**Service Science and Management** 服务科学和管理, 2024, 13(1), 29-36
Published Online January 2024 in Hans. https://www.hanspub.org/journal/ssem
https://doi.org/10.12677/ssem.2024.131005

Hans汉斯

# 基于流体方法的排队系统容量设置研究

**余滔滔*，戴　韬**

东华大学旭日工商管理学院，上海

收稿日期：2023年11月28日；录用日期：2023年12月28日；发布日期：2024年1月8日

## 摘　要

**在日益多变的市场环境中，提升服务质量对于各大企业来说至关重要。迫于资源有限、需求不定，服务系统常常超负荷运转，导致顾客无效等待严重、服务质量不佳。因此，本文基于容量可变的拥挤型服务系统，考虑顾客多行为模式的存在，构建了基于流体方法的排队系统容量设置模型，以最优化服务质量。通过算法求解及数值分析，验证了该模型对于服务质量提升的价值，并为企业运营提供了指导方向。**

## 关键词

**流体方法，排队系统，容量设置，服务质量**

# Research on Capacity Setting of Queuing System Based on the Fluid Method

**Taotao Yu*, Tao Dai**

Glorious Sun School of Business and Management, Donghua University, Shanghai

Received: Nov. 28th, 2023; accepted: Dec. 28th, 2023; published: Jan. 8th, 2024

## Abstract

**During the increasingly volatile market environment, improving service quality is crucial for major enterprises. Due to limited resources and uncertain demand, the service system is often overloaded, resulting in serious customer waiting and poor service quality. In this paper, a capacity setting model based on the fluid method is constructed to improve the service quality for queuing systems with variable capacity, while considering the multi-behavioral patterns of customers. Through algorithm solving and numerical analysis, the value of the model for service quality improvement is verified, and a guiding direction is provided for enterprise operation.**

*通讯作者。

文章引用: 余滔滔, 戴韬. 基于流体方法的排队系统容量设置研究[J]. 服务科学和管理, 2024, 13(1): 29-36.
DOI: 10.12677/ssem.2024.131005

**Keywords**

**Fuid Method, Queuing Systems, Capacity Setting, Service Quality**



## 1. 引言

市场环境日趋激烈，服务质量提升已逐步成为各大企业的核心战略之一。迫于资源有限、需求不定，服务系统常常在拥挤环境中运行。由于部分排队队列不可见，顾客无法及时做出正确选择，从而导致无效等待严重、服务质量不佳[1]。

在拥挤环境下，顾客重试行为较为强烈[2]，传统排队方法已无法完成模型构建及分析。近年来，一类确定型流体方法得到广泛运用，用以准确刻画顾客多行为模式下的排队系统[3] [4]。但该方法在轻负载或轻度拥挤情况下准确度不高[5] [6]。因此，本文引入了平均资源利用率系数，以克服容量设置变动对流体方法的影响。

本文以容量可变的拥挤型服务系统为研究对象，考虑构建了基于流体方法的容量设置模型，利用蒙特卡洛搜索算法求解稳态方程组，验证了模型的价值并进行了敏感性分析，为企业实践提供了指导方向。

## 2. 文献综述

服务质量是影响顾客满意度及公司盈利的关键要素之一[7]。增加服务资源、提供延迟公告等都能在一定程度上提升服务质量[8] [9]。同时，探索容量设置及优化也是可行之策：2020 年，胡蓉等[10]以休假延迟和 Min (N, V) – 策略控制的 M/G/1 排队系统为例，推导稳态队长表达式，讨论了系统容量优化设计问题；2023 年，万思燕等人[11]研究具有耐烦服务员的 M/G/1 休假排队系统，推导稳态性能指标，并分析了系统容量优化设计问题。

近年来，关于拥挤型服务系统的研究多集中于顾客多行为模式下的重负载极限推导。2015 年，Ding 等[12]考虑客户重拨及重连接行为，使用流体方法近似排队模型，以推导重负载下的稳态到达率、服务水平和放弃率；2018 年，Yu 等人[13]利用流体近似方法构造了等待提示策略下考虑客户重拨行为的连续排队模型，通过数值分析验证了流体近似方法在重负载环境下的适用性和精确性。

## 3. 基础模型

### 3.1. 模型假设

考虑一个容量可变的拥挤型服务系统，如图 1 所示。系统内共有 $s$ 个代理，服务规则为先到先服务。假定顾客首次到达服从 $\lambda$ 的泊松分布，所有顾客接受服务过程均服从 $\mu$ 的指数分布。顾客初始耐心值 $T_k$ 服从 $\gamma$ 的指数分布，其中，$k$ 代表顾客在系统中所处位置。假定放弃顾客存在 $p$ 的概率选择重试，其余则退出系统。为便于分析，假定顾客重试到达服从 $\delta$ 的指数分布，首次到达与重试到达相互独立，且系统所提供的服务完全相同，由于顾客到达方式分为首次和重试，定义 $\lambda_0$ 为系统可观测的顾客总到达率。

当顾客到达系统时，若系统总顾客数未超出固定容量 $N$，则该顾客可进入系统。否则，将被迫选择放弃(即被阻塞出系统)，假设 $p^a(m)$为稳态下的阻塞概率，其中 $m$ 为稳态下系统总顾客数；进入系统后，

若存在任意空闲服务台，则该顾客可被立即服务；否则将进入等候区等待；此时，“极端不耐烦”顾客由于耐心程度十分有限，将选择直接放弃；剩余顾客则进入队列开始等待。一旦实际等待时间超出耐心时间，顾客将会选择中途放弃，假定概率为 $r(n)$，其中 $n$ 为系统中排队等候的顾客数。

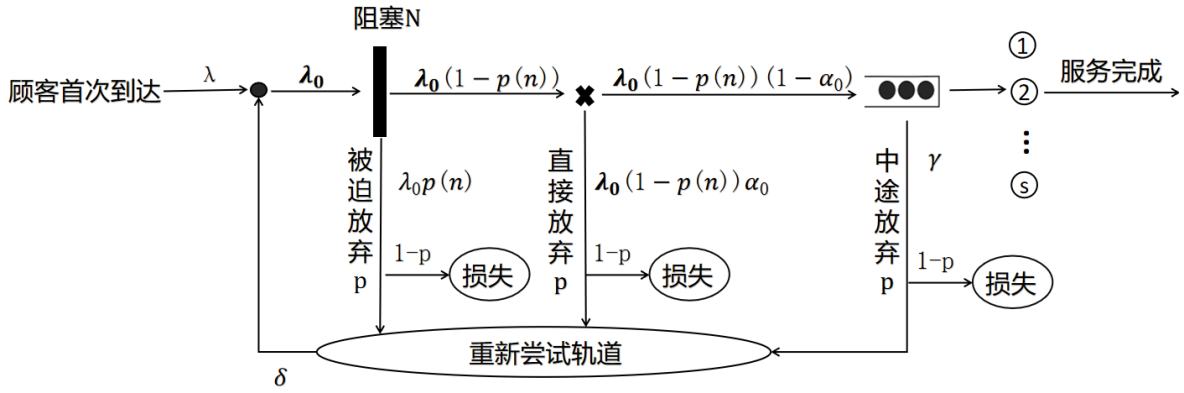


**Figure 1.** Diagram of the base queuing model considering system capacity settings
**图 1.** 考虑系统容量设置的基础排队模型图示

## 3.2. 流体模型构建

由于系统负载及顾客的多重行为，使用流体方法近似上述排队模型。考虑系统负载随系统容量变动，本文引入平均资源利用率参数 $c(0\le c\le 1)$，以确保系统容量设置决策的准确性。

定义 $x_1$， $x_2$ 分别为稳态下真实顾客人数及重试轨道顾客人数。根据排队分析，可推导出上述排队模型的稳态方程组：

$$\begin{cases}\lambda_0\left(1-p^a\left(x_1\right)\right)\left(1-\alpha_0\right)r\left(x_1-s\times c\right)=\gamma\left(x_1-s\times c\right)\\ \lambda_0\left(1-p^a\left(x_1\right)\right)\left(1-\alpha_0\right)-\gamma\left(x_1-s\times c\right)=\mu s\times c\\ p\left(\lambda_0\left(p^a\left(x_1\right)+\left(1-p^a\left(x_1\right)\right)\alpha_0\right)+\gamma\left(x_1-s\times c\right)\right)=\delta x_2\\ s.t.x_1\ge s\times c\end{cases} \tag{3-1}$$

然后，根据 Erlang-A 近似公式[14]可推导出相关参数：

$$p^a(m)=\sum_{i=N}^{\infty}\pi_i \tag{3-2}$$

$$\pi_i=\begin{cases}\pi_s\dfrac{s!}{i!\left(\lambda_0^{'}/\mu\right)^{s-i}} & 0\le i\le s\\ \pi_s\dfrac{\left(\lambda_0^{'}/\gamma\right)^{s-i}}{\prod_{k=1}^{i-s}\left(\frac{s\mu}{\gamma}+k\right)} & i\ge s+1\end{cases} \tag{3-3}$$

$$\pi_s=\frac{E_{1,s}}{1+\left[A\left(\frac{s\mu}{\gamma},\frac{\lambda_0^{'}}{\gamma}\right)-1\right]E_{1,s}} \tag{3-4}$$

其中，$\lambda_0^{'}=\left(1-p(m)\right)\left(1-\alpha_0\right)\lambda_0$ 代表双轨道环境下顾客的实际到达率，$\rho=\lambda_0^{'}/s\mu$ 代表系统实际负载，$E_{1,s}$ 是指经典 Erlang-B 模型下的阻塞概率。

$$A(x,y)=1+\sum_{i=1}^{\infty}\frac{y^i}{\prod_{k=1}^{i}(x+k)} \qquad x>0,y>0$$

$$E_{1,s}=\frac{\frac{\left(\lambda_0^{'}/\mu\right)^s}{s!}}{\sum_{i=0}^{s}\frac{\left(\lambda_0^{'}/\mu\right)^i}{i!}} \tag{3-5}$$

定义实际等待时间为 $d_n$，可近似为虚拟等待时间的平均值 $D_n$，则有：

$$d_n=E(D_n)=\sum_{i=0}^{n}\frac{1}{s\mu+i\gamma} \tag{3-6}$$

$$\begin{aligned} r(n)&=P(T_k<d_n)=1-e^{-\gamma d_n}\\ &=1-e^{-\gamma E(D_n)}=1-e^{-\gamma\sum_{i=0}^{n}\frac{1}{s\mu+i\gamma}} \end{aligned} \tag{3-7}$$

$$x_1=L=\sum_{i=0}^{\infty}i\pi_i \tag{3-8}$$

$$x_2=\frac{\lambda_0-\lambda}{\delta} \tag{3-9}$$

$$c=\frac{\sum_{i=0}^{s-1}i\pi_i+s\times\left(1-\sum_{i=0}^{s-1}i\pi_i\right)}{s} \tag{3-10}$$

以限定时间内服务顾客数为服务质量评定的标准，定义限定时间内完成服务的顾客为优质服务顾客。令 $T$ 为系统限定的服务完成时间，$P_w$ 为顾客在限定时间 $T$ 内被服务的概率，$P\{Ab\}$ 为顾客中途退出的概率，$P(W\le T)$ 为顾客等待时间小于等于 $T$ 的概率，则：

$$P_w=P\{W\le T;Sr\}=\left(1-P\{Ab\}\right)\times P(W\le T) \tag{3-11}$$

$$\begin{aligned} P\{Ab\}&=P\left[Ab\middle|W>0\right]\times P\{W>0\}\\ &=\left(\frac{1}{\rho A\left(s\mu/\gamma,\lambda_0^{'}/\gamma\right)}+1-\frac{1}{\rho}\right)\times\frac{A\left(s\mu/\gamma,\lambda_0^{'}/\gamma\right)E_{1,s}}{1+\left(A\left(s\mu/\gamma,\lambda_0^{'}/\gamma\right)-1\right)E_{1,s}} \end{aligned} \tag{3-12}$$

$$P\{W\le T\}=\sum_{i=0}^{x}\pi_i \tag{3-13}$$

其中，$x=s+n_1$。根据式(3-6)，通过遍历法可求出 $n_1$ 的值。

定义 $P_B$，$P_R$，$P_S$，$P_L$，顾客的直接放弃概率，中途放弃概率，服务顾客概率，以及损失顾客概率。其表达式推导如下：

$$P_B=p^a(m)+\left(1-p^a(m)\right)\alpha_0 \tag{3-14}$$

$$P_R=\left(1-p^a(m)\right)(1-\alpha_0)r(n) \tag{3-15}$$

$$P_S=1-P_B-P_R \tag{3-16}$$

$$P_L=(1-p)\left(p^a(m)+\left(1-p^a(m)\right)\alpha_0+\left(1-p^a(m)\right)(1-\alpha_0)r(n)\right) \tag{3-17}$$

## 4. 数值分析

以容量可变的拥挤型呼叫中心为例进行数值分析，设计蒙特卡洛算法并利用 PyCharm 求解计算。全

局参数固定为：系统服务人员数目 $s=20$，顾客接受服务服从 $\mu=1$ 的指数分布，“极端不耐烦顾客”的比例 $\alpha=0.05$，顾客初始耐心值服从 $\gamma=0.8$ 的指数分布，重试行为服从 $\delta=0.5$ 的指数分布。对比分析主要性能指标：优质服务顾客数目(Well-Served)、总服务顾客数(Total-Served)、总损失顾客数(Total-Lost)、直接放弃顾客数(Balking)、中途退出顾客数(Reneging)、重试顾客数(Retrying)。

首先，令系统负载 $\lambda=30$，顾客选择重试的概率 $p=0.5$，限定完成服务时间 $T=0.5$，探索容量设置策略对于服务质量的影响。

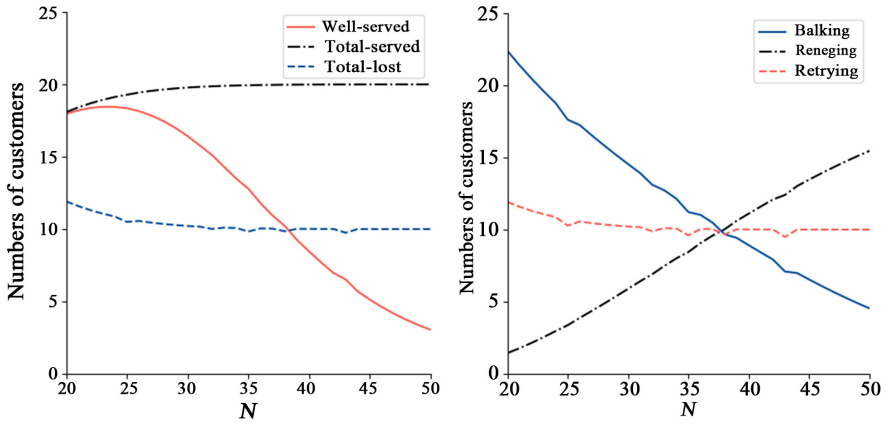


**Figure 2.** Impact of system capacity settings on service quality and customer behavior
**图 2.** 系统容量设置对于服务质量和顾客行为的影响

如图 2，在实验数据范围内，随着容量设置值增大，优质服务顾客数目呈现先增后减的趋势，顾客行为也会受到明显影响。随着 N 的不断增大，直接放弃顾客数逐步减少，中途放弃顾客数逐步增加，重试顾客数略微减少。同时，由于蒙特卡洛的仿真搜索特性，各指标曲线并不完全光滑。

## 4.1. 限定服务时间的敏感性分析

**Table 1.** Results under different limited times
**表 1.** 不同限定完成时间下的结果

| T | 0.1 | 0.3 | 0.5 | 0.8 | 1.0 |
|---|---|---|---|---|---|
| 最佳系统容量值 $N^*$ | 20 | 20 | 23 | 31 | 37 |
| 直接退出顾客数 | 22.41 | 22.41 | 19.60 | 13.92 | 10.48 |
| 中途退出顾客数 | 1.44 | 1.44 | 2.54 | 6.42 | 9.59 |
| 重试顾客数 | 11.92 | 11.92 | 11.07 | 10.17 | 10.04 |
| 优质服务顾客数 | 13.10 | 16.54 | 18.45 | 19.68 | 19.94 |
| 总服务顾客数 | 18.08 | 18.08 | 18.93 | 19.83 | 19.97 |
| 总损失顾客数 | 11.92 | 11.92 | 11.07 | 10.17 | 10.03 |

考虑到不同系统的需求，分析不同限定完成服务时间下的容量设置策略。固定 $\lambda=30$， $p=0.5$，限定时间 $T\in\{0.1,0.3,0.5,0.8,1.0\}$。结果如表 1 所示。

观察表格数据可知：随着系统限定完成服务时间的增大，最佳系统容量值 $N^*$ 逐步增大，系统服务质量也越来越好。当服务时间非常紧迫时，适当减小系统容量，可以引导更多耐心不足的顾客尽早离开，为进入系统的顾客争取足够的服务资源；而当时间较为宽松时，服务系统可适当扩大系统容量，让更多顾客进入系统等待，以系统争取更多的服务机会。

### 4.2. 初始到达率下的敏感性分析

令限定时间 $T=0.5$，顾客重试概率 $p=0.5$，初始到达率 $\lambda \in \{22,25,30,35,40\}$。

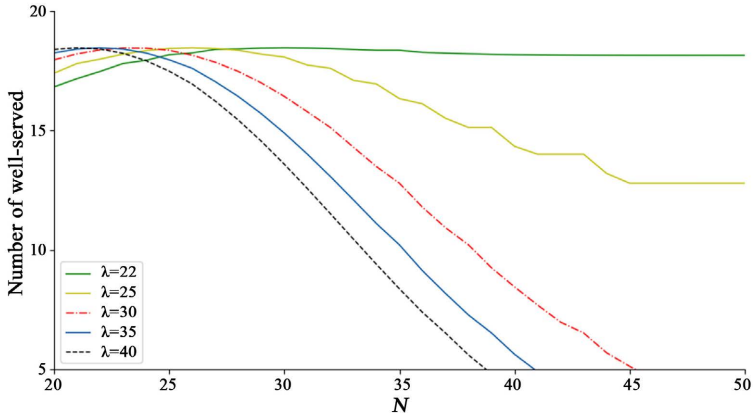


**Figure 3.** Optimal system capacity value for different initial arrival rates
**图 3.** 不同初始到达率下的最佳系统容量值设置

根据上图 3 可知：1) 无论顾客的初始到达率为何值，该服务质量优化模型都能计算出最佳的系统容量值 $N^*$，使得系统的服务质量最佳；2) 对于系统负载越大的系统而言，系统容量的设置尤为重要，如 $\lambda=40$ 时，N 的变动使得优质服务顾客数变化非常明显；3) 随着顾客初始到达率的不断增大，最佳系统容量设置值 $N^*$ 应适当减小。由于顾客到达率的不断增大，系统非常拥挤，因此限制系统容量可以劝退顾客，以免产生无效等待和不满情绪。

### 4.3. 重试行为的敏感性分析

分析最佳系统容量值对于顾客选择重试的概率的敏感程度，固定限定时间 $T=0.5$，初始到达率 $\lambda=30$，顾客重试概率 $p \in \{0,0.3,0.5,0.8\}$。

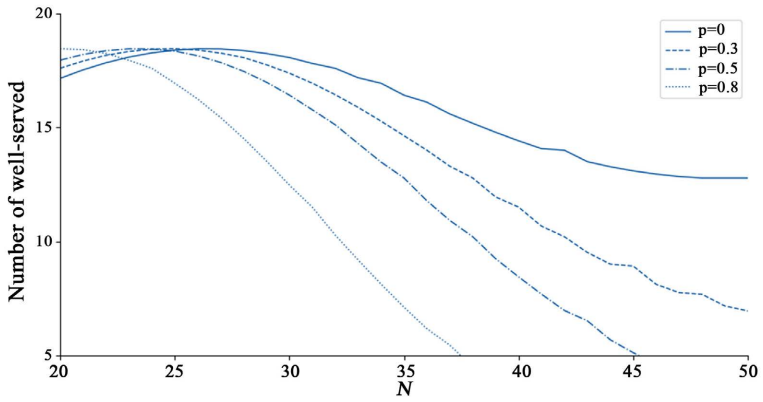


**Figure 4.** Effect of retry probability on the optimal system capacity setting value
**图 4.** 重试概率对最佳系统容量设置值的影响

如图 4，随着重试概率的增大，最佳系统容量设置值 $N^*$ 逐步变小。当顾客选择重试的概率越大时，即便顾客因为阻塞放弃掉，也会极大概率选择重试，因此，系统容量值应设置得更小，以便系统资源能更好地服务于耐心充足的顾客；而如果顾客选择重试的概率越小，意味着其一旦放弃便大概率退出系统，

此时，将系统容量设置得更大以免顾客流失从而造成服务资源的浪费。

**Table 2.** Impact of retry behavior on optimal system capacity settings under different loads
**表 2.** 不同负载下重试行为对最佳系统容量设置值的影响

| | 最佳系统容量设置值 $N^*$ | 直接退出顾客数 | 中途退出顾客数 | 重试顾客数 | 优质服务顾客比例 | 总服务顾客比例 | 总损失顾客比例 |
|---|---|---|---|---|---|---|---|
| | | | $\lambda = 22$ | | | | |
| p = 0 | 42 | 1.11 | 2.21 | 0.05 | 98.18% | 84.94% | 15.06% |
| p = 0.3 | 34 | 1.59 | 2.69 | 1.28 | 97.16% | 86.36% | 13.64% |
| p = 0.5 | 30 | 3.27 | 2.71 | 2.99 | 97.11% | 86.41% | 13.59% |
| p = 0.8 | 25 | 12.22 | 2.71 | 11.94 | 97.11% | 86.41% | 13.59% |
| | | | $\lambda = 25$ | | | | |
| | 最佳系统容量设置值 $N^*$ | 直接退出顾客数 | 中途退出顾客数 | 重试顾客数 | 优质服务顾客比例 | 总服务顾客比例 | 总损失顾客比例 |
| p = 0 | 30 | 3.28 | 2.71 | 0.00 | 97.11% | 76.04% | 23.96% |
| p = 0.3 | 28 | 5.69 | 2.80 | 2.55 | 96.85% | 76.24% | 23.76% |
| p = 0.5 | 26 | 9.25 | 2.72 | 5.99 | 97.11% | 76.04% | 23.96% |
| p = 0.8 | 22 | 27.58 | 2.61 | 24.15 | 97.36% | 75.84% | 24.16% |
| | | | $\lambda = 30$ | | | | |
| | 最佳系统容量设置值 $N^*$ | 直接退出顾客数 | 中途退出顾客数 | 重试顾客数 | 优质服务顾客比例 | 总服务顾客比例 | 总损失顾客比例 |
| p = 0 | 26 | 8.49 | 2.57 | 0.00 | 97.41% | 63.13% | 36.87% |
| p = 0.3 | 25 | 12.84 | 2.80 | 4.69 | 96.90% | 63.50% | 36.50% |
| p = 0.5 | 23 | 19.60 | 2.55 | 11.07 | 97.46% | 63.10% | 36.90% |
| p = 0.8 | 20 | 52.56 | 2.62 | 44.14 | 97.31% | 63.23% | 36.77% |
| | | | $\lambda = 40$ | | | | |
| | 最佳系统容量设置值 $N^*$ | 直接退出顾客数 | 中途退出顾客数 | 重试顾客数 | 优质服务顾客比例 | 总服务顾客比例 | 总损失顾客比例 |
| p = 0 | 24 | 18.01 | 2.89 | 0.00 | 96.60% | 47.75% | 52.25% |
| p = 0.3 | 22 | 27.46 | 2.60 | 9.02 | 97.36% | 47.40% | 52.60% |
| p = 0.5 | 21 | 39.33 | 2.68 | 21.00 | 97.16% | 47.50% | 52.50% |
| p = 0.8 | 20 | 99.33 | 3.72 | 82.44 | 94.02% | 48.48% | 51.53% |

考虑到不同负载下顾客选择重试的概率，进一步探索不同负载下重试行为对最佳系统容量设置值的影响：固定参数 $T = 0.5$， $p \in \{0, 0.3, 0.5, 0.8\}$， $\lambda \in \{22, 25, 30, 40, 50\}$。模型结果如表 2 所示。

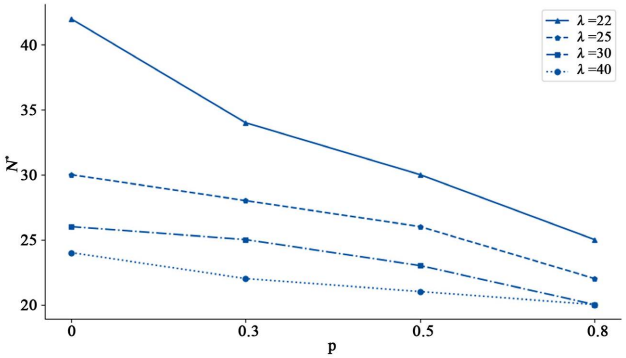


**Figure 5.** Impact of retry probability under different loads
**图 5.** 不同负载下重试概率的影响分析

结合表 2 和图 5 分析可知：1) 无论 p 值如何变动，该流体模型都能帮助计算出最佳的系统容量值，

从而使得系统服务质量水平较为平稳。2) 相较于重拥挤型服务系统而言，系统容量有限的轻拥挤型服务系统需更加关注顾客重试概率的变化：以 $\lambda = 22$ 为例，当顾客重试概率 $p$ 从 0 增加到 0.8 时，最佳系统容量值 $N^*$ 的变化幅度较大。

## 5. 结论与展望

基于容量可变的拥挤型服务系统，本文利用 Fluid 方法构建了容量设置优化模型，证实了该模型对于服务质量提升的意义。通过分析，得出以下几点结论：1) 一旦系统参数固定，该容量设置模型均能帮助获取最佳系统容量值，从而提升服务质量；2) 当限定完成时间较短、顾客的初始到达率较大、顾客重试行为更明显时，最佳系统容量值更小；3) 相较于重拥挤型服务系统而言，系统容量有限的轻拥挤型服务系统需更加注重顾客重试概率的变化。在未来研究中，可进一步优化服务评估方法、改进流体模型，并追求服务质量和服务效率的统一。

## 参考文献


[1] Jouini, O., Akin, O.Z., Karaesmen, F., *et al.* (2014) Call Center Delay Announcement Using a Newsvendor-Like Performance Criterion. *Production and Operations Management Society*, **24**, 587-604. https://doi.org/10.1111/poms.12259

[2] Aguir, M.S., Akşin, O.Z., Karaesmen, F. and Dallery, Y. (2008) On the Interaction between Retrials and Sizing of Call Centers. *European Journal of Operational Research*, **191**, 398-408. https://doi.org/10.1016/j.ejor.2007.06.051

[3] Dai, J.G. and He, S.C. (2012) Many-Server Queues with Customer Abandonment: A Survey of Diffusion and Fluid Approximations. *Journal of Systems Science and Systems Engineering*, **21**, 1-36. https://doi.org/10.1007/s11518-012-5189-y

[4] Whitt, W. (20069) Fluid Models for Multiserver Queues with Abandonments. *Operations Research*, **54**, 37-54. https://doi.org/10.1287/opre.1050.0227

[5] Randhawa, R.S. (2013) Accuracy of Fluid Approximations for Queueing Systems with Congestion-Sensitive Demand and Implications for Capacity Sizing. *Operations Research Letters*, **41**, 27-31. https://doi.org/10.1016/j.orl.2012.10.009

[6] Bassamboo, A. and Randhawa, R.S. (2010) On the Accuracy of Fluid Models for Capacity Sizing in Queueing Systems with Impatient Customers. *Operations Research*, **58**, 1398-1413. https://doi.org/10.1287/opre.1100.0815

[7] Gong, T. and Yi, Y. (2018) The Effect of Service Quality on Customer Satisfaction, Loyalty, and Happiness in Five Asian Countries. *Psychology & Marketing*, **35**, 427-442. https://doi.org/10.1002/mar.21096

[8] Sarkar, A., Mukhopadhyay, A.R. and Ghosh, S.K. (2011) Improvement of Service Quality by Reducing Waiting Time for Service. *Simulation Modelling Practice and Theory*, **19**, 1689-1698. https://doi.org/10.1016/j.simpat.2011.03.004

[9] Yu, M., Gong, J. and Tang, J.F. (2016) Optimal Design of a Multi-Server Queueing System with Delay Information. *Industrial Management & Data Systems*, **116**, 147-169. https://doi.org/10.1108/IMDS-05-2015-0201

[10] 胡蓉, 唐应辉. 具有延迟休假和 Min(N, V)-策略控制的 M/G/1 排队系统容量的优化设计与最优控制策略 N~*[J]. 数学的实践与认识, 2020, 50(19): 107-118.

[11] 万思燕, 兰绍军, 唐应辉. 具有耐烦服务员的 M/G/1 休假排队系统性能分析与容量优化设计[J]. 系统科学与数学, 2023, 43(9): 2292-2309.

[12] Ding, S., Remerova, M., van der Mei, R.D. and Zwart, B. (2015) Fluid Approximation of a Call Center Model with Redials and Reconnects. *Performance Evaluation*, **92**, 24-39. https://doi.org/10.1016/j.peva.2015.07.003

[13] Yu, M., Tang, J.F., Kong, F.W. and Chang, C.G. (2018) Fluid Models for Call Centers with Delay Announcement and Retrials. *Knowledge-Based Systems*, **149**, 99-109. https://doi.org/10.1016/j.knosys.2018.02.040

[14] Mandelbaum, A. and Zeltyn, S. (2005) The Palm/Erlang-A Queue, with Applications to Call Centers. In: Spath, D. and Fähnrich, K.P., Eds., *Advances in Services Innovations*, Springer, Berlin, Heidelberg, 17-45.