

CS418 — Operating Systems

Lecture 14

Queuing Analysis

Textbook: Operating Systems
by William Stallings

1. Why Queuing Analysis?

- If the system environment changes (like the number of users is doubled), we need to evaluate the resulting system performance
 - 1. Do an after-the-fact analysis based on actual values.
 - 2. Make a simple projection (estimation) based on previous experience.
 - 3. Do a queuing analysis.
 - 4. Program and run a simulation model.

2. Basic concepts — single server case

- λ : arrival rate; or mean number of arrivals per second.
- w : average number of waiting jobs.
- T_w : mean time a job must wait.
- T_s : mean service time for each job.
- r : average number of jobs in system.
- T_r : mean residence time for each job. $T_r = T_s + T_w$.
- ρ : fraction of time the server is busy, or, utilization of the server.
 $\rho = \lambda \times T_s \leq 1$ and $\lambda \leq \lambda_{max} = 1/T_s$.

- **Input:** λ, T_s are given.
- **Objectives:** (w, T_w, r, T_r) and their corresponding standard deviations, $(\sigma_w, \sigma_{T_w}, \sigma_r, \sigma_{T_r})$, should be returned.
- **Assumption 1:** λ follows Poisson distribution, i.e.,

$$Pr[X = k] = \frac{\lambda^k}{k!} e^{-\lambda}, k = 0, 1, 2, \dots$$

$$E[X] = \lambda$$

$$Pr[k \text{ items arrive in interval } T] = \frac{(\lambda T)^k}{k!} e^{-\lambda T}$$
Expected number of items arrive in interval $T = \lambda T$
Mean arrival rate = λ
- **Assumption 2:** service time distribution is exponential with λ , i.e.,
distribution function $Pr[X \leq x] = F[x] = 1 - e^{-\lambda x}$.
 $E[X] = \sigma_X = 1/\lambda$

• **We have:**

- 1. $r = \frac{\rho}{1-\rho}$.
- 2. $w = \frac{\rho^2}{1-\rho}$.
- 3. $T_r = \frac{T_s}{1-\rho}$.
- 4. $T_w = \frac{\rho T_s}{1-\rho}$.
- 5. $m_{T_r}(y) = T_r \times \ln\left(\frac{100}{100-y}\right)$ — the value T_r occurs y percent of time.
- 6. $Pr[R = N] = (1 - \rho)\rho^N$ — the probability that the number of jobs in system is N .
- 7. $m_{T_w}(y) = \frac{T_w}{\rho} \times \ln\left(\frac{100\rho}{100-y}\right)$ — the value T_w occurs y percent of time.
- 8.

3. Applications for single-server system

- Given a LAN with 100 PC's and a server which maintains a common database for a query application. The average query time is 0.6s and the standard deviation is equal to the mean. At peak times, the query rate over LAN reaches 20 per minute.
 - **1.** What is the average response time (ignoring line overhead)?
 - **2.** If a 1.5-second response time is the acceptable maximum, what is the maximum of message load?
 - **3.** If 20% more utilization is experienced, what will be the corresponding response time?

- **Answers.**

- **1.** $\rho = \lambda T_s = (20 \text{ arrivals per minute}) * (0.6 \text{ s per query}) / (60 \text{ s/min}) = 0.2$

$$T_r = \frac{T_s}{1-\rho} = 0.6 / (1 - 0.2) = 0.75 \text{ s.}$$

- **2.** Much harder to answer. We can rephrase the question as: if we want 90% of all responses to be less than 1.5s, then,

$$m_{T_r}(90) = T_r \times \ln\left(\frac{100}{100-90}\right) = \frac{T_s}{1-\rho} \times 2.3 = 1.5 \text{ s,}$$

so $\rho = 0.08$.

- **3.** Old $T_r = \frac{T_s}{1-\rho} = 0.75s$.
 New $T_r = \frac{T_s}{1-\rho_{new}} = 1.0s$,
 which is increased by $(1.0 - 0.75)/0.75 \approx 33\%$.

- Given a LAN which is connected to the Internet with a router, packets arrive with a mean arrival rate of 5 per second. The average packet length is 144 bytes and it is assumed to be exponentially distributed. Line speed from the router to the Internet is 9600bps.
 - **1.** What is the mean residence time in the router?
 - **2.** What is the average number of packets in the router?
 - **3.** What is the average number of packets in the router, for 90% of the time?

- **Answers.**

- **1.** $\lambda = 5$ per second.
 $T_s = (144 \times 8 \text{ bits})/9600\text{bps} = 0.12$ second.
 $\rho = \lambda T_s = 5 \times 0.12 = 0.6$.

$$T_r = \frac{T_s}{1-\rho} = 0.12/(1 - 0.6) = 0.3s.$$

- **2.** $r = \frac{\rho}{1-\rho} = 0.6/(1 - 0.6) = 1.5(\text{packets})$.

– **3.** $Pr[R = N] = (1 - \rho)\rho^N.$

So, $Pr[R \leq N] = [\sum_{i=0}^N (1 - \rho)\rho^i].$

$$\frac{y}{100} = [\sum_{i=0}^N (1 - \rho)\rho^i] = 1 - \rho^{1+N}.$$

$$N = \frac{\ln(1 - \frac{y}{100})}{\ln \rho} - 1.$$

When $y = 90$, $N = \frac{\ln(1 - \frac{y}{100})}{\ln \rho} - 1 = 3.5.$

When $y = 95$, $N = \frac{\ln(1 - \frac{y}{100})}{\ln \rho} - 1 = 4.8.$

4. Basic concepts — multiple server case

- Most of the concepts in single server case can carry over.
- N : number of servers.
- ρ : utilization of each server.
- $u = N\rho$: utilization of the whole system (or, traffic intensity).
- $\lambda \leq N/T_s$.
- **Assumptions:**
 - **1.** $\lambda =$ follows Poisson distribution.
 - **2.** T_s follows exponential distribution, for all servers.
 - **3.** All servers are equally loaded.
 - **4.** FIFO dispatching.
 - **5.** No items are discarded from queue.

• **We have:**

- 1. Poisson ratio function $K = \left[\frac{\sum_{i=0}^{N-1} (N\rho)^i / i!}{\sum_{i=0}^N (N\rho)^i / i!} \right]$
- 2. Erlang-C function C , which is the probability that all servers are busy, satisfies $C = \frac{1-K}{1-K\rho}$.
- 3. $r = C \frac{\rho}{1-\rho} + N\rho$.
- 4. $w = C \frac{\rho}{1-\rho}$.
- 5. $T_r = \frac{C}{N} \frac{T_s}{1-\rho} + T_s$.
- 6. $T_w = \frac{C}{N} \frac{T_s}{1-\rho}$.
- 7. $m_{T_w}(y) = \frac{T_s}{N(1-\rho)} \times \ln\left(\frac{100C}{100-y}\right)$ — the value T_w occurs y percent of time.
- 8.

5. Application for multi-server system

- Given a system with 5 processors with average service time $T_s = 0.1s$. Let the standard deviation of T_s , σ_{T_s} , be 0.094s. Assume that jobs arrive at a rate of 40 per second.

- **1.** What is the average response time?
- **2.** What is the average waiting time?
- **3.** What is the average (maximum) waiting time for 90% of the time?

- **Answers.**

- **1.** As $\sigma_{T_s} = 0.094s \approx T_s = 0.1s$, we can assume that T_s follows exponential distribution.

- **1.1.** If no common queue is used,

$$\lambda = 40/5 = 8 \text{ per second.}$$

$$\rho = \lambda T_s = 8 \times 0.1 = 0.8.$$

$$T_r = \frac{T_s}{1-\rho} = \frac{0.1}{1-0.8} = 0.5 \text{ second.}$$

- **1.2.** If a common queue is used,

$$\lambda = 40/5 = 8 \text{ per second, same as before.}$$

$$C = 0.554, \text{ calculation is omitted.}$$

$$T_r = \frac{C}{N} \frac{T_s}{1-\rho} + T_s \approx \frac{0.554}{5} \frac{0.1}{1-0.8} + 0.1 \approx 0.1544.$$

So, when a common queue is used, T_r is reduced by a factor

of $\frac{0.5}{0.1544} > 3$.

– **2.** As $\sigma_{T_s} = 0.094s \approx T_s = 0.1s$, we can assume that T_s follows exponential distribution.

– **2.1.** If no common queue is used,

$\lambda = 40/5 = 8$ per second.

$\rho = \lambda T_s = 8 \times 0.1 = 0.8$.

$T_w = \frac{\rho T_s}{1-\rho} = \frac{0.8 \times 0.1}{1-0.8} = 0.4$ second.

– **2.2.** If a common queue is used,

$\lambda = 40/5 = 8$ per second, same as before.

$C = 0.554$, calculation is omitted.

$T_w = \frac{C}{N} \frac{T_s}{1-\rho} \approx \frac{0.554}{5} \frac{0.1}{1-0.8} \approx 0.0544$.

So, when a common queue is used, T_w is reduced by a factor of $\frac{0.4}{0.0544} > 7$.

– **3.** As $\sigma_{T_s} = 0.094s \approx T_s = 0.1s$, we can assume that T_s follows exponential distribution.

– **3.1.** If no common queue is used,

$\lambda = 40/5 = 8$ per second.

$\rho = \lambda T_s = 8 \times 0.1 = 0.8$.

$T_w = \frac{\rho T_s}{1-\rho} = \frac{0.8 \times 0.1}{1-0.8} = 0.4$ second.

$m'_{T_w}(y) = \frac{T_w}{\rho} \times \ln\left(\frac{100\rho}{100-y}\right)$ — the value T_w occurs y percent of time.

$$m'_{T_w}(90) = \frac{0.4}{0.8} \times \ln\left(\frac{80}{100-90}\right) = 0.5 \times \ln 8.$$

– **3.2.** If a common queue is used,

$$m_{T_w}(y) = \frac{T_s}{N(1-\rho)} \times \ln\left(\frac{100C}{100-y}\right) = \frac{0.1}{5(1-0.8)} \times \ln\left(\frac{55.4}{100-90}\right) = 0.1 \times \ln 5.54.$$

So, when a common queue is used, $m'_{T_w}(90)/m_{T_w}(90) = 5(\ln 8/\ln 5.54) > 5$.