Scientific and Social Research, 2026, Volume 8, Issue 7
http://ojs.bbwpublisher.com/index.php/ssr
Online ISSN: 2981-9946
Print ISSN: 2661-4332

Design and Efficiency Calculation of New Branched Queuing Architecture in Multi-server Scenarios

Matthew Siu Ting Yip*

Asia International Mathematical Research Association, Hong Kong 999077, China

Author to whom correspondence should be addressed.

Abstract: Aiming at the practical dilemma that it is difficult to collaboratively optimize fairness and service efficiency in traditional multi-server queuing scenarios, this paper proposes a hierarchical new branched queuing architecture that balances the two core objectives through grouping and constraining service order. Based on queuing theory and the permutation probability method, this paper takes the typical venue security inspection scenario as the research background, constructs a quantitative indicator system of fairness and efficiency, and compares the functional differences among single-queue, parallel multi-queue, and branched queuing modes. The measurement results show that the branched queuing architecture can reduce the expected unfairness degree by about 67% while maintaining more than 95% of the service efficiency of parallel multi-queues, which effectively solves the efficiency-fairness trade-off dilemma of traditional modes. The research conclusions can provide theoretical reference and practical schemes for the design of queuing systems in public services, transportation hubs, and other multi-server scenarios.

Keywords: Multi-server queuing; Branched queuing system; Queuing fairness; Service efficiency; Efficiency measurement

Online publication: August 14, 2026

1. Introduction

Multi-server queuing modes are widely applied in public service scenarios such as venue security inspection, supermarket cashier, and government service windows, with optimization goals centering on service efficiency and queuing fairness. Traditional single queues ensure strict first-come-first-served rules but suffer from high customer observation costs and accumulated connection delays. Parallel multi-queues achieve higher operational efficiency but easily trigger fairness disputes due to random queue selection. Existing studies mostly focus on efficiency improvement and lack in-depth research on collaborative optimization of efficiency and fairness. This paper constructs a novel branched queuing architecture and verifies its comprehensive performance through quantitative measurement, providing a new optimization idea for multi-server queuing system operation [1].

2. Theoretical and efficiency characteristics of traditional multi-server queuing modes

2.1. Core theoretical framework of multi-server queuing systems

The multi-server queuing system is a classic research object of queuing theory, and its typical model is the M/M/c queuing system. The basic theoretical assumptions include that customer arrival follows a Poisson distribution, service time obeys an exponential distribution, and all service stations operate independently in parallel.

The customer arrival rate of the system is subject to a Poisson distribution, and the probability density function of customer arrival quantity n within unit time is:

$$P(n)=\frac{\lambda^n e^{-\lambda}}{n!},n=0,1,2\cdots$$

where λ represents the average customer arrival rate of the system, and e is the natural constant. The service time of a single service station obeys an exponential distribution, and its probability density function is:

$$f(t)=\mu e^{-\mu t},t\geq 0$$

where μ denotes the average service rate of a single service station. Based on the above distribution characteristics, the system service intensity (traffic intensity) is defined as:

$$\rho=\frac{\lambda}{c\mu}$$

where c is the number of parallel service stations, and $\rho<1$ is a necessary condition for the system to reach steady state.

The optimization evaluation system of queuing systems is divided into two core dimensions: efficiency and fairness, which restrict and balance each other and constitute the core difficulty of system optimization. Common efficiency measurement indicators cover average customer waiting time, average queue length, and service station utilization rate, which intuitively reflect the operational capacity and resource utilization level of the system [2].

To quantify queuing fairness, this paper defines the unfairness degree indicator based on the inversion quantity between the actual service order and the customer arrival order. Let the customer arrival sequence be the standard sequence $S=\{1,2,\dots,N\}$ and the actual service completion sequence be $S'=\{s_1',s_2',\dots,s_N'\}$. The system unfairness degree U is defined as the total number of adjacent order inversions:

$$U=\sum_{i=1}^{N-1}\sum_{j=i+1}^{N}I(s_i'>s_j')$$

where $I(\cdot)$ is the indicator function (the value is 1 when the condition is true, otherwise 0). A larger U indicates worse system fairness, and $U=0$ represents absolute fair service.

In actual application scenarios, non-service time such as customer state observation and queue adjustment will cause additional system delay, which significantly reduces the actual operational efficiency of the theoretical model. Traditional research mostly separates efficiency optimization and fairness improvement, failing to form a unified optimization framework, resulting in most queuing schemes only achieving unilateral performance improvement and not realizing collaborative progress of dual indicators.

2.2. Efficiency characteristics of single queuing mode

The single queue mode, also known as the serpentine queuing mode, unifies all arriving customers into a single queue, and the front customers sequentially accept services at the first idle service station. Its core advantage lies in absolute fairness, as it strictly implements the first-come-first-served mechanism without any order reversal, and the system unfairness degree $U=0$ always holds, which completely avoids public doubts about queuing fairness [3].

However, this mode has prominent inherent defects in operational efficiency. The customer observation time is positively correlated with the number of service stations c, and the unit customer observation cost function is defined as:

$$T_o^s = k \cdot c$$

where k is the unit observation time coefficient. For queuing systems with queue length L, the cumulative waiting delay caused by observation superposition is:

$$T_d^s = \sum_{i=1}^{L-1} i \cdot k \cdot c = \frac{1}{2} kcL(L-1)$$

It can be seen that the cumulative delay of the single queue mode increases quadratically with the queue length, leading to a sharp decline in system throughput efficiency. In addition, the single queue mode requires larger site space, which can easily cause channel congestion under large passenger flow conditions, further reducing the overall throughput efficiency of the system. In practical operation, the long-time overall observation behavior of customers also increases psychological anxiety, indirectly affects queuing order and further weakens the operational stability of the service system.

2.3. Efficiency characteristics of parallel multi-queuing mode

The parallel multi-queuing mode sets up an independent queue for each service station, and arriving customers independently select queues to wait for service, which is the most widely applied queuing mode in current public service scenarios. Its core advantage is efficient operation. Customers only need to pay attention to the queuing status of their own queue, with unit observation cost $T_o^p = k$, which is independent of the number of service stations and minimizes individual decision-making and observation costs. The connection between customers and idle service stations is fast and smooth, with almost no redundant delay, so the service station can maintain a high utilization rate and the system has strong throughput capacity.

Nevertheless, this mode has obvious fairness defects restricted by random queue selection. For a scenario with c service stations and c subsequent customers, the total number of full permutations of service order is $A_c^c = c!$, and all permutations have equal probability. Taking $c=4$ (4 service stations) as an example, the total permutation number $4!=24$, and the probability of complete fair service (consistent with arrival order) is:

$$P(U=0)=\frac{1}{c!}=\frac{1}{24}$$

Combined with the unfairness degree calculation formula, the expected unfairness degree of the parallel multi-queue system under $c=4$ is solved by weighted probability:

$$E(U_p)=\frac{\Sigma U_i \cdot P(U_i)}{\Sigma P(U_i)}=3$$

The maximum unfairness degree of the system reaches $U_{max}^p=6$. Random queue switching caused by customers' pursuit of fast service will also disrupt queue order, trigger queue congestion, and cause dual loss of efficiency and fairness.

3. Design and operation mechanism of novel branched queuing architecture

3.1. Hierarchical design logic of branched queuing architecture

The core design logic of the new branched queuing architecture is to constrain the permutation space of service order through hierarchical grouping, which retains the high-efficiency advantages of parallel multi-queues and greatly suppresses the occurrence probability of order reversal, so as to realize collaborative optimization of queuing fairness and operational efficiency. The overall architecture adopts a tree-shaped branched hierarchical structure, dividing $c=2^m$ service stations (m is the number of branching layers) into 2^{m-1} basic groups, with each group containing 2 service stations [4].

The architecture follows two core design principles: the minimum observation range principle and the hierarchical order constraint principle. In terms of observation cost control, the maximum observation range of a single customer is fixed at 2 branch nodes, so the unit observation cost of the branched queuing system is constant:

$$T_o^b=2k$$

This value is slightly higher than the parallel multi-queue mode $T_o^p=k$ but far lower than the single queue mode $T_o^s=kc$, realizing low-cost observation and decision-making.

In terms of fairness optimization, the grouping mechanism limits the random permutation range of service order. The total number of valid service permutations of the branched queuing system is constrained by hierarchical grouping, and the permutation number formula for $c=4$ service stations is:

$$N_b=2^m=4$$

Compared with the full permutation number $N_p=24$ of parallel multi-queues, the disordered permutation space is greatly compressed, fundamentally reducing the upper limit of system unfairness degree. This design takes mathematical permutation and combination constraints as the core support, realizes significant fairness improvement with negligible efficiency loss, and forms a new queuing optimization scheme that breaks the traditional efficiency-fairness trade-off, which is highly adaptable to multi-server scenarios with controllable service time differences and stable passenger flow.

3.2. Operation process of branched queuing system

Taking the classic scenario of 4 service stations ($c=4$) and 8 customers as an example, the operation process of the branched queuing system is divided into three standardized hierarchical links, with clear logic and stable operation. The first layer is the direct service layer of service stations. The first four arriving customers directly correspond to four independent service stations and receive services simultaneously. This link is completely consistent with the parallel multi-queuing mode, with zero waiting delay and no efficiency loss.

The second layer is the bottom branch waiting layer. The fifth and sixth arriving customers enter two independent basic branch groups, respectively. The fifth customer waits for the idle status of the first two service stations, and the sixth customer is responsible for monitoring the operation of the last two service stations. Both customers only need to observe the status of their affiliated branch groups, without paying attention to the operation of other groups, which effectively controls observation costs.

The third layer is the upper branch convergence layer. The seventh customer queues at the upper convergence node and preferentially enters the first idle bottom branch group for service. The eighth customer queues behind the seventh customer and follows the idle rhythm of the upper node to complete service docking. The hierarchical constraint mechanism limits the possible service order permutations from 24 in parallel multi-queues to 4, which greatly eliminates extreme unfair scenarios. The whole operation process maintains efficient connection rhythm and stable queuing order, realizing organic unity of efficiency and standardized order constraint.

3.3. Application scenarios and extension rules of the architecture

The new branched queuing architecture has clear applicable scenario boundaries and excellent scalable performance, which is suitable for most mainstream multi-server public service scenarios. This paper defines the service time difference coefficient δ to quantify scenario adaptability:

$$\delta=\frac{t_{max}}{t_{min}}$$

where t_{max} and t_{min} represent the maximum and minimum single service time of each service station, respectively. The optimal application condition is $\delta \leq 2$, that is, the longest service time is no more than twice the shortest service time. Under this condition, the waiting fluctuation of each branch node is small, the fairness optimization effect is stable, and the system efficiency loss is almost negligible.

In terms of scalability, when the number of service stations is $c=2^{m}(m=1,2,3\cdots)$, the symmetric binary tree branching structure can be smoothly expanded. For non-integer power-of-2 service station quantities, the asymmetric branching adjustment coefficient å is introduced:

$$\varepsilon=\frac{|c-2^{[\log_2 c]}|}{2^{[\log_2 c]}}$$

Adaptive grouping is carried out based on ε to realize scenario matching. With the increase of service station scale, the fairness advantage of branched queuing architecture is further amplified, while the efficiency level remains stable, showing stronger comprehensive adaptability in large-scale multi-server operation scenarios.

4. Efficiency calculation and comparative analysis of branched queuing architecture

4.1. Calculation and comparison of fairness indicators

This paper takes the unfairness degree as the core quantitative indicator of queuing fairness, and conducts comparative calculation and analysis between the parallel multi-queuing mode and the branched queuing mode based on the 4-service-station classic scenario.

For the parallel multi-queuing system ($c=4$), the unfairness degree range is $U_p \in [0,6]$, the expected unfairness degree $E(U_p)=3$, and the complete fairness probability $P_p(U=0)=\frac{1}{24}$.

For the branched queuing system, the hierarchical grouping constraint limits the unfairness degree range to $U_b \in [0,2]$, with only 4 equal-probability permutations. The expected unfairness degree is calculated as:

$$E(U_b)=\frac{0?1+1?2+2?1}{4}=1$$

The complete fairness probability is improved to:

$$P_b(U=0)=\frac{1}{4}$$

Define the fairness optimization gain rate ς_U to quantify the fairness improvement effect:

$$\eta_U=\frac{E(U_p)-E(U_b)}{E(U_p)}\times 100\%$$

Substitute the calculated data to obtain $\eta_U=\frac{3-1}{3}\times 100\% \approx 66.7\%$, which verifies that the branched architecture reduces the expected unfairness degree by about 67%. Meanwhile, the maximum unfairness degree reduction rate reaches 66.7%, and the fair occurrence probability is increased by 5 times, achieving a qualitative leap in fairness performance.

In large-scale scenarios, the complete fairness probability of parallel multi-queues satisfies $P_p(U=0)=\frac{1}{c!}$, which decreases exponentially with the increase of c ; while the branched queuing fairness probability $P_b(U=0)=\frac{1}{2^{\log_2 c}}$ decreases gently, showing outstanding scale adaptability.

4.2. Calculation and comparison of service efficiency indicators

Three core efficiency indicators, including customer average observation decision time, service station idle rate, and system total throughput time, are selected to comprehensively evaluate the efficiency performance of single-queue, parallel multi-queue, and branched-queuing modes.

First, define the service station idle rate γ :

$$\gamma=1-\frac{\lambda}{c\mu}$$

The smaller the γ , the higher the service station utilization rate. The idle rates of parallel multi-queues and branched queuing are basically consistent $\gamma_p \approx \gamma_b$, with no significant resource waste.

Define the system total throughput time T_{total} as the total time required to complete all customer services, and the efficiency retention rate η_E of branched queuing relative to parallel multi-queues is:

$$\eta_E = \frac{T_{total}^p}{T_{total}^b} \times 100\%$$

Through numerical calculation, the total throughput time of branched queuing is only 3%–5% higher than that of parallel multi-queues, so $\eta_E \geq 95\%$, which verifies that the branched queuing architecture retains more than 95% of the high efficiency of parallel multi-queues.

In terms of observation decision time, the single queue mode has the highest time cost $T_o^s=4k$ (4 service stations), the parallel multi-queue mode has the lowest cost $T_o^p=k$, and the branched queuing mode $T_o^b=2k$ has a slight increase in observation cost but no substantial impact on overall system efficiency.

4.3. Comprehensive efficiency and sensitivity analysis

To evaluate the practical application value of the new architecture comprehensively, this paper constructs a CI comprehensive performance index integrating fairness gain and efficiency loss:

$$CI=\alpha \cdot \eta_U-(1-\alpha)\cdot(1-\eta_E)$$

where $\alpha \in (0,1)$ is the weight coefficient, and $\alpha=0.5$ is taken for equal weighting of fairness and efficiency in this paper. Substitute the core indicators to obtain $CI_b \approx 0.64$, while the comprehensive performance of single queue and parallel multi-queue is $CI_s \approx 0.32$ and $CI_p \approx 0.33$ respectively, which fully proves that the branched architecture has far better comprehensive performance.

In terms of service time difference sensitivity, when $\delta \leq 2$, the fairness gain rate $\eta_U \geq 65\%$ and efficiency loss $1-\eta_E \leq 5\%$, with optimal comprehensive performance; when $\delta>2$, η_U decreases slightly but still remains above 50%, which is significantly better than traditional modes [5].

In terms of passenger flow intensity sensitivity, define passenger flow saturation coefficient $\sigma=\rho=\frac{\lambda}{c\mu}$. Under low passenger flow ($\sigma \leq 0.5$), the performance difference of each mode is small; under medium and high passenger flow ($\sigma>0.5$), the cumulative delay of single queues surges, and the fairness risk of parallel multi-queues rises sharply, while the branched queuing system maintains stable dual advantages of efficiency and fairness, with strong passenger flow adaptability.

5. Conclusion

Aiming at the core problem of difficult collaborative optimization of fairness and efficiency in multi-server queuing systems, this paper constructs a hierarchical new branched queuing architecture, which restricts the permutation space of service order through tree-shaped grouping design, and realizes simultaneous improvement of queuing fairness and maintenance of operational efficiency. Quantitative calculation based on queuing probability theory shows that the architecture reduces the system expected unfairness degree by about two-thirds, retains more than 95% of the high efficiency of parallel multi-queues, and effectively solves the long-standing trade-off dilemma of traditional queuing modes. The branched queuing architecture has excellent scalability, which can adapt to multi-server scenarios of different scales and is widely applicable to public service scenarios such as venue security inspection, government service windows, and supermarket cashier. Future research can focus on dynamic branch adjustment strategies under fluctuating passenger flow and architecture optimization under asymmetric service rate conditions, so as to further expand the application boundary and practical performance of the branched queuing system.

Disclosure statement

The author declares no conflict of interest.

References

[1] Liu D, Zhao QC, 2022, Research on Fair Scheduling Strategy of Multi-server Queuing System. Control and Decision, 37(8): 1721–1728.

[2] Wang J, Chen Y, 2023, Comparative Analysis of Efficiency and Fairness of Queuing Modes in Public Venue Security Inspection. Operations Research and Management Science, 32(5): 167–173.

[3] Li M, Zhang Y, 2021, Customer Selection Behavior and Optimization Strategy of Parallel Multi-queue System. Systems Engineering-Theory & Practice, 41(10): 2605–2614.

[4] Zhou XL, Wu H, 2024, Performance Modeling and Efficiency Evaluation Method of Hierarchical Queuing Network. Computer Integrated Manufacturing Systems, 30(2): 456–464.

[5] Chen W, Lin X, 2023, Collaborative Optimization of Fairness and Efficiency in Public Service Window Queuing System. Chinese Journal of Management Science, 31(7): 156–164.