Buffer Size and Quality of Service

Candidate Number: 1081225

April 27, 2024

1 Introduction

1.1 Background and Problem Recap

A buffer is a key component that serves as a temporary storage area for data transferred from one node to another in modern communication systems. In particular, when there is a mismatch between the speed of data arrival and transmission, the existence of a buffer can help balance these data transfer rates and ensure the stability and smoothness of the whole transmission process. It can be seen that the setting of buffers is crucial for improving the quality of service of the system. Generally, buffers with large volumes allow more jobs to be transmitted without being rejected, which usually indicates better service quality. While in practice, increasing the buffer size may also result in higher costs and technical challenges. A model that quantifies the relationship between buffer size and quality of service is therefore vital for practitioners to achieve a more cost-effective design for the system.

This report aims to develop a model which studies the relationship between buffer size and the general performance of the data transmission system, namely the quality of service (QoS). In Section 1, we will introduce some preliminary settings relevant to this article, including the general assumptions made about the model and the notations used throughout the paper. The basic model we used will be discussed in Section 2 along with an analytical solution, as well as a comparison with numerical simulations. Section 3 will further introduce some possible extensions of the model. Finally, the conclusion and possible directions for improvements will be given in Section 4. All codes in this paper are available through the link given in Appendix A.1.

1.2 Model Formulation and Assumptions

We will, from now on, consider the following data transmission system given in Fig. 1.2.1.

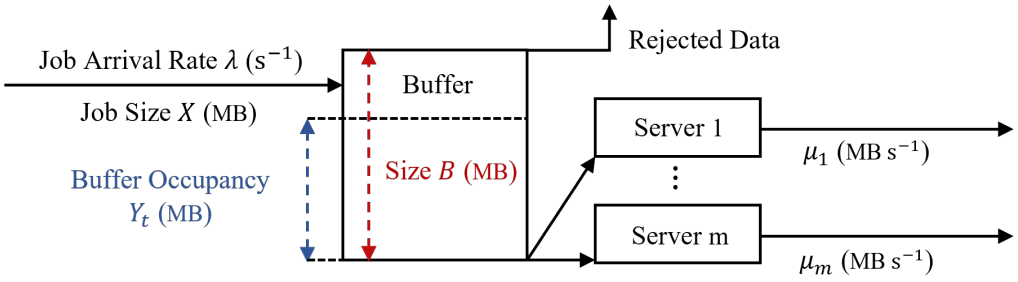

Fig. 1.2.1 The buffer system.

The system is mainly composed of three parts: the jobs, the buffer and the servers.

We assume the jobs arrive at the buffer via a compound Poisson process $\mathrm{CPP}(\lambda, f_X)$ with rate λ and size $X \sim f_X$, where f_X is the probability density function of the job size. The current occupancy of the buffer is represented by Y_t, and if there is space available, the jobs are then accepted by the buffer and assigned to the queue waiting to be transmitted via the servers, each with a transmission rate μ_i.

The ways how jobs are accepted depend on the rejection mode of the buffer. We will investigate two cases where in the first case, the jobs are **partially rejected**, meaning that for each job with a size X, only the part equals to $\min(X, B - Y_t)$ can enter into the buffer, and the rest part of the data is rejected and will need to be re-sent later. In the second case, the jobs are **fully rejected**, indicating a job can either be completely accepted by the buffer, if there is enough space for it, i.e., $X \leq B - Y_t$, or be completely rejected if the space in the buffer is limited.

Finally, the following assumptions are made on the model to continue our discussion. In general, we assume:

- (**Job**) the users of the system are acting independently; thus the one-off events involving large numbers of simultaneous transmissions are neglectable;
- (**Job**) the jobs arrive at the buffer via a compound Poisson process $\mathrm{CPP}(\lambda, f_X)$ with constant rate λ and size $X \sim f_X$;
- (**Job**) the job size distribution f_X is time-invariant, heavy-tailed and includes the information of sizes for re-sent jobs, which can be modelled by a **mixture of exponential distributions** $\mathrm{EXP}(\mathbf{a}, \mathbf{p})$ [1], such that

$$f_X(t) = \sum_{i=1}^{|\mathbf{a}|} \mathbf{p}_i \cdot \mathbf{a}_i \exp(-\mathbf{a}_i t) \quad \text{where} \quad \sum_{i=1}^{|\mathbf{p}|} \mathbf{p}_i = 1 \ ; \tag{1.2.1}$$

- (**Job**) job data are continuous quantities, which can be split up for partial rejection;
- (**Buffer**) the buffer is homogeneous, i.e., the speed of accessing the buffer remains constant;
- (**Server**) the transmission rate for each server remains constant, and, when there are multiple servers, jobs in the queue are assigned to the fastest server available;
- (**Server**) each job in the queue can only be transmitted by one server.

When there is no confusion raised, we will use T_i to denote the jump time of the latent Poisson process modelling the job entering procedure and use $Z_i = T_i - T_{i-1}$ to denote the time interval between two job arrivals. By the definition of the Poisson process, we therefore have $Z_i \sim \mathrm{EXP}(\lambda)$ are independent and identically distributed random variables with expectation $1/\lambda$.

2 The Basic Model

2.1 Preliminary Settings

Our fundamental model considers the specific case when there is only one server working in the system with rate μ and the jobs entering into the buffer are partially rejected. The jobs enter into the system with rate λ via a Poisson process, and their sizes follow a distribution of mixture exponential $X \sim \mathrm{EXP}(\mathbf{a}, \mathbf{p})$ with a probability density function f_X as defined in formula (1.2.1). We further assume that the steady-state distribution of the buffer occupancy Y_t admits the following form

$$f_Y(y) = p_0\delta(y) + p_1(y). \tag{2.1.1}$$

Here $\delta(y)$ is the Dirac delta function; p_0 is the probability for the buffer to be empty at time t and $p_1(y)$ can be interpreted as

$$p_1(y) = \mathbb{P}(Y_t = y \mid Y_t \neq 0) \cdot \mathbb{P}(Y_t \neq 0), \tag{2.1.2}$$

where $\mathbb{P}(Y_t = y \mid Y_t \neq 0)$ is the conditional density of Y_t given it is non-zero. Clearly in our model, f_Y shall have a compact support within the interval $[0, B]$ and satisfies

$$\int_{\mathbb{R}} f_Y(y)\ \mathrm{d}y = p_0 + \int_0^B p_1(y)\ \mathrm{d}y = 1. \tag{2.1.3}$$

2.2 Steady-state Analyses

2.2.1 The Balance Equation

We now start to derive the key equation used in our model to solve for the steady-state distribution f_Y by balancing the probability flow. To begin with, let N_t be the total number of jobs arriving into the system till time t, and consider a small time interval $I = [t, t + \Delta t]$ where the system reaches the steady state. By the property of the Poisson process, we have $N_{t+\Delta t} - N_t$, i.e., the number of jobs entering into the system within the time interval I, follows a Poisson distribution with mean $\lambda \cdot \Delta t$. Thus, within such a short time interval Δt, the probability for more than one job to arrive at the buffer is neglectable in order $\mathcal{O}(\Delta t^2)$

$$\begin{aligned}
q_0 &= \mathbb{P}(N_{t+\Delta t} - N_t = 0) = e^{-\lambda\Delta t} = 1 - \lambda\Delta t + \mathcal{O}(\Delta t^2),\\
q_1 &= \mathbb{P}(N_{t+\Delta t} - N_t = 1) = e^{-\lambda\Delta t} \cdot \lambda\Delta t = \lambda\Delta t + \mathcal{O}(\Delta t^2),\\
q_2 &= \mathbb{P}(N_{t+\Delta t} - N_t \geq 2) = 1 - e^{-\lambda\Delta t} - e^{-\lambda\Delta t} \cdot \lambda\Delta t = \mathcal{O}(\Delta t^2).
\end{aligned}$$

Let $\mathrm{P}(t, y) = \mathbb{P}(Y_t \geq y)$ for some $y \in [0, B]$ and when only considering the probability change caused by the jobs coming in, we have

$$\begin{aligned}\mathrm{P}(t+\Delta t, y) &= q_0 \cdot \mathrm{P}(t, y) + q_1 \cdot \mathrm{P}(t, y - X) + \mathcal{O}(\Delta t^2) \\ &= (1 - \lambda\Delta t) \cdot \mathrm{P}(t, y) + \lambda\Delta t \cdot \mathrm{P}(t, y - X) + \mathcal{O}(\Delta t^2) \\ &= \mathrm{P}(t, y) + \lambda\Delta t \cdot \mathbb{P}(y - X \leq Y_t \leq y) + \mathcal{O}(\Delta t^2). \end{aligned} \tag{2.2.1.1}$$

Recall that X is the job size. By moving the term $\mathrm{P}(t, y)$ in equation (2.2.1.1) from right to left and dividing both sides with Δt, we have

$$\frac{\mathrm{P}(t+\Delta t, y) - \mathrm{P}(t, y)}{\Delta t} = \lambda \cdot \mathbb{P}(y - X \leq Y_t \leq y) + \mathcal{O}(\Delta t).$$

Let $\Delta t \to 0$, the probability change caused by jobs coming in is

$$\left[\frac{\partial \mathrm{P}}{\partial t}\right]^+ = \lambda \cdot \mathbb{P}(y - X \leq Y_t \leq y). \tag{2.2.1.2}$$

Similarly, when only considering the probability change caused by data transmission, we have

$$\mathrm{P}(t+\Delta t, y) = \mathrm{P}(t, y + \mu\Delta t).$$

Thus, for $y > 0$, there is

$$\begin{aligned}\frac{\mathrm{P}(t+\Delta t, y) - \mathrm{P}(t, y)}{\Delta t} &= \frac{\mathrm{P}(t, y+\mu\Delta t) - \mathrm{P}(t, y)}{\Delta t} = -\frac{1}{\Delta t}\mathbb{P}(y \leq Y_t \leq y + \mu\Delta t) \\ &= -\frac{1}{\Delta t}\int_y^{y+\mu\Delta t} f_Y(\tilde{y})\ \mathrm{d}\tilde{y} = -\frac{1}{\Delta t}\int_y^{y+\mu\Delta t} p_1(\tilde{y})\ \mathrm{d}\tilde{y}.\end{aligned}$$

Letting $\Delta t \to 0$, and the probability change led by data transmission is given by

$$\left[\frac{\partial \mathrm{P}}{\partial t}\right]^- = -\lim_{\Delta t \to 0}\frac{1}{\Delta t}\int_y^{y+\mu\Delta t} p_1(\tilde{y})\ \mathrm{d}\tilde{y} = -\mu \cdot p_1(y). \tag{2.2.1.3}$$

In terms of the steady state, the probability distribution of the buffer occupancy Y_t should be invariant with time. We therefore conclude

$$\left[\frac{\partial \mathrm{P}}{\partial t}\right]^+ + \left[\frac{\partial \mathrm{P}}{\partial t}\right]^- = 0. \tag{2.2.1.4}$$

Namely, $\lambda \cdot \mathbb{P}(y - X \leq Y_t \leq y) = \mu \cdot p_1(y)$. By the law of total probability,

$$\begin{aligned}\mathbb{P}(y - X \leq Y_t \leq y) &= \int_0^B \mathbb{P}(y - X \leq Y_t \leq y \mid Y_t = \tilde{y}) \cdot f_Y(\tilde{y})\ \mathrm{d}\tilde{y} \\ &= \int_0^y \mathbb{P}(X \geq y - \tilde{y}) \cdot f_Y(\tilde{y})\ \mathrm{d}\tilde{y} \quad \stackrel{\Delta}{=} \int_0^y F_c(y - \tilde{y}) \cdot f_Y(\tilde{y})\ \mathrm{d}\tilde{y},\end{aligned}$$

the final form of the balance equation is given as

$$\int_0^y F_c(y - \tilde{y}) \cdot f_Y(\tilde{y})\ \mathrm{d}\tilde{y} = \frac{\mu}{\lambda} \cdot p_1(y), \tag{2.2.1.5}$$

where F_c is the complementary cumulative distribution function of the job size X.

2.2.2 The Analytical Solution

In the previous section, we end up with two integral equations which uniquely determine the steady-state distribution of the buffer occupancy

$$\int_0^y F_c(y-\tilde{y}) \cdot f_Y(\tilde{y}) \, \mathrm{d}\tilde{y} = \frac{\mu}{\lambda} \cdot p_1(y), \tag{2.2.2.1}$$

and

$$p_0 + \int_0^B p_1(y) \, \mathrm{d}y = 1. \tag{2.2.2.2}$$

Further substitute $f_Y(y) = p_0\delta(y) + p_1(y)$ into the equation (2.2.2.1), we obtain

$$p_0 \cdot F_c(y) + \int_0^y F_c(y-\tilde{y}) \cdot p_1(\tilde{y}) \, \mathrm{d}\tilde{y} = \frac{\mu}{\lambda} \cdot p_1(y). \tag{2.2.2.3}$$

Dividing both (2.2.2.2) and (2.2.2.3) with p_0, and denoting $p_1(y)/p_0$ with $g(y)$, we have

$$F_c(y) + \int_0^y F_c(y-\tilde{y}) \cdot g(\tilde{y}) \, \mathrm{d}\tilde{y} = \frac{\mu}{\lambda} \cdot g(y), \tag{2.2.2.4}$$

$$p_0 = \left(1 + \int_0^B g(y) \, \mathrm{d}y\right)^{-1}. \tag{2.2.2.5}$$

We note there is a convolution term $\int_0^y F_c(y-\tilde{y}) \cdot p_1(\tilde{y}) \, \mathrm{d}\tilde{y}$ in equation (2.2.2.4), which motivates us to apply the Laplace transform and obtained

$$\tilde{F}_c(s) + \tilde{g}(s)\tilde{F}_c(s) = \frac{\mu}{\lambda}\tilde{g}(s). \tag{2.2.2.6}$$

Here, when there is no confusion raised, we use $\tilde{f}(s)$ to represent the Laplace transform of the function $f(y)$. Recall by the definition of the Laplace transform, we have

$$\tilde{f}(s) \triangleq \mathcal{L}[f] = \int_0^\infty e^{-sy} \cdot f(y) \, \mathrm{d}y \quad \text{for} \quad s = \beta + \mathbf{i}\omega \quad \text{and} \quad \beta > 0,$$

$$f(y) \sim \mathcal{L}^{-1}[\tilde{f}] = \frac{1}{2\pi\mathbf{i}} \int_{\Re(s)=\beta} e^{sy} \cdot \tilde{f}(s) \, \mathrm{d}s.$$

The transform of the solution $g(y)$ can be solved easily through the equation (2.2.2.6), which is

$$\tilde{g}(s) = \frac{\lambda\tilde{F}_c(s)}{\mu - \lambda\tilde{F}_c(s)}. \tag{2.2.2.7}$$

The problem then becomes finding the inverse transform of the expression (2.2.2.7), which can be achieved using the residue theorem in complex analysis.

Consider integrating $\tilde{g}(s)e^{sy}$ along a positively oriented semicircular contour $\partial D_R = \partial D_R^1 + \partial D_R^2$, where $\partial D_R^1 = \{s : \Re(s) = \beta \ , \ |\Im(s)| \leq R\}$ and $\partial D_R^2 = \{s : |s - \beta| = R \ , \ \Re(s) \leq \beta\}$. We have

$$\oint_{\partial D_R} \tilde{g}(s) \cdot e^{sy} \, \mathrm{d}s = \int_{\partial D_R^1} \tilde{g}(s) \cdot e^{sy} \, \mathrm{d}s + \int_{\partial D_R^2} \tilde{g}(s) \cdot e^{sy} \, \mathrm{d}s. \tag{2.2.2.8}$$

If $\tilde{g}(s)$ vanishes as $s \to \infty$, by taking the limit $R \to +\infty$, the above equation (2.2.2.8) then turns into

$$\lim_{R \to +\infty} \int_{\partial D_R} \tilde{g}(s) \cdot e^{sy} \, \mathrm{d}s = \int_{\Re(s)=\beta} \tilde{g}(s) \cdot e^{sy} \, \mathrm{d}s, \tag{2.2.2.9}$$

as $\lim_{R \to +\infty} \int_{\partial D_R^2} \tilde{g}(s) \cdot e^{sy} \, \mathrm{d}s = 0$ (see Appendix A.2). By the residue theorem, if $\tilde{g}(s)$ has singularities s_k located within the half plane $\{s : \Re(s) < \beta\} \subset \mathbb{C}$, the inverse transform can then be calculated as

$$\begin{aligned} g(y) &\sim \frac{1}{2\pi \mathbf{i}} \int_{\Re(s)=\beta} \tilde{g}(s) \cdot e^{sy} \, \mathrm{d}s = \frac{1}{2\pi \mathbf{i}} \lim_{R \to +\infty} \int_{\partial D_R} \tilde{g}(s) \cdot e^{sy} \, \mathrm{d}s \\ &= \textstyle\sum_{\Re(s_k) < \beta} \operatorname{Res}\left[\tilde{g}(s)e^{sy}; s_k\right] . \end{aligned} \tag{2.2.2.10}$$

In our case, for $X \sim \mathrm{EXP}(\mathbf{a}, \mathbf{p})$ such that $F_c(t) = \sum_{i=1}^n \mathbf{p}_i e^{-\mathbf{a}_i t}$,

$$\begin{aligned} \tilde{g}(s) &= \frac{\lambda}{\mu - \lambda \sum_{i=1}^n \frac{\mathbf{p}_i}{\mathbf{a}_i + s}} \cdot \sum_{i=1}^n \frac{\mathbf{p}_i}{\mathbf{a}_i + s} \\ &= \frac{\lambda \prod_{j=1}^n (\mathbf{a}_j + s) \cdot \sum_{i=1}^n \frac{\mathbf{p}_i}{\mathbf{a}_i + s}}{\mu \prod_{j=1}^n (\mathbf{a}_j + s) - \lambda \sum_{i=1}^n \mathbf{p}_i \prod_{j \neq i}^n (\mathbf{a}_j + s)} \\ &= \frac{\lambda \sum_{i=1}^n \mathbf{p}_i \prod_{j \neq i}^n (\mathbf{a}_j + s)}{\mu \prod_{j=1}^n (\mathbf{a}_j + s) - \lambda \sum_{i=1}^n \mathbf{p}_i \prod_{j \neq i}^n (\mathbf{a}_j + s)} \ . \end{aligned}$$

Clearly, $\tilde{g}(s) \to 0$ as $s \to \infty$, so formula (2.2.2.10) is applicable. The singularities are just roots of the n_{th} order polynomial

$$P_n(s) = \mu \prod_{j=1}^n (\mathbf{a}_j + s) - \lambda \sum_{i=1}^n \mathbf{p}_i \prod_{j \neq i}^n (\mathbf{a}_j + s) = \mu \prod_{k=1}^n (s - s_k) \tag{2.2.2.11}$$

in the denominator. It can also be shown that $P_n(s) = \mu \prod_{k=1}^n (s - s_k)$ has exactly n real roots less than β, which can be found using regular rooting finding algorithms. Finally, since $\{s_k\}_{k=1}^n$ corresponds to n simple poles, the residues can be evaluated as

$$\operatorname{Res}\left[\tilde{g}(s)e^{sy}; s_k\right] = \frac{\lambda e^{s_k y} \prod_{j=1}^n (s_k + \mathbf{a}_j)}{\mu \prod_{j \neq k}^n (s_k - s_j)} \cdot \sum_{i=1}^n \frac{\mathbf{p}_i}{\mathbf{a}_i + s_k} \tag{2.2.2.12}$$

Plug the above expression (2.2.2.12) into (2.2.2.10) will give us the explicit formula for the inverse transform of $\tilde{g}(s)$

$$g(y) \sim \sum_{k=1}^{n} \left(\frac{\lambda e^{s_k y} \prod_{j=1}^{n} (s_k + \mathbf{a}_j)}{\mu \prod_{j \neq k}^{n} (s_k - s_j)} \cdot \sum_{i=1}^{n} \frac{\mathbf{p}_i}{\mathbf{a}_i + s_k} \right), \tag{2.2.2.13}$$

where $\{s_k\}_{k=1}^{n}$ are roots of the polynomial (2.2.2.11). Assuming $g(y)$ is smooth enough so that (2.2.2.13) admits the equality, the formula then gives an analytical solution for the model described by equations (2.2.2.4) and (2.2.2.5).

2.3 Quality of Service Measures

After solving the model and getting the steady-state distribution for the buffer occupancy $f_Y(y)$, we define the following five quality of service measures:

- **The Job Retransmission Proportion $\boldsymbol{\delta_{QoS}}$**, which is the average proportion of jobs requiring retransmission

$$\delta_{QoS} = \frac{1}{\mathbb{E}[N_t]} \mathbb{E}\left[\sum_{i=1}^{N_t} 1_{\{X_{T_i} + Y_{T_i} > B\}} \right] = \mathbb{P}(X + Y > B), \tag{2.3.1}$$

  where, as previously defined, N_t is the total number of jobs that arrived till time t, and T_i is the arrival time of the i_{th} job;

- **The Data Retransmission Proportion $\boldsymbol{\epsilon_{QoS}}$**, which is the average proportion of data requiring retransmission (under partial rejection)

$$\epsilon_{QoS} = \frac{\mathbb{E}\left[\sum_{i=1}^{N_t} (X_{T_i} + Y_{T_i} - B)^+ \right]}{\mathbb{E}\left[\sum_{i=1}^{N_t} X_{T_i} \right]} = \frac{\mathbb{E}\left[(X + Y - B)^+ \right]}{\mathbb{E}\left[X \right]}; \tag{2.3.2}$$

- **The Average Waiting Time $\boldsymbol{\bar{t}_{QoS}}$**, which is the average time a job has to wait before being transmitted (by a single buffer)

$$\bar{t}_{QoS} \;=\; \frac{1}{\mu} \cdot \bar{Y} \;, \tag{2.3.3}$$

  where $\bar{Y} = \int_0^B y \cdot p_1(y)\, \mathrm{d}y$ represents $\mathbb{E}[Y_t]$ under the steady-state distribution;

- **The Proportion of Average Buffer Occupancy $\boldsymbol{\gamma_{QoS}}$**, defined as

$$\gamma_{QoS} \;=\; \frac{1}{B} \cdot \bar{Y} \;; \tag{2.3.4}$$

- **The Average Buffer Occupancy $\boldsymbol{\bar{B}_{QoS}}$**, defined as the expectation of the buffer occupancy when it is steady

$$\bar{B}_{QoS} \;\stackrel{\Delta}{=}\; \bar{Y}. \tag{2.3.5}$$

It is also worth mentioning that under our simplified model of partial rejection with a single server, the expression of δ_{QoS} and ϵ_{QoS} can be further simplified. For the former, by the law of total probability and our equation (2.2.2.1),

$$\begin{aligned}\delta_{QoS} = \mathbb{P}(X+Y>B) &= \int_0^B f_Y(\tilde{y}) \cdot \mathbb{P}(X > B - Y | Y = \tilde{y}) \, \mathrm{d}\tilde{y} \\ &= \int_0^B f_Y(\tilde{y}) \cdot F_c(B-\tilde{y}) \, \mathrm{d}\tilde{y} = \frac{\mu}{\lambda} \cdot p_1(B). \end{aligned} \tag{2.3.6}$$

For the latter, first notice by our equation (2.2.2.1), we have

$$\begin{aligned}\mathbb{P}(X+Y>a) &= \int_0^a f_Y(\tilde{y}) \cdot F_c(a-\tilde{y}) \, \mathrm{d}\tilde{y} + \int_a^B p_1(\tilde{y}) \cdot F_c(a-\tilde{y}) \, \mathrm{d}\tilde{y} \\ &= \frac{\mu}{\lambda} \cdot p_1(a) + \int_a^B p_1(\tilde{y}) \cdot 1 \, \mathrm{d}\tilde{y} = \frac{\mu}{\lambda} \cdot p_1(a) + \mathbb{P}(Y>a) \end{aligned} \tag{2.3.7}$$

for all $0 \le a \le B$. Since $\delta_{QoS} = \mathbb{P}(X+Y>B)$, we have

$$\begin{aligned}&\mathbb{E}[(X+Y-B)^+] = \mathbb{E}[\, X+Y-B \mid X+Y>B \,] \cdot \delta_{QoS} + 0 \cdot (1-\delta_{QoS}) \\ &= \mathbb{E}[\, X+Y-B \,] - \mathbb{E}[\, X+Y-B \mid X+Y \le B \,] \cdot (1-\delta_{QoS}) \\ &= \bar{X} + \bar{Y} - \delta_{QoS} \cdot B - \mathbb{E}[\, X+Y \mid X+Y \le B \,] \cdot (1-\delta_{QoS}). \end{aligned} \tag{2.3.8}$$

And,

$$\begin{aligned}&\mathbb{E}[\, X+Y \mid X+Y \le B \,] \cdot (1-\delta_{QoS}) = \mathbb{E}[\, (X+Y) \cdot 1_{\{X+Y \le B\}} \,] \\ &= \int_0^\infty \mathbb{P}\{(X+Y) \cdot 1_{\{X+Y \le B\}} > a\} \, \mathrm{d}a = \int_0^B \mathbb{P}(a < X+Y \le B) \, \mathrm{d}a \\ &= \int_0^B \mathbb{P}(X+Y>a) \, \mathrm{d}a - \delta_{QoS} \cdot B \, . \end{aligned} \tag{2.3.9}$$

Substitute formula (2.3.7) for $\mathbb{P}(X+Y>a)$ into expression (2.3.9), we obtain

$$\begin{aligned}&\mathbb{E}[\, X+Y \mid X+Y \le B \,] \cdot (1-\delta_{QoS}) \\ &= \frac{\mu}{\lambda} \cdot \int_0^B p_1(a) \, \mathrm{d}a + \int_0^B \mathbb{P}(Y>a) \, \mathrm{d}a - \delta_{QoS} \cdot B \\ &= \frac{\mu}{\lambda} \cdot (1-p_0) + \bar{Y} - \delta_{QoS} \cdot B, \end{aligned} \tag{2.3.10}$$

indicating $\mathbb{E}[(X+Y-B)^+] = \bar{X} - \frac{\mu}{\lambda} \cdot (1-p_0)$, and

$$\epsilon_{QoS} = \frac{\bar{X} - \frac{\mu}{\lambda} \cdot (1-p_0)}{\bar{X}} = 1 - \frac{\mu(1-p_0)}{\lambda \bar{X}}. \tag{2.3.11}$$

2.4 Numerical Results

2.4.1 Steady-state Distribution

Fig. 2.4.1.1 below gives the analytical solution for the steady-state distribution f_Y under two different circumstances.

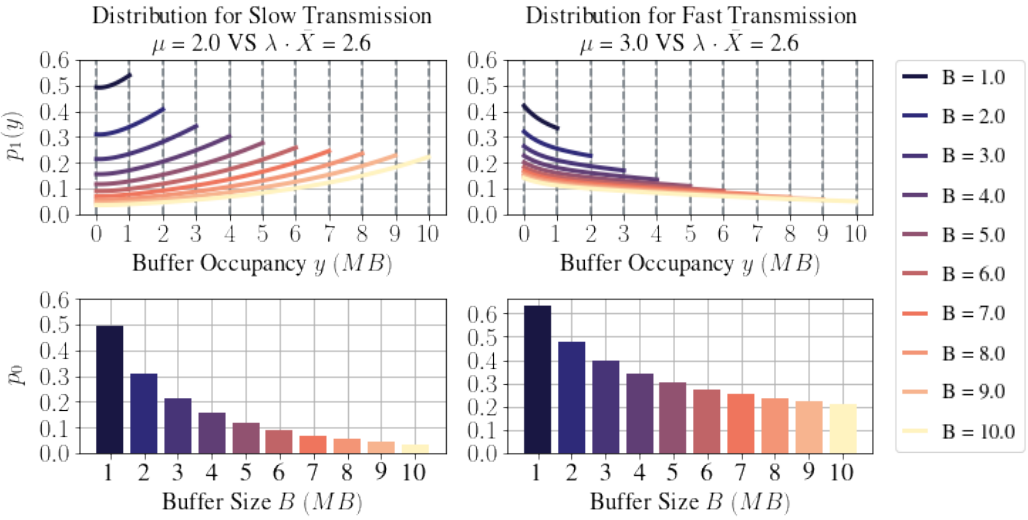

Fig. 2.4.1.1 Analytical solution for the basic model (partial rejection and single server) under slow and fast transmissions.

The first column represents a case of **slow transmission**, where the transmission rate of the server is less than the product of the job entering rate λ and the average job size $\bar{X} := \mathbb{E}(X)$, i.e., $\mu < \lambda\bar{X}$. This means the data are leaving the buffer at a slower rate than the average rate at which they are entering. Thus, it is not surprising that the buffer is more likely to be occupied: the density p_1 is monotonically increasing, and the probability for the buffer to be empty is, on average, below 0.5.

Conversely, the two plots in the second column represent a case of **fast transmission**, where $\mu > \lambda\bar{X}$. This time, the density p_1 is monotonically decreasing and the probabilities for the buffer to be empty are significantly higher than those in the former case.

2.4.2 Quality of Service and Comparison with Simulations

The analytical quality of service measures are calculated with respect to different buffer sizes B and transmission rates μ under a fixed job size distribution and are displayed in Fig. 2.4.2.1. For the retransmission measures δ_{QoS} and ϵ_{QoS}, both values can be decreased by increasing either the buffer size or the transmission rate. The behaviours of the occupancy measures γ_{QoS} and $\bar{B}_{QoS}$ are also reasonable: first, the values tend to decay as the transmission rate μ increases, meaning jobs are less likely

to get stuck in the buffer; also, larger buffers usually have higher average occupancies as they can take in more jobs compared to those with smaller sizes.

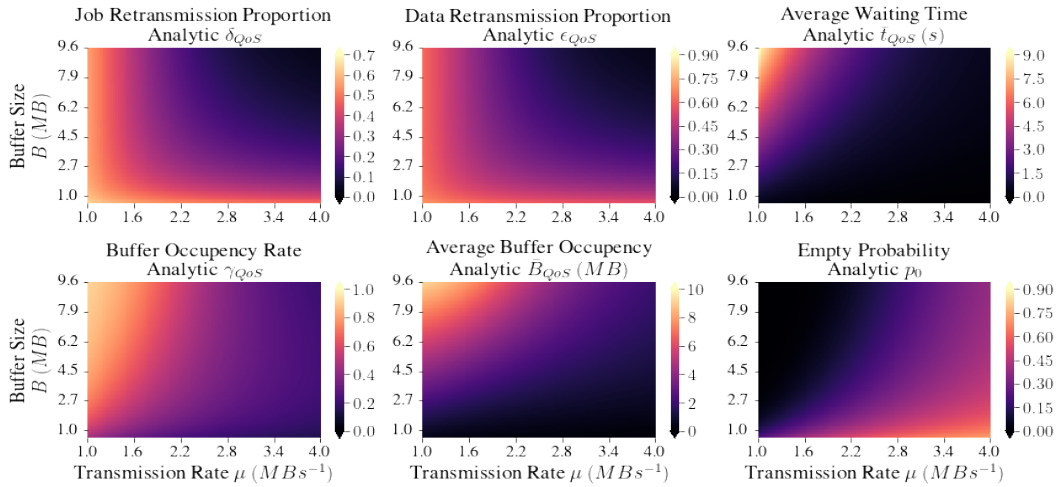

Fig. 2.4.2.1 Analytical solution for the quality of service measures under the basic model.

However, increasing the buffer size may also lead to longer waiting times $\bar{t}_{QoS}$ as a higher average occupancy naturally implies having more jobs in the queue. Therefore, in a system where there is only a single server and the jobs are partially rejected, it is necessary to increase the transmission rate in order to achieve a shorter waiting time.

We also compared our analytical results with the simulations and calculated the relative error, which is generally around 10^{-2} (Fig. 2.4.2.2), indicating a relatively good fit of our model for real-world predictions.

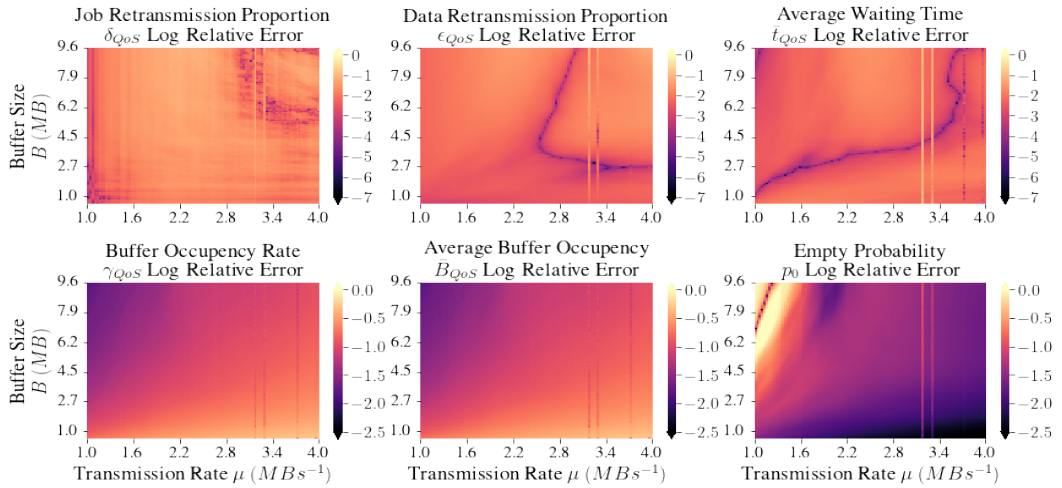

Fig. 2.4.2.2 $\log_{10}$ Relative error of the analytical quality of service measures with respect to the simulation result.

3 Extensions

3.1 A Model for Full Rejection

3.1.1 The Balance Equation under Full Rejection

Now change the model to consider the case of full rejections. Recall that this means the buffer will only accept a job when the space left is enough to fit the job fully. To model this case with a new balance equation, we only need to change the derivation of step (2.2.1.1) slightly. Again, consider a small time interval $I = [t, t + \Delta t]$, and let $\mathrm{P}(t, y) = \mathbb{P}(Y_t \geq y)$ for some $y \in [0, B]$. When only considering the probability change caused by the jobs coming in, we have

$$\mathrm{P}(t+\Delta t, y) = q_0 \cdot \mathrm{P}(t, y) + q_1 \cdot \{\mathrm{P}(t, y) + \mathbb{P}(y - X \leq Y_t < y, X + Y_t \leq B)\} + \mathcal{O}(\Delta t^2),$$

namely,

$$\frac{\mathrm{P}(t+\Delta t, y) - \mathrm{P}(t, y)}{\Delta t} = \lambda \cdot \mathbb{P}(y - X \leq Y_t < y, X + Y_t \leq B) + \mathcal{O}(\Delta t),$$

where $q_0 = 1 - \lambda\Delta t$ (resp., $q_1 = \lambda\Delta t$) is the probability of having no (resp., one) job coming in. Since with the law of total probability we have

$$\begin{aligned}\mathbb{P}(y - X \leq Y_t < y, X + Y_t \leq B) &= \int_0^B \mathbb{P}(y - X \leq \tilde{y} < y, X + \tilde{y} \leq B) \cdot f_Y(\tilde{y})\ \mathrm{d}\tilde{y} \\ = \int_0^y \mathbb{P}(y - \tilde{y} \leq X \leq B - \tilde{y}) \cdot f_Y(\tilde{y})\ \mathrm{d}\tilde{y} &= \int_0^y \{F_c(y - \tilde{y}) - F_c(B - \tilde{y})\} \cdot f_Y(\tilde{y})\ \mathrm{d}\tilde{y},\end{aligned}$$

by letting $\Delta t \to 0$, the probability change caused by jobs coming in is

$$\left[\frac{\partial \mathrm{P}}{\partial t}\right]^+ = \lambda \cdot \int_0^y \{F_c(y - \tilde{y}) - F_c(B - \tilde{y})\} \cdot f_Y(\tilde{y})\ \mathrm{d}\tilde{y}. \tag{3.1.1.1}$$

Following a similar argument in Section 2.2.1, the final balance equation is given as

$$\int_0^y \{F_c(y - \tilde{y}) - F_c(B - \tilde{y})\} \cdot f_Y(\tilde{y})\ \mathrm{d}\tilde{y} = \frac{\mu}{\lambda} \cdot p_1(y), \tag{3.1.1.2}$$

where F_c is the complementary cumulative distribution function of the job size.

3.1.2 Solutions via the Fixed Point Iteration

Expanding the equation 3.1.1.2 and dividing both sides with p_0, we get

$$g(y) = \frac{\lambda}{\mu}\left(F_c(y) - F_c(B) + \int_0^y g(\tilde{y})\left[F_c(y - \tilde{y}) - F_c(B - \tilde{y})\right]\mathrm{d}\tilde{y}\right), \tag{3.1.2.1}$$

where $g(y) = p_1(y)/p_0$. This time, however, the method of Laplace transform is no longer applicable due to the existence of the term $\int_0^y g(\tilde{y}) \cdot F_c(B-\tilde{y})\mathrm{d}\tilde{y}$. Nevertheless, we can still find a reasonable solution via fixed point iterations according to the contraction mapping theorem. Let

$$\mathcal{F}[g] = \frac{\lambda}{\mu}\left(F_c(y) - F_c(B) + \int_0^y g(\tilde{y}) \left[F_c(y-\tilde{y}) - F_c(B-\tilde{y})\right] \mathrm{d}\tilde{y} \right), \qquad (3.1.2.2)$$

be a map defined over the space $C[0, B]$ equipped with the norm $||\cdot||_\infty$ such that $||\phi||_\infty = \max_{y\in[0,B]} |\phi(y)|$. When the condition of $\lambda < \mu/B \ll \mu$ $(*)$ is strictly satisfied, we can show $\mathcal{F}[\cdot]$ is a contraction mapping (see Appendix A.3) and thus the iteration

$$g^{(k+1)}(y) := \mathcal{F}[g^{(k)}]$$

converges globally to a unique fixed point $g^*(y) = \mathcal{F}[g^*]$, which is exactly the solution of our balance equation (3.1.2.1).

The condition $(*)$ means that the method is only well-performed when the transmission rate is sufficiently fast. Indeed, as the numerical result displayed below in Fig. 3.1.2.1, the method provides quite robust solutions for cases involving fast transmissions, even when the condition $(*)$ is slightly violated, while in the case of slow transmissions, the results tend to be far less stable.

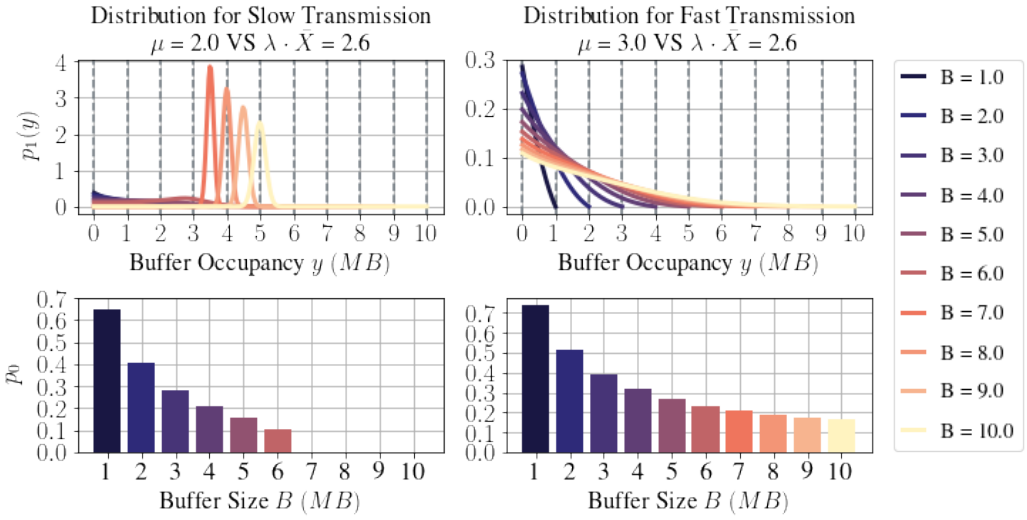

Fig. 3.1.2.1 Analytical solution for the model of full rejection obtained via fixed point iteration.

Simulations are also applied to investigate the influence of full rejection on the changes in the quality of service (Fig. 3.1.2.2). In general, by doing full rejection, the job

retransmission marked by δ_{QoS} is generally mitigated, while the data retransmission proportion ϵ_{QoS} deteriorates. The buffer tends to be less occupied according to changes in γ_{QoS} and $\bar{B}_{QoS}$, leading to a satisfactory decline in the average waiting time $\bar{t}_{QoS}$.

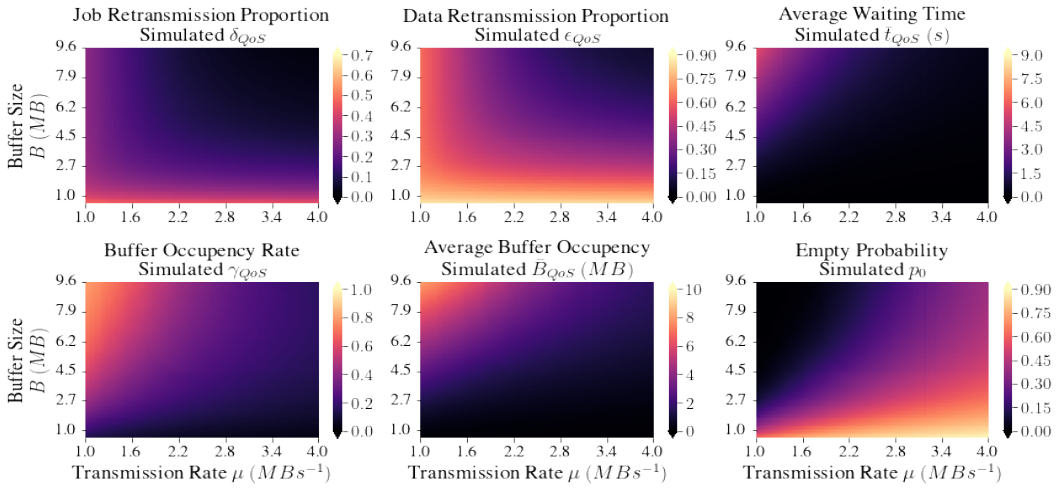

Fig. 3.1.2.2 Simulation for the quality of service measures with full rejection.

3.2 A Model for Multiple Servers

3.2.1 Solvability Discussion

Here, we consider a system with more than one server and jobs are partially rejected. Assume the simplest case where all servers share the same transmission rate μ. Let S_t be the number of servers occupied at time t; within a short time interval $[t, t+\Delta t]$, the change in the probability flow caused by the data transmission now depends on the state of the buffer

$$\mathbb{P}(Y_{t+\Delta t} \geq y) = \sum_{k=0}^{m} \mathbb{P}(S_t = k \mid Y_t \geq y + k\mu\Delta t) \cdot \mathbb{P}(Y_t \geq y + k\mu\Delta t),$$

which makes it extremely hard to derive an explicit formulation as we did before in Sections 2.2.1 and 3.1.1. Therefore, pursuing an analytical solution for such a problem is theoretically unrealistic. One motivation to address the problem is to learn a so-called "effective transmission rate" via regression, which can be plugged into the basic model in Section 2 and produce "close" approximates for the corresponding quality of service measures. However, this method seems to be less reliable as it may sometimes generate contradictory predictions, especially for $\bar{t}_{QoS}$ according to our experiments. A more advanced and accurate alternative to the above method is to learn the quality of service measures directly via deep neural networks.

3.2.2 Neural Network Approximation

In this section, we trained four grid-based neural networks to predict the value of quality of service measures δ_{QoS}, ϵ_{QoS}, $\bar{t}_{QoS}$ and γ_{QoS}. The input features of each network $\mathbf{x}$ is a vector of length 12,

$$\mathbf{x} = \begin{bmatrix} \mathbf{a} \\ \mathbf{p} \\ \lambda \\ m \end{bmatrix}, \text{ where } \mathbf{a}, \mathbf{p} \in \mathbb{R}^5, \text{ and } \lambda, m \in \mathbb{R},$$

for the distribution of mixture exponential $\text{EXP}(\mathbf{a}, \mathbf{p})$, job entering rate λ and m servers with a same transmission rate. The networks allow for fewer than 5 components in the mixture exponential distribution, as long as the parameters at fixed positions are set to 0. Before entering into the network, the i_{th} parameter $\mathbf{x}_i \in \{\mathbf{a}_j, \mathbf{p}_k, \lambda, m\}$ in the input $\mathbf{x}$ will be scaled to $[-1, 1]$ with the following scaling law

$$\text{scale}(\mathbf{x})_i = \frac{2\mathbf{x}_i - [ub(\mathbf{x}_i) + lb(\mathbf{x}_i)]}{ub(\mathbf{x}_i) - lb(\mathbf{x}_i)} \in [-1, 1] \tag{3.2.2.1}$$

using their corresponding lower and upper bounds $lb(\mathbf{x}_i)$ and $ub(\mathbf{x}_i)$ specified in Table 3.2.2.1

$\mathbf{x}_i$	$\mathbf{a}_j$	$\mathbf{p}_k$	λ	m
lower	0.20	0.00	0.10	1
upper	10.0	1.00	5.00	10

Table 3.2.2.1 Lower and upper bounds for network input parameters.

The output of the networks $\mathbf{y}$ are values of corresponding quality of service measures over a 5×10 μ-B grid stored in a flattened vector of length 50.

In each hidden layer, we use the ELU activation function. The output of networks predicting dimensionless measures δ_{QoS}, ϵ_{QoS} and γ_{QoS} are activated with the Sigmoid activation, while the output of the $\bar{t}_{QoS}$ network is activated with the ReLU activation. The training features are sampled from uniform distributions and the learning labels are generated through simulations. Once the approximating function for the quality of service is learned, we can evaluate the value of the measure at any point within the range of the grid through proper spline interpolations.

3.2.3 Numerical Results

We trained each net with a batch size of 128 for 5000 epochs and recorded their changes in the average square loss at each grid point with respect to the epoch number as displayed in Fig. 3.2.3.1 below.

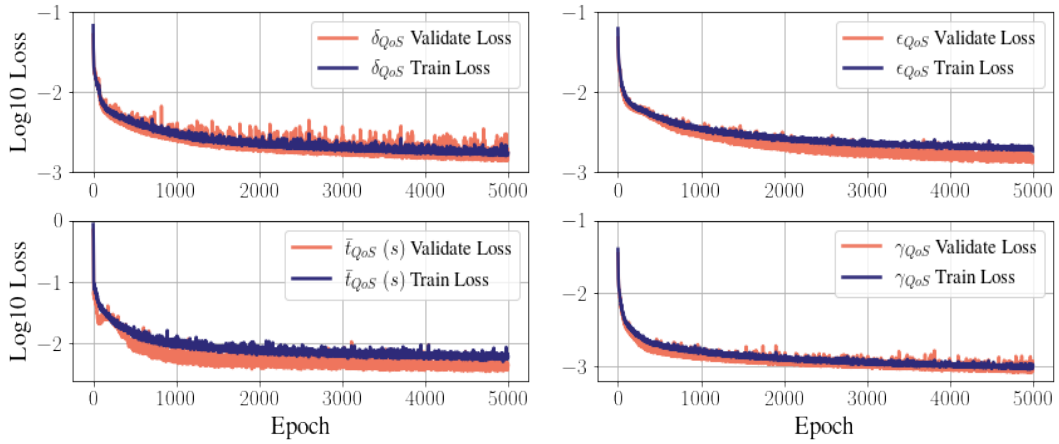

Fig. 3.2.3.1 Change in the log-mean-square-loss at each grid point during training, calculated as $\mathbf{log}_{10}\frac{\mathbb{E}|\tilde{\mathbf{y}}-\mathbf{y}|^2}{50}$, where $\mathbf{y}$ denotes the network's output and $\tilde{\mathbf{y}}$ stands for the label.

The calculated (absolute) error statistics of each network's outputs over the discretization grid are present in Fig 3.2.3.2, which shows that most of the average relative errors are around 10^{-2} across all possible parameter combinations.

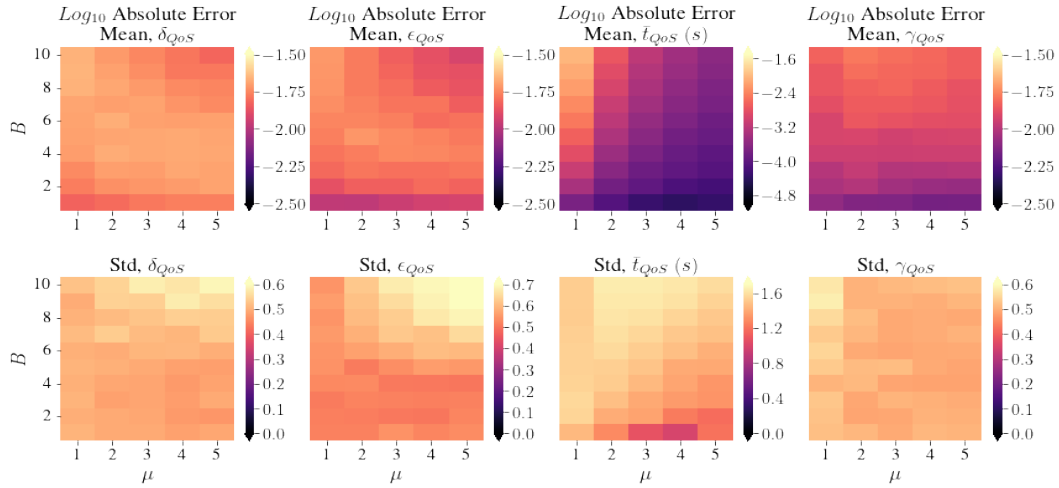

Fig. 3.2.3.2 Mean and standard deviation of the $\log_{10}$ absolute error of network outputs compared to the label over the grid.

Finally, we provide the predictions generated by the networks for each quality of service measure in Fig 3.2.3.3. We can see that the network successfully replicates the pattern displayed in Fig 2.4.2.1 for the singer server when $m = 1$. By increasing the number of servers, all measures are successfully decreased, especially for the average waiting time.

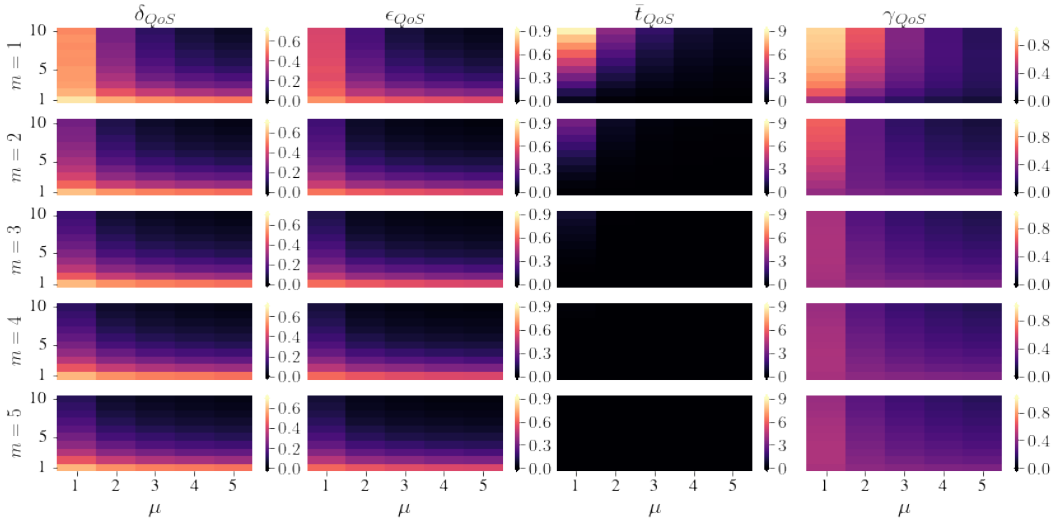

Fig. 3.2.3.3 Grid-based predictions of neural networks for each quality of service measure. The y-axis represents the buffer size; m is the server number.

4 Conclusion

This paper mainly proposed three models to study the quality of service of a specific buffer system. We start from the simplest case involving only one server and partial rejection and provide a model together with its analytical solution which can be used to calculate the five quality of service measures. We then move on to study the influence when jobs entering are fully rejected by the buffer. In this case, an analytical model can be proposed with a robust solution found using fixed point iterations for fast transmissions. When multiple servers get involved, the formulation of an analytical model is theoretically infeasible. Nevertheless, we can still get predictions for the quality of service measures precisely enough via grid-based neural networks.

In control of the quality of service, our results show that retransmission can be alleviated by either enlarging the buffer size or strengthening the transmission capability, which includes increasing both the transmission rate and the number of servers. For fixed buffer sizes, the average waiting time and the average buffer occupancy can be decreased by applying full rejection or strengthening the transmission capability.

Finally, our study did not include the influence of job parameters on service quality, especially the influence of various job size distributions. Note that both analytical models proposed in Sections 2.2.1 and 3.1.1 are applicable to different job size distributions, although changing the distribution may add additional challenges in solving the model. One potential direction for future efforts could be the development of more general or targeted methods for solving the model with various job size distributions.

5 References

[1] A. Feldmann and W. Whitt. "Fitting mixtures of exponentials to long-tail distributions to analyze network performance models". In: *Performance Evaluation* 31.3 (1998), pp. 245–279. ISSN: 0166-5316. DOI: https://doi.org/10.1016/S0166-5316(97)00003-5. URL: https://www.sciencedirect.com/science/article/pii/S0166531697000035.

A Appendix

A.1 Codes

All codes can be found via: https://github.com/abaaba337/MMSC-Modelling-Case-Study-Buffer.

A.2 Proof of Convergence in Formula 2.2.2.9

Let $s_\theta = \beta + Re^{\mathbf{i}\theta} \in \partial D_R^2$ for $\theta \in [\frac{\pi}{2}, \frac{3\pi}{2}]$ and $M_R = \sup_{s\in\partial D_R^2} |\tilde{g}(s)|$:

$$\begin{aligned}
\left| \int_{\partial D_R^2} \tilde{g}(s)\cdot e^{sy}\ \mathrm{d}s \right| &= \left| \int_{\pi/2}^{3\pi/2} \tilde{g}(s_\theta)\cdot e^{y(\beta+R\cos\theta)+\mathbf{i}Ry\sin\theta} Re^{\mathbf{i}\theta}\mathbf{i}\ \mathrm{d}\theta \right| \\
&\leq \int_{\pi/2}^{3\pi/2} \left| \tilde{g}(s_\theta)\cdot e^{y(\beta+R\cos\theta)+\mathbf{i}Ry\sin\theta} Re^{\mathbf{i}\theta}\mathbf{i} \right| \mathrm{d}\theta \\
&= R\int_{\pi/2}^{3\pi/2} |\ \tilde{g}(s_\theta)\ |\cdot e^{y(\beta+R\cos\theta)}\ \mathrm{d}\theta \\
&\leq M_R e^{y\beta} R\int_{\pi/2}^{3\pi/2} e^{yR\cos\theta}\ \mathrm{d}\theta = M_R e^{y\beta} R\int_0^{\pi} e^{-yR\sin\theta}\ \mathrm{d}\theta \\
&= 2M_R e^{y\beta} R\int_0^{\pi/2} e^{-yR\sin\theta}\ \mathrm{d}\theta \\
&\leq 2M_R e^{y\beta} R\int_0^{\pi/2} e^{-yR\frac{2\theta}{\pi}}\ \mathrm{d}\theta = \frac{\pi M_R e^{y\beta}}{y}(1-e^{-yR}) \\
&\leq \frac{\pi M_R e^{y\beta}}{y} \to 0 \quad \text{as} \quad R\to+\infty \quad \text{if} \quad M_R \to 0.
\end{aligned}$$

Therefore the integral $\int_{\partial D_R^2} \tilde{g}(s)\cdot e^{sy}\ \mathrm{d}s$ vanishes for $R\to+\infty$ as long as $\tilde{g}(s)$ converges to 0 for $s\to\infty$.

A.3 Proof of Contraction Mapping

We show here that the mapping (3.1.2.2) is a contraction if $\lambda < \mu/B$:

$$\begin{aligned}
\| \ \mathcal{F}[g] - \mathcal{F}[h] \ \|_\infty &= \frac{\lambda}{\mu} \cdot \left\| \int_0^y [g(\tilde{y}) - h(\tilde{y})] \cdot [F_c(y-\tilde{y}) - F_c(B-\tilde{y})] \, \mathrm{d}\tilde{y} \right\|_\infty \\
&\leq \frac{\lambda}{\mu} \cdot \max_{y \in [0,B]} \int_0^y |g(\tilde{y}) - h(\tilde{y})| \cdot |F_c(y-\tilde{y}) - F_c(B-\tilde{y})| \, \mathrm{d}\tilde{y} \\
&\leq \frac{\lambda}{\mu} \cdot \max_{y \in [0,B]} \int_0^y |g(\tilde{y}) - h(\tilde{y})| \cdot 1 \, \mathrm{d}\tilde{y} \\
&\leq \frac{\lambda}{\mu} \cdot \max_{y \in [0,B]} \int_0^y \|g - h\|_\infty \, \mathrm{d}\tilde{y} = \frac{\lambda B}{\mu} \cdot \|g - h\|_\infty .
\end{aligned}$$

Clearly, $\frac{\lambda B}{\mu} < 1$ if and only if $\lambda < \mu/B$, and when the condition is satisfied, $\mathcal{F}[\cdot]$ is a contraction with respect to the space $C[0, B]$ under the norm $|| \cdot ||_\infty$.