Lecture 3:
Introduction to Queuing Theory

PAMS’18

Zsolt István
zsolt.istvan@imdea.org
What is a queuing system?

- Jobs arriving to a queue
- One/multiple servers dealing with the jobs from the queue

- When modeling queuing systems, it is important to talk about their properties
  - Some rules apply to all systems
Properties of queuing systems

1. What is the arrival rate?
2. What is the service time?
3. What is the service discipline?
4. What is the system capacity?
5. What is the number of servers?
6. What is the population size?
Interarrival time

• $\tau_{1..N}$: Independent and Identically Distributed random variables

• Events come from a “process”, Often assumed to be Poisson:
  • (1) the event is something that can be counted in whole numbers
  • (2) occurrences are independent, so that one occurrence neither diminishes nor increases the chance of another
  • (3) the average frequency of occurrence for the time period in question is known
  • (4) it is possible to count how many events have occurred, but we are not interested in how many have not occurred
Mean Arrival rate

- Interarrival time $\tau$ is a random variable
  - Mean value: $E[\tau]$

- Mean arrival rate: $\lambda = 1/E[\tau]$
  - Does it look similar to something we discussed previously?

- Useful to assume fixed $\lambda$ for modeling
  - Do real systems have fixed $\lambda$?

- Examples:
  - A new pizza order is received on average every 3 minutes. The arrival rate is 20/hour.
  - The printer receives a new job to print on average every 100ms, the arrival rate is 10jobs/second.
Service time

• Time to process a job ("useful work", no queueing)
  • Random variable: $s$

• Mean service rate $\mu = 1 / E[s]$
  • What if we have $m$ servers?
  • Not a random variable

• Example: pizza oven bakes pizza on average in 6 minutes. $\mu = 10$/hour
A word on throughput

• The service rate of a system is rarely the measured throughput!
  • Throughput is client and workload dependent
  • Throughput only counts successful operations

• Is arrival rate the same as throughput?
  • In open systems?
  • In closed systems?
    • If there are no failed jobs?
Default assumptions for other properties

• What is the service discipline?
  • First come, first served (FIFO)

• What is the system capacity?
  • Large enough buffers → Infinite buffers

• What is the population size?
  • Very large → Infinite size
Equations valid for all queueing systems

• Load on system (traffic intensity): $\rho = \frac{\lambda}{m\mu}$

• Stability condition: $\rho < 1$ because this meant that $\lambda < (m\mu)$
  • What if $\rho = 1$? Can the system still be considered stable?
  • Remember arrival time is a random variable!
  • Once queueing starts, it never empties...
Traffic intensity example

• A USB thumb drive is serving 5k I/O ops/s
• The average time spent on the I/O operation is 0.1ms
• What is its utilization?

• \( \rho = \frac{\lambda}{m\mu} \)
• \( \rho = \frac{5k}{\mu} \)
• \( \mu = \frac{1}{0.1} = 10k \)
• \( \rho = \frac{5k}{10k} = 0.5 \ (50\%) \)
More detailed metrics

• Number of jobs in the system is the sum of the jobs in the queue and the ones in service
  • \( n = n_s + n_q \)
• Total time spent in system (response time) is the sum of time spent queuing and that in service
  • \( r = w_q + s \)

• Remember these are random variables, we’ll speak of their expected value.
Little’s law

• Remember: \( n = n_s + n_q \)
• The system as a whole: \( E[n] = \lambda \times E[r] \)
• Only the queue part: \( E[n_q] = \lambda \times E[w_q] \)

• \( E[n_s] = \lambda \times E[s] \) – looks familiar?
  • If \( m=1 \) and \( E[n_s]=1 \), the system is unstable!

• \( \rho = \frac{\lambda}{m\mu} \Rightarrow \rho = \frac{E[n_s]}{m} \)
Quick overview of M/M/1 queues

• Interarrival times and service times Poisson
• Single server
• FIFO processing

• Parameters:
  • Mean arrival rate
  • Mean service rate

• Please look at the book for more detail and explanations. Have a look at the list of formulas.
Response time in M/M/1

• The mean number of jobs in the system is computed using the probabilities of having 0..infinity jobs in the system.

\[ E[n] = \sum_{n=1}^{\infty} n \cdot p^n = \sum_{n=1}^{\infty} n(1-\rho)\rho^n = \frac{\rho}{1-\rho} \]

• Using Little’s law \((E[n] = \lambda \ast E[r])\), we get

\[ E[r] = \frac{\rho}{(\lambda \ast (1-\rho))} = \left(\mu \ast (1-\rho)\right)^{-1} \]
Quick overview of M/M/m queues

• Interarrival times and service times Poisson
• Single server
• FIFO processing
• m parallel servers (no queueing if the number of jobs ≤ m)

• Parameters:
  • Mean arrival rate
  • Mean service rate
  • Server parallelism

• Please look at the book for more detail and explanations. Have a look at the list of formulas.
All formulas for M/M/1 and M/M/m

Box 3.1 M/M/1 Queue
1. Parameters:
   - $\lambda$ = arrival rate in jobs per unit time
   - $\mu$ = service rate in jobs per unit time
2. Traffic intensity: $\rho = \lambda / \mu$
3. Stability condition: Traffic intensity $\rho$ must be less than 1.
4. Probability of zero jobs in the system: $p_0 = 1 - \rho$
5. Probability of n jobs in the system: $p_n = (1 - \rho)p_n$, $n = 0, 1, ..., \infty$
6. Mean number of jobs in the system: $E[N] = \rho/(1 - \rho)$
7. Variance of number of jobs in the system: $Var[N] = \rho^2/(1 - \rho)$
8. Probability of $k$ jobs in the queue: $P(n_q = k) = \frac{\rho^k}{k!}(1 - \rho)^{1-k}$
   - $k > 0$
9. Mean number of jobs in the queue: $E[N_q] = \rho^2/(1 - \rho)$
10. Variance of number of jobs in the queue: $Var[N_q] = (\rho^2/(1 - \rho) - \rho^2/(1 - \rho)^2)
11. Cumulative distribution function of the response time: $F(r) = 1 - e^{-\rho r}$
12. Mean response time: $E[R] = 1/(\mu(1 - \rho))$
13. Variance of the response time: $Var[R] = (1 - \rho)/\mu^2$
14. $q$-Percentile of the response time: $E[R] = 100/(100 - q)$
15. 90-Percentile of the response time: $2.3E[R]$
16. Cumulative distribution function of waiting time: $F(w) = 1 - e^{-w/\mu(1 - \rho)}$
17. Mean waiting time: $E[W] = 1/\mu - 1/\rho$
18. Variance of the waiting time: $Var[W] = (2 - \rho)/(\mu^2(1 - \rho)^2)$
19. $q$-Percentile of the waiting time: $max\left(0, E[W] + 100/(100 - q)\right)$
20. 90-Percentile of the waiting time: $max\left(0, E[W] + 100\ln(100)/100 - q\right)$
21. Probability of finding n or more jobs in the system: $p^n$
22. Probability of serving n jobs in one busy period: $\frac{1}{n}\left(\frac{2\lambda n - 1}{2\lambda n - 1 - \rho}\right)^n$

Box 3.2 M/M/m Queue
1. Parameters:
   - $\lambda$ = arrival rate in jobs per unit time
   - $\mu$ = service rate in jobs per unit time
   - $m$ = number of servers
2. Traffic intensity: $\rho = \lambda / (m\mu)$
3. The system is stable if the traffic intensity $\rho$ is less than 1.
4. Probability of zero jobs in the system: $p_0 = \left[1 + \left(\frac{m\mu}{\lambda}\right)^n + \sum_{n = 1}^{m-1} \left(\frac{m\mu}{\lambda}\right)^n\right]^{-1}$
5. Probability of n jobs in the system: $p_n = \left(\frac{m\mu}{\lambda}\right)^n/n!$, $n < m$
   - $p_n = \left(\frac{m\mu}{\lambda}\right)^m/m!$, $n \geq m$
6. Probability of queuing: $\theta = P(N > m) = (\frac{m\mu}{\lambda})^m p_0$
7. In the remaining formulas below we will use $\phi$ as defined here.
8. Mean number of jobs in the system: $E[N] = m\rho + \rho/(1 - \rho)$
9. Variance of number of jobs in the system: $Var[N] = \rho(\rho + \rho^2 - \rho)/(1 - \rho)^2$
10. Mean waiting time: $E[W] = (m\mu)^n/p_n + m\rho/(1 - \rho)$
11. Variance of number of jobs in the system: $Var[N] = \phi(1 + \rho - \phi)/(\phi(1 - \rho)^2)$
12. Average utilization of each server: $U = \phi/(m\mu)^n/p_n$
13. Cumulative distribution function of response time: $F(r) = \begin{cases} 1 - e^{-\rho r}, & 0 \leq r < 1 - m \rho, \\ 1 - e^{-\rho r} - q e^{-\rho r}, & m \rho \leq r < (m - 1) \rho/m, \\ \rho e^{\rho (r - 1)/m} - (m - r)/m, & \rho e^{\rho (r - 1)/m} \geq (m - 1) \rho/m. \\ \end{cases}$

Box 3.2 Continued
13. Mean response time: $E[R] = \frac{1}{\rho} \left[1 + \frac{\phi}{m(1 - \rho)}\right]$
14. Variance of the response time: $Var[R] = \frac{1}{\rho^2} \left[1 + \frac{\phi^2}{m(1 - \rho)^2}\right]$
15. Cumulative distribution function of waiting time: $F(w) = 1 - e^{-\phi w}$
16. Mean waiting time: $E[W] = E[N] / \lambda = \phi/(m\mu(1 - \rho))$
17. Variance of the waiting time: $Var[W] = \phi^2/(m^2\mu^2(1 - \rho)^2)$
18. $90$-Percentile of the waiting time: $max\left(0, E[W] + \frac{100}{100 - q}\right)$
19. 90-Percentile of the waiting time: $E[W] = \frac{100}{100 - q}$

Once again, $\phi$ in these formulas is the probability of $m$ or more jobs in the system: $\phi = (\frac{m\mu}{\lambda})^m/p_n$. For $m = 1$, $\phi$ is equal to $\rho$ and all of the formulas become identical to those for M/M/1 queues.
Exercise

• $\mu = 250/s$
  • $E[s] = ?$

• $\lambda = 1200/s$

• What is better for clients? $5x \text{ M/M/1}$ or $1x \text{ M/M/m}$?

• For M/M/1: $E[r] = 0.1s$

• For M/M/m: $E[r] = 0.022s$
  • The jobs wait in a single queue and can go to any available server. In the other case they need to wait for their pre-chosen server to become available...

Hint: there are many tools/websites that help with computing the outputs of the models (e.g., https://www.supositorio.com/rcalc/rcalclite.htm)
Network of queues

- A collection of queue/server pairs
- Jobs “flow” through the network
- Can represent arbitrarily complex systems
- Open and closed variants

PAMS18
Properties of NoQ devices

• Service discipline
  • FIFO (e.g., M/M/1 and M/M/m)
  • Delay center (imagine M/M/∞)
  • ...

• Job classes
  • All jobs are equal

• Job flow balance
  • Number of arrivals at each device equals number of leaving jobs

• One-step behavior
  • The state of the network changes only as a result of a job entering the system (a device)
Operational laws

• Valid for all devices
  • Arrival rate $\lambda_i = \frac{\text{(number of arrivals)}}{\text{time}} = \frac{A_i}{T}$
  • Throughput $X_i = \frac{\text{(number of completions)}}{\text{time}} = \frac{C_i}{T}$
  • Utilization $U_i = \frac{\text{(busy time)}}{\text{time}} = \frac{B_i}{T}$
  • Mean service time = $\frac{\text{(busy time)}}{\text{(number of completions)}} = \frac{B_i}{C_i}$

• Utilization Law
  • $U_i = B_i / T = \frac{C_i}{T} * \frac{B_i}{C_i}$
  • $U_i = X_i * S_i$ (Device with highest utilization is the bottleneck device)

• Forced flow law
  • $A_i = C_i$

• System throughput $X = \frac{\text{(jobs completed)}}{\text{time}}$
  • Device throughput $X_i = X * V_i$
  • $V_i$ is the visit ratio; how many times a job is handled by the device $i$
Mean Value Analysis (MVA)

- Algorithm to compute the behavior of a NoQ with increasing clients
  - Might have to map throughput levels to number of clients!

![Graph showing throughput vs. clients]

**Box 34.2 MVA Algorithm**

**Inputs:**
- $N$ = number of users
- $Z$ = think time
- $M$ = number of devices (not including terminals)
- $S_i$ = service time per visit to the $i$th device
- $V_i$ = number of visits to the $i$th device

**Outputs:**
- $X$ = system throughput
- $Q_i$ = average number of jobs at the $i$th device
- $R_i$ = response time of the $i$th device
- $U_i$ = utilization of the $i$th device

**Initialization:** FOR $i$ = 1 TO $M$ DO $Q_i = 0$

**Iterations:**
FOR $n$ = 1 TO $N$ DO BEGIN
  FOR $i$ = 1 TO $M$ DO $R_i = \left\{ \begin{array}{ll}
  S_i(1 + Q_i) & \text{Fixed capacity} \\
  S_i & \text{Delay centers}
\end{array} \right.$
  $R = \sum_{i=1}^{M} R_i V_i$
  $X = \frac{N}{Z + R}$
  FOR $i$ = 1 TO $M$ DO $Q_i = XV_i R_i$
END

Device throughputs: $X_i = XV_i$
Device utilizations: $U_i = XS_i V_i$