Queuing Theory

Kaixin Hu 11129417

November 2015

1 Introduction of basic models and notations

A queueing system description looks as follows:

$$A/B/n$$

where A denotes the distribution of the inter-arrival time, B denotes the distribution of the service, n denotes the number of servers.

The common formations of A and B are M(Markov), D (Deterministic) and G (General). M denotes the exponential distribution, while D denotes that the arriving or service time are constant. For G, general, it means that the distribution is not further specified.

For M/M/n and M/G/n models, parameters are as following.

λ: the arrival rate into the system as a whole.

μ: the capacity of each of n equal servers

ρ: the system load; occupation load; server utilization; the fraction time the server is working.

To avoid that the queue eventually grows to infinity, we have to require that $\rho < 1$. According to Little's law states, in a single-server system G/G/1 with arrival rate λ and mean service time E(B) the amount of work arriving per unit time equals λE(B), i.e. λ/μ .In a multi-server system, $\rho = \lambda E(B)/n = \lambda/(\mu n)$.

2 Theoretical analysis of Average waiting times for M/M/1 queue and M/M/n queue[1]

2.1 M/M/1 queue

This part will derive the theory result of the waiting time distribution and mean waiting time for the M/M/1 queue, FIFO scheduling, with arrive rate λ and server capacity μ.

2.1.1 Time-dependent behaviour of customers

According to the memoryless property of exponential random distribution of the inter-arrive time and service time: $P(X > x + t|X > t) = P(X > x) = exp(-\mu x)$

so

$P(X < t + \Delta t|X > t) = 1 - exp(-\lambda x) = \lambda \Delta t + o(\Delta t); (\Delta t \to 0)$

$P(Y < t + \Delta t|Y > t) = 1 - exp(-\mu x) = \mu \Delta t + o(\Delta t); (\Delta t \to 0)$

$p_n(t)$: the probability that at time t there are n customers in the system; Since the exponential distribution is memoryless, it doesn't need to know when the last customer entered or left, so for $\Delta t \to 0$:

$p_0(t + \Delta t) = (1 - \lambda \Delta t)p_0(t) + \lambda \Delta t p_1(t) + o(\Delta t)$;

$p_n(t + \Delta t) = \lambda \Delta t p_{n-1}(t) + (1 - (\lambda + \mu)\Delta t)p_n(t) + \mu \Delta t p_{n+1}(t) + o(\Delta t)$; n = 1; 2; .....

Hence,

$p_0'(t) = -\lambda p_0(t) + \mu p_1(t)$

$p_n'(t) = \lambda p_{n-1}(t) - (\lambda + \mu)p_n(t) + \mu p_{n+1}(t)$; n = 1; 2; .....

2.1.2 Limiting behavior of customers

As t $\to$ 1, then $p_n(t) \to 0$ and $p_n(t) \to p_n$. [2]

so,

$0 = -\lambda p_0(t) + \mu p_1(t)$

$0 = \lambda p_{n-1}(t) - (\lambda + \mu)p_n(t) + \mu p_{n+1}(t)$; n = 1; 2; .....

Through the method of continuously substituting the state information of time t-2 and t-1 to time t:

$p_1 = \rho p_0$;

$p_2 = \rho^2 p_0$;

$p_n = \rho^n p_0$; n = 1; 2; .....

Considering, the property that $\sum_{i=1}^{\infty} p_n = 1$,$p_0 = 1-\rho$

Hence,$p_n = (1 - \rho)\rho^n$; n = 1; 2; .....

2.1.3 Service behavior

Define B_k : the service time of the k-th customer, which is independent and exponentially distributed with mean $1/\mu$.

So, $\sum_{k=1}^{n+1} B_k$ is a Erlang-n+1 distribution with mean (n+1)/μ.

$$P(\sum_{k=1}^{n+1} B_k > t) = \sum_{k=0}^{n} \frac{(\mu t)^k}{k!} e^{-\mu t}$$

2.1.4 Distribution of waiting time

Define L(t) : the number of customers in the system at time t

$L^{\alpha}(t)$: the number of customers in the system just before the arrival of a customer

S_n : the sojourn time of the n-th customer in the system

B_k : the service time of the k-th customer, which is independent and exponentially distributed with mean $1/\mu$.

Then,

$S = \sum_{k=1}^{L^\alpha+1} B_k$

Since L^α and B_k are independent events variables, according to the Total Probability Formula:

$$P(S > k) = P(\sum_{k=1}^{L^\alpha+1} B_k > t) = \sum_{n=0}^{\infty} P(\sum_{k=1}^{n+1} B_k > t) P(L^\alpha = n)$$

$$= \sum_{n=0}^{\infty} \sum_{k=0}^{n} \frac{(\mu t)^k}{k!} e^{-\mu t} \rho^n (1-\rho)$$

$$= \sum_{k=0}^{\infty} \sum_{n=k}^{\infty} \frac{(\mu t)^k}{k!} e^{-\mu t} \rho^n (1-\rho)$$

$$= \sum_{k=0}^{\infty} \frac{(\mu \rho t)^k}{k!} e^{-\mu t}$$

$$= e^{-\mu(1-\rho)t}, t >= 0$$

Since S = W+B, W and B are independent,

$$\widetilde{S}(s) = \widetilde{W} * \widetilde{B} = \widetilde{W} * \frac{\mu}{\mu + s}$$

Thus,

$$\widetilde{W} = \frac{(1-\rho)(\mu+s)}{\mu(1-\rho)+s} = (1-\rho) * 1 + \rho * \frac{(1-\rho)\mu}{\mu(1-\rho)+s}$$

So W is with probability (1- ρ) equal to zero, and with probability ρ equal to an exponential random variable with parameter $\mu(1-\rho)$. Hence,

$$P(W > t) = \rho e^{-\mu(1-\rho)t}, t \leq 0$$

$$P(W > t | W > 0) = e^{-\mu(1-\rho)t}$$

The conditional waiting time$W|W > 0$ is exponentially distributed with parameter $\mu(1-\rho)$.

2.1.5 Mean waiting time

$E(W|W > 0) = 1/\mu(1-\rho) = 1/(\mu - \lambda)$

$E(W) = \rho/\mu(1-\rho)) = \lambda/\mu(\mu-\lambda)$.

2.2 M/M/n queue

This section will derive the theory result of the waiting time distribution and mean waiting time for the M/M/n queue, FIFO scheduling, with arrive rate λ and the capacity of each of n equal servers μ.

2.3 Equilibrium probabilities

p_m : the equilibrium probability that there are m customers in the system.

$\prod_W$: the probability a customer has to wait.

By analysing the flow diagram, the flow into and out of the set of states $\{0, 1, \ldots\ldots, n-1\}$ should be equal:

$\lambda p_{m-1} = min(m, n)\mu p_m$; m=1,2,.....

Iterating gives:

$$p_m = \frac{(c\rho)^m}{m!} p_0; m = 1, \ldots\ldots, n$$

$$p_{m+n} = \rho^n p_n = \rho^n \frac{(c\rho)^m}{m!} p_0; m = 1, 2, \ldots\ldots$$

Considering:

$$\sum_{i=1}^{\infty} p_m = 1$$

yielding:

$$p_0 = (\sum_{m=0}^{n-1} \frac{(n\rho)^m}{m!} + \frac{(n\rho)^n}{n!} * \frac{1}{1-\rho})^{-1}$$

So,

$$\prod_W = P_n + p_{n+1} + p_{n+2} + ... = \frac{p_n}{1-\rho} = \frac{(n\rho)^n}{n!}((1-\rho)\sum_{m=0}^{n-1} \frac{(n\rho)^m}{m!} + \frac{(n\rho)^n}{n!})^{-1}$$

2.4 Mean waiting time

The mean queue length:

$$E(L^\alpha) = \sum_{m=0}^{\infty} n p_{n+m} = \prod_W * \frac{\rho}{1-\rho}$$

Then, according to the Little's Law:

$$E(W) = \prod_W * \frac{1}{1-\rho} * \frac{1}{n\mu}$$

2.4.1 Comparison between M/M/1 and M/M/n

For a M/M/n queue, $E(W) = \prod_W * \frac{1}{1-\rho} * \frac{1}{n\mu}$.

When n=2,

$$E_{n=2}(W) = 2\rho^2(1+\rho) * \frac{1}{1-\rho} * \frac{1}{2\mu} = \frac{\rho^2}{\mu(1-\rho^2)}$$

For a single M/M/1 queue with the same load characteristics and customer arrive rate (and thus an n-fold lower arrival rate), $E(W) = \rho/\mu(1-\rho) > E_{n=2}(W)$

This is also reasonable in a non-mathematical sense. To explain it in the case of n=2, the M/M/2 queue can be seen as two 'M/M/1 queue' systems, each with the same server capacity and half of the arrive rate, except that the two servers can accept the customer in the other queue system when he is idle. So this means that the two servers can help each other when one of them is idle, and this improve the efficiency, making the waiting time less than that of the single, i.e. true, M/M/1 queue, with the same ρ and μ, and half μ. In the latter system, the only server has to be idle and wait for the random generator to generate a customer when there is no one in the system. And this also applies to the M/M/n queue where $n > 2$. What's more, as n increases, efficiency improves, thus the waiting time decreases, keeping the load characteristics and customer arrive rate the same.

3 DES programs for M/M/1, M/M/2, and M/M/4

In this part, M/M/1, M/M/2, and M/M/4 will be simulated to compare the mean waiting time, with same system load and processor capacity. To ensure the statistical significance, Method of Batch Means is adopted in this report.

3.1 Methods

The steady state simulation in this report for queue system is achieved by implementing the Method of Batch Means to exclude the transient period.

This method involves one very long simulation run which is suitably subdivided into an initial transient period k and the rest period (n-k) is broken up into s batches, with each batch interval size equal to r=(n-k)/s. Each of the batch is then treated as an independent run of the simulation experiment while no observation are made during the transient period which is treated as warm-up interval.

In each batch, an estimator Y_j, j=1,2,...,s. is calculated:

$$Y_j = \frac{1}{r} \sum_{i \text{ in batch } j}^{r} W_i; j = 1, 2, ..., s$$

The mean waiting time is calculated:

$$W = \frac{1}{s} \sum_{j=1}^{s} Y_j$$

The sample variance is calculated:

$$\hat{\sigma_W}^2 = \frac{1}{s-1} \sum_{j=1}^{s} (Y_j - \bar{Y})^2$$

The estimate of MSE is $\hat{\sigma_W}^2/s$.

Following is the value set for the simulation:

k: 10000;

r: 100;

d: 0.2 for M/M/n;

As for k, since Batch Means is a method of estimating the steady-state characteristic from a single-run simulation, it is important that the warm-up period, k, is set large enough to ensure that the bias due to initial conditions is removed to achieve at least a covariance stationary waiting time process. Abate and Whitt [3] related the k required for an M/M/1/ queue system with the system load ρ, and came out that k=838.10 for $\rho = 0.9$ to reach and remain 99 percent limits of the steady-state value. But in this report, the system will be analysed with both the load ρ and number of servers n varies, so in order to ensure the k is set large enough, k is set to 10000.

As for r, since in Batch means method, each of the batch is treated as an independent run of the simulation experiment, so the length is the batch, r, should not be too small to avoid ensure independent batches, nor too large to ensure the number of batches, s, to ensure that central limit theorem can be applied to construct the needed confidence interval. So in this report, t is set to 100.

Besides, in order to control the mean square error of the mean waiting time, d is set to 0.2, so in the program, it will continue to generate more batches until $\hat{\sigma_W}/\sqrt{s} < d$. Hence, the calculated mean waiting time is 95 percent certain that will not differ from the mean by 1.96 MSE, i.e. 0.392. In the long tail distribution, d is not controlled, instead graphs with error bar will be plotted.

3.2 M/M/1 and M/M/n queue

	Mean Waiting Time		
System load	n=1	n=2	n=4
0.1	0.09	0.01	0.00
0.2	0.18	0.03	0.00
0.3	0.29	0.06	0.01
0.4	0.46	0.13	0.02
0.5	0.73	0.21	0.05
0.6	0.97	0.39	0.11
0.7	1.67	0.67	0.28
0.8	3.35	1.48	0.43
0.9	8.26	3.84	1.82

Figure 1: The mean waiting time for M/M/n, with $\mu = 1$

Figure 1 shows the mean waiting times for M/M/n, n=1,2,4, with $\mu = 1$, and system load varies from 0.1 to 0.9. As expected, with the same system load

and arrive rate, the waiting times are shorter for n larger;And the system load increases, the waiting time increases.

Figure 2 and **??**shows the number of batches needed to attain the confidence interval [average waiting time-0.196,average waiting time +0.196] at the 95 percent confidence level. For each batch there are 100 measurements included.

It shows that, with a certain M/M/n. i.e. with n set, the number of measurements needed to assure statistic significance increases as system load increases.With a equal system load and server capacity, the the number of measurements needed decreases as n increases. Figure **??** shows that the number increases sharply as ρ closes to 1.

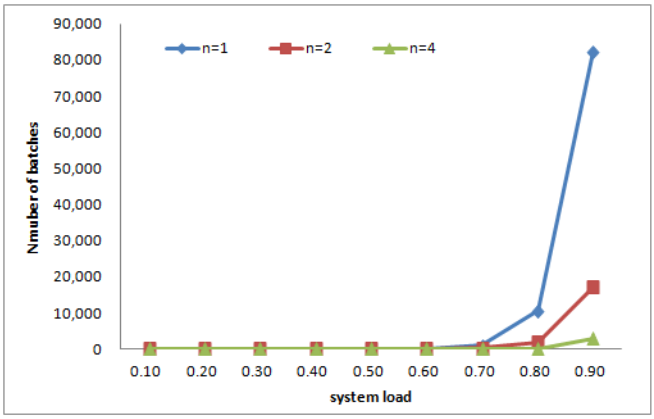

Figure 2: The number of batches needed to attain statistical significance for various ρ in M/M/n, $\mu = 1$

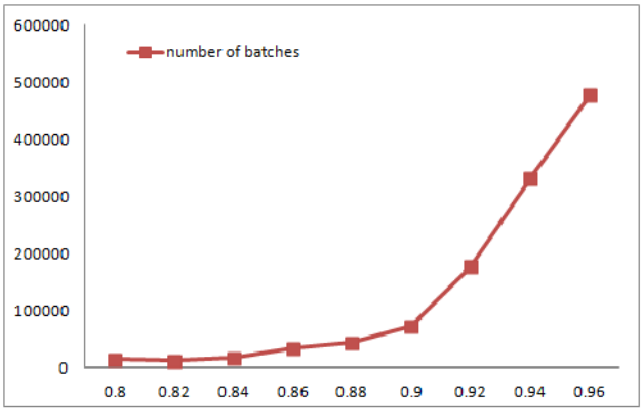

Figure 3: The number of batches needed to attain statistical significance for various ρ in M/M/1, $\mu = 1$

3.3 FIFO scheduling and SJF scheduling for M/M/1 queue

Figure 4 shows the mean waiting times of First In First Out scheduling and Shortest Job First scheduling method for M/M/1.

In theory, it should be that for the SJF method should have less average waiting time. For the customer with shortest job can first get service and get out of the system, so the customer with longer jobs then wait a relative shorter time compared to the other way around in which the the customer with shortest job wait for the longer one, reducing the average waiting time. But the simulating outcomes seems more or less equal, due to the mean error.

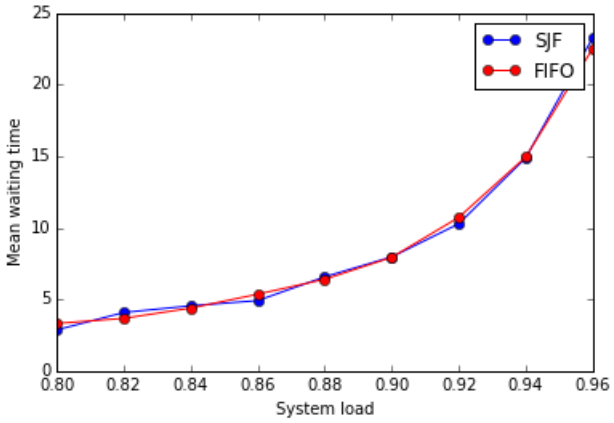

Figure 4: M/M/1,with $\mu = 1$, and M/D/1, with the deterministic service time equal to 1

3.4 M/D/1 and M/D/n queue

Figure 5 shows the M/M/1 and M/D/1 queue, in the latter the service time for each customer is equal and deterministic. With D Set equal to 1.

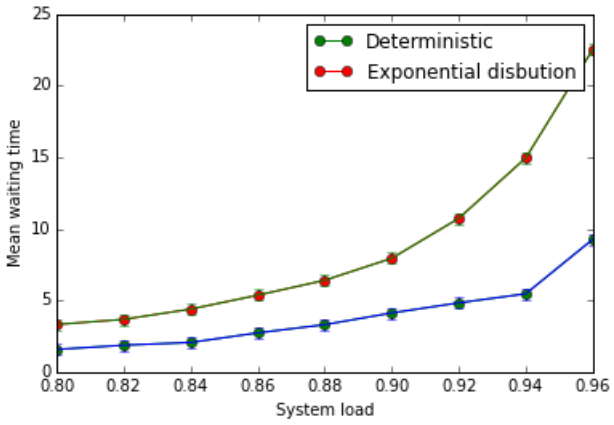

Figure 5: M/M/1,with $\mu = 1$, and M/D/1, with the deterministic service time equal to 1

3.5 M/G/1 and M/G/n

In this part, the service time is a long-tail distribution random variable. Specifically, this report analyses a hyper-exponential distribution, with 75 percent probability taking on the form of the exponential distribution with rate 1,and 25 percent probability taking on the form of the exponential distribution with rate 0.2. That is the serve system is such that with 75 percent of the jobs have an exponential distribution with an average service time of 1.0 and the remaining 25 percent an exponential distribution with an average service time of 5.0. So the expected service time of each of the equal servers E(B)= 0.75*1.0+0.25*5.0=2. $\lambda = n * \rho / E(B)$

Figure 6 illustrates the mean waiting time for M/G/n queue system with ρ varies from 0.8 to 0.9, with G denote the hyper-exponential distribution mentioned above.

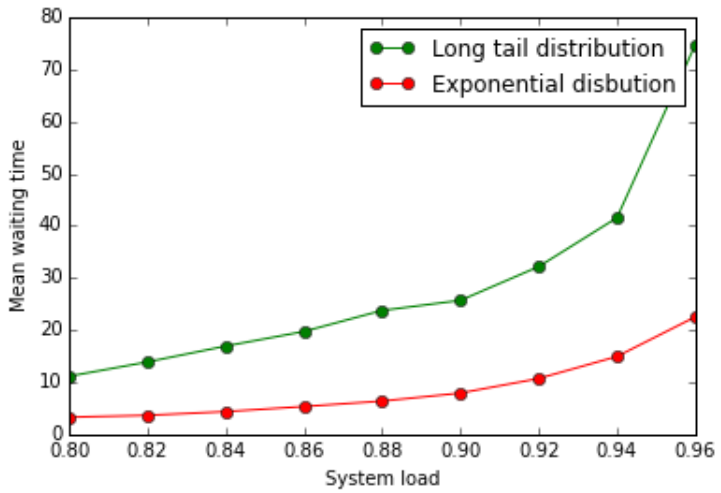

Figure 6: M/M/1,with $\mu = 1$, and M/G/1,with G denotes a long tail distribution

References

[1] Ivo Adan and Jacques Resing. Queueing theory, 2002.

[2] Jacob Willem Cohen. *The single server queue*. Elsevier, 2012.

[3] Joseph Abate and Ward Whitt. Transient behavior of regulated brownian motion, i: starting at the origin. *Advances in Applied Probability*, pages 560–598, 1987.