CA' FOSCARI UNIVERSITY - VENICE

Department of Enviroment sciences, Informatics and Statistics

[CM0481] SOFTWARE PERFORMANCE AND SCALABILITY (CM90)

Assignment 1: Queuing Systems

Student:
Michele Lotto 875922

a.y. 2023-24

Contents

Chapter 1

Assignment task

We are tasked with assessing the performance of three distinct queuing system models based on their expected response time. Our objective is to draw conclusions regarding the results obtained; in particular we are asked to compare these systems under varying load conditions.

1. **Single processor** : a single processor with speed 2μ, Poisson arrival rate λ and exponentially distributed service time.

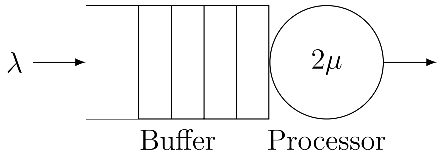

Figure 1.1: Queuing System 1

2. **Two processors with random dispatching** : two processors with speed μ, Poisson arrival rate λ and exponentially distributed service time. Each processor is equipped with a different buffer. Each incoming job is randomly dispatched (with probability 50%) to one of the two buffers.

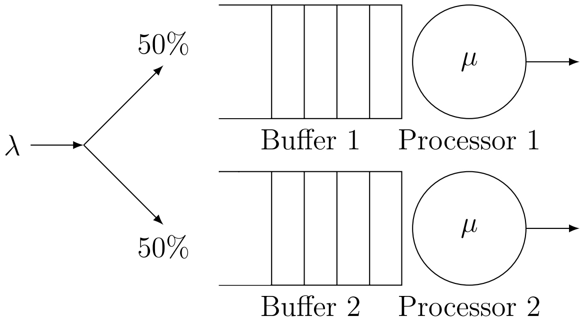

Figure 1.2: Queuing System 2

3. **Two processor with shared buffer** : two processors with speed μ, Poisson arrival rate λ, exponentially distributed service time and shared buffer.

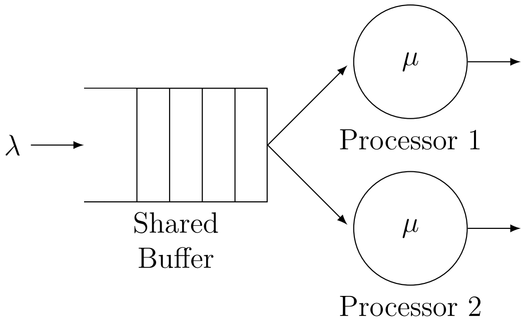

Figure 1.3: Queuing System 3

Chapter 2

Queuing Systems Analysis

In this chapter, I will examine each provided queuing system structure and compute the average response times for later comparison.

2.1 Single processor

The "single processor" queuing system (1.1) has a Poisson Arrival rate λ and exponentially distributed service times $(2\mu)^{-1}$. Thus it can be modelled as a M/M/1 queuing system. The average response time is calculated as:

$$\overline{R}_1 = \frac{1}{2\mu - \lambda}$$

Recalling that one of the assumptions for the M/M/1 system is that the system is stable, for this system we have the constraint $\lambda < 2\mu$.

In the following sections I will refer to this system as "system 1".

2.2 Two processors with random dispatching

The "two processors with random dispatching" queuing system (1.2) has a Poisson Arrival rate λ and each of its processors has exponentially distributed service times $(\mu)^{-1}$. Using the thinning (or splitting) property of Poisson processes, this system can be modelled by two M/M/1 queuing systems with arrival rate $\lambda/2$ and service rate μ. The average response time for each processor is calculated as:

$$\overline{R}_{P_1} = \overline{R}_{P_2} = \frac{1}{\mu - \frac{\lambda}{2}}$$

The overall average response time is thus the average of $\overline{R}_{P_1}$ and $\overline{R}_{P_2}$. Since they are equal:

$$\overline{R}_2 = \frac{\overline{R}_{P_1} + \overline{R}_{P_2}}{2} = \frac{1}{\mu - \frac{\lambda}{2}}$$

Another way to calculate the overall average response time is to calculate the overall average number of jobs in the system ($\overline{N}_2$) and then applying Little's theorem. First calculate the average number of jobs in each M/M/1 queuing system:

$$\overline{N}_{P_1} = \overline{N}_{P_2} = \frac{\rho}{1-\rho}$$

where $\rho = \frac{\lambda}{2\mu}$ (utilization factor) since we have arrival rate $\lambda/2$ for each M/M/1 system. Then calculate the overall average number of jobs in the system:

$$\begin{aligned}\overline{N}_2 &= \overline{N}_{P_1} + \overline{N}_{P_2} \\ &= \frac{\frac{\lambda}{2\mu}}{1-\frac{\lambda}{2\mu}} + \frac{\frac{\lambda}{2\mu}}{1-\frac{\lambda}{2\mu}} \\ &= \frac{\lambda}{\mu - \frac{\lambda}{2}}\end{aligned}$$

Finally using Little's theorem:

$$\begin{aligned}\overline{R}_2 &= \frac{\overline{N}_2}{\lambda} \\ &= \frac{\frac{\lambda}{\mu-\frac{\lambda}{2}}}{\lambda} \\ &= \frac{1}{\mu - \frac{\lambda}{2}}\end{aligned}$$

Since this system is modelled using two M/M/1 systems we have the constraint $\lambda < 2\mu$. In the following sections I will refer to this system as "system 2".

2.3 Two processors with shared buffer

The "two processors with shared buffer" queuing system (1.3) has a Poisson Arrival rate λ and each of its processors has exponentially distributed service times $(\mu)^{-1}$. Thus this system can be modelled with a M/M/2 queuing system. The average response time is calculated as:

$$\overline{R}_3 = \frac{C(2, \frac{\lambda}{\mu})}{2\mu - \lambda} + \frac{1}{\mu}$$

where $C(2, \frac{\lambda}{\mu})$ is the Erlang-C formula which is the probability of fining all servers busy:

$$\begin{aligned}C(2, \frac{\lambda}{\mu}) &= \frac{1}{1 + (1-\rho)(\frac{2}{(2\rho)^2})\sum_{k=0}^{2-1}\frac{(2\rho)^k}{k!}} \\ &= \frac{2\rho^2}{\rho+1}\end{aligned}$$

where $\rho = \frac{\lambda}{2\mu}$ (utilization factor of M/M/2 system).

After some calculations we obtain the average response time:

$$\overline{R}_3 = \frac{2\mu}{2\mu^2 - \lambda^2/2}$$

Since this system is modelled using two M/M/2 systems we have the constraint $\lambda < 2\mu$. In the following sections I will refer to this system as "system 3".

Chapter 3

Comparing the systems

In this chapter I will compare the average response times previously calculated with inequalities, then I will compare the response times using the notion of low, medium and high loads.

3.1 Comparing the systems with inequalities

3.1.1 Comparing systems 1 and 2

$$\begin{aligned}
&\overline{R}_1 \leq \overline{R}_2 \\
\implies &\frac{1}{2\mu - \lambda} \leq \frac{1}{\mu - \frac{\lambda}{2}} \\
\implies &\frac{1}{\mu - \frac{\lambda}{2}} \leq \frac{2}{\mu - \frac{\lambda}{2}} \\
\implies &\frac{-1}{\mu - \frac{\lambda}{2}} \leq 0
\end{aligned}$$

which is true if and only if:

$$\begin{aligned}
&\mu - \frac{\lambda}{2} > 0 \\
\implies &\lambda < 2\mu
\end{aligned}$$

Notice that the fraction cannot be equal to zero, thus the response times are always different; furthermore since $\lambda < 2\mu$ for both the systems by assumption, I conclude the comparison saying that queuing system 1 is always faster than queuing system 2 ($\overline{R}_1 < \overline{R}_2$).

3.1.2 Comparing systems 1 and 3

$$\begin{aligned} &\overline{R}_3 \geq \overline{R}_1 \\ \implies &\frac{2\mu}{2\mu^2 - \lambda^2/2} \geq \frac{1}{2\mu - \lambda} \end{aligned}$$

after a brief computation we obtain:

$$\frac{(\lambda - 2\mu)^2}{(2\mu - \lambda)^2(2\mu - \lambda)} \geq 0$$

which is true if and only if:

$$\begin{aligned} &2\mu - \lambda > 0 \\ \implies &\lambda < 2\mu \end{aligned}$$

Since $\lambda < 2\mu$ for both the systems by assumption:

- the fraction cannot be equal to zero, thus the response times are always different.
- queuing system 1 is always faster than queuing system 3 ($\overline{R}_1 < \overline{R}_3$).

3.1.3 Comparing systems 2 and 3

$$\begin{aligned} &\overline{R}_3 \leq \overline{R}_2 \\ \implies &\frac{2\mu}{2\mu^2 - \lambda^2/2} \leq \frac{1}{\mu - \frac{\lambda}{2}} \end{aligned}$$

after a brief computation we obtain:

$$\frac{\frac{\lambda^2}{2} - \mu\lambda}{2(\mu - \frac{\lambda}{2})^2(\mu + \frac{\lambda}{2})} \leq 0$$

The denominator is always positive:

- $\mu + \frac{\lambda}{2} < 0 \implies \lambda \leq -2\mu$ which is never true since $\lambda > 0$.
- $(\mu - \frac{\lambda}{2})^2 < 0$ which is never true since it is a square.
- $2 < 0$ which is never true.

The numerator instead is always negative and never zero, since $\lambda < 2\mu$ by assumption:

$$\frac{\lambda}{2} - \mu\lambda \leq 0 \implies \lambda < 2\mu$$

Thus, the initial fraction is always negative which implies that:

- the response times are always different.
- queuing system 3 is always faster than queuing system 2 ($\overline{R}_3 < \overline{R}_2$).

3.1.4 Comparing all the systems

Given the results obtained in the previous sections, I conclude this initial comparison by saying that $\overline{R}_1 < \overline{R}_3 < \overline{R}_2$. Thus, the faster system is queuing system 1 and the slower is queuing system 2.

3.2 Comparing the systems with load

In Figure 3.1 is plotted the average response time as a function of lambda for each system with $\mu = 1$ for simplicity.

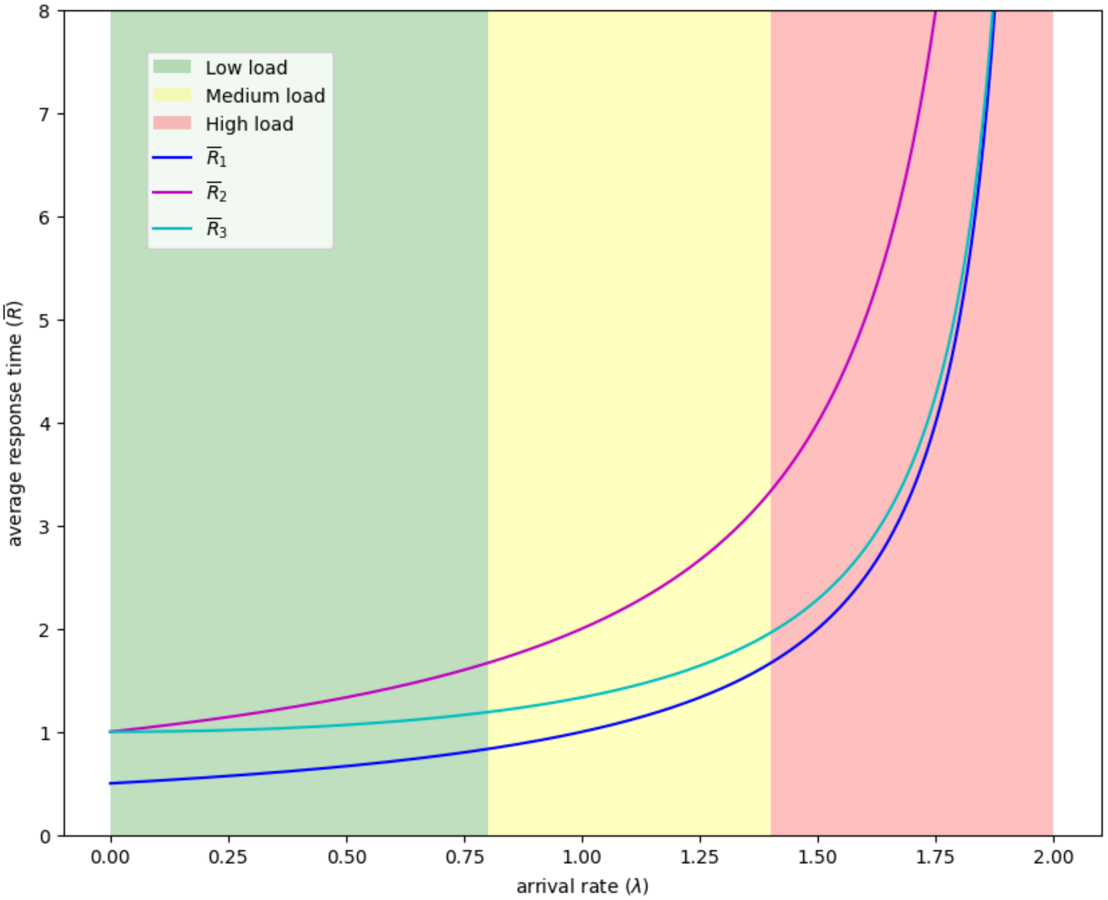

Figure 3.1: Average Response time as a function of lambda for each system with $\mu = 1$.

3.2.1 Low load

The graph tells us that in low load system 2 and 3 are quite similar in terms of average response time: when λ approaches zero both the average response times tend to be 1;

furthermore system 1 average response time corresponds exactly to $\frac{1}{2}$ when λ approaches zero, as system 1 is two times faster in terms of service time compared to the other two systems.

3.2.2 Medium load

The graph tells us that in medium load, system 1 and 3 are more similar in terms of average response time compared to the average response time of system 2. In this area, system 2 average response time starts diverging greatly from the other two system response times.

3.2.3 High load

The graph tells us that in high load, system 1 and 3 are quite similar in terms of average response time: when λ approaches 2μ both the average response times grow to infinity with the same rate; furthermore notice how system 2 grows to infinity much faster when λ approaches 2μ compared to the other two systems.

Chapter 4

Conclusions

The analysis reveals that the expected response time of the first system consistently outperforms that of the second and third systems. This discrepancy arises from the first model's speed, which is two times the other system speeds.