Cluster Computing (2024) 27:2901–2915
https://doi.org/10.1007/s10586-023-04111-w

Performance analysis of P2p network content delivery based on queueing model

Zhanyou Ma[1] · Miao Yan[1] · Rong Wang[1] · Shunzhi Wang[1]

Received: 12 December 2022 / Revised: 24 June 2023 / Accepted: 20 July 2023 / Published online: 25 August 2023

Abstract

In peer-to-peer(P2P) networks, content delivery is very critical, but there are not many relevant research results in content delivery technology. In this paper, in order to simulate content delivery in P2P mode, the number of online players is abstracted into the number of servers that can provide services, the peers write content to buffer is abstracted into the arrival of customers, and the players read content from buffer is abstracted into the service process of the servers. Based on the consideration of the effect of the number of P2P online players on energy consumption, an M/M/c queueing model with random variation in the number of servers is developed, and negative customers and preemptive priority policies are introduced. The matrix–geometric solution method and Gauss-Seidel iterative method are used to derive the performance measures of the system at steady state for two types of contents. And some numerical examples are given using Matlab for analyzing the trend of performance measures in P2P networks with parameters. The benefit function is established to obtain the parameter values that make the social benefit optimal by Nash equilibrium and social optimal strategy, and provide the theoretical basis for the scheduling of P2P peers.

Keywords P2P networks · Content delivery · Preemptive priority · M/M/c queue · Matrix–geometric solution method

1 Introduction

P2P networks break the traditional client/server model of the Internet, where each peer in P2P networks is equal and there is no central peer, it can serve as a requester of network services and respond to requests from other computers to provide resources and services. Therefore, the implementation of P2P applications has high scalability and low deployment costs, giving a huge space for Internet publishing and sharing (Loo et al. [1], Luo et al. [2], Chen et al. [3]).

In fact, P2P networks have already occupied a dominant position in the early Internet, and with the rapid increase of users, the resources and service capabilities of the system are expanding simultaneously, which makes P2P networks have an "inherent" scalability. Zhou et al. [4] proposed a hybrid P2P strategy of Rarest First and Greedy in order to balance start-up latency and continuity, and to ensure scalability, an algorithm was designed to dynamically adapt to the size of the peer population while retaining some peers to ensure real-time playback requirements and low start-up latency. Most P2P applications use multiple chunks when sharing files, Susitaival and Aalto [5] proposed a detailed Markov model based on how files are divided into chunks and how the chunk selection strategy used affects the performance of the P2P file sharing system, and compared different chunk selection strategies. The results show that splitting the file from one block to two blocks significantly improves the performance of the system. Guo et al. [6] proposed an adaptive queue-based chunk scheduling method, which can achieve high bandwidth utilization and optimal streaming rate in P2P

Rong Wang and Shunzhi Wang have contributed equally to this work.

✉ Zhanyou Ma
mzhy55@ysu.edu.cn

Miao Yan
1277148778@qq.com

Rong Wang
wangrong9327@163.com

Shunzhi Wang
wdsysu@163.com

[1] School of Science, Yanshan University, Qinhuangdao 066004, China

streaming media, and evaluated the scheme in a real network, the results show that the adaptive queue-based chunk scheduling method can achieve near-optimal streaming rate, and can well adapt to the changes of peers and the dynamic changes of the underlying network.

Load optimization plays a vital role in cloud computing, which represents the system performance, and the best optimization technology goal is to effectively meet the needs of users with the least resources and processing time. Priya and Gnanasekaran [7] provided a data center resource allocation method based on load prediction in order to optimize different hybrid P2P cloud data center regions and different users in the cloud environment, it enhances load optimization by maintaining the reliability and stability between the user group and the data center during data transmission. Compared with the traditional methods, the proposed algorithm also reduces the resource utilization and response time. Du et al. [8] proposed an interference-constrained routing scheme on a P2P-shared multi-hop device-to-device(D2D) network, which accelerate the average download rate of sub users within the limits of interference in the mobile network. To provide better performance, a routing scheme with coverage-based P2P-share(R-CPS) scheme was also developed, which can effectively exploit the broadcast characteristics of the wireless channel to ensure that each route can cover other users. A routing scheme with a location-aware P2P sharing mechanism was also proposed, whose core idea was to find a route that detours through or directly passes through other users, and simulation evaluations showed that both schemes have the ability to effectively improve data download rates.

Nowadays, P2P networking technologies are gaining more and more attention, such as in file sharing and streaming live media systems. P2P has become an almost indispensable technology as user requirements for file download rates and streaming quality continue to increase. Sun et al. [9] proposed a P2P transaction mechanism based on carbon emission flow and utilized the alternating direction method of multipliers (ADMM) algorithm to protect the privacy of users in the P2P transaction process. It is also becoming popular to use P2P technology to provide security, high availability, and persistent data access services between servers for load balancing. Among the various P2P file sharing protocols, BitTorrent is the most common and popular one, which focuses on illegal sharing of some copyrighted files. Nowadays, many anti-P2P companies have started to launch attacks on BitTorrent networks. Fattaholmanan and Rabiee [10] analyzed the fragmentation attacks on BitTorrent and set up the validity of the attacks, which were found to be effective in combating illegal BitTorrent sharing through experiments. Sankar et al. [11] verified the blocks in the network and achieved data security by using private blockchain in software-defined networks(SDN) and peer-to-peer communication in public blockchain. To increase confidentiality, the sender signs specific operations while transferring data from one user to another. The results showed that security was enhanced, throughput and response time were improved, end-to-end latency and expenses were reduced during data transfer. Bradai et al. [12] proposed a layered P2P streaming mechanism which relies on an algorithm that enables each peer to select the right streaming layer and to find the right peer to serve them. The mechanism also efficiently utilized network resources and provided high system throughput, and the results showed that the mechanism improved the video quality while reducing the number of layer changes and useless blocks.

With the continuous development of queueing theory, classical queueing systems have been extensively researched and their results have received wide attention and have been applied to many fields. Xu et al. [13] analyzed the multi-server working vacation queueing system with negative customers, its quasi-birth and death (QBD) process and generating matrix were derived from the multiple vacation model, and then the steady-state distribution expression of the system queue length and certain system indicators were given by using the matrix–geometric solution method. Ma et al. [14] studied the discrete working vacation queue with non-preemptive priority and variable service rate, gave the stationary distribution of the system and analyzed the equilibrium behavior of two types of customers. Wang et al. [15] established a synchronous multiple vacation queueing model based on the sleep mechanism of the standby module and studied a MEC task offloading strategy incorporating virtual machine clustering and sleep mechanisms. By setting up the main module which was always active and the standby module which can sleep in time in the server, the energy saving rate of the system was derived by using quasi-birth and death process and matrix–geometry solution method, and the effectiveness of the task offloading strategy was verified. Chen and Chen [16] established a queuing model for stopping and dropping out in the queuing process based on the large number of people and service interruptions that exist during manual customer service consultation, derived the steady-state distribution through matrix analysis, and gave control suggestions for system optimization through numerical experiments. GnanaSekar and Kandaiyan [17] studied a single-server retry queuing system with delayed repair and feedback in a work vacation queue, derived the steady-state probability function of the system using the complementary variables method, and analyzed the impact of some key metrics of the system on the performance metrics. Singla and Kaur [18] studied a two-stage queueing model with impatient customers and feedback, where customers arrive and depart with a certain probability, and analyzed the parameter changes through numerical results. Ye and Chen [19] studied a M/M/1 retrial queuing system with working breakdowns and compared the spectral expansion method with the matrix geometric solution method.

In real life, queue with preemptive priority strategies is a broader and more complex class of queueing systems, which is an extension and expansion of classical queueing systems. In addition, many scholars at home and abroad have applied feedback strategies, thresholds and working vacations to multi-server systems. Pandey and Pal [20] studied the discrete-time queueing system with non-preemptive priority, and obtained the steady-state distribution and the average waiting queue length of the system. Aibatov [21] introduced unreliable servers and preemptive priority in the queueing system and obtained the conditions that make the system reach steady state. Ma et al. [22] studied the M/M/c queueing model with preemptive priority and synchronous multiple working vacations, gave the steady-state indexes of two types of customers and the disappearance rate of high-priority customers, and finally described the impact of parameter changes on the system performance measure. Valentina [23] studied a single server queueing model with a finite buffer that described the operation of the system using a multi-dimensional Markov chain, calculated the smooth distribution of the system and several performance characteristics. Bian et al. [24] applied the Laplace transform and inverse transform to find the density function and distribution function of the system in a queueing system with priority queueing but non-preemptive service after M/M/1 feedback, and finally verified the obtained results by using the properties of the Laplace transform. Zhao et al. [25] set up a certain number of channels based on the idle time in the system, introduced variable transmission rates in discrete queue, and investigated the effect of buffer capacity on each index.

In daily life, the feature of variable number of servers also has many applications. For example, Zhao et al. [26] proposed a cloud service system with variable number of servers based on the feature that virtual machines in the cloud service system can be migrated, gave the approximate lower bound of virtual machines needed to be turned on in the steady state, and analyzed the impact of parameter changes on the number of virtual machines in the on state. Based on the fact that the number of customers may exceed the set queue threshold during busy periods, Zang and Li [27] introduced concepts such as customer patience and found that the introduction of queue thresholds effectively improved the system efficiency when compared with the traditional queuing model (Table 1).

Currently, almost all of the global providers of broadband content, that is, high-traffic content, are resorting to P2P technology to solve the problem of resource delivery. With the popularity of these bandwidth-intensive P2P applications, P2P traffic continues to grow rapidly, consuming a large amount of network bandwidth resources and even causing network traffic congestion, which has a negative effect on the normal operation of the Internet. It is particularly important to model the P2P content delivery process to improve the service performance, reduce energy consumption and analyze how to optimize the social benefit. The above literatures have studied P2P networks in terms of scalability, system performance, bandwidth allocation, and load optimization, but almost no literature has modeled and studied the feature that P2P peers change randomly. In this paper, based on the P2P network content delivery technology, the negative customer, preemptive priority and the number of servers randomly varying strategies are introduced into the classical M/M/c queueing to model and study content delivery in P2P networks. There are three main contributions of this paper.

- In order to ensure the stability of the P2P content delivery process and optimize the social benefits of the system, the behavior of peers is analyzed and a preemptive priority M/M/c queuing model with negative customer and randomly varying number of servers is established. The main purpose of this paper is to investigate how the arrival rate, service rate and the number of online servers in P2P systems affect the energy consumption of the content delivery process and how to make the delivery process maximally socially beneficial.
- A three-dimensional Markov model is established, and performance measures such as the average queue length of the two types of customer, and the utilization rate of the server are obtained using the matrix–geometric solution method and the Gauss-Seidel iteration method.
- The individual and social benefit functions are defined, the effects of parameter changes on performance measures as well as Nash equilibrium and socially optimal strategies are derived through numerical analysis.

The rest of this paper is organized as follows. In Sect. 2, real-time streaming delivery techniques are combined with queuing models to model the P2P networks. In Sect. 3, the steady-state system is analyzed and some performance measures at steady state are obtained. In Sect. 4, the impact of some parameters varying on system performance measures is analyzed and the optimal arrival rate in Nash equilibrium is analyzed by constructing individual and social benefit for two types of customer. Section 5 are the conclusions.

Table 1 Comparison of researches work on scheduling strategy

Authors	Approach	Priority	Load balancing	Reliability	Resource minimization	Refs
Loo et al. (2004)	An alternate approach based on Distributed Hash Tables (DHTs) engine	No	Yes	Yes	No	[1]
Priya et al. (2020)	Load optimization and resource Minimization (ELORM) algorithm	No	Yes	Yes	Yes	[7]
Du et al. (2017)	The interference-constrained routing Schemes over peer-to-peer (P2P) Share enabled multi-hop D2D networks	No	Yes	Yes	No	[8]
Sankar et al. (2021)	Implements data security by Employing private blockchain in SDN and public blockchain For peer to peer communication	No	Yes	Yes	No	[10]
Ma et al. (2018)	Matrix–geometric solution method and Gauss-Seidel iterative method	Yes	Yes	No	Yes	[13]
Wang et al. (2021)	Quasi-birth-and-death process And matrix–geometric solution	No	Yes	No	Yes	[14]
Chen et al. (2022)	Matrix analytical method	No	No	Yes	Yes	[15]
GnanaSekar et al. (2022)	The approach of supplementary variables	No	Yes	No	No	[16]
Singla et al. (2021)	The difference-differential equations Laplace transform	Yes	Yes	No	No	[17]
Aibatov et al. (2016)	Laplace-stieltjes transform	Yes	No	No	No	[19]
Ma et al. (2018)	Quasi-birth-and-death Process and matrix–geometric Solution method	Yes	No	Yes	Yes	[20]
Valentina et al. (2022)	Multidimensional markov chain	No	Yes	No	No	[21]
Zang et al. (2021)	Simulate the overall performance Of the contact center queueing Model using ProModel software	No	Yes	No	Yes	[25]

2 P2P network model description

In P2P networks, the transmission of real-time content is generally one-way, in line with the principles of real-time streaming transmission systems. In real-time streaming applications, there are some storage locations called buffer, which are used to store peers. Peers keep writing to buffer through P2P networks, while players keep reading from buffer. The organization of buffer can be described by a circular queue, the length of the queue is the number of blocks it can hold. As shown in Fig. 1, there are two pointers in this queue, one is pointing to the storage location of the latest arriving content, called the latest content block pointer; the other is pointing to the storage location of the content currently read by the player, called the current playing block. The content between the current playing block along the counterclockwise direction to the latest content block is called the playable content. When a new content block arrives, the latest content block pointer

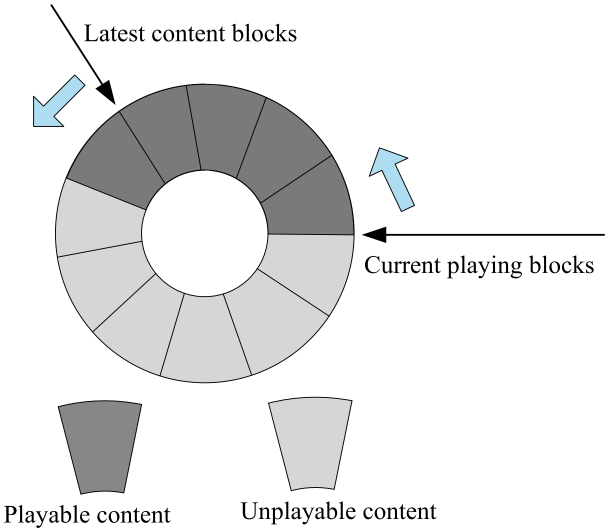


Fig. 1 Content delivery mechanism diagram

moves counterclockwise by one grid, and when the player finishes playing a content block, the current playing block pointer moves counterclockwise by one grid.

To analyze this delivery process, two relevant parameters are introduced: one is the code stream speed of the real-time content, which corresponds to the service rate of the server; the other is the speed at which the peer acquires the content, which corresponds to the speed at which the customer arrives. In P2P networks, the number of online peers is abstracted into the number of servers that can provide services, the peer writes to buffer is abstracted into the arrival of customers, and the player reads from buffer is abstracted into the service process. Thus build a preemptive queueing model with negative customer and random variation in the number of servers. Assuming that a negative customer is an insecure element in the network, its arrival causes corruption of the content being played at the top of the queue. In order to facilitate readers' better understanding, in the following section denoting the player in P2P network by the server, and Type I customer and Type II customer denoting Type I content and Type II content.

1. Based on the particularity that P2P peers can arrive and depart at any time, a time-continuous queueing model with the number of online servers in P2P networks varying randomly is established. When the number of online servers at one moment is $j(1 \leq j \leq c-1)$, the number of online servers at the next moment may be $j+1, j-1$ and j. Specially, when the number of online servers at one moment is 0, the number of online servers at the next moment may be 0 and 1; when the number of online servers at one moment is c, the number of online servers at the next moment may be $c-1$ and c. The online process of servers in P2P networks follows a Poisson process with the parameter λ_s, and the online time follows an exponential distribution with parameter μ_s. The time when the server turns from offline to online is random, and the online time is also random. Therefore, the number of online servers at any moment is a random variable.
2. Assuming that there are two types of customer in P2P networks, Type I customer is a common event in real-time streaming, Type II customer is a sudden emergency or major event in real-time streaming, so Type II customer has preemptive priority over Type I customer, which means that when Type II customer enters the buffer, if some online servers are found to be serving Type I customer and there is no idle online server, Type I customer in front of the line is interrupted at this time, and the interrupted Type I customer will be replayed only when there is an idle server in P2P networks again. Both types of customer are arranged in a single queue upon arrival, and the storage space for Type I customer is infinite, and the maximum storage space for Type II customer is c. When all online servers are serving Type II customer, the latest arrival of Type II customer no longer enters into the system, and the server goes offline while the corresponding Type II customer being served stops serving. The arrival intervals of Type I customer and Type II customer follow exponential distributions with parameters λ_1 and λ_2 respectively. The service rate of Type I customer and Type II customer follow exponential distributions with parameters μ_1 and μ_2, respectively.
3. The arrival interval of the negative customer follows an exponential distribution with parameter λ^-, the negative customer priority damage Type I customer being served at the head of the queue. When there is no customer to serve in P2P networks, the arriving negative customer disappear automatically and the negative customer only attack the customer and don't attack the server.
4. Both types of customer in the system follow the First-Come-First-Served (FCFS) rule. The arrival interval, the service time, and the process of serving the customer by any two online servers are independent of each other, and the maximum number of online servers is c. The P2P network operation diagram is shown in Fig. 2.

3 Model analysis

3.1 State transfer rate matrix

Let $L_1(t) = i, \quad i \geq 0$ denotes that there are i Type I customers in the P2P network at moment t, $L_2(t) = l, \quad 0 \leq l \leq c$ denotes that there are lType II customers in the P2P network at moment t, and $J(t) = j, \quad 0 \leq j \leq c$ denotes that there are j online servers in the P2P network at moment t, thus $\{(L_1(t), L_2(t), J(t)), \quad t \geq 0\}$ is a three-dimensional Markov process, and there is a state space

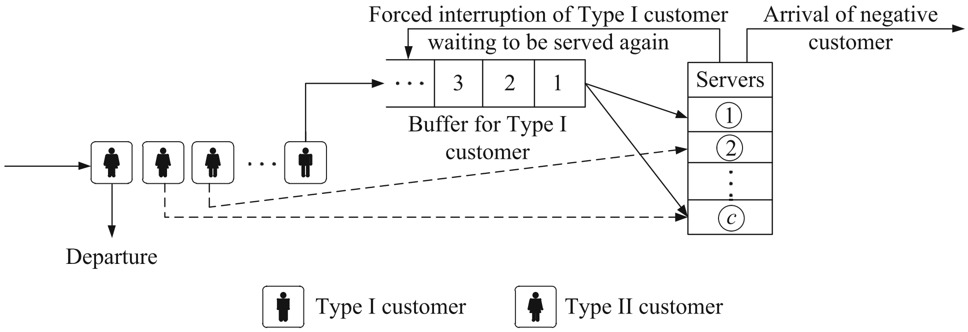
$$\Omega = \{(i,l,j), \quad i \geq 0, \quad 0 \leq l \leq c, \quad l \leq j \leq c\}. \tag{1}$$

the state set $(i,0,0), (i,0,1), \cdots, (i,0,c), (i,1,1), (i,1,2), \cdots, (i,1,c), (i,2,2), (i,2,3), \cdots, (i,2,c), \cdots, (i,c-1,c-1), (i,c-1,c), (i,c,c)$ is called level i, where $i \geq 0$.

Specifically, the state transfer diagram when $c = 3$ is shown in Fig. 3.

Arranging the states of the system in lexicographic order, the transfer rate matrix can be expressed in the following form

Fig. 2 P2P network operation mechanism



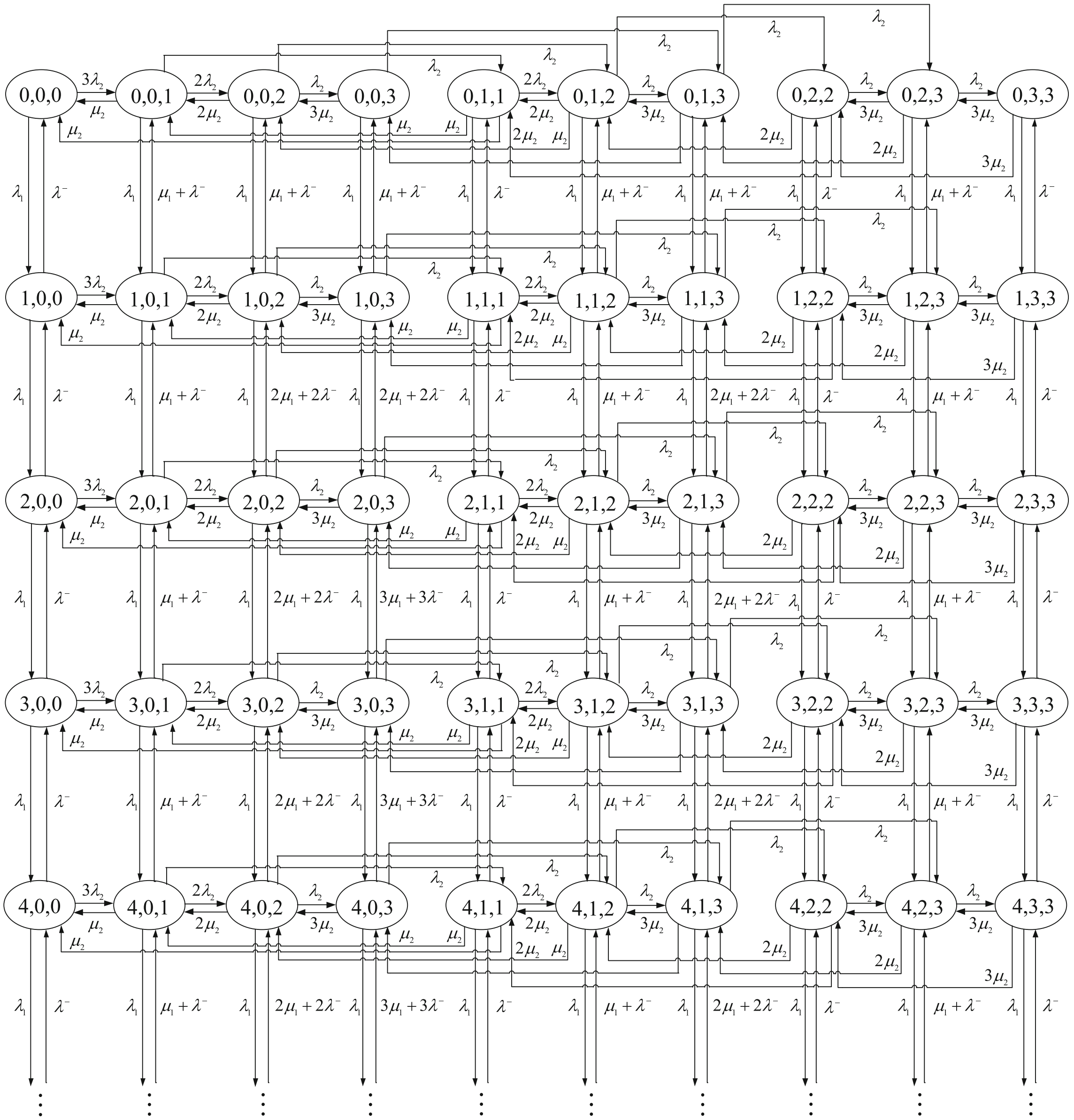


Fig. 3 State transfer diagram of queueing model

$$\boldsymbol{Q}=\begin{bmatrix} \boldsymbol{A}_0 & \boldsymbol{C} & & & & & \\ \boldsymbol{B}_1 & \boldsymbol{A}_1 & \boldsymbol{C} & & & & \\ & \boldsymbol{B}_2 & \boldsymbol{A}_2 & \boldsymbol{C} & & & \\ & & \ddots & \ddots & \ddots & & \\ & & & \boldsymbol{B}_{c-1} & \boldsymbol{A}_{c-1} & \boldsymbol{C} & \\ & & & & \boldsymbol{B} & \boldsymbol{A} & \boldsymbol{C} \\ & & & & & \ddots & \ddots & \ddots \end{bmatrix}. \tag{2}$$

where $\boldsymbol{A}_0,\boldsymbol{A}_i(1\le i\le c-1),\boldsymbol{A},\boldsymbol{B}_i(1\le i\le c-1),\boldsymbol{B},\boldsymbol{C}$ denote the inter-level transfer rate matrix respectively. And $\boldsymbol{A}_0,\boldsymbol{A}_i(1\le i\le c-1),\boldsymbol{A},\boldsymbol{B}_i(1\le i\le c-1),\boldsymbol{B},\boldsymbol{C}$ are all $(c+1)\times(c+2)/2$ dimensional block square matrix. To express the matrix easily, the following symbols are defined.

When $1\le l\le c$,

$$\delta_{0,l,j}=\begin{cases} -((j-1)\lambda_s+(c-j+1)\mu_s+(c-l)\mu_2+\lambda_1+\lambda_2), & 1\le j\le l, \\ -(l\lambda_s+(c-l)\mu_s+(c-l)\mu_2+\lambda_1), & j=l+1. \end{cases} \tag{3}$$

When $1\le l\le c,\quad 1\le j\le l$,

$$\delta_{i,l,j}=\begin{cases} -((j-1)\lambda_s+(c-j+1)\mu_s+(c-l)\mu_2+\lambda_1+\lambda_2+ \\ \quad \min(i,l-j+1)(\mu_1+\lambda^-)), \quad 1\le i\le c-1, \\ -((j-1)\lambda_s+(c-j+1)\mu_s+(c-l)\mu_2+\lambda_1+\lambda_2+ \\ \quad (l-j+1)(\mu_1+\lambda^-)), \quad i=c. \end{cases} \tag{4}$$

When $1\le i\le c,\quad 1\le l\le c$,

$$\delta_{i,l,l+1}=-(l\lambda_s+(c-l)\mu_s+(c-l)\mu_2+\lambda_1+\lambda^-). \tag{5}$$

When $1\le l\le c$,

$$\xi_{i,l,j}=\begin{cases} \min(i,l-j+1)(\mu_1+\lambda^-), & 1\le i\le c-1,\quad 1\le j\le l, \\ \lambda^-, & 1\le i\le c,\quad j=l+1, \\ (l-j+1)(\mu_1+\lambda^-), & i=c,\quad 1\le j\le l. \end{cases} \tag{6}$$

In order to represent the sub-block matrix of matrix $\boldsymbol{Q}$, the following matrix is defined according to the above mentioned notations

$$\boldsymbol{D}_{i,l+1}=\begin{cases} -(c\mu_s+c\mu_2+\lambda_1), & i=0,\quad l=0, \\ -(c\mu_s+c\mu_2+\lambda_1+\lambda^-), & 1\le i\le c,\quad l=0, \\ \boldsymbol{T}_{l+1}+\mathrm{diag}(\delta_{i,l,l+1},\delta_{i,l,l},\cdots;\delta_{i,l,1}) & ,0\le i\le c,\ 1\le l\le c, \end{cases} \tag{7}$$

where the $l+1$ dimensional square matrix $\boldsymbol{T}_{l+1}(1\le l\le c)$ is represented as follows

$$\boldsymbol{T}_{l+1}=\begin{bmatrix} 0 & l\lambda_s & & & & \\ (c-l+1)\mu_s & 0 & (l-1)\lambda_s & & & \\ & (c-l+2)\mu_s & 0 & (l-2)\lambda_s & & \\ & & \ddots & \ddots & \ddots & \\ & & & (c-1)\mu_s & 0 & \lambda_s \\ & & & & c\mu_s & 0 \end{bmatrix}. \tag{8}$$

$\boldsymbol{\varphi}_{l+1,l},\boldsymbol{\psi}_{l,l+1}$ and $\boldsymbol{F}_{i,l+1}$ are defined as follows:

$$\boldsymbol{\varphi}_{l+1,l}=\begin{bmatrix} 0 & 0 & \cdots & 0 \\ \lambda_2 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ 0 & 0 & \ddots & 0 \\ 0 & 0 & \cdots & \lambda_2 \end{bmatrix}_{(l+1)\times l}, \quad 1\le l\le c, \tag{9}$$

$$\boldsymbol{\psi}_{l,l+1}=\begin{bmatrix} l\mu_s & l\mu_2 & 0 & \cdots & 0 \\ 0 & 0 & l\mu_2 & \cdots & 0 \\ \vdots & 0 & 0 & \ddots & 0 \\ 0 & 0 & 0 & 0 & l\mu_2 \end{bmatrix}_{l\times(l+1)}, \quad 1\le l\le c, \tag{10}$$

when $1\le i\le c,\quad 1\le l\le c$,

$$\boldsymbol{F}_{i,l+1}=\mathrm{diag}(\xi_{i,l,l+1},\xi_{i,l,l},\cdots,\xi_{i,l,2},\xi_{i,l,1}).$$

Thus, the sub-block matrix of matrix $\boldsymbol{Q}$ is expressed as follows:

$$\boldsymbol{A}_0=\begin{bmatrix} \boldsymbol{D}_{0,c+1} & \boldsymbol{\varphi}_{c+1,c} & & & \\ \boldsymbol{\psi}_{c,c+1} & \boldsymbol{D}_{0,c} & \boldsymbol{\varphi}_{c,c-1} & & \\ & \ddots & \ddots & \ddots & \\ & & \boldsymbol{\psi}_{2,3} & \boldsymbol{D}_{0,2} & \boldsymbol{\varphi}_{2,1} \\ & & & \boldsymbol{\psi}_{1,2} & \boldsymbol{D}_{0,1} \end{bmatrix}, \tag{11}$$

$$\boldsymbol{A}_i=\begin{bmatrix} \boldsymbol{D}_{i,c+1} & \boldsymbol{\varphi}_{c+1,c} & & & \\ \boldsymbol{\psi}_{c,c+1} & \boldsymbol{D}_{i,c} & \boldsymbol{\varphi}_{c,c-1} & & \\ & \ddots & \ddots & \ddots & \\ & & \boldsymbol{\psi}_{2,3} & \boldsymbol{D}_{i,2} & \boldsymbol{\varphi}_{2,1} \\ & & & \boldsymbol{\psi}_{1,2} & \boldsymbol{D}_{i,1} \end{bmatrix}, \quad 1\le i\le c-1, \tag{12}$$

$$\boldsymbol{A}=\begin{bmatrix} \boldsymbol{D}_{c,c+1} & \boldsymbol{\varphi}_{c+1,c} & & & \\ \boldsymbol{\psi}_{c,c+1} & \boldsymbol{D}_{c,c} & \boldsymbol{\varphi}_{c,c-1} & & \\ & \ddots & \ddots & \ddots & \\ & & \boldsymbol{\psi}_{2,3} & \boldsymbol{D}_{c,2} & \boldsymbol{\varphi}_{2,1} \\ & & & \boldsymbol{\psi}_{1,2} & \boldsymbol{D}_{c,1} \end{bmatrix}, \tag{13}$$

$$\boldsymbol{B}_i = \begin{bmatrix} \boldsymbol{F}_{i,c+1} & & & & \\ & \boldsymbol{F}_{i,c} & & & \\ & & \ddots & & \\ & & & \boldsymbol{F}_{i,2} & \\ & & & & \lambda^- \end{bmatrix}, \quad 1 \le i \le c-1, \tag{14}$$

$$\boldsymbol{B} = \begin{bmatrix} \boldsymbol{F}_{c,c+1} & & & & \\ & \boldsymbol{F}_{c,c} & & & \\ & & \ddots & & \\ & & & \boldsymbol{F}_{c,2} & \\ & & & & \lambda^- \end{bmatrix}, \tag{15}$$

$$\boldsymbol{C} = \begin{bmatrix} \lambda_1 & & & \\ & \lambda_1 & & \\ & & \ddots & \\ & & & \lambda_1 \end{bmatrix}. \tag{16}$$

3.2 System steady-state analysis

The structure of the matrix $\boldsymbol{Q}$ indicates that the Markov process $\{(L_1(t), L_2(t), J(t)),\ \ t \ge 0\}$ is QBD. When the process is positive recurrent, the steady-state distribution is defined as follows:

$$\pi_{i,l,j} = \lim_{t\to\infty} P\{L_1(t) = i, L_2(t) = l, J(t) = j\}, \quad (i,l,j) \in \Omega$$

$$\begin{aligned}\boldsymbol{\pi}_i =& (\pi_{i,0,0}, \pi_{i,0,1}, \cdots, \pi_{i,0,c}, \pi_{i,1,1}, \pi_{i,1,2}, \cdots, \pi_{i,1,c}, \pi_{i,2,2}, \pi_{i,2,3},\\ & \cdots, \pi_{i,2,c}, \cdots, \pi_{i,c-1,c-1}, \pi_{i,c-1,c}, \pi_{i,c,c}), \quad i \ge 0,\end{aligned} \tag{17}$$

$$\boldsymbol{\Pi} = (\boldsymbol{\pi}_0, \boldsymbol{\pi}_1, \boldsymbol{\pi}_2, \cdots). \tag{18}$$

The sufficient and necessary conditions that QBD $\{(L_1(t), L_2(t), J(t)),\ \ t \ge 0\}$ is positive recurrent is that the matrix quadratic equation

$$\boldsymbol{R}^2\boldsymbol{B} + \boldsymbol{R}\boldsymbol{A} + \boldsymbol{C} = \boldsymbol{0} \tag{19}$$

has a minimum non-negative solution, and the spectral radius $SP(\boldsymbol{R}) < 1$, $(c+1)^2 \times (c+2)/2$ dimensional stochastic matrix

$$B[\boldsymbol{R}] = \begin{bmatrix} \boldsymbol{A}_0 & \boldsymbol{C} & & & & \\ \boldsymbol{B}_1 & \boldsymbol{A}_1 & \boldsymbol{C} & & & \\ & \boldsymbol{B}_2 & \boldsymbol{A}_2 & \boldsymbol{C} & & \\ & & \ddots & \ddots & \ddots & \\ & & & \boldsymbol{B}_{c-1} & \boldsymbol{A}_{c-1} & \boldsymbol{C} \\ & & & & \boldsymbol{B} & \boldsymbol{R}\boldsymbol{B} + \boldsymbol{A} \end{bmatrix} \tag{20}$$

has left-zero vector. When the process is positive recurrent, the steady-state distribution satisfies the following equations

$$\begin{cases} (\boldsymbol{\pi}_0, \boldsymbol{\pi}_1, \cdots, \boldsymbol{\pi}_c) B[\boldsymbol{R}] = \boldsymbol{0}, \\ \sum_{i=0}^{c-1} \boldsymbol{\pi}_i \mathbf{e} + \boldsymbol{\pi}_c (\boldsymbol{I} - \boldsymbol{R})^{-1} \mathbf{e} = 1, \\ \boldsymbol{\pi}_i = \boldsymbol{\pi}_c \boldsymbol{R}^{i-c}, \quad i \ge c \end{cases} \tag{21}$$

where $\boldsymbol{e}$ denotes a $(c+1) \times (c+2)/2$ dimensional column vector which all elements are 1, and $\boldsymbol{I}$ denotes a $(c+1) \times (c+2)/2$ dimensional unit matrix.

The proof process of the above conclusion uses matrix–geometric solution method, and the specific proof process of matrix–geometric solution form of the steady-state distribution can refer to Neuts [26] and Vinod [29]. Since the matrices $\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}$ are complex, the solution $\boldsymbol{R}$ of the equation $\boldsymbol{R}^2\boldsymbol{B} + \boldsymbol{R}\boldsymbol{A} + \boldsymbol{C} = \boldsymbol{0}$ is expressed in implicit form, and the Gauss-Seidel iterative method is used here to solve the approximate solution of the rate matrix $\boldsymbol{R}$, the main steps of the algorithm are shown in Algorithm 1, and the accuracy ε of the algorithm is given here, and the approximate solution of the rate matrix $\boldsymbol{R}$ is obtained when $\|\boldsymbol{R}_n - \boldsymbol{R}_{n-1}\| < \varepsilon$. The specific steps of the iterative algorithm are as follows.

```
Algorithm1.   Itreative algorithm for rate matrix $\boldsymbol{R}$

Step 1.   Input the error precision $\varepsilon(\varepsilon = 10^{-5}), c, \lambda_1, \lambda_2,$
          $\lambda_s, \lambda^-, \mu_1, \mu_2, \mu_s$ and rate matrix $\boldsymbol{R} = \boldsymbol{0}$.
Step 2.   Input $\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}$,
Step 3.   Define $n = 1$,
          $\boldsymbol{R}_{n-1} = \boldsymbol{R}$,
          $\boldsymbol{R}_n = -(\boldsymbol{R}_{n-1}{}^2\boldsymbol{B} + \boldsymbol{C})\boldsymbol{A}^{-1}$.
Step 4.   While $\|\boldsymbol{R}_n - \boldsymbol{R}_{n-1}\|_\infty > \varepsilon$,
          $n = n + 1$,
          Go to Step 3,
          else,
          Go to Step 5.
Step 5.   $\boldsymbol{R} = \boldsymbol{R}_n$.
```

3.3 System performance measures

Based on the above analysis, the following expressions of the system can be obtained under the condition that the steady-state distribution of the P2P network exists.

1. The average queue length of Type I customer is given by

$$E(L_1) = \sum_{i=0}^{\infty} iP(L_1 = i) = \sum_{i=1}^{\infty} i\left(\sum_{l=0}^{c}\sum_{j=l}^{c}\pi_{i,l,j}\right). \tag{22}$$

2. The average queue length of Type II customer is given by

$$E(L_2) = \sum_{l=0}^{c} lP(L_2 = l) = \sum_{l=1}^{c} l\left(\sum_{i=0}^{\infty}\sum_{j=l}^{c}\pi_{i,l,j}\right). \tag{23}$$

3. The average delay of Type I customer is given by

$$E(W_1) = \frac{1}{\lambda_1}E(L_1) = \frac{1}{\lambda_1}\sum_{i=1}^{\infty} i\left(\sum_{l=0}^{c}\sum_{j=l}^{c}\pi_{i,l,j}\right). \tag{24}$$

4. The average delay of Type II customer is given by

$$E(W_2) = \frac{1}{\lambda_2}E(L_2) = \frac{1}{\lambda_2}\sum_{l=1}^{c} l\left(\sum_{i=0}^{\infty}\sum_{j=l}^{c}\pi_{i,l,j}\right). \tag{25}$$

5. The probability that Type II customer is lost is given by

$$P_d = \sum_{i=0}^{\infty}\pi_{i,c,c}. \tag{26}$$

6. The utilization rate of the server is given by

$$P_u = \frac{\min\{E(L_1) + E(L_2), c\}}{c}. \tag{27}$$

7. The activation rate of the server (Jin et al. [30]) (the probability that a server is online in steady state) is given by

$$\beta = \frac{\lambda_s}{\lambda_s + \mu_s}. \tag{28}$$

4 Numerical experiments

The minimum non-negative solution R of the matrix quadratic equation is obtained by Gauss-Seidel iterative method, and the numerical results of $\pi_{i,l,j}$ are obtained by solving the equation of steady-state distribution, and then the P2P system performance measures are obtained. In this section, Matlab is used to program the image of the system performance measures with parameters variation and analyze the influence of system parameters on the performance measures, construct individual benefit functions for two types of customer and social benefit function, analyze the Nash equilibrium between individual benefit and performance measures, and the parameter values that make social benefit optimal, so as to provide a theoretical basis for the scheduling strategy of P2P peers and improve the performance of the system.

4.1 Impact of parameter variation on P2P systematic performance measures

Based on the expressions of the system performance measures obtained above, assuming some parameters $\lambda_1 = 6MB/s, \lambda_3 = 1MB/s, \lambda_s = 2MB/s, \mu_1 = 3MB/s, \mu_s = 1MB/s$.

Assuming $\lambda_2 = 3MB/s$, Fig. 4 reflects the trend of the average queue length $E(L_1)$ of Type I customer in the system with the service rate μ_2 of Type II customer and the number c of online servers. When the number c of online servers is constant, as the service rate μ_2 increases, the speed of the server serving Type I customer becomes larger, so the average queue length $E(L_1)$ of Type I customer decreases. When the service rate μ_2 is constant, as the number c of online servers increases, the chance of the server serving Type I customer becomes larger, so the average queue length $E(L_1)$ of Type I customer decreases.

Assuming $c = 8$, Fig. 5 reflects the trend of the average queue length $E(L_2)$ of Type II customer in the system with the arrival rate λ_2 of Type II customer and the service rate μ_2 of Type II customer. When the arrival rate λ_2 is constant, the average queue length $E(L_2)$ of Type II customer decreases as the service rate μ_2 increases, mainly because the average queue length $E(L_2)$ of Type II customer

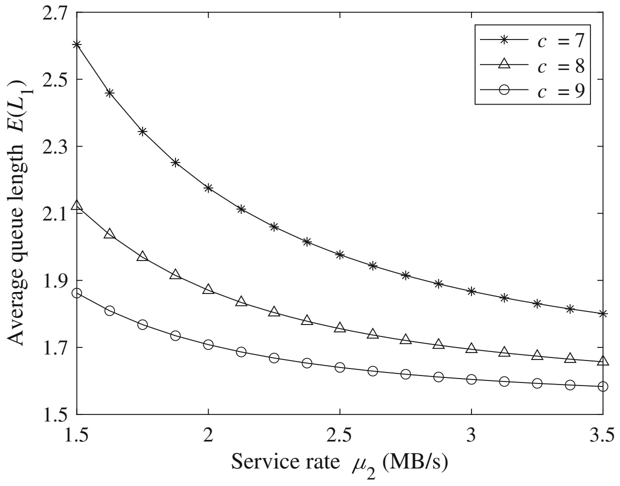


Fig. 4 The relationship between $E(L_1)$ and μ_2, c

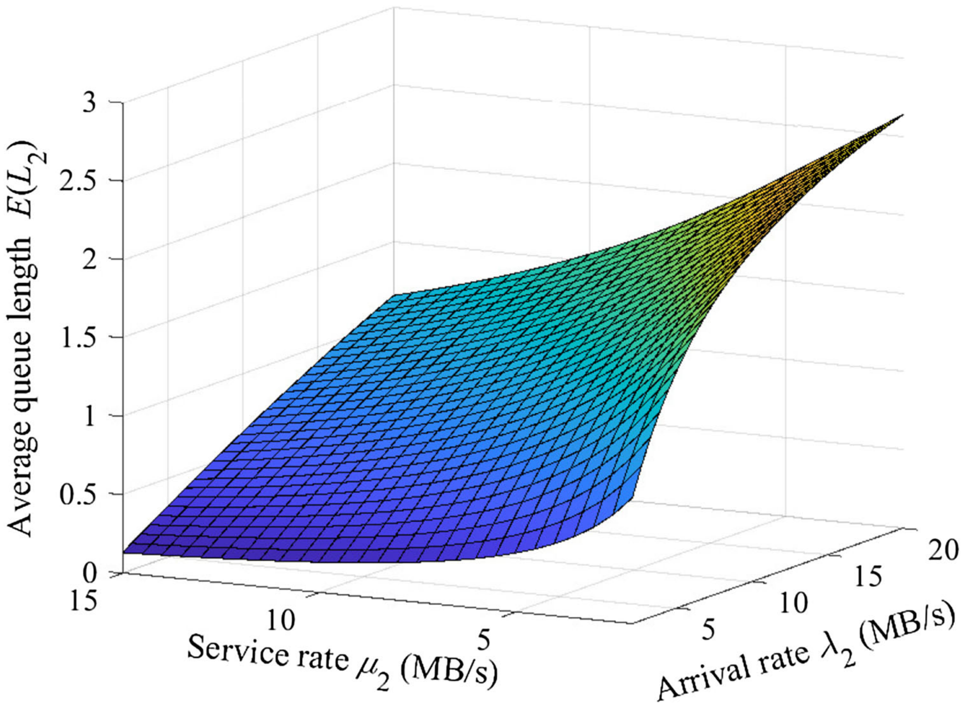


Fig. 5 The relationship between $E(L_2)$ and λ_2, μ_2

decreases when its service rate becomes faster. When the service rate μ_2 is constant, the average queue length $E(L_2)$ of Type II customer keeps increasing as the arrival rate λ_2 increases.

Figure 6 reflects the trend of the average delay $E(W_1)$ of Type I customer in the system with the arrival rate λ_2 of Type II customer, the service rate μ_2 of Type II customer, and the number c of online servers. When the number c of online servers and service rate μ_2 are constant, the average delay $E(W_1)$ of Type I customer increases with the increase of arrival rate λ_2. The main reason is that as the arrival rate λ_2 increases, more Type II customer in the system will preempt the server that is serving Type I customer, which leads to the increase of the average queue length of Type I customer, and therefore the average delay $E(W_1)$ of Type I customer also increases. When the arrival rate λ_2 and service rate μ_2 are constant, as the number c of online servers increases, the system has more opportunities to serve Type I customer, and thus the average delay $E(W_1)$ of Type I customer decreases. When the arrival rate λ_2 and the number c of online servers are constant, as the service rate μ_2 increases, the system can serve Type I customer faster, and thus the average delay $E(W_1)$ of Type I customer decreases. When the arrival rate of Type II customer increases by 20 times, the average delay of the system increases by about 1 to 7 percent; when the service rate of Type II customer increases by 50 percent, the average delay of the system decreases by about 3 to 8 percent; when the number of online servers increases by one, the average delay of the system decreases by about 2 to 6 percent. The numerical results of the system's average delay regarding the arrival and service rates of Type II customer and the number of online servers are shown in Table 2.

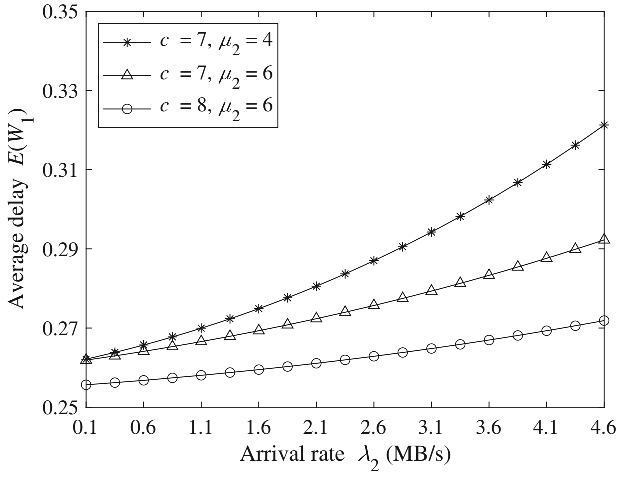


Fig. 6 The relationship between $E(W_1)$ and λ_2, c, μ_2

Assuming $\lambda_2 = 3MB/s$, Fig. 7 reflects the trend of the utilization P_u of the server with the service rate μ_2 and the number c of online servers for Type II customer. When the number c of online servers is constant, the utilization P_u of the server shows a decreasing trend as the service rate μ_2 increases. When the service rate μ_2 is constant, the utilization P_u of the server tends to decrease as the number c of online servers increases. The main reason is that when

Table 2 The values of $E(W_1)$ on parameters λ_2, c, and μ_2

$E(W_1)$						
	$\lambda_2 = 0.1$	$\lambda_2 = 1.1$	$\lambda_2 = 2.1$	$\lambda_2 = 3.1$	$\lambda_2 = 3.6$	$\lambda_2 = 4.1$
$c = 7, \mu_2 = 4$	0.2621	0.2700	0.2806	0.2942	0.3023	0.3113
$c = 7, \mu_2 = 6$	0.2619	0.2666	0.2724	0.2793	0.2833	0.2876
$c = 8, \mu_2 = 6$	0.2557	0.2581	0.2611	0.2648	0.2670	0.2693

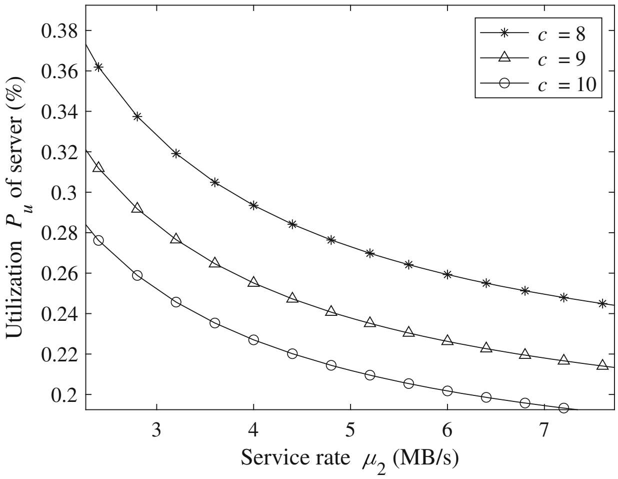


Fig. 7 The relationship between P_u and μ_2, c

the number c of online servers becomes larger, the number c of online servers in the system is more likely to be idle, and thus the utilization P_u of the server decreases. The numerical results of the utilization rate of Type II customer and the number of online servers are shown in Table 3.

Assuming $c = 6$, Fig. 8 reflects the trend that the probability P_d of losing Type II customer in the system varies with the arrival rate λ_2 of Type II customer and the service rate μ_2 of Type II customer. When the service rate μ_2 is constant, the probability P_d of losing Type II customer gradually increases as the arrival rate λ_2 increases. The main reason is that when the service rate μ_2 is constant and the arrival rate λ_2 increases, more Type II customers in the system occupy the server, so the probability of the number of Type II customers reaching the maximum number of servers increases, thus the probability P_d of losing Type II customers increases. When the arrival rate λ_2 is constant, the probability P_d of losing Type II customer decreases as the service rate μ_2 increases. This is mainly because as the service rate μ_2 becomes faster, the average delay of Type II customer decreases, therefore the probability P_d of losing Type II customer decreases.

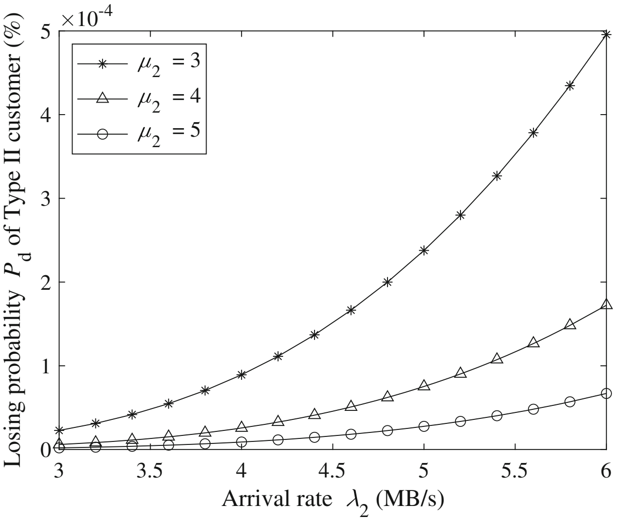


Fig. 8 The relationship between P_d and λ_2, μ_2

Table 3 The values of P_u on parameters μ_2 and c

P_u						
	$\mu_2 = 2$	$\mu_2 = 4$	$\mu_2 = 5.2$	$\mu_2 = 6$	$\mu_2 = 7.2$	$\mu_2 = 8$
$c = 10$	0.4091	0.3302	0.3031	0.2896	0.2739	0.2657
$c = 11$	0.3591	0.2908	0.2670	0.2551	0.2414	0.2341
$c = 12$	0.3255	0.2626	0.2407	0.2298	0.2172	0.2107

4.2 Nash equilibrium and social optimum

By constructing the individual and social benefit functions of Type I and Type II customer, it is analyzed that the optimal arrival rate of two types of customer under the Nash equilibrium strategy and the arrival rate and sevice rate of Type II customer under the social optimum. Assuming R_1, R_2 denote the benefits after serving Type I customer and Type II customer respectively, C_1, C_2 denote the unit waiting cost produced by Type I customer and Type II customer in the system due to delay respectively, C_3 denotes the cost produced per unit time when the server is online, and β denotes the activation rate of the server. According to the above assumptions, the individual benefit of Type I customer and Type II customer are defined as U_1 and U_2 respectively. Then there have

$$U_1 = R_1 - C_1 E(W_1) - C_3\beta, \tag{29}$$

$$U_2 = R_2 - C_2 E(W_2) - C_3\beta. \tag{30}$$

Assuming $c = 7,\ R_1 = 15,\ C_1 = 12,\ C_3 = 1,$ Fig. 9 reflects the trend of the individual benefit U_1 of Type I customer in the system with the arrival rate λ_2 of Type II customer and the service rate of Type II customer. When the service rate μ_2 is constant, the individual benefit U_1 of Type I customer decreases as the arrival rate λ_2 increases. When the arrival rate λ_2 is constant, the individual benefit U_1 of Type I customer keeps increasing as the service rate μ_2 increases. The main reason is that when the arrival rate λ_2 increases, more servers will serve Type II customer first, and thus the average delay of Type I customer becomes larger, so the individual benefit U_1 of Type I customer shows a decreasing trend. When the arrival rate λ_2 is constant and the service rate μ_2 increases, more servers will serve Type I customer, and thus the average delay of Type I customer becomes smaller, so the individual benefit U_1 of Type I customer keeps increasing as the service rate μ_2 increases. The point when the value of the individual benefit U_1 of Type I customer is 0 is the Nash equilibrium point, and the arrival rate at this point is the Nash equilibrium arrival rate. From Table 4,it can be seen that when $\mu_2 = 3.0MB/s$, the Nash equilibrium arrival rate is between $\lambda_2 = 8.7MB/s$ and $\lambda_2 = 9MB/s$; when $\mu_2 = 3.5MB/s$, the Nash equilibrium arrival rate is between $\lambda_2 = 10.2MB/s$ and $\lambda_2 = 10.5MB/s$; and when $\mu_2 = 4.0MB/s$, the Nash equilibrium arrival rate is between $\lambda_2 = 11.7MB/s$ and $\lambda_2 = 12MB/s$.

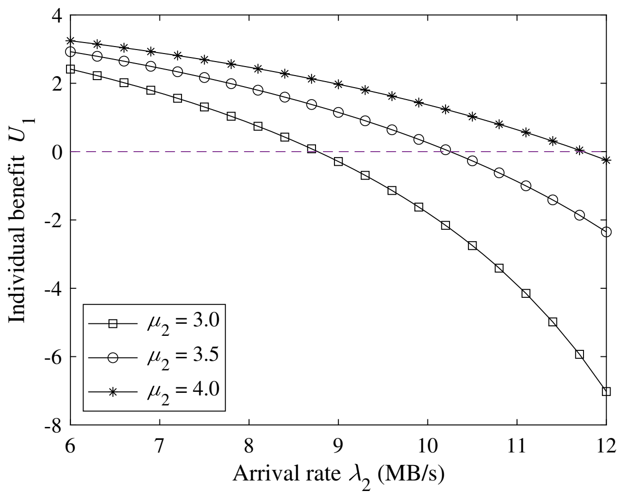


Fig. 9 The relationship between U_1 and λ_2, μ_2

Assuming $c = 8,\ R_2 = 36,\ C_2 = 2,\ C_3 = 1,$ Fig. 10 reflects the trend of the individual benefit U_2 of Type II customer in the system with the arrival rate λ_2 and the service rate μ_2 of Type II customer. When the service rate μ_2 is constant, the individual benefit U_2 of Type II customer decreases with the increases of arrival rate λ_2. When the arrival rate λ_2 is constant, the individual benefit U_2 of Type II customer increases with the increase of service rate μ_2. The main reason is that when the arrival rate λ_2 increases, the individual benefit U_2 of Type II customer increases with the increase of service rate μ_2. The main reason is that when the arrival rate λ_2 increases, the average delay of Type II customer becomes larger, and thus the individual benefit U_2 of Type II customer decreases. When the service rate μ_2 increases, the average delay of Type II customer becomes smaller, and thus the individual benefit U_2 of Type II customer increases. From Fig. 10, it can be known that when $\lambda_2 = 2MB/s$, $\mu_2 = 15MB/s$, there exists a maximum value of individual benefit $U_2 = 268.0662$.

To discuss the optimal social benefit of the system, assuming that R_0 denotes the average benefit gained by the server after serving two types of customer, C_4 denotes the cost of the server serving one customer per unit of time, and C_d denotes the loss of the system when a Type II customer lost, the social benefit U_s is defined as

$$U_s = \lambda_m(R_0 - C_4 E(W_m) - C_3\beta) - C_d P_d. \tag{31}$$

where $\lambda_m = (\lambda_1 + \lambda_2)/2$, $E(W_m) = (E(W_1) + E(W_2))/2$.

Assuming $c = 8$, $R_0 = 14$, $C_3 = 1$, $C_4 = 1.2$, $\beta = 0.7$, $C_d = 26$, Fig. 11 reflects the trend of the social benefit U_s in the system with the arrival rate λ_2 of Type II customer and the service rate μ_2 of Type II customer. When the service rate μ_2 is constant, the social benefit U_s shows a trend of increasing and then decreasing with the increase of the arrival rate λ_2. When the arrival rate λ_2 is constant, the

Table 4 The values of U_1 on parameters λ_2 and μ_2

U_1						
	$\lambda_2 = 8.7$	$\lambda_2 = 9$	$\lambda_2 = 10.2$	$\lambda_2 = 10.5$	$\lambda_2 = 11.7$	$\lambda_2 = 12$
$\mu_2 = 3.0$	0.0822	−0.2898	−2.1582	−2.7507	−5.9330	−7.0221
$\mu_2 = 3.5$	1.3784	1.1480	0.0552	−0.2701	−1.8606	−2.3505
$\mu_2 = 4.0$	2.1302	1.9706	1.2365	1.0253	0.0388	−0.2501

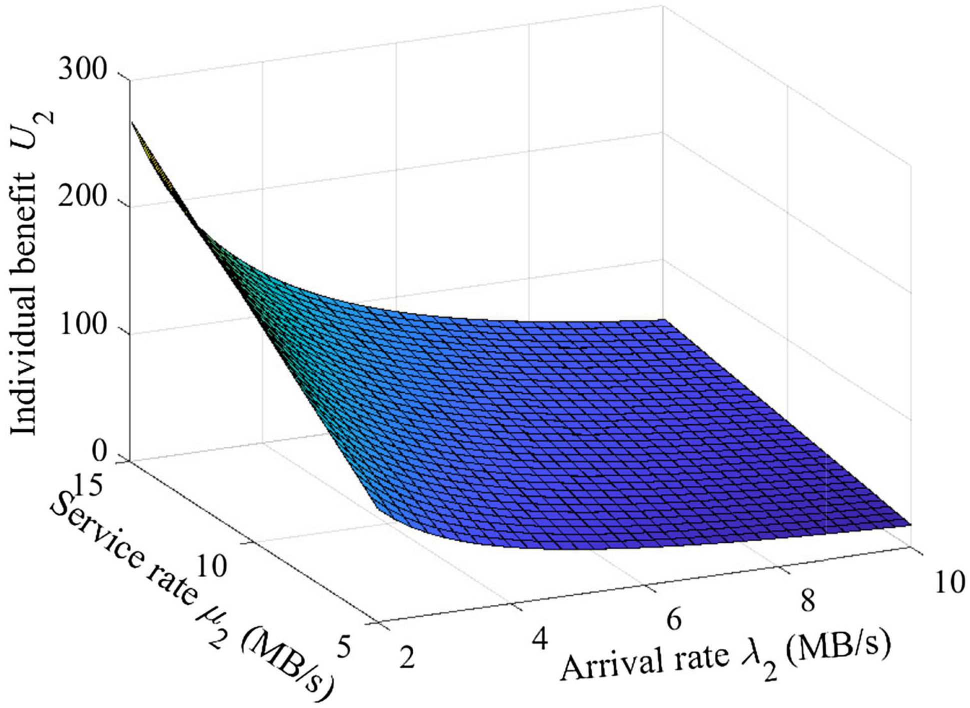


Fig. 10 The relationship between U_2 and λ_2, μ_2

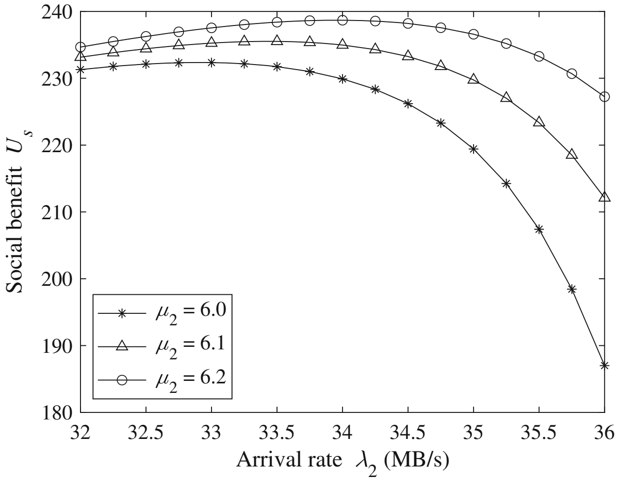


Fig. 11 The relationship between U_s and λ_2, μ_2

social benefit U_s shows a trend of increasing and then decreasing with the increase of the arrival rate λ_2. When the arrival rate λ_2 is constant, the corresponding social benefit U_s increases as the service rate μ_2 increases. The main reason is that the increase of arrival rate λ_2 will make λ_m also increase, and thus U_s will keep increasing, when the arrival rate λ_2 increases to a certain level, it will make the average delay of Type II customer increase, and thus the social benefit U_s will show a trend of increasing and then decreasing. When the service rate μ_2 increases, the average delay of Type II customer decreases, and thus the social benefit U_s keeps increasing. By comparing the data in the diagram, it is obtained that when $\mu_2 = 6MB/s$, $\lambda_2 = 33MB/s$, U_s reaches the maximum; when$\mu_2 = 6.1MB/s$, $\lambda_2 = 33.5MB/s$, U_s reaches the maximum; when $\mu_2 = 6.2MB/s$, $\lambda_2 = 34MB/s$, U_s reaches the maximum.

5 Conclusion and future work

Real-time content delivery in P2P networks is analyzed based on the player online mechanism in P2P networks, and a preemptive queueing model with random changes in the number of servers and negative customer is established. A three-dimensional Markov chain is constructed based on the model assumptions, and the matrix–geometric form of the rate matrix and the steady-state distribution of the system is obtained by using quasi-birth and death process and the matrix–geometric solution method, then the performance measures such as the average queue length of two types of customer in the system and the average delay are obtained. The effect of parameter variation on each measure is analyzed by numerical experiments. Finally, the Nash equilibrium between the arrival rate and individual benefit of Type II customer and the parameter values that make the social benefit optimal are analyzed by constructing the individual benefit function and the social benefit function of the two types of customer.

In the future, it can be considered to expand the application of queueing system in content delivery by extending the arrival interval and playback time of content during

P2P network content delivery and the process of playing content by any two online players to discrete time distribution. In the numerical experimental part, analyzing the impact on performance measures using the throughput of peers, while simulation experiments on system performance measures are conducted to provide more accurate theoretical guidance for the scheduling of P2P nodes.

Acknowledgements This work was supported in part by the National Natural Science Foundation of China under Grant Nos. 61973261, 61872311, Natural Science Foundation of Hebei Province under Grant Nos. A2020203010, A2018203088, Project of Hebei Key Laboratory of Software Engineering, No. 22567637 H.

References

1. Loo, B.T., Hellerstein, J.M., Huebsch, R., Shenker, S., Stoica, I.: Enhancing P2P file-sharing with an internet-scale query processor. In: Proceedings of the Thirtieth Interbational Conference on Very Large Databases, Toronto, Canada, pp. 432-443 (2004)
2. Luo, J.G., Zhang, M., Zhao, L., Yang, S.: A large-scale live video streaming system based on P2P networks. J. Softw. **18**(2), 391–399 (2007)
3. Chen, K., Choffnes, D.R., Potharaju, R., Chen, Y., Bustamante, F., Pei, D., Zhao, Y.: Where the sidewalk ends: Extending the Internet aS graph using traceroutes from P2P users. IEEE Trans. Comput. **63**(4), 1021–1036 (2014)
4. Zhou, Y.P., Chiu, D.M., John, C.S.: A simple model for analyzing P2P streaming protocols. In: IEEE International Conference on Network Protocols, Beijing, China, pp. 226-235 (2007)
5. Susitaival, R., Aalto, S.: Analyzing the file availability and download time in a P2P file sharing system. In: IEEE Next Generation Internet Networks, Trondheim. Norway, pp. 88–95 (2007)
6. Guo, Y., Liang, C., Liu, Y.: AQCS: Adaptive queue-based chunk scheduling for P2P live streaming. In: Proceedings of the Twenty-Fifth International Conference, Helsinki, Finland, pp. 433-444 (2008)
7. Priya, B., Gnanasekaran, T.: To optimize load of hybrid P2P cloud data-center using efficient load optimization and resource minimization algorithm. Peer-to-Peer Netw. Appl. **13**(2), 717–728 (2020)
8. Du, Q.H., Liu, M., Xu, Q., Song, H., Sun, L., Ren, P.: Interference-constrained routing over P2P-share enabled multi-hop D2D networks. Peer-to-Peer Netw. Appl. **10**(6), 1354–1370 (2017)
9. Sun, G.Q., Zhang, Z.Y., Zhou, Y.Z., Han, H.T., Zang, H.X., Wei, Z.N.: Bi-level model for integrated energy service providers in joint electricity and carbon P2P market. J. Clean. Prod. **393**, 136303 (2023)
10. Fattaholmanan, A., Rabiee, H.R.: A large-scale active measurement Study on the effectiveness of piece-attack on bitTorrent networks. IEEE Trans. Depend. Secur. Comput. **13**(5), 509–518 (2016)
11. Sankar, S.P., Subash, T.D., Vishwanath, N., Geroge, D.: Security improvement in block chain technique enabled peer to peer network for beyond 5G and internet of things. Peer-to-Peer Netw. Appl. **14**(1), 392–402 (2021)
12. Bradai, A., Abbasi, U., Landa, R., Ahmed, T.: An efficient playout smoothing mechanism for layered streaming in P2P networks. Peer-to-Peer Netw. Appl. **7**(2), 101–117 (2014)
13. Xu, Z.R., Li, M.J., Zhu, Y.J.: M/M/c queue with negative customers and multiple working vacations. Math. Pract. Theory **41**(11), 91–98 (2011)
14. Ma, Z.Y., Wang, W.B., Wang, Z., Cao, J.: The discrete time working vacation queue with non-preemptive priority and variable service rates. J. Henan Normal Univ. (Natural Science Edition) **46**(1), 23–28 (2018)
15. Wang, Y., Li, W., Chen, M.P., Chen, L., Jin, S.F.: Task offloading strategy with clustered virtual machine and sleep mechanism in MEC. J. Yanshan Univ. **45**(4), 343–351 (2021)
16. Chen, Y.T., Chen, J.Y.: Task offloading strategy with clustered virtual machine and sleep mechanism in MEC. J. Yanshan Univ. **42**(12), 3253–3272 (2022)
17. GnanaSekar, M.M.N., Kandaiyan, I.: Analysis of an M/G/1 retrial queue with delayed repair and feedback under working vacation policy with impatient customers. Symmetry **14**(10), 2024–2042 (2022)
18. Singla, N., Kaur, H.: A two-state retrial queueing model with feedback having two identical parallel servers. Indian J. Sci. Technol. **14**(11), 915–931 (2021)
19. Ye, Q.Q., Chen, Y.: Performance analysis of M/M/1 retrial queuing system with working breakdomns. J. Henan Normal Univ. **51**(3), 82–89 (2023)
20. Pandey, D.C., Pal, A.K.: Delay analysis of a discrete-time non-preemptive priority queue with priority jumps. Appl. Appl. Math. **9**(1), 1–12 (2014)
21. Aibatov, S.Z.: System with priority queue and unreliable server. Moscow Univ. Math. Bull. **71**(3), 111–114 (2016)
22. Ma, Z.Y., Wang, W.B., Zheng, X.M.: The M/M/c queueing model with preemptive priority and multiple synchronous working vacation. J. Chongqing Normal Univ. (Natural Science) **35**(3), 96–100 (2018)
23. Valentina, I.K.: A queueing system with a batch Markovian arrival process and varying priorities. J. Belarusian State Univ. **2**, 47–56 (2022)
24. Bian, J.H., Yin, W.Y., Pang, X.M.: Distribution function of system time of M/M/1 feedback queueing system under nonpreemptive priority discipline. J. Jianghan Univ. (Natural Science) **37**(3), 25–28 (2009)
25. Zhao, Y., Jin, S.F., Yue, W.Y.: An adjustable channel bonding strategy in centralized cognitive radio networks and its performance optimization. Qual. Technol. Quant. Manag. **12**(3), 293–312 (2015)
26. Zhao, G.X., Jin, S.F., Ma, C.B., Cao, J., Xu, L.: Analysis of cloud service systems with variable number of servers. J. Beijing Univ. Posts Telecommun. **42**(4), 114–120 (2019)
27. Zang, W.B., Li, J.X.: Queue model of contact center for customer abandon and variable number of reception desks. J. Appl. Sci. **39**(3), 419–432 (2021)
28. Neuts, M.F.: Matrix-Geometric Solutions in Stochastic Models, pp. 81–141. Johns Hokpins University Press, Baltimore (1981)
29. Vinod, B.: Exponential queues with server vacations. J. Oper. Res. Soc. **37**(10), 1007–1014 (1986)
30. Jin, S.F., Li, Y., Liu, J.P., Huo, Z.: Strategies of nash equilibrium and social optimization for online mechanisms of P2P nodes. J. Jilin Univ. (Engineering and Technology Edition) **46**(1), 296–302 (2016)

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Zhanyou Ma received the B.Sc. degree in Mathematics from Jilin Normal University, Siping, China, the M.Sc. degree in operations research and Ph.D. in Management Science and Engineering from Yanshan University, Qinhuangdao, China. Now Dr. Ma is a professor at School of Sciences, Yanshan University, Qinhuangdao, China. His research interests include queueing systems with vacations and performance evaluation models in communication networks.

Miao Yan received her Bachelor's in Information and Computing Science from the Lvliang University, Lvliang, China. Currently, she is a postgraduate at the School of Science, Yanshan University, Qinhuangdao, China. Her research directions include performance analysis of P2P networks, queuing theory and its application.

Rong Wang received her Bachelor's degree in Mathematics and Applied Mathematics from Hengshui University, Hengshui, China. Currently, she is a postgraduate at the School of Science, Yanshan University, Qinhuangdao, China. Her research interests include P2P networks, queueing systems with vacations and performance evaluation models in communication networks.

Shunzhi Wang received the bachelor degree in statistics from Yanshan University, Qinhuangdao, China. He is currently a postgraduate in School of Science, Yanshan University, Qinhuangdao, China. His research interests include P2P networks and applications, queueing theory and its application.