1 Review of M/M/k from Chpt 14

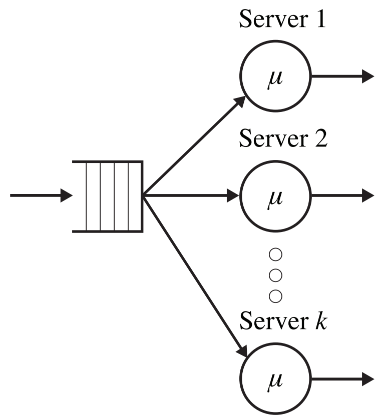

Figure 1: $M/M/k$

Recall:

$$\pi_i = \begin{cases} \frac{(k\rho)^i}{i!} \cdot \pi_0 & \text{if } i \le k \\ \frac{\rho^i}{k!} \cdot k^k \cdot \pi_0 & \text{if } i > k \end{cases}$$

where

$$\pi_0 = \left[\sum_{i=0}^{k-1} \frac{(k\rho)^i}{i!} + \frac{(k\rho)^k}{k!(1-\rho)} \right]^{-1}$$

Recall also that P_Q represents the probability that an arrival queues, where

$$\begin{aligned} P_Q &= \sum_{i=k}^{\infty} \pi_i \\ &= \frac{(\rho k)^k}{k!} \cdot \frac{1}{1-\rho} \cdot \pi_0 \end{aligned}$$

Question: What is ρ? Why does this represent the utilization of a single server?

Question: What is R and what are two things that R represents?

Question: Express $\mathbf{E}[N_Q]$ as a simple function of P_Q, then derive $\mathbf{E}[T_Q]$.

2 Warmup: Mean Delay for Delayed Customers

Note that the formula for $\mathbf{E}[T_Q]$ is quite complicated. As a warmup, let's instead look at a simpler quantity:

$$\mathbf{E}[T_Q \mid \text{delayed}] .$$

Question: What is $\mathbf{E}[T_Q \mid \text{delayed}]$?

Question: What happens to λ as k goes up, given fixed ρ and $\mathbf{E}[S]$?

Question: What happens to $\mathbf{E}[T_Q \mid \text{delayed}]$?

FIX $\rho = 0.99$, $\mathbf{E}[S] = 1$.

Case 1: $k = 1$: $\mathbf{E}[T_Q \mid \text{delayed}] =$ ____________________

Case 2: $k = 10$: $\mathbf{E}[T_Q \mid \text{delayed}] =$ ____________________

Case 3: $k = 100$: $\mathbf{E}[T_Q \mid \text{delayed}] =$ ____________________

LESSON 1: More servers at the same fixed load lead to __________________________.

3 Asymptotic Regimes

Main Purpose of Asymptotic Regimes Analysis:

- Understand queueing behavior in ________________________________
- Allow us to analyze quantities with messy expressions, such as $\mathbf{E}[T_Q]$ and P_Q.

Question: What is an asymptotic regime?

Question: What is asymptotic analysis trying to answer?

Prior Example: $\underline{\text{FIX } \rho = 0.99}$, $\mathbf{E}[S] = \frac{1}{\mu}$.

Question: Fill in the arrival rates and number of servers in system n as function of ρ and μ. Set $k_n = n$.

$$\lambda_n = \_\_\_\_\_ \times n, \qquad R_n = \_\_\_\_\_, \qquad k_n = n = \_\_\_\_\_ \times R_n$$

Question: Does the load ρ_n change in this regime?

This regime is called the **Mean-Field Regime**. The load remains constant, and the arrival rate and number of servers grow proportionally.

4 Mean-Field Regime Analysis

In each system n: $\qquad k_n = n, \quad \rho_n = \rho = 0.99$

The queueing probability for an M/M/k system is

$$P_Q = \frac{\dfrac{(\rho k)^k}{k!(1-\rho)}}{\displaystyle\sum_{i=0}^{k-1} \frac{(\rho k)^i}{i!} + \frac{(\rho k)^k}{k!(1-\rho)}}, \qquad P_Q(n) = \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_.$$

Fill in Sterling's approximation: $n! \approx$ ____________________.

$$\begin{aligned}
P_Q(n) &\approx \frac{\dfrac{(\rho n)^n}{n!(1-\rho)}}{e^{\rho n} + \dfrac{(\rho n)^n}{n!(1-\rho)}} && \text{By}____________________ \\
&= \frac{1}{e^{\rho n}\dfrac{n!(1-\rho)}{(\rho n)^n} + 1}. \\
&\approx \frac{1}{e^{\rho n}\dfrac{\sqrt{2\pi n}\, n^n e^{-n}(1-\rho)}{(\rho n)^n} + 1} && \text{By}____________________ \\
&= \frac{1}{\sqrt{2\pi n}\,(1-\rho)\left(\dfrac{e^{1-\rho}}{\rho}\right)^n + 1}.
\end{aligned}$$

Since $\rho < 1$,

$$P_Q(n) \overset{n\to\infty}{\longrightarrow} \_\_\_.$$

Summary:

1. Mean-Field Regime is defined as ____________________
2. In MF regime, $P_Q \to$ ____________________
3. In MF regime, $\mathbf{E}[T_Q] \to$ ____________________

5 Square-Root Staffing (Halfin-Whitt Regime)

Recall before/after provisioning example from day 1 of class:

Main Theorem of Chpt 15: [Square-Root Staffing Theorem]
Given an M/M/k with arrival rate λ and average job size $\mathbf{E}[S] = 1/\mu$, if we set $k = R + \sqrt{R_n}$, where $R = \lambda/\mu$, then we will always have $P_Q = 16\%$.

Question: How many servers do we need according to Square-Root Staffing?

Definition of the Halfin–Whitt Regime:

$$\lambda_n = n\lambda, \qquad k_n = R_n + \beta\sqrt{R_n} = \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_,$$

where β is a fixed constant (for now, $\beta = 1$).

Question: Write down the expression for the load ρ_n.

$$\rho_n = \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_$$

Question: Is ρ_n a constant?

Question: Which value does ρ_n converge to as $n \to \infty$?

LESSON 2: Having a large number of servers allows you to operate your system at much higher load, while keeping delay low.

6 Halfin-Whitt regime v.s. Heavy traffic

Question: What is P_Q in both regimes?

Question: What is $\mathbf{E}[T_Q]$ in both regimes?

7 Intuition behind Square-Root Staffing

8 Other Regimes

We can write

$$k_n = R_n + \alpha_n.$$

Question: In the Mean-Field Regime and Halfin–Whitt Regime, what are the corresponding values of α_n in terms of R_n?

<u>Mean Field:</u> $\alpha_n =$ ____________________

<u>Halfin–Whitt:</u> $\alpha_n =$ ____________________

Recall that R_n is __________________. Thus α_n represents ____________________

Question: What if $\alpha_n = o(\sqrt{R_n})$? What about $\alpha_n = \Omega(\sqrt{R_n})$? How will this affect P_Q and $\mathbf{E}[T_Q]$?

9 Intuitions for More Regimes

$\alpha_n = o(\sqrt{R_n}) \ll \sqrt{R_n}$:

$\alpha_n = \Omega(\sqrt{R_n}) \gg \sqrt{R_n}$:

10 Summary Tables for P_Q and $\mathbf{E}[T_Q]$

Question: Fill in the following table for P_Q as $n \to \infty$.

Regime	**Number of spare servers**	$P_Q(n)$ **Behavior**
Mean Field	$\alpha_n = \Theta(R)$	$P_Q(n) \to 0$
Sub HW	$\alpha_n = \Omega(\sqrt{R_n})$	
Halfin–Whitt	$\alpha_n = \Theta(\sqrt{R_n})$	$P_Q(n) \to c$
Super HW	$\alpha_n = o(\sqrt{R_n})$	

Question: What about $\mathbf{E}[T_Q]$? When does $\mathbf{E}[T_Q]$ converge to a constant?

This is called the **Non-Degenerate Slowdown** regime.

Regime	**Scaling of** α_n	$\mathbf{E}[T_Q]$ **Behavior**
Mean Field	$\alpha_n = \Theta(R_n)$	$\mathbf{E}[T_Q] \to 0$
Halfin–Whitt	$\alpha_n = \Theta(\sqrt{R_n})$	$\mathbf{E}[T_Q] \to 0$
Sub NDS		
NDS		
Super NDS		