SCIENCE CHINA
Information Sciences

• LETTER •

April 2025, Vol. 68, Iss. 4, 149301:1–149301:2
https://doi.org/10.1007/s11432-024-4279-1

Delay sensitive user association strategy in massive machine-type communications

Qinwen JI[1] & Yongxu ZHU[1,2*]

[1]*National Mobile Communications Research Laboratory, Southeast University, Nanjing 210096, China*
[2]*Purple Mountain Laboratories, Nanjing 211111, China*

Received 23 October 2024/Accepted 17 January 2025/Published online 14 February 2025

Citation Ji Q W, Zhu Y X. Delay sensitive user association strategy in massive machine-type communications. Sci China Inf Sci, 2025, 68(4): 149301, https://doi.org/10.1007/s11432-024-4279-1

The upcoming 6G mobile network is anticipated to meet the performance requirements of a vast number of terminal devices while providing simultaneous service [1]. However, ultra-dense user environments may lead to network congestion and increased latency. Effective user association is critical in addressing challenges such as load balancing, interference mitigation, and seamless handover [2, 3]. Existing user association strategies often focus on offloading tasks from overloaded access points (APs) through interlayer balancing mechanisms [4, 5]. Nevertheless, these strategies typically overlook queueing delays and do not account for reassociation based on the real-time queue states of the APs. To address this gap, this study proposes a minimum delay user association strategy tailored for massive machine-type communications (mMTC). The proposed strategy allows dynamic reassociation to prevent users from blindly waiting for service in their current queue when the network changes dynamically.

System model and problem formulation. In an mMTC scenario with K tiers of APs, the locations of users and the k-th tier APs are modeled as independent homogeneous Poisson point processes denoted by Φ_U with density λ_U for users and Φ_k with density λ_k for the k-th tier APs, where $k = 1, \ldots, K$. The network operates in discrete slots, with each slot τ representing the interval $[t, t+1)$ for $t \in \mathbb{N}^+$ [5]. Let $\Phi_k^a(t) \in \Phi_k$ and $\Phi_U^a(t) \in \Phi_U$ denote the sets of the k-th tier active APs and active users in slot t, respectively, where active means currently transmitting or receiving data. The b_u^t represents the AP associated with user u in slot t, the $\boldsymbol{b}_u = [b_u^1, b_u^{T_0}, \ldots, b_u^{m_u T_0}]$ and $m_u \in \mathbb{N}^+$ denote the set and the number of all APs associated with user u, respectively, and T_0 is the reassociation period. Then, the optimization problem for user u in the downlink is

$$\mathbb{P}_1: \ \hat{\mathcal{D}}_u^{\mathrm{D}} = \arg\min_{\boldsymbol{b}_u \in \mathcal{N}} \left[\mathcal{D}_u^{\mathrm{D}} + \sum\nolimits_{b_u^t \in \boldsymbol{b}_u} \tilde{\mathcal{D}}_{b_u^t, u}^{\mathrm{D}} \right], \tag{1}$$

where $\mathcal{N} = \{\Phi_k\}_{\forall k}$, and $\tilde{\mathcal{D}}_{b_u^{T_0}, u}^{\mathrm{D}}$ represents the queueing delay of user u when associated with $b_u^{T_0}$. The total receiving data delay of user u is given by $\mathcal{D}_u^{\mathrm{D}} = \sum_{t=1}^{m_u T_0} \min[\tau, \frac{a_u^t \rho_u^{\mathrm{D}}(t)}{W \log_2(1+\gamma_{b_u^t, u}^{\mathrm{D}}(t))}]$, where $a_u^t = 1$ if user u is active, and $a_u^t = 0$ otherwise. The $\rho_u^{\mathrm{D}}(t)$ denotes the number of data bits still required by user u in slot t, and $\gamma_{b_u^t, u}^{\mathrm{D}}(t)$ is the signal-to-interference-plus-noise ratio (SINR) between user u and its associated AP b_u^t in slot t. If b_u^t is a k-th tier AP, the $\gamma_{b_u^t, u}^{\mathrm{D}}(t)$ can be written as $\gamma_{b_u^t, u}^{\mathrm{D}}(t) = \frac{P_k h_{b_u^t}^u |X_{b_u^t}^u|^{-\alpha_k}}{\sum_{j=1}^K \sum_{i \in \Phi_j^a(t) \setminus b_u^t} P_j h_{j,i}^u |X_{j,i}^u|^{-\alpha_j} + \sigma^2}$, where $h_{b_u^t}^u \sim \exp(1)$ and $|X_{b_u^t}^u|$ are the small-scale fading channel power gain and the distance between user u and its associated AP, respectively, α_k represents the path loss exponent, and σ^2 stands for the additive white Gaussian noise. Specifically, if $\gamma_{b_u^t, u}^{\mathrm{D}}(t) \geqslant \gamma_0$, then $\rho_u^{\mathrm{D}}(t+1) = \max\{0, \rho_u^{\mathrm{D}}(t) - W \log_2(1 + \gamma_{b_u^t, u}^{\mathrm{D}}(t))\}$, if $\gamma_{b_u^t, u}^{\mathrm{D}}(t) < \gamma_0$, then $\rho_u^{\mathrm{D}}(t) = 0$ with γ_0 is the SINR threshold. In the uplink, the delay $\hat{\mathcal{D}}_u^{\mathrm{U}}$ can be derived similarly by replacing downlink terms with uplink equivalents, e.g., $\gamma_{b_u^t, u}^{\mathrm{D}}(t) \to \gamma_{b_u^t, u}^{\mathrm{U}}(t) = \frac{P_U h_{b_u^t}^u |X_{b_u^t}^u|^{-\alpha_k}}{\sum_{i \in \Phi_U^a(t) \setminus \{u\}} P_U h_{b_u^t}^i |X_{b_u^t}^i|^{-\alpha_k} + \sigma^2}$, where P_U representing the users' transmit power.

Minimum delay user association strategy. The proposed minimum delay strategy is summarized in Algorithm 1, where the superscripts for downlink and uplink are omitted for simplicity. The minimum delay strategy assigns each user to the AP with the minimum sum of queueing delay plus data transmission delay. Specifically, the queueing information $\boldsymbol{\mathcal{M}}_{k,i}(\tilde{\tau})$ of the i-th AP ($i \in \Phi_k$) in the k-th tier (denoted as the (k,i)-th AP) in slot $\tilde{\tau} \in \{\boldsymbol{\varpi}_u\}_{u \in \Phi_U}$ is

$$\boldsymbol{\mathcal{M}}_{k,i}(\tilde{\tau}) = \left\{ \left(Q_{k,i}(\tilde{\tau}), L_{k,i}(\tilde{\tau}), \boldsymbol{r}_{k,i}(\tilde{\tau}) \right) \right\}, \tag{2}$$

where $(\cdot)$ denotes a tuple, and $\boldsymbol{\varpi}_u = \{1, T_0, \ldots, m_u T_0\}$ represents the set of reassociation slots for user u. Here, $L_{k,i}(\tilde{\tau}) = ||\mathcal{U}_{k,i}(\tilde{\tau})||_0$ and $\boldsymbol{r}_{k,i}(\tilde{\tau}) = \{r_{k,i}^u\}_{u \in \mathcal{U}_{k,i}(\tilde{\tau})}$ are the total number of associated users and the rank of users within the (k,i)-th AP in slot $\tilde{\tau}$, respectively. The queuing delay of the (k,i)-th AP in slot $\tilde{\tau}$ is

$$Q_{k,i}(\tilde{\tau}) = \sum\nolimits_{u \in \mathcal{U}_{k,i}(\tilde{\tau})} \frac{\rho_u(\tilde{\tau})}{W \log_2(1 + \gamma_{b_u^{\tilde{\tau}}, u}(\tilde{\tau}))}, \tag{3}$$

* Corresponding author (email: yongxu.zhu@seu.edu.cn)

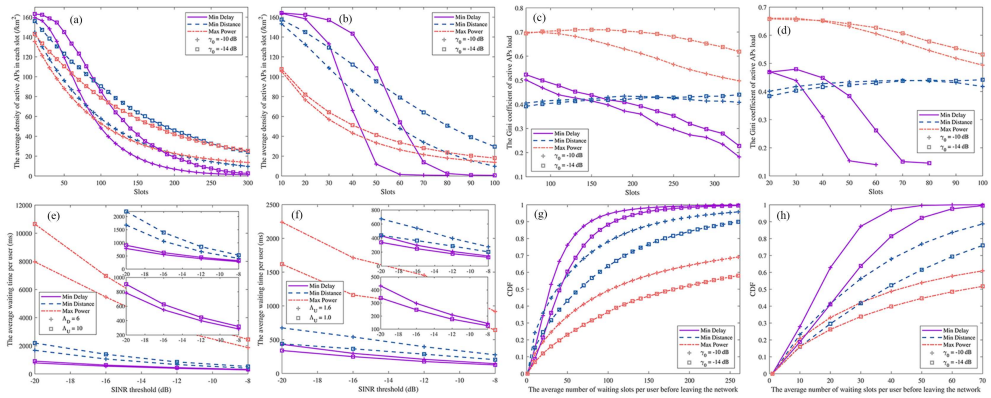

Figure 1 (Color online) The average densities of active APs in (a) the downlink and (b) the uplink; the Gini coefficients of active APs load in (c) the downlink and (d) the uplink; the average waiting time per user before leaving the network in (e) the downlink and (f) the uplink; the cumulative distribution function (CDF) of the average number of waiting slots per user before leaving the network in (g) the downlink and (h) the uplink.

where $\mathcal{U}_{k,i}(\tilde{\tau})$ is the set of users associated with the (k,i)-th AP in slot $\tilde{\tau}$. The value of $\gamma_{b_u^{\tilde{\tau}},u}(\tilde{\tau})$ can either be updated using pilots in slot $\tilde{\tau}$ to enhance the accuracy of $Q_{k,i}(\tilde{\tau})$, or it can use the SINR value measured when the user entered the queue, thereby reducing pilot overhead. The optimal association for user u in $\tilde{\tau} \in \boldsymbol{\varpi}_u$ is

$$b_u^{\tilde{\tau}} = \arg \min_{b_u^{\tilde{\tau}} \in \mathcal{N}} \left[\frac{\rho_u(\tilde{\tau})}{W \log_2(1+\gamma^p_{b_u^{\tilde{\tau}},u}(\tilde{\tau}))} + Q_{b_u^{\tilde{\tau}}}(\tilde{\tau}) \right], \quad (4)$$

where $\gamma^p_{b_u^{\tilde{\tau}},u}(\tilde{\tau})$ denotes the SINR measured by user u in slot $\tilde{\tau}$ using pilot signals. To avoid the scenario where a user at the front of one AP's queue moves to the back of another queue after reassociation, we use queue length and position ranking whether to execute the reassociation operation. Specifically, user u will only update its associated AP only if a random number is smaller than $r^u_{b_u^{\tilde{\tau}}}(\tilde{\tau})/L_{b_u^{\tilde{\tau}}}(\tilde{\tau})$.

Algorithm 1 Minimum delay user association strategy.

Input: $\boldsymbol{\mathcal{M}}_{k,i}(\tilde{\tau})$, $k \in [1,K]$, $i \in \Phi_k$, $\tilde{\tau} \in \boldsymbol{\varpi}_u$.
Output: Latest associated AP $b_u^{\tilde{\tau}}$.
1: Each AP broadcasts its queuing information;
2: u acquires the $r^u_{b_u^{\tilde{\tau}}}(\tilde{\tau})$ and $L_{b_u^{\tilde{\tau}}}(\tilde{\tau})$ in (2) from the queuing information $\boldsymbol{\mathcal{M}}_{b_u^{\tilde{\tau}}}(\tilde{\tau})$ of the associated AP $b_u^{\tilde{\tau}}$;
3: Generate a random number $p \in [0,1]$;
4: **if** $p < \frac{r^u_{b_u^{\tilde{\tau}}}(\tilde{\tau})}{L_{b_u^{\tilde{\tau}}}(\tilde{\tau})}$ **then**
5: u acquires the $Q_{k,i}(\tilde{\tau})$ in (2) from the queuing information $\boldsymbol{\mathcal{M}}_{k,i}(\tilde{\tau})$, $k \in [1,K]$, $i \in \Phi_k$;
6: u measures the SINR by using pilot signals;
7: u updates the associated AP by (4);
8: **else**
9: u does not update the associated AP;
10: **end if**

Numerical results. We consider an mMTC scenario with two tiers of APs and conduct simulations for both downlink and uplink. Data packets for each user follow independent Poisson distributions with means Λ^{D} and Λ^{U}, respectively, and packet lengths follow exponential distributions with means ξ^{D} and ξ^{U}, respectively. The system parameters are set as follows: $\lambda_1 = 15/\mathrm{km}^2$, $\lambda_2 = 150/\mathrm{km}^2$, $\lambda_U = 3000/\mathrm{km}^2$, $\alpha_1 = 3$, $\alpha_2 = 3.5$, $P_1 = 46$ dBm, $P_2 = 38$ dBm, $P_U = 23$ dBm, $\Lambda^{\mathrm{D}} = 6$–10, $\Lambda^{\mathrm{U}} = 1.0$–$1.6$, $\xi^{\mathrm{D}} = 1$ Mb, $\xi^{\mathrm{U}} = 0.5$ Mb, $\sigma^2 = -90$ dBm, $W = 10$ MHz, $\tau = 10$ ms, and $T_0 = 10$. The simulation results are shown in Figure 1. It is evident that the density of active APs in the minimum delay strategy is initially high, yet decreases more rapidly, eventually falling below that of the comparison strategies in later slots. This is because the minimum delay strategy allows users to re-associate with APs that can provide faster service periodically, enabling idle APs to be selected during reassociation slots. This results in a higher initial density of active APs, serving more users simultaneously. As the remaining users decrease, the active AP density quickly drops to zero. Reassociation also balances AP load, preventing overload in some APs while others remain idle. The decreasing Gini coefficient shows a more balanced AP load due to periodic reassociations. Additionally, the minimum delay strategy significantly reduces the average waiting time per user, with CDF curves showing fewer waiting slots before users leave the network compared to other strategies.

Complexity and convergence analysis. In each slot $\tilde{\tau}$, at most $o(||\Phi^a_U(\tilde{\tau})||_0 \times ||\mathcal{N}||_0)$ operations are required, while APs broadcast queue information using $o(||\mathcal{N}||_0)$ operations. Thus, the complexity per reassociation is $o(||\Phi^a_U(\tilde{\tau})||_0 \times ||\mathcal{N}||_0 + ||\mathcal{N}||_0)$, which simplifies to $o(||\Phi_U||_0 \times ||\mathcal{N}||_0 + ||\mathcal{N}||_0)$ since $||\Phi^a_U(\tilde{\tau})||_0 < ||\Phi_U||_0$. Assuming I reassociations, the total complexity is $o(I(||\Phi_U||_0 \times ||\mathcal{N}||_0 + ||\mathcal{N}||_0))$. The proposed minimum delay strategy does not involve an iterative process, then the convergence performance is verified by assessing whether the proposed algorithm can serve all users within a limited number of slots, i.e., reducing the number of active APs to zero, as seen in Figures 1(a) and (b).

Conclusion. This study proposed a minimum delay strategy for mMTC, where users dynamically selected APs based on queueing and transmission delays, effectively balancing AP loads and reducing overall service time.

Acknowledgements This work was supported by Major Science and Technology Project of Jiangsu Province (Grant No. BG2024002).

References

1 You X H. 6G extreme connectivity via exploring spatiotemporal exchangeability. Sci China Inf Sci, 2023, 66: 130306
2 You L, Huang Y F, Zhong W, et al. Robust online energy efficiency optimization for distributed multi-cell massive MIMO networks. Sci China Inf Sci, 2023, 66: 132302
3 Zhu Y, Wang L, Wong K K, et al. Wireless power transfer in massive MIMO-aided HetNets with user association. IEEE Trans Commun, 2016, 64: 4181–4195
4 Haghgoy S, Mohammadi M, Mobini Z, et al. Decoupled UL/DL user association in wireless-powered HetNets with full-duplex small cells. IEEE Trans Veh Technol, 2023, 72: 14341–14355
5 Kong F, Sun X, Guo Y J, et al. Queue-aware power consumption minimization in two-tier heterogeneous networks. IEEE Trans Veh Technol, 2018, 67: 8875–8889