# FairHealth: Long-Term Proportional Fairness-Driven 5G Edge Healthcare in Internet of Medical Things

Xi Lin , Jun Wu , *Member, IEEE*, Ali Kashif Bashir, *Member, IEEE*, Wu Yang, Aman Singh , and Ahmad Ali AlZubi

***Abstract*—Recently, the Internet of Medical Things (IoMT) could offload healthcare services to 5G edge computing for low latency. However, some existing works assumed altruistic patients will sacrifice quality of service for the global optimum. For priority-aware and deadline-sensitive healthcare, this sufficient and simplified assumption will undermine the engagement enthusiasm, i.e., unfairness. To address this issue, we propose a long-term proportional fairness-driven 5G edge healthcare, i.e., FairHealth. First, we establish a long-term Nash bargaining game to model the service offloading, considering the stochastic demand and dynamic environment. We then design a Lyapunov-based proportional-fairness resource scheduling algorithm, which decouples the long-term fairness problem into single-slot subproblems, realizing a tradeoff between service stability and fairness. Moreover, we propose a block-coordinate descent method to iteratively solve nonconvex fair subproblems. Simulation results show that our scheme can improve 74.44% of the fairness index (i.e., Nash product), compared with the classic global time-optimal scheme.**

This work was supported in part by NSFC under Grant 61972255 and Grant U21B2019 and in part by the Researchers Supporting Project under Grant RSP-2021/395, King Saud University, Riyadh, Saudi Arabia.

Xi Lin is with the School of Electronic Information and Electrical Engineering, Shanghai Jiao Tong University, Shanghai 200240, China (e-mail: linxi234@sjtu.edu.cn).

Jun Wu is with the Graduate School of Information, Production, and Systems, Waseda University, Fukuoka 808-0135, Japan (e-mail: jun.wu@ieee.org).

Ali Kashif Bashir is with the Department of Computing and Mathematics, Manchester Metropolitan University, M15 6BH Manchester, U.K. (e-mail: a.bashir@mmu.ac.uk).

Wu Yang is with the College of Computer Science and Technology, Harbin Engineering University, Harbin 200240, China (e-mail: yangwu@hrbeu.edu.cn).

Aman Singh is with the Higher Polytechnic School, Universidad Europea del Atlántico, 39011 Santander, Spain, with the Department of Project Management, Universidad Internacional Iberoamericana, Campeche 24560, México, and also with the Department of Engineering, Universidad Internacional Iberoamericana, Arecibo 00613, Puerto Rico (e-mail: amansingh.x@gmail.com).

Ahmad Ali AlZubi is with the Computer Science Department, Community College, King Saud University, Riyadh 11437, Saudi Arabia (e-mail: aalzubi@ksu.edu.sa).

***Index Terms*—5G, edge healthcare, Internet of Medical Things (IoMT), proportional fairness.**

## I. Introduction

NOWADAYS, the emerging Internet of Medical Things (IoMT)-based healthcare has achieved remote disease diagnosis and patient monitoring, which makes up for the scarcity of medical resources in local clinics, especially in pandemic scenarios, e.g., COVID-19. IoMT-based healthcare is commonly dubbed Medicine 4.0 or Health 2.0 [1], relying on real-time generated medical data from the exponentially adopted diagnostic tools, such as physician's notebooks, nurses' smartphones, and sensor-based patient-monitoring tools. For network delay and data privacy concerns, cloud infrastructure-based IoMT data analysis methods are impractical, while 5G multiaccess edge computing (MEC) equipped with cellular infrastructures (e.g., base stations) has enabled IoMT data proximal storage and processing [2]. By offloading computing-intensive IoMT data services to MEC servers for handling, IoMT devices can break through the bottleneck of their constrained storage, computing, communication, and energy resources [3], thereby realizing low-latency and energy-saving edge healthcare.

However, realizing such a 5G edge healthcare in IoMT still faces considerable challenges from the perspective of service providers (SPs) and requesters (SRs). On the one hand, the randomly generated IoMT data brings stochastic demands, while the dynamic 5G edge environment includes time-varying channel parameters, available resources, etc. For SPs, considering the stochastic demands and network dynamics [4], how to scheduling communications and computing resources for various healthcare services with a high quality of service (QoS) (i.e., lower service delay)? A dynamic scheduling policy provision is required to ensure long-term service stability for 5G edge healthcare, i.e., avoiding long-term service backlog. On the other hand, when serving IoMT-based healthcare to a large number of requesters (i.e., patients), the fairness requirements among them urgently need to be satisfied. Some existing works [5], [6] assume altruistic patients will sacrifice QoS for global optimum coordination, which is not often feasible in the absence of a fairness guarantee. For instance, some patients who request emergency healthcare services (e.g., remote surgery) cannot tolerate

the QoS reduction. Thus, fairness-driven 5G edge healthcare should comprehensively customize offloading policies for requesters with different priorities and deadlines. In summary, it is significant to design a long-term stable and fairness-guaranteed edge healthcare service scheme in 5G-enabled IoMT.

We investigate the fairness topic in the view of utilitarianism. In 5G edge healthcare, there are various distinct stakeholders, including different patients and SPs. Intuitively, patients would like to engage in the 5G edge healthcare if they obtain higher QoS than if they were not to participate, which implies some fairness concepts. The patients' payoffs are fairly distributed during the service process, and individuals are not needed to sacrifice their own QoS in service offloading. The individuals are not needed to sacrifice their own QoS in service offloading. In fact, rational individuals who observe their own QoS consistently detrimental will cease the engagement [7]. Meanwhile, for almost all high-priority and delay-sensitive healthcare services, the life-related QoS should not be allowed to reduce. We consider that 5G edge healthcare should be established on fairness, even if it may result in a global performance reduction. Although the global optimal performance seems very attractive, from a practical point of view, it will undermine the engagement enthusiasm and service feasibility. Moreover, considering the long-term dynamics of edge networks, fairness also needs to achieve long-term dynamic adaptation.

Despite this, most existing MEC task offloading algorithms [5], [6] consider altruistic nodes just struggle to achieve a global optimum, which could not be applied in scenarios where multiple parties participate. There are also some works [8]–[10] focusing on general fair resource allocation in MEC but do not consider priority-aware and deadline-sensitive service characteristics in healthcare sectors. Consequently, we propose a long-term proportional fairness-driven edge healthcare scheme, referred to as FairHealth. By supplying deadline-sensitive proportional-fairness service configuration for priority-aware healthcare applications, our FairHealth could effectively improve long-term patients' QoS, acting as exogenous incentives for patients' engagement in healthcare. The contributions of our work are summarized as follows.

1) We propose a novel FairHealth scheme, which realizes a long-term proportional fairness-guaranteed 5G edge healthcare in IoMT.
2) We establish and analyze a long-term dynamic Nash bargaining game for priority-aware and deadline-sensitive healthcare services, jointly considering service stability and proportional fairness.
3) We design a dynamic fairness-aware resource online scheduling algorithm via Lyapunov optimization technology and block coordinate descent (BCD) method.
4) We conduct extensive simulations to show the QoS improvement in our scheme. And we demonstrate the proportional fairness could be ensured compared to the existing works.

The rest of this article is organized as follows. We discuss the related works in Section II. We then present the motivation and system model of our work in Section III. The long-term proportional fair healthcare problem is formulated as a dynamic Nash bargaining game in Section IV. Besides, we propose a Lyapunov-based resource scheduling algorithm with proportional fairness in Section V. In Section VI, through extensive simulations, we show the advanced performance of our scheme. Finally, Section VII concludes this article.

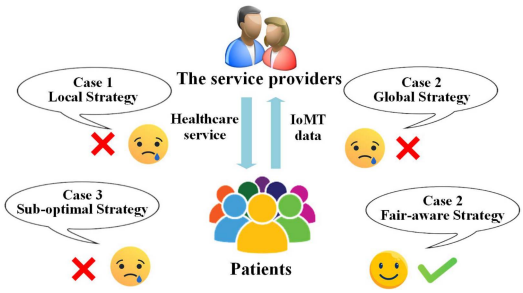
## II. Related Works

### A. IoMT-Enabled Smart Healthcare

IoMT is a new paradigm of Internet of Things (IoT) that provides an intelligent healthcare service. The IoMT system can realize remote disease monitoring for patients. Therefore, it can provide patients with timely diagnoses and save their lives in emergency situations. However, it needs further advancement. The framework, challenges, and future issues were discussed and analyzed in [11]. Philip *et al.* [11] also illustrated the factors that drive the development of IoT-based in-home health monitoring systems. They claimed that the private medical records of the patients should be carefully processed and stored for modern society. Some work focuses on the collection and analysis of IoMT data [12], [13]. For instance, Usman *et al.* [12] established a general data collection and analysis framework for IoMT applications. The proposed framework divided the underlying wireless multimedia sensor network into multiple clusters. Each cluster is responsible for aggregating IoMT data and extracting meaningful information on the cloud. Peng *et al.* [13] designed a highly concurrent and massive IoMT data collection algorithm. The ability of IoMT data-parallel collection can be well realized in [13]. However, these works [12], [13] have not considered the dynamic nature of IoMT data collection, that is, time-series and stochastic data generation. Some other works are concerned about IoMT security. Ghubaish *et al.* [14] comprehensively overviewed IoMT systems' potential physical and network attacks during data collection, communication, and storage. Deebak *et al.* [15] proposed a secure and privacy-preserving cloud-based medical healthcare framework. In addition, the blockchain technique was adopted in [16] to realize secure and decentralized medical data sharing among entities in the IoMT system.

### B. 5G MEC-Based Smart Healthcare

The low latency feature of 5G will greatly facilitate the development of healthcare services. Existing research work [17]–[21] mainly focuses on how to effectively jointly optimize the computing, communication, and storage resources of the MEC system to provide delay-sensitive healthcare services. Ning *et al.* [17] constructed an MEC-enabled 5G IoMT which can minimize the system-wide cost. And the utility-optimal wireless channel resource strategies in [17] were derived by the decentralized noncooperative game. They also theoretically calculated the upper bound of the time complexity of this framework. Computing resource allocation is also an issue that needs paying attention to. Lin *et al.* [18] designed a three-tier edge network framework for smart healthcare applications in terms of communication, computing, and service. They also proposed an algorithm that optimizes the resource allocation and




| Patients | Case 1 | Case 2 | Case 3 | Case 4 |
|---|---|---|---|---|
| Patient 1 | 0.50 | 0.80 | 0.50 | 0.47 |
| Patient 2 | 1.00 | 1.09 | 1.00 | 0.93 |
| Patient 3 | 1.50 | 1.31 | 1.50 | 1.39 |
| Patient 4 | 2.00 | 1.50 | 1.75 | 1.84 |
| Patient 5 | 2.50 | 1.66 | 1.94 | 2.30 |
| Total delay | 7.50 | 6.36 | 6.69 | 6.93 |

Fig. 1. Motivational example of 5G edge healthcare.

computation offloading in the MEC scenario. Thus, the system can realize efficient medical data storing and requesting. For the lack of global information, the resource allocation for 5G MEC-based healthcare faces challenges. Zhou *et al.* [19] proposed a learning-based task offloading approach under the constraints of ultralow-latency communications (URLLC). There are also some studies focusing on intelligent resource allocation. The integration of edge computing and artificial intelligence gives birth to edge intelligence. Hayyolalam *et al.* [20] have proposed an improved edge intelligence framework for IoMT-based healthcare systems. Rahman *et al.* [21] proposed a B5G network architecture that leverages the features of 5G like low-latency and high-bandwidth to realize a better remote diagnosis. Distributed deep learning was also integrated into the framework proposed by the authors.

In summary, in terms of IoMT, most existing works [12], [13] often ignore the dynamic time-series IoMT data generation process, making it difficult to meet the stochastic services demand. In terms of edge computing, most existing MEC resource allocation schemes [17]–[21] ignore the dynamic fair service configuration of priority-aware healthcare for patients. Unlike existing works, we focus on dynamic proportional fairness-powered 5G edge healthcare, while guaranteeing stochastic healthcare service stability.

## III. Motivation and System Model

In this section, to show our design purpose more clearly, we start with a simple motivational example of 5G edge healthcare in Fig. 1, which presents the service delay of four cases in the mini edge healthcare system. After that, we then introduce our system model, i.e., the long-term dynamic service offloading model for 5G edge healthcare.

### A. Motivational Example of Fair Healthcare

We employ a simple edge healthcare service offloading system as a motivating instance. Considering one MEC server and five nearby patients requesting IoMT-based service offloading. It is desirable that both parties collaborate to determine how to allocate computation resources, e.g., to improve QoS for all patients. Intuitively, each patient will selfishly focus on some concept of individual payoff. In 5G edge healthcare, we measure this payoff as the QoS they obtain during the offloading, i.e., to greatly reduce the service delay. The computation of the MEC server and local IoMT devices is 30 and 5 GHz, respectively. The healthcare data size of five patients is [5,10,15,20,25] Mbit, respectively. The required CPU cycles of each data packet are 500 cycles/bit. The communication time of all the patients is simply denoted as 0.1 s. We compare four resource scheduling strategies for four optimization goals as follows.

*Case 1:* Local strategy, where the patients just perform data processing locally. Even in the absence of communication time, the constrained local computation cannot enable patients to obtain higher QoS. As the benchmark, the local strategy means that patients will not resort to the 5G edge healthcare system for service offloading.

*Case 2:* Global optimal strategy, where the SP assumes that each patient participates in the 5G edge healthcare system spontaneously. Therefore, the SP will take the total service delay of all patients as the global optimization goal. The total delay reaches the theoretical minimum, i.e., 6.36 s. However, if we check the individual QoS and compare to the local strategy, we observe that patients 1 and 2 will increase in service delay causing a lower QoS.

*Case 3:* Clobal suboptimal strategy, where the SP introduces some restrictions with the total service delay minimization, i.e., the service time of each patient should not exceed the benchmark time in Case 1. The total delay (i.e., 6.69 s) has increased compared to Case 2, while each patient will not reduce QoS compare to Case 1. However, it can be clearly noticed that the QoS of patients 4 and 5 is obviously improved compared to patients 1, 2, and 3.

*Case 4:* Fair-aware strategy, where the SP struggles to fairly improve the QoS of each patient rather than pursue pure global optimum. Fair-aware strategy increases the total service delay from 6.36 to 6.93 s, while compared to Cases 1 and 2, it guarantees that all the patients fairly increase QoS to the same extent, reaching the Pareto-optimal efficiency. All the patients will be encouraged to join such a fairness-guaranteed edge healthcare system.

The above four cases illustrate a serious but easily overlooked issue. When we consider the global optimum, it will always go against the QoS of some participants. Time-sensitive edge healthcare requires a prompt response, such as emergency treatment and remote surgery, where the life-critical QoS must not be reduced. The SP should provide QoS that exceeds the local baseline, and improve the QoS equally among patients. Otherwise, patients who are unfairly treated for global optimum would cancel edge healthcare services. 5G edge healthcare needs to find the optimal tradeoff between overall performance and individual fairness.

### B. System Model of 5G Edge Healthcare

The above motivational example is relatively simple, while the actual system model of 5G edge healthcare is more complicated. In the 5G ultradense heterogeneous network, MEC servers are

TABLE I
MAIN TERMS REFERRED IN OUR ARTICLE

| Symbol | Explanations |
|---|---|
| $\mathcal{L}$ | Patient set requesting edge healthcare |
| $\mathcal{K}$ | MECs set providing healthcare service |
| $\alpha$ | Priority level of healthcare services |
| $s_l[t]$ | Data size of generating IoMT data |
| $c_l[t]$ | Required CPU cycles of the unit data |
| $y_{lk}[t]$ | Edge association choice of patients |
| $\lambda_{lk}[t]$ | Computation allocation rate of MECs |
| $\omega_k[t]$ | Computation capabilities of MECs |
| $T_{lk}^{local}[t]$ | Local computing time of patients |
| $T_{lk}^{cmp}[t]$ | Computing time by service offloading |
| $D_k[t]$ | Data backlog of the MEC servers |
| $\Theta[t]$ | Queue vector of the MEC servers |
| $\Delta(\Theta[t])$ | Lyapunov drift function |

densely deployed, requiring resource collaboration to break single MEC server resource constraints. The interaction between multiple MEC servers and multiple patients will be investigated, thereby introducing the patient association challenge, i.e., where to offload the IoMT-based healthcare service. In addition, there are some dynamic uncertainties over edge networks. For patients, the generated IoMT data are time-varying, resulting in dynamic service demands. For MEC servers, the wireless network status is unstable (e.g., communication bandwidth and channel parameters), and the available resources are also dynamic. All the dynamics are difficult to accurately predict. We employ time slicing to capture the dynamics. The main symbols and explanations in our article are presented in Table I.

In Fig. 2, we show the system model of 5G edge healthcare, including the $L$ service requesters (patients) and $K$ providers (MEC servers). The $L$ patients are denoted as $\mathcal{L}=\{1,2,\ldots,l,\ldots,L\}$. And the $K$ MEC servers located in different regions, which are denoted as $\mathcal{K}=\{1,2,\ldots,k,\ldots,K\}$. Similar to the previous work [4], [5], cloud-based centralized resource control and management of MEC servers could be realized by softwarized networking paradigms, i.e., software-defined networks (SDN). With the information flow tables, the SDN controller can easily obtain the system parameters and achieve more efficient network management. As shown in Fig. 2, we could show the interaction process among the three layers of IoMT, MEC, and cloud. IoMT generates medical data for 5G healthcare, then MEC processes the uploaded IoMT data to provide medical services for patients. And the cloud center is responsible for managing and scheduling resources of multiple MEC servers via the SDN controller.

*1) Service Requesters (patients) in IoMT Layer:* The IoMT will continuously produce healthcare data from the patients. The entire timeline is divided into a series of time slots, i.e., $\Gamma=\{1,2,\ldots,t,\ldots,T-1,T\}$. Considering the Age of Information, healthcare data are required to analyze and process in real time. During time slot $t$, IoMT data will generate from the patients, and the data size is $s_l[t]$, $(\forall l\in\mathcal{L})$. $c_l[t]$ is the required CPU cycles for the unit data packet, which presents the different types of healthcare services. The dynamic $s_l[t]$ and $c_l[t]$ are still upperbounded, i.e., $s_l[t]\le\bar{s}$, $c_l[t]\le\bar{c}$. In time slot $t$, the patients will offload the IoMT data to one MEC server

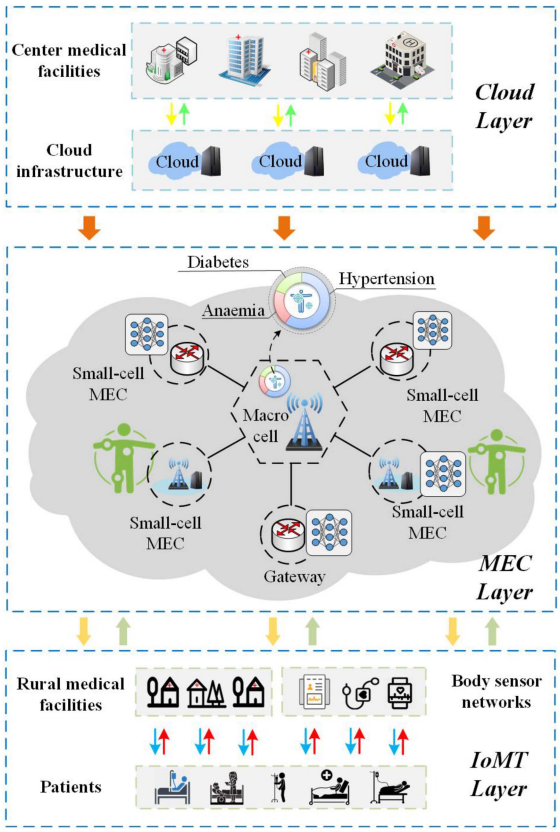



Fig. 2. System model of 5G edge healthcare.

through wireless uplinks. The patient association is denoted as $y_{lk}\in\{0,1\}$, where $\sum_{k=1}^{K}y_{lk}[t]=1,(l=1,2,\ldots,L)$. In 5G ultradense networks [5], [22], [23], we could model the wireless channel gain $H_{lk}[t]$ as follows:

$$H_{lk}[t]=127+30\log_{10}\left(d_{lk}[t]\right)+20\left(1-\delta_{lk}[t]\right) \tag{1}$$

where $\delta_{lk}[t]\in\{0,1\}$, and $\delta_{lk}[t]=1$ is for indoor scene, while $\delta_{lk}[t]=0$ is for outdoor scene. In our work, we mainly consider in-home healthcare, thus $\delta_{lk}[t]=1$. $d_{lk}[t]$ (km) presents the geographic distance between patients and MEC servers. And we consider the distance $d_{lk}[t]$ keeps constant in time slot $t$, thus $H_{lk}[t]$ is also constant in a time period. The transmission rate of the wirelss uplink between the patient $l$ and MEC server $k$ could be formulated as

$$r_{lk}[t]=B_{lk}[t]\cdot\log_2\left(1+\frac{H_{lk}[t]\cdot P_{lk}[t]}{\sigma_{lk}^2[t]+\theta_{lk}[t]}\right) \tag{2}$$

where $B_{lk}[t]$ is channel bandwidth, $P_{lk}[t]$ presents the transmit power, $\sigma_{lk}^2[t]$ denotes the environmental noise, and $\theta_{lk}[t]$ shows the intercell interference noise. Therefore, the uplink communication time $T_{lk}^{\text{com}}[t]$ is presented as

$$T_{lk}^{\text{com}}[t]=\sum_{k=1}^{K}\frac{y_{lk}[t]\cdot s_l[t]}{r_{lk}[t]},\quad y_{lk}[t]\in\{0,1\}. \tag{3}$$

*2) Service Requesters (MEC Servers) in MEC Layer:* For each MEC server, the arriving IoMT data are considered

as a single server sequence. That is, the arriving healthcare service follows the first-in–first-out processing principle. We consider the computation capacities of the MEC servers as $\Omega[t]=\{\omega_1[t],\omega_2[t],\ldots,\omega_k[t],\ldots,\omega_K[t]\}$, while the local IoMT devices of patients are $V[t]=\{\nu_1[t],\nu_2[t],\ldots,\nu_l[t],\ldots,\nu_L[t]\}$. The computation capacity shows the CPU cycles that can be executed in unit time. Thus the computing time $T_l^{\text{local}}[t]$ and $T_{lk}^{\text{cmp}}[t]$ could be formulated as follows:

$$T_l^{\text{local}}[t]=\frac{c_l[t]s_l[t]}{\nu_l[t]}, T_{lk}^{\text{cmp}}[t]=\sum_{k=1}^{K}\frac{y_{lk}[t]\cdot c_l[t]\cdot s_l[t]}{\omega_k[t]\cdot\lambda_{lk}[t]} \quad (4)$$

where $\lambda_{lk}[t]$ presents the computation allocation rate, and $\sum_{l=1}^{L}\lambda_{lk}[t]\leq 1,(k=1,2,\ldots,K)$. And the downlink communication time is ignored due to the smaller feedback data and faster downlink speed, thus we have

$$T_{lk}^{\text{offload}}[t]=\sum_{k=1}^{K}y_{lk}[t]\cdot s_l[t]\cdot\left(\frac{1}{r_{lk}[t]}+\frac{c_l[t]}{\omega_k[t]\cdot\lambda_{lk}[t]}\right). \quad (5)$$

The amount of data and the number of services reaching to the MEC server $k$ can be expressed as

$$Z_k[t]=\sum_{l=1}^{L}y_{lk}[t]\cdot c_l[t]\cdot s_l[t], M_k[t]=\sum_{l=1}^{L}y_{lk}[t]. \quad (6)$$

Let $D_k[t]$ be the data backlog of the MEC server $k$. Thus, the data backlog update rule is formulated as

$$D_k[t+1]=\max\{D_k[t]-t\omega_k[t],0\}+\sum y_{lk}[t]c_l[t]s_l[t]. \quad (7)$$

Thus, $D_k[t]$ are considered as dynamic service queues. Considering the dynamic congestion control of the healthcare service, we build the virtual queues $X_k[t]$ to model the dynamic number of healthcare services [4], [24], i.e.,

$$X_k[t+1]=\max\{X_k[t]-m_k,0\}+\sum y_{lk}[t] \quad (8)$$

where $m_k$ indicates the upper-bound number of services at MEC server $k$. According to the Lyapunov queue theory [4], [24], if service queues $D_k[t]$ are stable, the long-term average arriving rate $\sum y_{lk}[t]c_l[t]s_l[t]$ will not exceed the queue serving rate $t\omega_k[t]$. Also, if virtual queues $X_k[t]$ are stable, the long-term average arriving rate $\sum y_{lk}[t]$ will not exceed the queue serving rate $m_k$. We could realize both the service stability and congestion avoidance of each MEC server via the control of service and virtual queues, respectively. Now, we introduce the mean-rate stability of queues as follows:

$$\lim_{t\to\infty}\frac{\mathbb{E}\{|D_k[t]|\}}{t}=0, \lim_{t\to\infty}\frac{\mathbb{E}\{|X_k[t]|\}}{t}=0, (\forall k\in\mathcal{K}). \quad (9)$$

If the above (9) is satisfied, the long-term average queue arriving rate will not exceed the queue serving rate, i.e.,

$$\lim_{T\to\infty}\sup\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\{Z_k[t]\}\leq t\omega_k[t], (\forall k\in\mathcal{K}) \quad (10)$$

$$\lim_{T\to\infty}\sup\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\{M_k[t]\}\leq m_k[t], (\forall k\in\mathcal{K}). \quad (11)$$

TABLE II
COMPARISON OF NBS AND KS

| Items | Axioms | NBS | KS |
|---|---|---|---|
| (1) | Pareto optimality | ✓ | ✓ |
| (2) | Scale invariance | ✓ | ✓ |
| (3) | Symmetry | ✓ | ✓ |
| (4) | Independence | ✓ | × |
| (5) | Monotonicity | × | ✓ |

## IV. Dynamic Nash Bargaining Game Formulation

In this section, we aim to find a fair solution for 5G edge healthcare services. In specific, we model a Nash bargaining game to realize a balance of high system performance and patient fairness. The Nash bargaining game could enable a multiparities collaboration via fair utility value allocation.

Fairness should strictly satisfy the five axioms [7], [25] in Table II. Nash has proved that the unique equilibrium in the multiparities Nash bargaining game could satisfy axioms (1)–(4), which is defined as Nash bargaining solution (NBS). We consider $\mathrm{u}^*$ as a NBS in $\mathcal{P}$ for $\mathrm{u}^{\min}$, i.e., $\mathrm{u}^*=g(\mathcal{P},\mathrm{u}^{\min})$, the axioms of the NBS could be explained as follows.

1) Pareto optimality: $\mathrm{u}^*$ is Pareto-optimal; no individual can be better off without making at least one individual worse off or without any loss thereof.
2) Scale invariance: If $\mathrm{u}^*\in\mathcal{P}'\subset\mathcal{P}$, $\mathrm{u}^*=g(\mathcal{P},\mathrm{u}^{\min})$, then we have $\mathrm{u}^*=g(\mathcal{P},'\mathrm{u}^{\min})$.
3) Symmetry: If $\mathcal{P}$ is invariant under all exchanges of players (patients), $g_l(\mathcal{P},\mathrm{u}^{\min})=g_k(\mathcal{P},\mathrm{u}^{\min})\ \forall l,k\in\mathcal{L}$.
4) Independence: For any linear scale transformation $\Upsilon$, we have $\Upsilon(g(\mathcal{P},\mathrm{u}^{\min}))=g(\Upsilon(\mathcal{P}),\Upsilon(\mathrm{u}^{\min}))$.

And the Kalai–Smorodinsky solution (KS) could satisfy axioms (1)–(3) and (5). Table II presents the comparison of NBS and KS. Both the NBS and KS have the corresponding fairness metrics. In our work, we pay more attention to NBS and corresponding proportional fairness, for NBS commonly has a closed-form expression with equal treatment among multiple players (i.e., patients), compared with KS.

### A. Dynamic Bargaining Game and Proportional Fairness

The patients act as players in the bargaining game, which try to maximize their own utility. In 5G edge healthcare, the utility of the patient could be concretized as the QoS during the service offloading, which is inversely proportional to service time. All the patients would struggle to obtain the lowest possible service time. Thus, selfish patients strive to optimize their QoS via the combination of utility-optimal edge association $y_{lk}[t]$ and computation schedule $\lambda_{lk}[t]$. Obviously, if QoS can be maximized by only local strategy, the patient will not turn to the edge healthcare system. We now define the NBS to address the fairness issue as **P1**, i.e.,

$$\max_{y_{lk}[t],\lambda_{lk}[t]}\prod_{l=1}^{L}\left(\alpha T_l^{\text{local}}[t]-T_l^{\text{offload}}\left(y_{lk}[t],\lambda_{lk}[t]\right)\right). \quad (12)$$

As discussed in Section III-B, we denote $T_l^{\text{local}}[t]$ to present the initial disagreement value for patients $l$, which is also the worst

QoS a patient would join the 5G healthcare service offloading. In other words, the service time $T_l^{\text{offload}}[t]$ for each patient should be no higher than the local service time. This ensures that each patient can improve their QoS by participating in 5G edge healthcare. In practice, it is necessary to consider the priorities of different IoMT services in 5G healthcare, and it will not harm the fairness of patients. For example, the priority of emergency services is higher than ordinary medical report analysis. As is shown in (12), we introduced $\alpha$ to represent the priority parameter. When $\alpha$ is smaller, $\alpha T_l^{\text{local}}[t]$ is smaller, which represents a smaller initial disagreement value. A smaller and stricter initial disagreement value (or serving deadline) means that patients have a stronger priority for the service. Therefore, healthcare services with a smaller $\alpha$ have a higher priority. In this article, we consider the priority-sensitive service delay for edge healthcare, Nash product optimization goal in **P1** tries to proportionally reduce the service delay of each patient in order to achieve fairness, i.e., NBS. We then formulate the formal Nash bargaining game and define the fairness for 5G edge healthcare.

*Definition 1:* The service offloading bargaining game is formulated as a tuple $\mathcal{P}[t] = \{\Lambda[t], \Pi[t]\}$, where $\Lambda[t]$ includes all the serving time values via service offloading. And $\Pi[t]$ consists of all the disagreement values (i.e., local time) as a breakdown point in bargaining.

The fair service offloading game is accompanied by a Nash bargaining solution. We are required to find Pareto-efficient edge association and resource schedule strategies to uniquely maximize the Nash product. By taking the logarithm and negation of (12), **P1** is equivalently transformed to **P2**

$$\min U^{\text{total}}[t] = -\sum_{l=1}^{L} \ln\left(\alpha T_l^{\text{local}}[t] - T_l^{\text{offload}}[t]\right). \tag{13}$$

*Definition 2: Proportional fairness*: The service offloading strategies $\{y_{lk}^*[t], \lambda_{lk}^*[t]\}$ satisfy the proportional fairness if and only if $\forall (y_{lk}[t], \lambda_{lk}[t]) \neq (y_{lk}^*[t], \lambda_{lk}^*[t])$, we have

$$\sum \frac{T_l^*(y_{lk}^*[t], \lambda_{lk}^*[t]) - T_l^{\text{offload}}(y_{lk}[t], \lambda_{lk}[t])}{T_l^{\text{local}}[t] - T_l^*(y_{lk}^*[t], \lambda_{lk}^*[t])} < 0. \tag{14}$$

We believe that each patient needs to improve QoS compared to local computing during the service offloading, and the improvement among patients is equal, following some concepts of fairness. If patients are treated unfairly, they will further go away from the edge healthcare system. There are three classic and clearly defined fairness, that is, egalitarian, max-min, and proportional fairness. Egalitarian fairness cannot achieve Pareto efficiency, which is not commonly adopted in practice. Proportional fairness is built on NBS, while max-min fairness is defined by KS. In our work, we concentrate on NBS, thus we just consider proportional fairness in Definition 2.

We could consider a service offloading strategy is proportional fairness if the reassociation or allocation of any patients would increase the proportional service time of a patient by less than the aggregated reduced service time for others. For instance, a patient adopts a reassociation or allocation strategy. If such an operation increases the service time of patient A by 30%, causing patient B to decrease the service time by just 1%, it is not considered fair. But if patient B could decrease the service time by 70%, it is fair. Proportional fairness could improve the QoS of all the patients over their local computing by themselves. Otherwise, the service offloading will intuitively break down, replaced by local computing. **P2** describes the optimization problem for the static Nash bargaining game, considering the unpredictable system dynamics, we further consider the long-term dynamic Nash bargaining game **P3**, which is also shown as follows:

$$\min \lim_{T\to\infty} \inf \frac{1}{T} \sum\nolimits_{t=0}^{T-1} \mathbb{E}\left\{-\sum_{l=1}^{L} \ln(\alpha T_l^{\text{local}}[t] - T_l^{\text{offload}}[t])\right\}$$

$$s.t. \quad T_l^{\text{offload}}[t] < \alpha T_l^{\text{local}}[t], (\forall l \in \mathcal{L})$$

$$\sum\nolimits_{l=1}^{L} \lambda_{lk}[t] \leq 1, \sum_{k=1}^{K} y_{lk}[t] = 1, (\forall k \in \mathcal{K}, \forall l \in \mathcal{L})$$

$$\lim_{T\to\infty} \sup \frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}\{Z_k[t]\} \leq t\omega_k[t], (\forall k \in \mathcal{K})$$

$$\lim_{T\to\infty} \sup \frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}\{M_k[t]\} \leq m_k[t], (\forall k \in \mathcal{K}).$$

## V. Proposed Fair Edge Healthcare Scheme

To guarantee long-term proportional fairness for 5G edge healthcare, we next find the optimal solutions of the **P2** and **P3**, i.e., our proposed FairHealth scheme. We first consider the static proportional fairness in **P2**, so as to in-depth reveal the essential difference of the cases in Section III. For long-term dynamic fairness in **P3**, we propose a Lyapunov-based online algorithm to derive the optimal patient association and resource schedule in each time slot $t$.

### A. Motivational Bases of FairHealth Scheme

We consider an MEC server $k$, and $L$ patients in 5G healthcare system. The optimization goals of global optimum and proportional fairness are expressed as follows:

$$\min_{\lambda_{lk}} T^{\text{total}} = \sum_{l=1}^{L} \left(T_l^{\text{com}} + \frac{c_l s_l/\omega_k}{\lambda_{lk}}\right), s.t. \sum_{l=1}^{l=1} \lambda_{lk} \leq 1$$

$$\min_{\lambda_{lk}} U^{\text{total}} = \sum_{l=1}^{L} -\ln\left(\alpha T_l^{\text{local}} - T_l^{\text{com}} - \frac{c_l s_l/\omega_k}{\lambda_{lk}}\right)$$

$$s.t. \quad T_l^{\text{offload}} < \alpha T_l^{\text{local}}, \sum_{l=1}^{l=1} \lambda_{lk} \leq 1.$$

*1) Global Optimum Case:* According to classical Cauchy's inequality, we could obtain

$$\sum_{l=1}^{L} \left(\frac{\sqrt{c_l s_l/\omega_k}}{\sqrt{\lambda_{lk}}}\right)^2 \sum_{l=1}^{L} \left(\sqrt{\lambda_{lk}}\right)^2 \geq \left(\sum_{l=1}^{L} \sqrt{c_l s_l/\omega_k}\right)^2. \tag{15}$$

For the global optimum strategies, we should obtain the resource allocation to minimize the total time, we could obtain

$$T^{\text{total}} \geq \sum_{l=1}^{L} T_l^{\text{com}} + \frac{\sum \sqrt{c_l s_l/\omega_k}}{\sum \lambda_{lk}} \geq \sum_{l=1}^{L} T_l^{\text{com}} + \sum_{l=1}^{L} \sqrt{c_l s_l/\omega_k}. \quad (16)$$

The inequality takes equal conditions as $\sqrt{c_1 s_1/\omega_k}/\lambda_{1k} = \ldots = \sqrt{c_l s_l/\omega_k}/\lambda_{lk} = \sqrt{c_L s_L/\omega_k}/\lambda_{lk}$, The global optimal resource allocation for the patients $l$ is $\omega_k \sqrt{c_l s_l}/\sum \sqrt{c_l s_l}$, which is proportional to the required computing amount $\sqrt{c_l s_l}$. Although the global optimum seems the most attractive, it is not suitable for priority-aware and deadline-sensitive 5G edge healthcare as we discussed in Section III.

*2) Proportional Fairness Case:* To simplify, let $A_l = \alpha T_l^{\text{local}} - T_l^{\text{com}}$, $B_l = c_l s_l/\omega_k$, $(\forall l \in \mathcal{L})$, and we investigate the concave of the optimization goal, i.e., the Hessian matrix of $U^{\text{total}}$, for $\forall i, j \in \mathcal{L}$ and $i \neq j$

$$\frac{\partial^2 U^{\text{total}}}{\partial \lambda_{ik} \partial \lambda_{jk}} = \frac{B_i^2/\lambda_{ik}^4}{(A_i - B_i/\lambda_{ik})^2} + \frac{2B_i/\lambda_{ik}^3}{(A_i - B_i/\lambda_{ik})} > 0. \quad (17)$$

And $\partial^2 U^{\text{total}}/\partial \lambda_{ik} \partial \lambda_{jk} = 0$ for $i = j$. It is concluded that the Hessian matrix is a symmetrical positive definite matrix. According to the theorems in [26], it is concluded $U^{\text{total}}$ is convex. Thus, the uniqueness of the NBS will exist in our proposed game. For the restricted convex optimization problem for fairness case, we employ the Lagrangian dual function to obtain the unique NBS, i.e.,

$$\begin{aligned} \mathbb{L}(\lambda_{lk}, \mu_l, \eta) = & -\sum_{l=1}^{L} \ln(A_l - B_l/\lambda_{lk}) \\ & + \eta \left( \sum_{l=1}^{L} \lambda_{ik} - 1 \right) \\ & + \sum_{l=1}^{L} \mu_l (B_l/\lambda_{lk} - A_l) \end{aligned} \quad (18)$$

where $\mu_l$ and $\eta$ are dual variables. We employ Karush–Kuhn–Tucher (KKT) condition to derive the optimal solution, i.e.,

$$\begin{cases} \nabla \mathbb{L}(\lambda_{lk}, \mu_l, \eta) = 0 \\ \mu_l (B_l/\lambda_{lk} - A_l) = 0 \\ \eta((\sum_{l=1}^{L} \lambda_{lk}) - 1) = 0 \\ \mu_l \geq 0, \eta \geq 0 \end{cases} \Rightarrow \begin{cases} \lambda_{lk} = \frac{2\eta^{-1}}{\sqrt{1 + \frac{4A_l}{\eta B_l}} - 1}. \\ \sum_{l=1}^{L} \lambda_{lk} = 1 \end{cases} \quad (19)$$

According to (18), the proportional fair resource allocation is related to $A_l B_l^{-1}$, which depends not only on the computing amount $c_l s_l$ but also on service priority $\alpha$ and local deadline $T_l^{\text{local}}$. When the service priority is higher (or lower deadline), and when the computing amount is larger, the patient will be allocated more service resources. Compared with global optimization, this proportionally fair scheme is much more practical for 5G edge healthcare systems.

## B. Long-Term Dynamic FairHealth Scheme

We further consider the long-term dynamic proportional fairness via the Lyapunov optimization theory. The collaboration of multiple MECs needs to be studied, thus introducing the edge association problem. And the tradeoff between dynamic network stability and fairness also needs to be explored. For each MEC server $k$, there are $K$ service queues and $K$ virtual queues. We describe $\Theta[t] = (D_1[t], \ldots, D_k[t], \ldots, D_K[t], X_1[t], \ldots, X_k[t] \ldots, X_K[t])$ as the queue vector in time slot $t$. And the Lyapunov function $I(\Theta[t])$ and Lyapunov drift $\Delta(\Theta[t])$ is shown as follows:

$$I[t] = \sum_{l=1}^{L} \frac{D_l^2[t] + Z_l^2[t]}{2}, \Delta(\Theta[t]) = I[t+1] - I[t]. \quad (20)$$

In **P3**, we aim to reduce Lyapunov drift to ensure network stability. Meanwhile, we still need to minimize the proportional fair optimization goal. We resort to the drift-plus-penalty (DPP) expression to integrate the network stability and fairness, i.e., $\Delta(\Theta[t]) + VU^{\text{total}}[t]$, where $V > 0$ is the adjustable parameter, to weight service stability and fairness. Then, **P3** is transformed to **P4**, i.e.,

$$\begin{aligned} & \min \mathbb{E}\{\Delta[t] - V \sum_{l=1}^{L} \ln(\alpha T_l^{\text{local}}[t] - T_l^{\text{offload}}[t]) | \Theta[t]\} \\ & s.t. \quad T_l^{\text{offload}}[t] < \alpha T_l^{\text{local}}[t], (\forall l \in \mathcal{L}) \\ & \sum_{l=1}^{L} \lambda_{lk}[t] \leq 1, \sum_{k=1}^{K} y_{lk}[t] = 1, (\forall k \in \mathcal{K} \forall l \in \mathcal{L}). \end{aligned}$$

We cannot directly solve the above **P4**, for we need to obtain $Y(\Theta[t+1])$ in future time slot $(t+1)$. According to the Lyapunov optimization techniques, we could solve **P4** by online minimizing the upper bound of the DPP, without knowing future system information. We then derive the upper bound of DPP expression as follows.

*Theorem 1 :* For all the queues $\Theta[t]$ and $V$, the expected DPP expression is upperbounded as

$$\begin{aligned} & \mathbb{E}\left\{\Delta[t] - V \sum\nolimits_{l=1}^{L} \ln(\alpha T_l^{\text{local}}[t] - T_l^{\text{offload}}[t]) | \Theta[t]\right\} \\ & \leq \mathbb{E}\left\{\sum\nolimits_{l=1}^{L} \sum\nolimits_{k=1}^{K} y_{lk}[t](D_k[t]c_l[t]s_l[t] + X_k[t]) | \Theta[t]\right\} \\ & \quad + G - \mathbb{E}\left\{\sum\nolimits_{k=1}^{K} tD_k[t]\omega_k[t] + \sum\nolimits_{k=1}^{K} m_k X_k[t] | \Theta[t]\right\} \\ & \quad - V\left\{\sum\nolimits_{l=1}^{L} \ln(\alpha T_l^{\text{local}}[t] - T_l^{\text{offload}}[t]\right\} \end{aligned} \quad (21)$$

$$G = \max\left\{\frac{1}{2}\sum_{k=1}^{K}(t^2\omega_k^2[t] + Z_k^2[t] + m_k^2[t] + M_k^2[t])\right\}. \quad (22)$$

*Proof:* Due to the limited space, the proof is omitted.

Based on Theorem 1, the upper bound of DPP expression is only related to the current system status. The highly coupled system control strategy includes $y_{lk}[t]$ and $\lambda_{lk}[t]$. Only when $y_{lk}[t]$ is determined, MEC servers could allocate the resources

**Algorithm 1:** Lyapunov-Based Long-Term Proportional Fairness Algorithm.

1: **Input:** $V$, $\alpha$, $K$, $L$, $P$, $\theta$;
2: **Output:** $y^*_{lk}[t]$ and $\lambda^*_{lk}[t]$;
3: **Initialization:** Initial queue vector $\Theta[0]=0$;
4: **for** $t \in \Gamma$ **do**
5: **for** $k \in \mathcal{K}$ **do**
6: Getting the current status: $\Theta[t]$, $s_l[t]$, $c_l[t]$, $\omega_k[t]$;
7: Setting $W[t] = -V\sum \ln(\alpha T_l^{\text{local}}[t] - T_l^{\text{offload}}[t])$;
8: Setting $Q[t] = \sum\sum y_{lk}[t](D_k[t]c_l[t]s_l[t]+X_k[t])$;
9: $(y^*_{lk}[t], \lambda^*_{lk}[t]) \leftarrow \arg\min(W[t]+Q[t])$;
10: **end for**
11: Updating queue status:
12: $D_k[t+1] = \max\{D_k[t] - t\omega_k[t], 0\} + \sum y_{lk}[t]c_l[t]s_l[t]$;
13: $X_k[t+1] = \max\{X_k[t] - m_k, 0\} + \sum y_{lk}[t]$;
14: **end for**

**Algorithm 2:** Block Coordinate Descent-Based Proportional Fairness Algorithm.

1: **Input:** $\Theta[t]$, $s_l[t]$, $c_l[t]$, $\omega_k[t]$;
2: **Output:** Block variables $\mathbf{Y}^*[t]$ and $\mathbf{\Phi}^*[t]$, $y^*_{lk}[t]$, $\lambda^*_{lk}[t]$;
3: **Initialization:** Iteration threshold $\kappa$, initial value $\mathbf{Y}^{(0)}[t]$ and $\mathbf{\Phi}^{(0)}[t]$, $\tau=0$;
4: Setting $R[t] = W[t] + Q[t]$;
5: **while** $|R^{(\tau+1)}[t] - R^{(\tau)}[t]| > \kappa$ **do**
6: $\mathbf{\Phi}^{(\tau+1)}[t] \leftarrow \arg\min R[t]$ with fixed $\mathbf{Y}^{(\tau)}[t]$;
7: $\mathbf{Y}^{(\tau+1)}[t] \leftarrow \arg\min R[t]$ with fixed $\mathbf{\Phi}^{(\tau+1)}[t]$;
8: **end while**
9: **for** all the patients $l \in \mathcal{L}$ **do**
10: $k^* = \arg\max_k y_{lk}[t]$;
11: $y^*_{lk^*}[t] \leftarrow 1$; $y^*_{lk}[t] \leftarrow 0$, $(k \neq k^*)$;
12: **end for**
13: $\lambda^*_{lk}[t] \leftarrow \arg\min R[t]$ with fixed $y^*_{lk}[t]$;

with a suitable $\lambda_{lk}[t]$. As presented in Theorem 1, only the first and fourth item on the right of (21) is controllable, other items are the state system parameters. That is, minimizing the DPP expression is equal to minimizing the first and fourth items on the right of (21). The first item can control the congestion level of the service system, i.e., the service stability, while the fourth item can achieve Pareto-optimal proportional fairness. Therefore, we propose the Lyapunov-based long-term proportional fairness algorithm, which is presented in Algorithm 1.

According to the Lyapunov optimization theory [4], [24], we can achieve a $O(V, 1/V)$ tradeoff between queue stability and proportional fairness. In specific, the long-term service stability is determined by $O(V)$, while the long-term proportional fairness is measured by $O(1/V)$. A smaller $V$ means we just need fewer time iterations to realize the queue stability. When $V$ is large enough, the solution of Algorithm 1 will be infinitely approaching the optimization goal in **P3**. In practice, we could employ suitable $V$ to configure the 5G edge healthcare

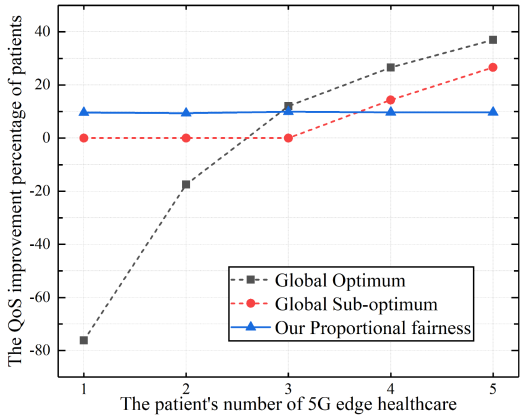



Fig. 3. QoS improvement percentage of patients.

system [4]. In Algorithm 1, lines (5–10) represent the online optimization problem solved in each time slot. The online fairness subproblem in line 9 is a mixed-integer nonlinear programming (MINLP) problem, i.e., a nonconvex NP-hard problem, which is hard to be solved in a polynomial time. Thus the global optimal solution in line 9 is hard to obtain. We first relax the integer variables $y_{lk}[t]$ into continuous variables [5], i.e., $y_{lk}[t] \in [0,1]$. Then, we could have two blocks of variables $\mathbf{Y}[t]$ and $\mathbf{\Phi}[t]$ as follows:

$$\mathbf{Y}[t] = \begin{vmatrix} y_{11}[t] & \cdots & y_{12}[t] & \cdots & y_{1K}[t] \\ \vdots & & \vdots & \vdots & \vdots \\ y_{l1}[t] & \cdots & y_{l2}[t] & \cdots & y_{lK}[t] \\ \vdots & & \vdots & \vdots & \vdots \\ y_{L1}[t] & \cdots & y_{L2}[t] & \cdots & y_{LK}[t] \end{vmatrix} \tag{23}$$

$$\mathbf{\Phi}[t] = \begin{vmatrix} \lambda_{11}[t] & \cdots & \lambda_{12}[t] & \cdots & \lambda_{1K}[t] \\ \vdots & & \vdots & \vdots & \vdots \\ \lambda_{l1}[t] & \cdots & \lambda_{l2}[t] & \cdots & \lambda_{lK}[t] \\ \vdots & & \vdots & \vdots & \vdots \\ \lambda_{L1}[t] & \cdots & \lambda_{L2}[t] & \cdots & \lambda_{LK}[t] \end{vmatrix}. \tag{24}$$

We could easily find the online fairness subproblem is convex with respect to $\mathbf{Y}[t]$ and $\mathbf{\Phi}[t]$. Thus, we employ a low complexity BCD and convex optimization method [27], [28] to iteratively solve an approximate solution in line 9. The detailed process of line 9 is presented in detail in Algorithm 2.

## VI. Simulation and Discussion

We simulate the real network topology with Shanghai Telecom's dataset [4], [29], which contains 3233 geographically distributed base stations (BSs). It is an open-source dataset [29]. The dataset contains the detailed base station latitude and longitude coordinates, and start time and end time of base station access for service requesters. We select five real geographic coordinates of the BSs located in a 300 × 300-m square area, and the number of IoMT devices is set as 20. The specific coordinates of the five BSs are presented in Table III.

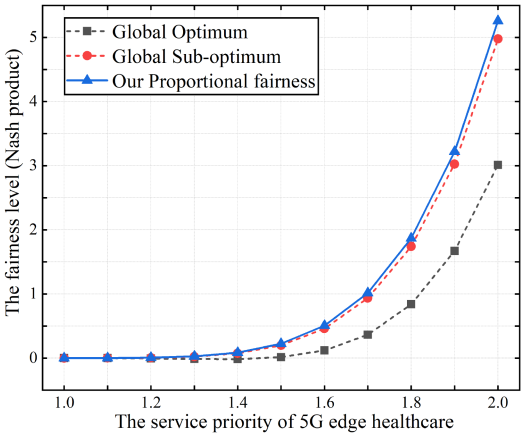



Fig. 4. Fairness level of 5G edge healthcare.

TABLE III
GEOGRAPHIC COORDINATES OF THE 5 BSS

| Number | Latitude | Longitude |
|---|---|---|
| BS 1 | 121.443586 | 31.203031 |
| BS 2 | 121.444401 | 31.200811 |
| BS 3 | 121.445984 | 31.203473 |
| BS 4 | 121.443811 | 31.200249 |
| BS 5 | 121.446511 | 31.200882 |

Each BS is equipped with MEC servers. The IoMT devices follow a homogeneous Poisson point process (PPP) distribution in the given square area. In addition, data amount $s_l[t]$ is uniformly chosen from [5, 25] Mbit and $c_l[t]$ follows uniform distribution from [500, 900] cycles/bit. The service priority $\alpha$ follows a uniform distribution on an interval [1.1, 2.0]. The wireless transmission bandwidth $B[t]$ is 20 MHz and the transmit power is 500 mW. And the Gaussian noise power $\sigma^2[t]$ is $2 \times 10^{-13}$ W. The CPU computation resource of the MEC servers in BSs and IoMT devices follows uniform distribution from [30, 35] GHz, [3, 5] GHz, respectively.

The comparison offloading schemes in our article are the global optimum offloading scheme (GOP) and global suboptimum offloading scheme (GSOP), which are commonly considered in previous works [5], [6]. The GOP and GSOP offloading schemes both optimize the service delay of the entire system from a global perspective, ignoring the requirement of fair resource allocation among users.

1) *Global optimum scheme:* GOP offloading scheme would take the total service offloading delay of all users as the optimization goal. Thus the total offloading delay will realize the theoretical minimum.
2) *Global suboptimum scheme:* GSOP offloading scheme would introduce a restriction with the total service delay minimization compared with GOP. That is, the service time of each user should be less than the benchmark time, i.e., local service time.

Fig. 3 investigates the QoS improvement in the three strategies. The baseline is the local strategy, where patients just perform data processing locally. Specifically, we analyze five patients associated with the same BS, and their increasing order is arranged by their service data amount. As shown in Fig. 3, GOP will cause the QoS of patients 1 and 2 to drop 76.18% and 17.52%, which is unacceptable in practice. Compared to GOP, GSOP could ensure that the QoS of each patient will not drop, while it is still not fair to patients 1 and 2. In GSOP, patients 1, 2, and 3 have almost no improvement in QoS, while patients 4 and 5 have increased by 14.34% and 26.63%, respectively. In our FairHealth, the QoS of each patient has been improved to a certain extent, and this improvement is fair and unbiased.

In Fig. 4, we also study the fairness level of 5G edge healthcare, i.e., Nash product. As shown in Fig. 4, our FairHealth can achieve the highest fairness level, GOS has the worst fairness, while GSOP achieves a tradeoff between fairness and total service delay. We could also quantitatively compare the three strategies: when $\alpha = 2.0$, FairHealth has increased by 74.44% and 5.66%, compared to existing GOS and GSOP, respectively. At the same time, the fairness level will increase as the service priority $\alpha$ increases, which is in accordance with (12).

The Pareto-optimal computation allocation is more significant for proportional fairness in FairHealth, compared with edge association strategies. Thus we focus on the computation allocation in Fig. 5. As shown in Fig. 5(a), with the improvement of service priority parameters $\alpha$, our FairHealth will gradually tend to provide more CPU computing resources to high-priority healthcare services. The two strategies, GOP and GSOP, will not take into account the different service priorities, while allocating resources evenly among services. Therefore, GOP and GSOP are hard to apply to priority-sensitive edge healthcare. In Fig. 5(b), the data amount of healthcare services increases, GOP and GSOP will supply more computation for services with more data. This is an optimization measure taken by GOP and GSOP to reduce the total service time, but this will lead to unreasonable resource allocation for patients. Imagine a sudden emergency treatment, whose pending data amount is usually lower than ordinary medical log analysis. Should we allocate too much computation for log analysis so that cause emergency treatment not be processed in time? Our FairHealth is sensitive to priorities rather than data amount, compared to GOP and GSOP.

The local task processing capabilities of IoMT devices also need to be analyzed, which is ignored in GOP and GSOP, as shown in Fig. 5(c). Note that, when the local computation is higher, patients can get better QoS by themselves. Thus, the SP needs to attract and motivate patients to participate by more resource provision. Because most healthcare services are delay-sensitive, patients concern more about QoS. If they cannot improve QoS by participating in 5G edge healthcare, they will only process data locally. As shown in Fig. 5(c), our FairHealth can offer more computation when the local computation is high, which will encourage patient engagement. In summary, our FairHealth is better than GOP and GSOP when considering the priority-sensitive and latency-critical 5G edge healthcare services.

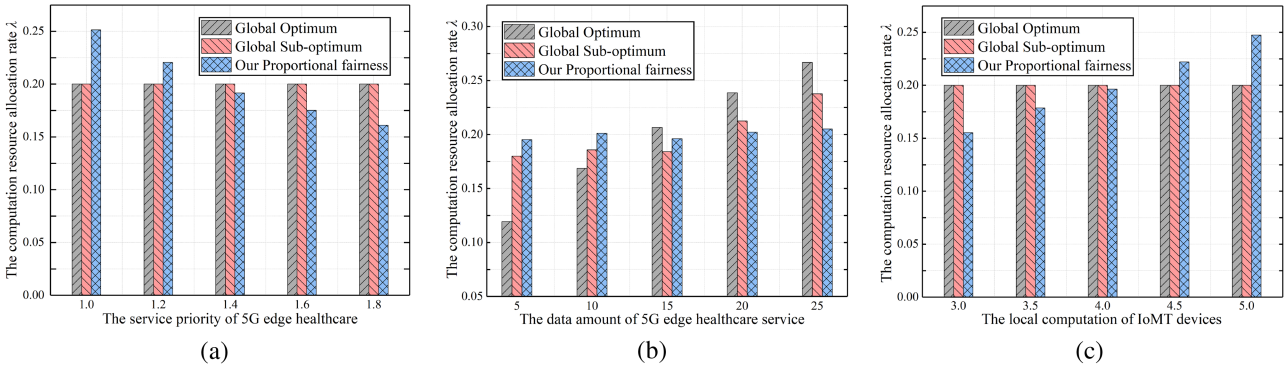



Fig. 5. Simulations of fair computation allocation. (a) Computation allocation ratio versus service popularity. (b) Computation allocation ratio versus data amount. (c) Computation allocation ratio versus local computation.

## VII. Conclusion

In this article, we proposed a long-term proportional fairness-driven 5G edge healthcare, named FairHealth. FairHealth provides long-term Pareto efficiency and proportional fairness in 5G edge healthcare, guaranteeing that all patients are incentivized to participate. Then, considering the priority-sensitive and delay-critical healthcare, we formulated a Nash bargaining game problem to jointly optimize edge associations and computation resources in FairHealth. The long-term dynamic game problem is a nonconvex stochastic network optimization problem. We employed the Lyapunov techniques to transform the original problem into a series of online fair subproblems. Moreover, we utilized BCD and convex optimization method to iteratively solve the online subproblems. Simulation results demonstrated that our FairHealth significantly outperforms the existing schemes. Our future work will focus on the intelligent edge healthcare framework. Through advanced deep reinforcement learning, we will design an adaptive multidimensional edge resource (i.e., communication, computing, and energy) scheduling scheme, which could achieve fairness-guaranteed intelligent edge healthcare.

## References


[1] Y. A. Qadri, A. Nauman, Y. B. Zikria, A. V. Vasilakos, and S. W. Kim, "The future of healthcare Internet of Things: A survey of emerging technologies," *IEEE Commun. Surv. Tut.*, vol. 22, no. 2, pp. 1121–1167, Apr./Jun. 2020.

[2] Y. Siriwardhana, P. Porambage, M. Liyanage, and M. Ylianttila, "A survey on mobile augmented reality with 5G mobile edge computing: Architectures, applications, and technical aspects," *IEEE Commun. Surv. Tut.*, vol. 23, no. 2, pp. 1160–1192, Apr./Jun. 2021.

[3] X. Huang, S. Leng, S. Maharjan, and Y. Zhang, "Multi-agent deep reinforcement learning for computation offloading and interference coordination in small cell networks," *IEEE Trans. Veh. Technol.*, vol. 70, no. 9, pp. 9282–9293, Sep. 2021.

[4] X. Lin, J. Wu, J. Li, W. Yang, and M. Guizani, "Stochastic digital-twin service demand with edge response: An incentive-based congestion control approach," *IEEE Trans. Mobile Comput.*, to be published, doi: 10.1109/TMC.2021.3122013.

[5] M. Chen and Y. Hao, "Task offloading for mobile edge computing in software defined ultra-dense network," *IEEE J. Sel. Areas Commun.*, vol. 36, no. 3, pp. 587–597, Mar. 2018.

[6] H. Guo, J. Liu, J. Zhang, W. Sun, and N. Kato, "Mobile-edge computation offloading for ultradense IoT networks," *IEEE Internet Things J.*, vol. 5, no. 6, pp. 4977–4988, Dec. 2018.

[7] L. Wang, G. Tyson, J. Kangasharju, and J. Crowcroft, "Milking the cache cow with fairness in mind," *IEEE/ACM Trans. Netw.*, vol. 25, no. 5, pp. 2686–2700, Oct. 2017.

[8] J. Zhou and X. Zhang, "Fairness-aware task offloading and resource allocation in cooperative mobile edge computing," *IEEE Internet Things J.*, vol. 9, no. 5, pp. 3812–3824, Mar. 2022.

[9] Y. Dong, S. Guo, J. Liu, and Y. Yang, "Energy-efficient fair cooperation fog computing in mobile edge networks for smart city," *IEEE Internet Things J.*, vol. 6, no. 5, pp. 7543–7554, Oct. 2019.

[10] X. Huang, S. Zeng, D. Li, P. Zhang, S. Yan, and X. Wang, "Fair computation efficiency scheduling in NOMA-Aided mobile edge computing," *IEEE Wireless Commun. Lett.*, vol. 9, no. 11, pp. 1812–1816, Nov. 2020.

[11] N. Y. Philip, J. J. P. C. Rodrigues, H. Wang, S. J. Fong, and J. Chen, "Internet of things for in-home health monitoring systems: Current advances, challenges and future directions," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 2, pp. 300–310, Feb. 2021.

[12] M. Usman, M. A. Jan, X. He, and J. Chen, "P2DCA: A. privacy-preserving-based data collection and analysis framework for IoMT applications," *IEEE J. Sel. Areas Commun.*, vol. 37, no. 6, pp. 1222–1230, Jun. 2019.

[13] J. Peng, K. Cai, and X. Jin, "High concurrency massive data collection algorithm for IoMT applications," *Comput. Commun.*, vol. 157, pp. 402–409, May 2020.

[14] A. Ghubaish, T. Salman, M. Zolanvari, D. Unal, A. Al-Ali, and R. Jain, "Recent advances in the Internet-of-Medical-Things (IoMT) systems security," *IEEE Internet Things J.*, vol. 8, no. 11, pp. 8707–8718, Jun. 2021.

[15] B. D. Deebak and F. Al-Turjman, "Smart mutual authentication protocol for cloud based medical healthcare systems using Internet of Medical Things," *IEEE J. Sel. Areas Commun.* vol. 39, no. 2, pp. 346–360, Feb. 2021.

[16] A. A. Abdellatif *et al.*, "ssHealth: Toward secure, blockchain-enabled healthcare systems," *IEEE Netw.*, vol. 34, no. 4, pp. 312–319, July/Aug. 2020.

[17] Z. Ning *et al.*, "Mobile edge computing enabled 5G health monitoring for Internet of Medical Things: A decentralized game theoretic approach," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 2, pp. 463–478, Feb. 2021.

[18] D. Lin and Y. Tang, "Edge computing-based mobile health system: Network architecture and resource allocation," *IEEE Syst. J.*, vol. 14, no. 2, pp. 1716–1727, Jun. 2020.

[19] Z. Zhou *et al.*, "Learning-based URLLC-aware task offloading for Internet of health things," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 2, pp. 396–410, Feb. 2021.

[20] V. Hayyolalam, M. Aloqaily, Ö. Özkasap, and M. Guizani, "Edge intelligence for empowering IoT-Based healthcare systems," *IEEE Wireless Commun.*, vol. 28, no. 3, pp. 6–14, Jun. 2021.

[21] M. A. Rahman, M. S. Hossain, N. A. Alrajeh, and N. Guizani, "B5G and explainable deep learning assisted healthcare vertical at the edge: COVID-I9 perspective," *IEEE Netw.*, vol. 34, no. 4, pp. 98–105, Jul./Aug. 2020.

[22] Y. Sun, S. Zhou, and J. Xu, "EMM: Energy-aware mobility management for mobile edge computing in ultra dense networks," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 11, pp. 2637–2646, Nov. 2017.

[23] C. Niu, Y. Li, R. Q. Hu, and F. Ye, "Fast and efficient radio resource allocation in dynamic ultra-dense heterogeneous networks," *IEEE Access*, vol. 5, pp. 1911–1924, 2017.

[24] M. J. Neely, "Stochastic network optimization with application to communication and queueing systems," *Synth. Lectures Commun. Netw.*, vol. 3, no. 1, 2010, Art. no. 1211.
[25] D. Xu *et al.*, "Peer-to-peer multienergy and communication resource trading for interconnected microgrids," *IEEE Trans. Ind. Informat.*, vol. 17, no. 4, pp. 2522–2533, Apr. 2021.
[26] S. Boyd and L. Vandenberghe, *Convex Optimization*. New York, NY, USA: Cambridge Univ. Press, 2004.
[27] Q. Liu, T. Han, and N. Ansari, "Joint radio and computation resource management for low latency mobile edge computing," in *Proc. IEEE Glob. Commun. Conf.*, 2018, pp. 1–7.
[28] T. Bai, C. Pan, Y. Deng, M. Elkashlan, A. Nallanathan, and L. Hanzo, "Latency minimization for intelligent reflecting surface aided mobile edge computing," *IEEE J. Sel. Areas Commun.*, vol. 38, no. 11, pp. 2666–2682, Nov. 2020.
[29] Shanghai Telecom, Shanghai, China., The distribution of 3233 base stations, 2019. [Online]. Available: http://www.sguangwang.com/dataset/telecom.zip