achieves a significant power gain compared with both the conventional no grouping CS based RA scheme [16] and the conventional no CS aided group paging strategy [39]. There are three major advantages of the proposed layered grouping RA scheme, i.e. (1) in the proposed system, only GH of an active group has to send access preamble, while in [16], [39], every active user has to send access preamble; (2) GHs also have better channel conditions than other group members owing to the Opt-EC based K-means grouping algorithm; (3) the DPS algorithm given in Algorithm E1 effectively reduces the preamble length used in phase-II. In more details, the energy consumed per active user is reduced as the number of clusters K is increased, as illustrated in Fig.H3, while the associated penalty is the increase of RA delay as shown in Fig.H6 per active user versus. On the other hand, as can be seen from Fig.H4, the energy consumed per active user can be reduced more significantly when the group size is increased. However, it may become impractical if the group size is too big due to the complexity and synchronization requirements.

The capability of DPS aided RA scheme to support the massive connectivity is depicted in Fig.H5, which is measured in the maximum number of coexisted users in a cell while fixing the preamble length, the sparsity, the transmit power, as well as the target level of pF and pM. As can be observed from Fig.H5, for the given condition, the proposed RA scheme is capable of supporting more than 10^5 users. In contrast, no grouping strategy [16] can only support approximately 2500 users for the same amount of physical resources.

Finally, the average time required by an active group head $\dot{u}_{k,m}$ for completing random access procedure is adopted as our delay metric. According to the proposed RA procedure in Fig.D1 (b) and in line with our RA energy analysis, the RA delay of $\dot{u}_{k,m}$ consists of six major components: (1) The time $\overline{T}_1$ required by all active GHs for transmitting their cluster preambles of $|\boldsymbol{s}_k^I|$, $k = 1, 2, \cdots, K$; (2) the waiting time $\overline{T}_{\mathrm{wait-I}}$ during "Phase-I"; (3) the time $\overline{T}_{\mathrm{pcs-I}}$ required for processing phase-II solution message; (4) the time $\overline{T}_2$ required by an active GH for transmitting its group access preamble; (5) the waiting time $\overline{T}_{\mathrm{wait-II}}$ during "Phase-II" and (6) the time $\overline{T}_{\mathrm{pcs-II}}$ required for processing PDTS message. Again, the number of symbols is employed as the time metric.

The random access delay performance of the proposed DPS aided RA scheme is demonstrated in Fig.H6, where $\lambda = 0.05$, $N = 20000$. Observe at Fig.H6 that the proposed RA scheme imposes a higher random access delay than the conventional no grouping CS based RA scheme, especially when the number of clusters K grows. But, benefiting from the DPS scheme, employing a large group size could slightly mitigate the latency. Consequently, our proposal may be more suitable for the mMTC devices which have a relative higher tolerance of time delay. Fortunately, benefiting from the DPS aided RA scheme, the random access delay will not linearly increase with respect to the number of clusters.

References

1 3GPP, IMT-2020 connection density evaluation results (mMTC), R1-1808866, Aug. 2018.
2 Bockelmann C, Pratas N K, Wunder G, et al. Towards massive connectivity support for scalable mMTC communications in 5g networks, IEEE Access, 2018, 6: 28969-28992.
3 3GPP, IMT-2020 self evaluation: mMTC coverage, data rate, latency & battery life, R1-1905187, Apr. 2019.
4 J. Choi, On fast retrial for two-step random access in mtc, IEEE Internet of Things Journal, 2021,8: 1428-1436.

7) The duration of "Phase-II" is determined by system configurations of N,K,λ etc. It can be calculated by experimental method as shown in Fig.H6.

5 3GPP, IMT-2020 revised work item proposal: 2-step RACH for NR, RP-190711, Mar. 2019.
6 3GPP, Discussion on grant-free transmission, R1-166095, Aug. 2016.
7 3GPP, Grant-based and grant-free multiple access for mMTC, R1-164268, May. 2016.
8 Shahab M B, Abbas R, Shirvanimoghaddam M, et al. Grant-free non-orthogonal multiple access for iot: A survey, IEEE Communications Surveys Tutorials, 2020, 22(3): 1805-1838.
9 Ding Z, Yang Z, Fan P, et al. On the performance of non-orthogonal multiple access in 5g systems with randomly deployed users, Signal Processing Letters, IEEE, 2014, 21(12): 1501-1505.
10 Bayesteh A, Yi E, Nikopour H, et al. Blind detection of scma for uplink grant-free multiple-access, in 2014 11th International Symposium on Wireless Communications Systems (ISWCS), 2014.
11 Wanwei, Tang, Shaoli, et al. Uplink grant-free pattern division multiple access (GF-PDMA) for 5g radio access, China Communications, 2018.
12 Yuan Z, Yan C, Yuan Y, et al. Blind multiple user detection for grant-free musa without reference signal, in 2017 IEEE 86th Vehicular Technology Conference (VTC-Fall), 2017.
13 3GPP, Contention-based non-orthogonal multiple access for ul mMTC, R1-164269, May 2016.
14 Fazel F, Fazel M, and Stojanovic M, Random access compressed sensing over fading and noisy communication channels, IEEE Transactions on Wireless Communications, 2013, 12(5): 2114-2125.
15 Hong J, Choi W, and Rao B D, Sparsity controlled random multiple access with compressed sensing, IEEE Transactions on Wireless Communications, 2015, 14(2): 998-1010.
16 Chen Z, Sohrabi F, and Yu W, Sparse activity detection for massive connectivity, IEEE Transactions on Signal Processing, 2018, 66: 1890-1904.
17 Liu L and Yu W, Massive connectivity with massive MIMO part -I: Device activity detection and channel estimation, IEEE Transactions on Signal Processing, 2018, 66: 2933-2946.
18 Senel K and Larsson E G, Grant-free massive MTC-enabled massive MIMO: A compressive sensing approach, IEEE Transactions on Communications, 2018, 66(12): 6164-6175.
19 Candes E J and Tao T, Decoding by linear programming, IEEE Transactions on Information Theory, 2005, 51:4203-4215.
20 Donoho D L and Elad M, Optimally sparse representation in general (nonorthogonal) dictionaries via l1 minimization, in Proceedings of the National Academy of Sciences of the United States of America, 2013, 100: 2197-2202.
21 3GPP, Considerations and evaluation results for IMT-2020 for mMTC connection density, R1-1903968, Apr. 2019.
22 Sharma S K and Wang X, Toward massive machine type communications in ultra-dense cellular iot networks: Current issues and machine learning-assisted solutions, IEEE Communications Surveys1 & Tutorials, 2020, 22(1): 426-471.
23 Si P, Yang J, Chen S, et al. Adaptive massive access management for QoS guarantees in M2M communications, IEEE Transactions on Vehicular Technology, 2015, 64: 3152-3166.
24 Hu X, Zhong C, Chen X, et al. Cluster grouping and power control for angle-domain MmWave MIMO NOMA systems, IEEE Journal of Selected Topics in Signal Processing, 2019, 13: 1167-1180.
25 Biral A, Centenaro M, Zanella A, et al. The challenges of M2M massive access in wireless cellular networks, Digital Communications and Networks, 2015, 1(1): 1-19.
26 Tsoukaneri G, Condoluci M, Mahmoodi T, et al. Group communications in narrowband-IoT: Architecture, procedures, and evaluation, IEEE Internet of Things Journal, 2018, 5: 1539-1549.
27 Tu C, Ho C, and Huang C, Energy-efficient algorithms and evaluations for massive access management in cellular based machine to machine communications, in 2011 IEEE Vehicular Technology Conference (VTC Fall), 2011: 1-5.
28 Zhan W, Sun X, Li Y, et al. Optimal group paging frequency for machine-to-machine communications in LTE networks with contention resolution, IEEE Internet of Things Journal, 2019, 6: 10534-10545.
29 Lau E and Chung W, Enhanced RSSI-based real-time user location tracking system for indoor and outdoor environments, in 2007 International Conference on Convergence Information Technology (ICCIT 2007),2007: 1213-1218.
30 Bhattacharjee A, Islam M R, and Ullah M A, Weather independent positioning of MS using power slope patterning algorithm via one BS, in 2019 International Conference on Electrical, Computer and Communication Engineering (ECCE), 2019,: 1-3.
31 Han B, Sciancalepore V, Holland O, et al. D2D-based grouped random access to mitigate mobile access congestion in 5G sensor networks, IEEE Communications Magazine, 2019, 57: 93-99.
32 Lianghai J, Han B, Liu M, et al. Applying device-to-device communication to enhance IoT services, IEEE Communications Standards Magazine, 2017, 1(2): 85-91.
33 FUTURE & TIAA V2X Working Group, Spectrum requirements in automatic driving 5G NR V2X direct communications, World 5G Convention, Nov. 2019, Beijing.
34 Donoho D L, Maleki A, and Montanari A, Message passing algorithms for compressed sensing: II. analysis and validation, in 2010 IEEE Information Theory Workshop on Information Theory (ITW 2010, Cairo), 2010: 1-5.
35 Donoho D L, Maleki A, and Montanari A, Message-passing algorithms for compressed sensing, Proceedings of the National Academy of ences of the United States of America, 2009, 106(45): 18914.
36 Kutyniok G, Compressed sensing: Theory and applications, Corr, 2012, 52(4): 1289-1306.
37 Aksoylar C, Atia G, and Saligrama V, Sparse signal processing with linear and non-linear observations: A uni edshannon theoretic approach, in 2013 IEEE Information Theory Workshop (ITW), 2013: 1-5.
38 Chen X, Chen T, and Guo D, Capacity of gaussian many-access channels, IEEE Transactions on Information Theory, 2017, 63(6): 3516-3539.
39 Arouk O, Ksentini A, and Taleb T, Group paging-based energy saving for massive mtc accesses in lte and beyond networks, IEEE Journal on Selected Areas in Communications, 2016, 34(5): 1086-1102.
40 Stefania S, Tou I, and Baker M, LTE - The UMTS Long Term Evolution: From Theory to Practice. John Wiley & Sons, Sept. 2009.