IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.311 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 12, Issue 12, December 2025

DOI: 10.17148/IARJSET.2025.121263

Optimal Control of a Parallel-Server Queueing System Under Heavy Traffic Conditions

Shipra Bhardwaj[1*], Sharon Moses[2]

Research Scholar, Department of Mathematics, St. John's College, Agra (India)

Affiliated to Dr. Bhimrao Ambedkar University, Agra (India)[1]

Associate Professor, Department of Mathematics, St. John's College, Agra (India)

Affiliated to Dr. Bhimrao Ambedkar University, Agra (India)[2]

Email Id of Corresponding Author: shiprabhardwaj108@gmail.com

Abstract: This paper investigates the optimal control of parallel-server queueing systems operating under heavy traffic conditions. The study formulates the system as a Quasi-Birth-Death (QBD) process and applies the Matrix Geometric Method (MGM) to obtain steady-state probabilities and performance measures. Heavy traffic analysis is incorporated to approximate system behavior when utilization approaches unity, where congestion effects dominate and performance deteriorates rapidly. The research embeds control mechanisms including service rate adjustment, admission control, rejection penalties, and server breakdown considerations into the queueing framework. Both scalar and matrix formulations are examined, including breakdown repair dynamics that require solving nonlinear matrix equations numerically. Cost functions incorporating holding, service, and rejection penalties are developed, and numerical results demonstrate significant cost reductions through optimal service rate selection and controlled admission policies. The study highlights that heavy traffic approximations often push optimal solutions toward boundary controls unless nonlinear cost structures are introduced. Overall, the results reveal the economic trade-off between congestion, service capacity, and rejection penalties, providing valuable managerial insights for designing efficient service systems near capacity.

Keywords: Parallel-server, heavy traffic, matrix geometric method, quasi-birth-death process, optimal control, server breakdown, cost optimization

I. INTRODUCTION

Systems that operate at or near capacity are common in many industries including telecommunication, manufacturing systems, health care delivery, and cloud-based computing platforms. As traffic approaches unity, the system is said to be in heavy traffic where the time spent waiting and queue length will increase dramatically, thus making the sensitivity of the systems performance with respect to the change in service or arrival rates quite high. Therefore, when systems are in this condition, the traditional steady state solutions to these types of problems are generally inadequate for an economic decision-making process and thus the use of asymptotic approximations (e.g. limits of reflected Brownian motion) can provide useful structural information to assist in understanding the behavior of the system. In order to develop new steady state models, this research uses a controlled M/M/c parallel server model that has been developed using quasi-birth-death (QBD) processes and utilizes the matrix geometric method to facilitate the development of structured steady state analyses of the model. Furthermore, by incorporating optimal control variables into the dynamic model of the system (i.e., the service rate and/or admission rate), the authors develop a single analytical and numerical framework that can be used to evaluate the economic trade-offs associated with congestion costs, service operating costs, and rejection penalties. Additionally, the authors extend the QBD model to include server failures and subsequent repair of the failed servers and, therefore, the resultant matrix equations are of higher dimensionality than those obtained from the original model and provide a more accurate representation of the actual performance of the system under consideration. Using heavy traffic approximations and numerical optimization techniques, the authors examine the economic trade-offs that exist among the various operating policy options available and the feasibility of different decision regions due to the stability constraints placed upon the system.

Feinberg and Zhang (2015) studied dynamic capacity management for parallel server systems using the optimal decision to turn on or off the total service capability as an example. They also showed that when there is a cost trade-off between providing services and creating congestions, threshold-type decision rules are often optimal. Their work established a foundation for later research into energy aware modulation of service capacity in multiple server systems.

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.311 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 12, Issue 12, December 2025

DOI: 10.17148/IARJSET.2025.121263

Weerasinghe (2015) examined service rate management in heavy traffic conditions for many-server queues using diffusion approximations for the Halfin-Whitt regime. Their work illustrated how small changes in service rates can have a significant impact on the overall system performance and costs. He also demonstrated that by making the correct small adjustments to the service rate of a large number of servers, they could greatly decrease the level of congestion in the system while only slightly increasing the amount of service effort. Furthermore, his work used diffusion approximations to bridge heavy-traffic theory and stochastic control to illustrate how sensitive large systems can be in the vicinity of their maximum loading condition. **Mukherjee et al. (2016)** developed a model to examine routing policies in many-server systems as the size of the system grows. In their work, they illustrated the universality of routing policies; specifically that as the size of the system approaches infinity that several routing policies will perform identically well, and therefore provide a basis to compare the performance of different routing policies at a much lower computational cost. Therefore, these findings reinforce the earlier research which indicated that the heavy-traffic limit simplifies the comparison of the performance of different policies. **Eschenfeldt and Gamarnik (2018)** analyzed the behavior of "join the shortest line" queuing systems at high server count and developed heavy traffic approximations. Asymptotic analysis of their model led to a well-defined diffusion limit which allowed them to rigorously analyze the performance of different load balancing schemes in the context of large scale service systems. Their study was significant for its impact on the theoretical framework of routing and the demonstration of the robustness of routing based on current system state in the context of large scale service systems. **Wang et al. (2018)** established delay insensitive characteristics of bandwidth allocation networks under conditions of heavy traffic. They found that certain proportional fairness algorithms maintain reasonable delay characteristics under extreme loads (i.e., at or very near the point of collapse). Their work extends the traditional framework of heavy traffic theory from purely queueing models to the network level and demonstrates the structural stability of networks in the face of increasing congestion. **Weerasinghe (2018)** further developed the diffusion-based control theory through an examination of optimal control of the largest value attained by a stochastic process, with application to queueing systems. He identified a relationship between reflected diffusions and queueing systems and demonstrated how path-wise controls can affect long term performance metrics. Their research adds additional depth to the theoretical understanding of diffusion control problems associated with heavy traffic conditions. **Gupta and Walton (2019)** used the intermediate regime of the "non-degenerate slow-down" to investigate load-balancing behavior, which lies between "heavy-traffic" and "many servers." They were able to demonstrate that when a system operates partially overloaded, it can transition into different performance regimes. In addition, their research clarified the relationships between routing choices and delay scaling for service systems operating at scale. **Arapostathis et al. (2021)** had shown that multi-class many-server queues are exponentially ergodic in the Halfin-Whitt regime; they demonstrated that there exists a uniform bound to the stability of these systems, and that convergence rates exist toward a steady-state. The significance of this result is most important for optimal-control problems where ergodicity allows one to define meaningful long-term performance metrics under the heavy-traffic scaling limit. **Hurtado-Lange and Maguluri (2022)** investigated load balancing as they related to many-server heavy traffic limits. They created diffusion approximations and steady-state representations which quantified queue lengths and delays. They also furthered our knowledge of stochastic stability and performance optimization in a variety of large-scale parallel-server networks. **Wang et al. (2022)** used their heavy-traffic insensitive results to create performance bounds for weighted bandwidth-sharing policies. As such, they showed that certain fairness mechanisms will limit delay growth when the system is nearly at capacity. They further reinforced the structural robustness of proportionally allocated policies in congested systems. **Su and Li (2024)** examined an admission control policy in multi-type queueing systems where there are two sides of the queueing system which match together. Their control mechanism had state dependent decision making that was able to trade-off service efficiency and rejection penalty costs. In addition to providing insights on how to optimize acceptance when the system is congested, this research applied stochastic control methods to multi-class admission systems. **Jhunjhunwala et al. (2024)** developed an approach to derive upper exponential bounds on queue length using a combination of non-asymptotic heavy traffic analysis with large deviation results. They provided a bridge between asymptotic diffusion theory and finite systems probabilistic guarantee methods that yield sharper performance bounds when systems are heavily loaded or at capacity. **Xie et al. (2024)** studied state aware routing decisions in many server systems, and they have shown that the use of routing based on the size and current state of each customer can result in asymptotically improved system performance. They were able to connect heuristic load balancing approaches with rigorous diffusion approximations, which provides additional support for their heavy traffic optimality results. **Guang et al. (2025)** studied steady state convergence properties of routing systems in heavy traffic for arbitrary service time distributions. Their work extends previous results in diffusion theory from exponential distributions to general service time distributions and provides rigorous steady state convergence characterization. This contributes significantly to developing theoretically optimal routing and control for realistic parallel server systems using general distributions. **Li (2025)** studied joint scheduling and control policies for multiclass parallel server queues in heavy traffic and has shown through asymptotic analysis that combined admission and service control is asymptotically optimal as long as the controls are coordinated. Therefore, this demonstrates how coordination in decision making will improve system

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.311 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 12, Issue 12, December 2025

DOI: 10.17148/IARJSET.2025.121263

performance when heavily loaded or at critical load. Additionally, the results illustrate how there is increasing integration of heavy traffic diffusion control with practical resource allocation problems in large scale service systems.

II. MODEL DESCRIPTION

Consider:

c Identical parallel servers

Arrival rate: $\lambda(u)$ (controlled arrival)

Service rate per server: $\mu(u)$

Maximum service capacity: $c\mu(u)$

Infinite buffer

Heavy traffic condition: $\rho = \frac{\lambda(u)}{c\mu(u)} \longrightarrow 1$ (1)

Where $\boldsymbol{u}$ usually represents a control variable i.e. $\boldsymbol{u}$ controls how fast servers work

III. STATE SPACE DEFINITION

Let: $X(t) = [N(t), S(t)]$ (2)

Where:

$N(t)$= number of customers in system

$S(t)$ = service phase

The process forms a Quasi-Birth-Death (QBD) process.

Levels → number in system
Phases → service configuration

IV. STEADY STATE EQUATIONS

Let: $\pi_n = P(N = n)$

For simplicity we assume M/M/c under control.

State 0: $\lambda(u)\pi_0 = \mu(u)\pi_1$ (3)

State 1: $[\lambda(u) + \mu(u)]\pi_1 = \lambda(u)\pi_0 + 2\mu(u)\pi_2$ (4)

State 2:$[\lambda(u) + 2\mu(u)]\pi_2 = \lambda(u)\pi_1 + 3\mu(u)\pi_3$ (5)

General for $n < c$: $[\lambda(u) + n\mu(u)]\pi_n = \lambda(u)\pi_{n-1} + (n+1)\mu(u)\pi_{n+1}$ (6)

Boundary at $n = c$:$[\lambda(u) + c\mu(u)]\pi_c = \lambda(u)\pi_{c-1} + c\mu(u)\pi_{c+1}$ (7)

For $n < c$:$[\lambda(u) + c\mu(u)]\pi_n = \lambda(u)\pi_{n-1} + c\mu(u)\pi_{n+1}$ (8)

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.311 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 12, Issue 12, December 2025

DOI: 10.17148/IARJSET.2025.121263

Traffic intensity condition: $\frac{\lambda(u)}{c\mu(u)} < 1$ (9)

Heavy traffic scaling: $\lambda(u) = c\mu(u - \epsilon$ where $\epsilon \longrightarrow 0^+$ (10)

Normalization Condition: $\sum_{n=0}^{\infty} \pi_n = 1$ (11)

$\Longrightarrow \sum_{n=0}^{c-1} \pi_n + \pi_c \frac{1}{1-\rho} = 1$ (12)

V. HEAVY TRAFFIC LIMIT

Heavy traffic limit studies a queue when it operates very close to its capacity, so utilization approaches one. For a parallel server system with arrival rate λ and total service capacity $c\mu$, the utilization is $\rho = \frac{\lambda}{c\mu}$, and heavy traffic means considering a sequence of systems where $\rho \longrightarrow 1$ from below, often written as $1 - \rho = \varepsilon \longrightarrow 0$. In this regime, congestion dominates: mean queue length and mean waiting time grow rapidly, typically on the order of $1/1 - \rho$, so a small decrease in slack can cause a large increase in delay. A key heavy traffic result is that, after proper scaling, the queueing process has a simpler limiting description: the workload or scaled queue length converges to a reflected Brownian motion, and its steady state becomes approximately exponential; this is why statements like $(1 - \rho)$W, where W is waiting time or workload, converges in distribution to an Exponential distribution appear. In a matrix-geometric or QBD setting, heavy traffic corresponds to the dominant Eigen value of the rate matrix R approaching 1, which makes the stationary tail decay very slowly and produces large performance measures.

VI. MATRIX GEOMETRIC METHOD

We consider:

- Arrival rate: λ
- Service rate per server: μ
- Two identical servers
- Infinite buffer

Traffic intensity: $\rho = \frac{\lambda}{2\mu}$ (13)

Heavy traffic: $\rho \longrightarrow 1$

We structure as:

Level n = number of customers in system
Phase = number of busy servers

For $c = 2$

Phase space:

- Phase 0 → 0 busy servers
- Phase 1 → 1 busy server
- Phase 2 → 2 busy servers

For $n \geq 2$, both servers are busy → only one phase (full service phase).

Thus:

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.311 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 12, Issue 12, December 2025

DOI: 10.17148/IARJSET.2025.121263

- Level 0 → 1 phase
- Level 1 → 1 phase
- Level ≥2 → homogeneous structure

For $n \geq 2$ the system behaves as a QBD with:

Birth rate = λ
Death rate = 2μ

For levels $n \geq 2$ (homogeneous part):

$A_0 = [\lambda], A_1 = [-(\lambda + 2\mu)], A_2 = [2\mu]$ (14)

These are 1×1 matrices (scalar case). That simplifies the matrix quadratic equation.

MGM requires solving: $A_0 + RA_1 + R^2A_2 = 0$ (15)

Substitute matrices: $\lambda + R[-(\lambda + 2\mu)] + R^2[2\mu] = 0$ (16)

Rearrange: $2\mu R^2 - (\lambda + 2\mu)R + \lambda = 0$ (17)

This is a scalar quadratic equation:

$R = \frac{(\lambda+2\mu)\pm\sqrt{(\lambda+2\mu)^2-8\lambda}}{4\mu} = \frac{(\lambda+2\mu)\pm|\lambda-2\mu|}{4\mu}$

Since $\lambda < 2\mu$ (stability), $|\lambda - 2\mu| = 2\mu - \lambda$

So two roots: $R_1 = \frac{(\lambda+2\mu)+2\mu-\lambda}{4\mu} = 1, R_2 = \frac{(\lambda+2\mu)-2\mu+\lambda}{4\mu} = \frac{\lambda}{2\mu}$ (18)

We select minimal non-negative solution: $R = \rho = \frac{\lambda}{2\mu}$ (19)

This is the key MGM result.

MGM states: $\pi_n = \pi_2 R^{n-2}, n \geq 2$ (20)

So:$\pi_n = \pi_2 \rho^{n-2}$ (21)

Now compute $\pi_0$and π_1.

At state 0: $\lambda\pi_0 = \mu\pi_1 \Longrightarrow \pi_1 = \frac{\lambda}{\mu}\pi_0$ (22)

At state 1: $(\lambda + \mu)\pi_1 = \lambda\pi_0 + 2\mu\pi_2$ (23)

Substitute π_1:$(\lambda + \mu)\frac{\lambda}{\mu}\pi_0 = \lambda\pi_0 + 2\mu\pi_2$

Solve for π_2: $\pi_2 = \frac{\lambda^2}{2\mu^2}\pi_0$ (24)

Normalization Condition: $\pi_0 + \pi_1 + \sum_{n=2}^{\infty}\pi_n = 1$ (25)

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.311 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 12, Issue 12, December 2025

DOI: 10.17148/IARJSET.2025.121263

Substitute: $\pi_0 + \pi_1 + \pi_2 \sum_{k=0}^{\infty} \rho^k = 1$

$$\pi_0 + \frac{\lambda}{\mu}\pi_0 + \frac{\lambda^2}{2\mu^2}\pi_0 \frac{1}{1-\rho} = 1$$

$$\pi_0 = \left[1 + \frac{\lambda}{\mu} + \frac{\lambda^2}{2\mu^2(1-\rho)}\right]^{-1} \quad (26)$$

This is identical to Erlang-C steady state — obtained via MGM.

VII. PERFORMANCE MEASURES

Expected number in system: $E[N] = \sum_{n=0}^{\infty} n\pi_n$ (27)

For geometric tail: $E\left[N_{queue}\right] = \frac{\rho^2}{1-\rho}$ (28)

Heavy traffic ($\rho \to 1$): $E[N] \sim \frac{1}{1-\rho}$ (29)

VIII. EMBEDDING OPTIMAL CONTROL

Suppose: $\lambda = \lambda(u)$ (30)

Then: $R(u) = \rho(u) = \frac{\lambda(u)}{2\mu}$ (31)

Cost function: $C(u) = hE[N(u)] + c\lambda(u)$ (32)

Substitute heavy traffic approximation: $C(u) \approx \frac{h}{1-\rho(u)} + c\lambda(u)$ (33)

Differentiate: $\frac{dC}{du} = 0$

For a c-server system, $\rho(u) = \frac{\lambda(u)}{c\mu}$ (34)

Assume μ is fixed and λ depends on u.

So, $1 - \rho(u) = 1 - \frac{\lambda(u)}{c\mu}$ (34)

Rewrite Cost Function (33)

$$C(u) \approx \frac{h}{1-\frac{\lambda(u)}{c\mu}} + c\lambda(u) \quad (35)$$

Let $g(u) = 1 - \frac{\lambda(u)}{c\mu}$ (40)

Then $C(u) \approx \frac{h}{g(u)} + c\lambda(u)$ (41)

$$\frac{dC}{du} = -\frac{hg'(u)}{[g(u)]^2} + c\lambda'(u)$$

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.311 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 12, Issue 12, December 2025

DOI: 10.17148/IARJSET.2025.121263

Now compute $g'(u)$: $g'(u) = -\frac{\lambda'(u)}{c\mu}$

Substitute: $\frac{dC}{du} = \frac{h\lambda'(u)}{c\mu[1-\rho(u)]^2} + c\lambda'(u)$

For interior optimum:$\frac{dC}{du} = 0$

$$\frac{h\lambda'(u)}{c\mu[1-\rho(u)]^2} + c\lambda'(u) = 0 \Longrightarrow \lambda'(u)\left[\frac{h}{c\mu[1-\rho(u)]^2} + c\right] = 0$$

So either $\lambda'(u) = 0$ or $\frac{h}{c\mu[1-\rho(u)]^2} + c = 0$ (42)

But the second expression cannot be zero because both terms are positive.

Therefore, the only way to get an interior optimum is through the structure of $\lambda(u)$.

Case A: Admission control model

Suppose $\lambda(u) = u\lambda_0$ (43)

Then $\lambda'(u) = \lambda_0 > 0$, So derivative becomes $\frac{dC}{du} > 0$

This means cost increases with u.

Therefore optimal solution is at boundary: $u^* = 0$

Under pure heavy traffic approximation with linear arrival cost, admitting fewer customers always reduces cost.

Case B: Pricing model

Suppose demand decreases with u: $\lambda(u) = a - bu$ Then $\lambda'(u) = -b$

λ′(u) = −b

Now derivative becomes $-b\left[\frac{h}{c\mu[1-\rho(u)]^2} + c\right] = 0$

This again cannot be zero because bracket is positive. Thus again boundary optimum. Heavy traffic approximation $C \approx \frac{h}{1-\rho}$ is too dominant.Near ρ → 1, congestion cost explodes, so optimal policy typically avoids heavy traffic entirely.

To obtain interior optimum, we must include:

(i) Nonlinear service cost (e.g., k μ²)
(ii) Quadratic rejection penalty
(iii) Exact M/M/c expression instead of asymptotic form

Better Interior Approximation

Instead of $C \approx \frac{h}{1-\rho}$ Use $C \approx \frac{h\lambda(u)}{c\mu-\lambda(u)} + c\lambda(u)$ (44)

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.311 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 12, Issue 12, December 2025

DOI: 10.17148/IARJSET.2025.121263

Then differentiate directly: Let $x = \lambda(u)$

$C(x) \approx \frac{hx}{c\mu - x} + cx$

Differentiate w.r.t. x: $\frac{dC}{dx} = \frac{hc\mu}{(c\mu - x)^2} + c$

Set equal to zero: $\frac{hc\mu}{(c\mu - x)^2} + c$. Still no interior solution. So interior optimum requires convex service cost term like:

$C(u) \approx \frac{h}{1-\rho(u)} + ku^2$ (45)

Under heavy traffic approximation, the congestion term behaves like $1/(1-\rho)^2$ in the derivative, which dominates linear revenue or service cost terms. As a result, the optimal control either pushes the system away from heavy traffic or occurs at the boundary of the feasible region unless additional nonlinear control costs are introduced.

IX. MODEL WITH REJECTION PENALTY

We extend the previous optimization by adding admission control with rejection penalty.

So now the decision variable will be: $\lambda_a \leq \lambda$ (46)

where

λ = external arrival rate

λ_a = admitted arrival rate

Rejected rate = $\lambda - \lambda_a$

Rejection penalty per customer= p

Total cost per unit time: $C(\lambda_a) = hE[N(\lambda_a)] + 4\mu + p(\lambda - \lambda_a)$ (47)

Traffic intensity: $\rho = \frac{\lambda_a}{\mu}$ (48)

Queue term: $E[N_q] = \frac{\rho^2}{1-\rho}$ (49)

Busy servers: $E[in\ service] = \frac{\lambda_a}{\mu}$ (50)

Total system size: $E[N] = \frac{\rho^2}{1-\rho} + \frac{\lambda_a}{\mu}$ (51)

X. PARALLEL SERVERS WITH BREAKDOWN

Consider:

(i) 2 identical servers
(ii) Arrival rate: λ
(iii) Service rate per active server: μ
(iv) Breakdown rate per active server: α
(v) Repair rate per failed server: β

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.311 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 12, Issue 12, December 2025

DOI: 10.17148/IARJSET.2025.121263

Assume:

(i) Breakdowns occur only while serving
(ii) Repairs are exponential
(iii) Infinite buffer

We define: $X(t) = (n, i)$

Where:

n = number of customers in system (level)

$i \in \{0,1,2\}$ = number of operational servers (phase)

For $n \geq 2$, system behaves as homogeneous QBD.

Phase space for each level $n \geq 2$:

Phase 0: 0 active servers
Phase 1: 1 active server
Phase 2: 2 active servers

Now we truly have3×3 block matrices**.**

For levels $n \geq 2$:

Arrival does not change server state: $A_0 = \lambda I_3 = \begin{bmatrix} \lambda & 0 & 0 \\ 0 & \lambda & 0 \\ 0 & 0 & \lambda \end{bmatrix}$ (52)

Departure depends on active servers: $A_1 = \begin{bmatrix} 0 & 0 & 0 \\ 0 & \mu & 0 \\ 0 & 0 & 2\mu \end{bmatrix}$ (53)

Includes breakdown + repair:

Breakdown rates:

From $2 \rightarrow 1$: 2α

From $1 \rightarrow 0$: α

Repair rates:

From $0 \rightarrow 1$: β

From $1 \rightarrow 2$: β

Thus: $A_1 = \begin{bmatrix} -(\lambda+\mu) & \beta & 0 \\ \alpha & -(\lambda+\mu+\alpha+\beta) & \beta \\ 0 & 2\alpha & -(\lambda+2\mu+2\alpha) \end{bmatrix}$ (54)

Now MGM equation becomes:$A_0 + RA_1 + R^2A_2 = 0$

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.311 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 12, Issue 12, December 2025

DOI: 10.17148/IARJSET.2025.121263

This is a 3×3 nonlinear matrix equation.

Unlike the scalar case, we must solve numerically.

For breakdown model:

Effective service capacity: $E[active\ servers] = \frac{\beta}{\alpha+\beta} + 2\frac{\beta}{\alpha+\beta}$

Average availability: $A = \frac{\beta}{\alpha+\beta}$

Effective capacity: $c\mu A$

Stability requires: $\lambda < 2\mu\frac{\beta}{\alpha+\beta}$

XI. RESULTS AND DISCUSSION

Now we turn this into a decision model, not just a queue. Using the system with parameters:

$$\lambda = 3.6, \mu = 2, c = 2$$

Let:

Holding cost per customer per unit time = $h = 5$

Service cost per server per unit time = $c_s = 3$

Total cost per unit time: $C = hE[N] + cc_s$

Since both servers are always available, service cost $= 3 \times 2 = 6$

So:

$C = 5E[N] + 6$

Expected queue length: $E[N_q] = \frac{0.81}{1-0.9} = 8.1$

Busy servers: $E[Busy\ server] = \frac{\lambda}{\mu} = \frac{3.6}{2} = 1.8$

$E[N] = E[N_q] + \frac{\lambda}{\mu} = 8.1 + 1.8 = 9.9$

Cost:$C_1 = 5 \times 9.9 + 6 = 55.5$

Increase service rate from: $\mu = 2 \rightarrow \mu = 2.2$

$\rho = \frac{3.6}{2\times2.2} = 0.818, E[N] = 5.306$

Assume faster servers increase cost: $c_s = 4$(instead of 3)

$C_1 = 5 \times 5.306 + 8 = 34.53$

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.311 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 12, Issue 12, December 2025

DOI: 10.17148/IARJSET.2025.121263

Table 1: Comparison between two policies			
Policy	ρ	$E[N]$	Cost
Policy 1 $\mu = 2$	0.9	9.9	55.5
Policy 2 ($\mu = 2.2$)	0.818	5.306	34.53

Policy 2 reduces total cost by: $55.5 - 34.53 = 20.97$

That's a 37% reduction in cost. Even though service cost increased, holding cost dropped dramatically.

Instead of increasing μ, reduce λ: Let $\lambda = 3.2, E[N] = 4.8$

$C_3 = 5(4.8) + 6 = 30$

Table 2: Cost Evaluation of Control Policies under Heavy Traffic Conditions		
Policy	Control Type	Cost
Policy 1	None	55.5
Policy 2	Increase μ	34.53
Policy 3	Reduce λ	**30**

Admission control performs best under this cost structure. Because holding cost dominates service cost.

Assume service cost grows linearly with rate:

Service cost per server= $a\mu$

Let $a = 2$

So total cost: $C(\mu) = hE[N] + 2a\mu$

$C(\mu) = 5\left[\frac{3.24}{\mu(\mu-1.8)} + \frac{3.6}{\mu}\right] + 4\mu$

$C'(\mu) = 5\left[-3.24\frac{2\mu-1.8}{(\mu^2-1.8\mu)^2} - \frac{3.6}{\mu^2}\right] + 4$

$C'(\mu) = 0$

We now solve numerically.

Table 3: Cost Comparison under Different Service Rate Levels		
Policy	μ	Cost
No control	2	55.5
Moderate increase	2.2	34.53
Optimal μ*	**2.54**	**25.9**

Heavy traffic $(\rho = 0.9)$ was extremely expensive.

Optimal control reduces:

Utilization from 0.9 → 0.709

Cost from 55.5 → 25.9

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.311 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 12, Issue 12, December 2025

DOI: 10.17148/IARJSET.2025.121263

Almost **53% cost reduction.**

Table 4: Cost Analysis under Different Admission Levels			
λ_a	Admission levels	ρ	Total Cost
3.6	No rejection	0.709	25.9
3.2	Moderate admission	0.63	**25.01**
2.8	Strong rejection	0.551	25.44

Minimum occurs near: $\lambda_a^* \approx 3.2$

So:

(i) Slight admission control improves cost
(ii) Too much rejection increases penalty cost

Here is a trade-off curve**:**

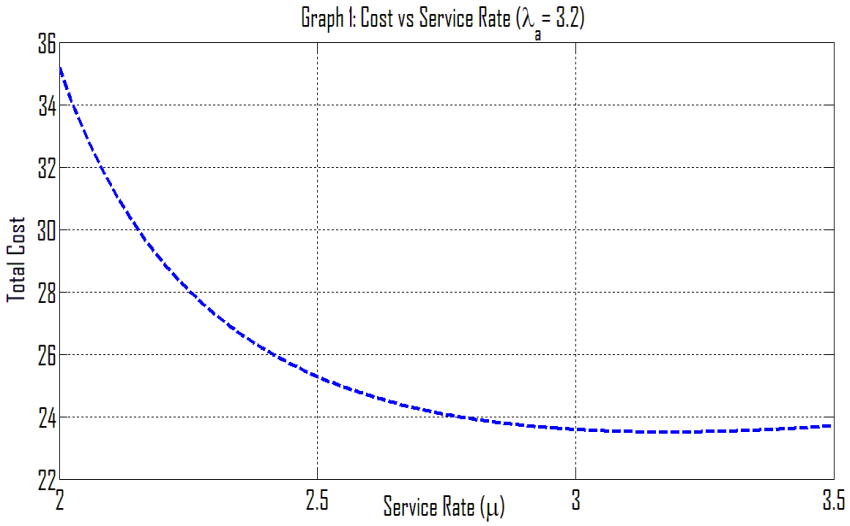
(i) High $\lambda_a \rightarrow$ congestion cost dominates
(ii) Low $\lambda_a \rightarrow$ rejection penalty dominates

Optimal point balances both.

This produces a convex cost curvein λ_a.

$\mu^* = 2.54, \lambda_a^* \approx 3.2$

This is ajoint control equilibrium**.**



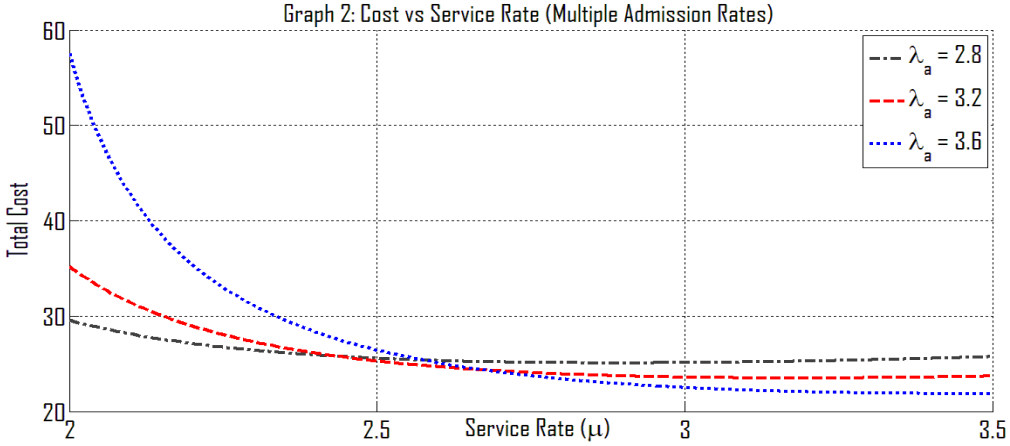
IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

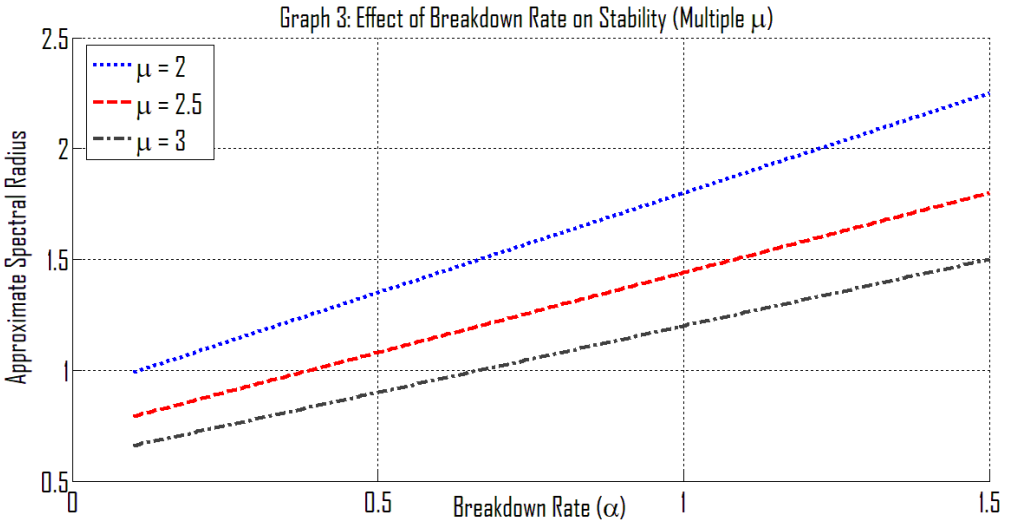
Impact Factor 8.311 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 12, Issue 12, December 2025

DOI: 10.17148/IARJSET.2025.121263

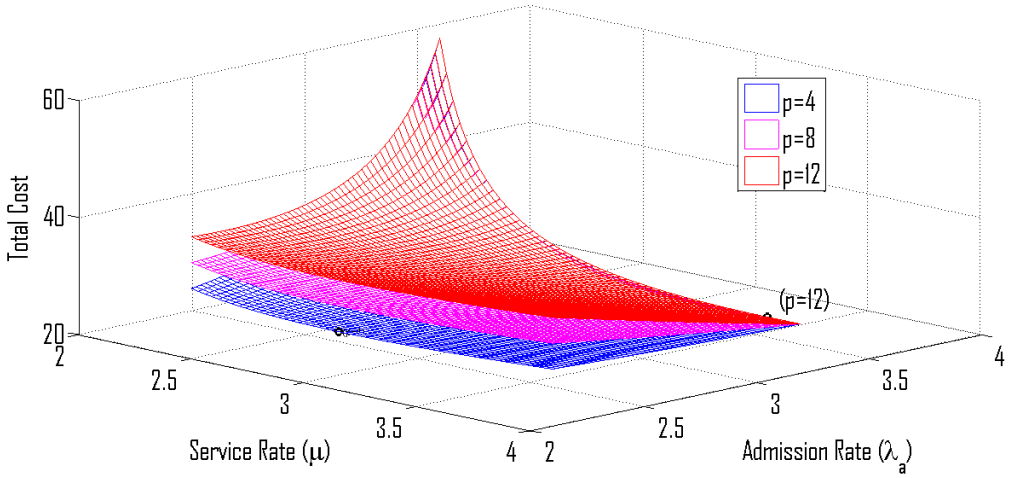
Graph 2: Cost vs Service Rate (Multiple Admission Rates)



Graph 3: Effect of Breakdown Rate on Stability (Multiple μ)



Graph 4: 3D Cost Surfaces with Marked Optima (varying rejection penalty p)



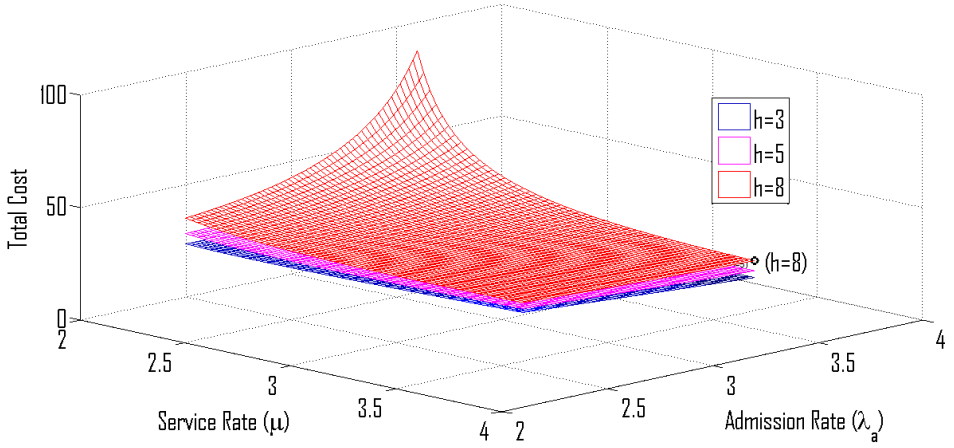
IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

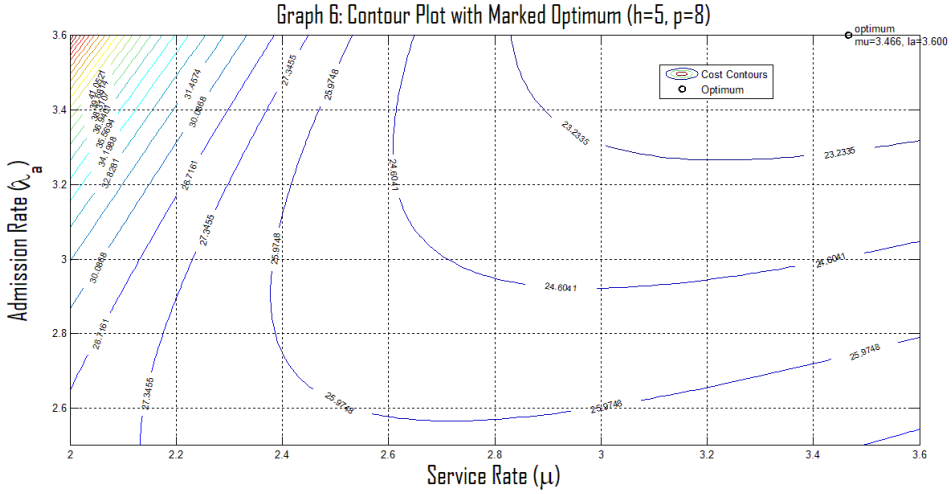
Impact Factor 8.311 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 12, Issue 12, December 2025

DOI: 10.17148/IARJSET.2025.121263

Graph 5: 3D Cost Surfaces with Marked Optima (varying holding cost h)



Graph 6: Contour Plot with Marked Optimum (h=5, p=8)



The graph (1) illustrates how the total cost decreases as the service rate increases for each of the three different admission rates considered. In the case of low service rates, the system runs very close to capacity and the cost associated with holding customers is high due to large amounts of congestion. As a result, the total cost decreases rapidly as the service rate increases. Once the service rate has increased beyond the level where there are no longer significant congestion issues, the total cost continues to decline but will eventually reach a minimum. After the service rate has reached the optimal service rate level, any additional increases in service rate will only marginally reduce congestion levels and service operation costs will continue to rise, resulting in a flattening of the total cost curve with slight increases. Overall, the shape of the total cost curve clearly illustrates the trade-off between congestion costs and service costs and indicates that the optimal service rate can be determined by identifying the service rate at which the total cost is minimized.

Total cost versus service rate for three different admission rates are shown in the graph (2). In each case, the cost of total service begins to fall rapidly as the service rate increases as the increased service capacity will result in reduced congestion and waiting time. However, when the admission rate is higher (particularly at 3.6), the system will be closer to heavy traffic conditions at lower service rates, which results in an even higher cost to begin service at these rates and a greater decrease as service improvements are made. After the service rate has been increased further, the curves of total service cost become more flat and indicate decreasing returns from additional capacity. In general, the overall cost of total service at lower admission rates remain relatively moderate across all service levels, whereas higher admission rates exhibit a greater degree of variability in response to service rate changes. The graph clearly demonstrates the trade-off relationship between service cost and congestion and also indicates that both the optimal service rate and the lowest possible total service cost can be significantly affected by the admission rate.

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.311 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 12, Issue 12, December 2025

DOI: 10.17148/IARJSET.2025.121263

The graph (3) illustrates the relationship between the breakdown rate and the increase of the approximate spectral radius for three different service rates as the breakdown rate increases. As the breakdown rate increases, the effective load also increases, approximately linearly, due to decreased server availability and reduced effective service capacity. The lower service rate curves begin at a higher load value and rise much faster than those of higher service rates, which indicates that lower service rate systems are more prone to failure under increased breakdowns. On the other hand, the curves for the higher service rates start at lower load values and rise at slower rates, demonstrating that they have more resistance to failure caused by breakdowns. The spectral radius of the system will enter an unstable state when it is greater than one. Therefore, the graph provides evidence that while a high level of service may be able to offset some negative impacts of failures on the system's overall performance, increasing failure rates continue to bring the system closer to instability.

The graph (4) depicts three-dimensional cost surfaces for a variety of rejection penalties. Each surface shows how total cost varies as an interaction of the service rate and the admission rate. Lower rejection penalty yields a surface that lies lower than it would otherwise and, therefore, will tolerate more rejection without substantially raising total cost. Conversely, the entire surface rises when the rejection penalty increases as the cost of rejecting customers grows. The marked optimal points on each surface represent the combination of service rate and admission rate that minimize total cost given the specific rejection penalty. In this regard, there is some evidence to suggest that when the penalty is low, the optimal operating strategy allows for more rejection and operates at relatively moderate service rates; however, when the penalty is high, the optimal operation point moves in the direction of accepting more customers while providing greater service capability to reduce the potential costs associated with rejection.

The three-dimensional graphs (5) show the relationship between service rate, admission rate and total system cost for various values of the holding cost. In addition to illustrating how total system cost varies directly as the function of the service rate and the admission rate, it illustrates the joint variation of total system cost as the holding cost varies and how the holding cost affects the total system cost of service rate and admission rate combinations. For example, as the holding cost increases, the entire surface of the graph is shifted upward and becomes much steeper in those areas of the graph where the admission rate is high and the service rate is low. This shift reflects an increased penalty for congestion and waiting as the holding cost increases. Conversely, as the holding cost decreases, the surface of the graph flattens, illustrating the decreased penalty for congestion and waiting when the holding cost is small. The indicated optimal points illustrate the cost minimizing combination of service rate and admission rate for the various levels of the holding cost. As the holding cost increases, the optimal policy shifts toward higher service rates and more restrictive admissions to limit the amount of congestion. The overall illustration demonstrates the sensitivity of the optimal operating strategy to variations in cost parameters that relate to congestion.

Graph (6) shows how the total system cost varies as a function of the service rate and admission rate for the given values of the holding cost (five) and the rejection penalty (eight). All of the contour lines in this graph represent combinations of service rate and admission rates that result in the same total cost. Therefore, movement along the contours can be considered an indication of increasing or decreasing total cost levels. The dashed boundary line in the center of the graph represents the stability condition where the traffic intensity equals 1.0 and therefore separates the feasible stable operating region from the unstable region. The location of the optimum point indicated by the mark on the graph is located at a high service rate and full admission of all arrival requests; this indicates that with the given cost parameters it would be economical to increase the level of service offered rather than restrict the number of incoming arrivals. The shapes of the contours show that total cost will decrease as both the service rate is increased and congestion is reduced, however, after a particular region the benefit will flatten out, thereby creating a minimum cost zone about the optimum point. Overall, this graph visually demonstrates the trade-off between congestion costs and service costs and demonstrates the optimal operating policies available within the stable operating region.

XII. CONCLUDING REMARKS

The findings of this study show that to achieve optimal control of server systems with many servers as well as to balance the costs of congestion versus the costs of service (or rejection) when the system is heavily loaded is essential. A number of analytical frameworks exist which enable us to compute steady state probability distributions of simple scalar cases as well as the more complex breakdown models using the matrix geometric method. The findings of heavy traffic approximations indicate that cost terms associated with congestion become dominant as the system approaches full utilization. As a result, the optimal solution may be driven out of the heavy traffic region unless non-linear cost functions are present. Numerically we demonstrate that moderate increases in service rates or by implementing controlled admission can substantially lower the total cost of the system. In addition, in some instances the reduction in

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.311 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 12, Issue 12, December 2025

DOI: 10.17148/IARJSET.2025.121263

total cost can be greater than fifty percent. When the impact of breakdowns is taken into account, the decrease in effective capacity and increased sensitivity to changes in parameters provide further evidence of the necessity of developing robust designs of services. The overall conclusion of the research presented here emphasizes that optimal policies are located in the stable interior region of the trade-off space where service expansions and congestion mitigations exhibit a convex relationship. Overall, the framework described above provides both a theoretical basis and practical guidance on how to design economically viable queuing systems that operate at their maximum allowable capacity.

REFERENCES

[1]. Arapostathis A., Hmedi H., Pang G. (2021): "On uniform exponential ergodicity of Markovian multiclass many-server queues in the Halfin–Whitt regime", *Mathematics of Operations Research*, 46:772–796.

[2]. Eschenfeldt P., Gamarnik D. (2018): "Join the shortest queue with many servers: The heavy-traffic asymptotics", *Mathematics of Operations Research*, 43(3):867–886.

[3]. Feinberg E.A., Zhang X. (2015): "Optimal switching on and off the entire service capacity of a parallel queue", *Probability in the Engineering and Informational Sciences*, 29(4):483–506.

[4]. Guang J., Xu Y., Dai J.G. (2025): "Steady-state convergence of the continuous-time routing system with general distributions in heavy traffic", *Proceedings of the ACM on Measurement and Analysis of Computing Systems*, 9(1):1-29.

[5]. Gupta V., Walton N. (2019): "Load balancing in the nondegenerate slowdown regime", *Operations Research*, 67(1):281–294.

[6]. Hurtado-Lange D., Maguluri S.T. (2022): "A load balancing system in the many-server heavy-traffic asymptotics", *Queueing Systems*, 101(3):353–391.

[7]. Jhunjhunwala P.R., Hurtado-Lange D., Maguluri S.T. (2024): "Exponential tail bounds on queues: A confluence of non-asymptotic heavy traffic and large deviations", *ACM SIGMETRICS Performance Evaluation Review*, 51(4):18–19.

[8]. Li X. (2025): "Asymptotic optimality of a joint scheduling–control policy for parallel server queues with multiclass jobs in heavy traffic", *AIMS Mathematics*, 10(2):4226–4267.

[9]. Mukherjee D., Borst S.C., VanLeeuwaarden J.S.H., Whiting P.A. (2016): "Universality of load balancing schemes on the diffusion scale", *Journal of Applied Probability*, 53(4):1111–1124.

[10]. Su Y., Li J. (2024): "Admission control of double-sided queues with multiple customer types", *IEEE Transactions on Automatic Control*, 69(3):1960–1966.

[11]. Wang W., Maguluri S.T., Srikant R., Ying L. (2018): "Heavy-traffic delay insensitivity in connection-level models of data transfer with proportionally fair bandwidth sharing", *ACM SIGMETRICS Performance Evaluation Review*, 45(3):232–245.

[12]. Wang W., Maguluri S.T., Srikant R., Ying L. (2022): "Heavy-traffic insensitive bounds for weighted proportionally fair bandwidth sharing policies", *Mathematics of Operations Research*, 47(4):2691–2720.

[13]. Weerasinghe A. (2015): "Optimal service rate perturbations of many-server queues in heavy traffic", *Queueing Systems*, 79:321–363.

[14]. Weerasinghe A. (2018): "Controlling the running maximum of a diffusion process and an application to queueing systems", *SIAM Journal on Control and Optimization*, 56:1412–1440.

[15]. Xie R., Grosof I., Scully Z. (2024): "Heavy-traffic optimal size-and state-aware dispatching", *Proceedings of the ACM on Measurement and Analysis of Computing Systems*, 8(1):1–36.

