8. Queueing systems

Contents

- Refresher: Simple teletraffic model
- Queueing discipline
- M/M/1 (1 server, ∞ waiting places)
- Application to packet level modelling of data traffic
- M/M/n (n servers, ∞ waiting places)

Simple teletraffic model

- Customers arrive at rate λ (customers per time unit)
  - $1/\lambda$ = average inter-arrival time
- Customers are served by n parallel servers
- When busy, a server serves at rate µ (customers per time unit)
  - $1/\mu$ = average service time of a customer
- There are $n + m$ customer places in the system
  - at least $n$ service places and at most $m$ waiting places
- It is assumed that blocked customers (arriving in a full system) are lost

Pure queueing system

- Finite number of servers ($n < \infty$), $n$ service places, infinite number of waiting places ($m = \infty$)
  - If all $n$ servers are occupied when a customer arrives, it occupies one of the waiting places
  - No customers are lost but some of them have to wait before getting served
- From the customer’s point of view, it is interesting to know e.g.
  - what is the probability that it has to wait “too long”?
8. Queueing systems

Contents

- Refresher: Simple teletraffic model
- Queueing discipline
  - M/M/1 (1 server, \(\infty\) waiting places)
  - Application to packet level modelling of data traffic
  - M/M/n (\(n\) servers, \(\infty\) waiting places)

Queueing discipline

- Consider a single server (\(n = 1\)) queueing system
- Queueing discipline determines the way the server serves the customers
  - It tells
    - whether the customers are served one-by-one or simultaneously
    - Furthermore, if the customers are served one-by-one, it tells
      - in which order they are taken into the service
      - And if the customers are served simultaneously, it tells
        - how the service capacity is shared among them
- Note: In computer systems the corresponding concept is scheduling
- A queueing discipline is called work-conserving if customers are served with full service rate \(\mu\) whenever the system is non-empty

Work-conserving queueing disciplines

- First In First Out (FIFO) = First Come First Served (FCFS)
  - ordinary queueing discipline ("queue")
    - arrival order = service order
    - customers served one-by-one (with full service rate \(\mu\))
    - always serve the customer that has been waiting for the longest time
    - default queueing discipline in this lecture
- Last In First Out (LIFO) = Last Come First Served (LCFS)
  - reversed queueing discipline ("stack")
  - customers served one-by-one (with full service rate \(\mu\))
  - always serve the customer that has been waiting for the shortest time
- Processor Sharing (PS)
  - "fair queueing"
  - customers served simultaneously
  - when \(i\) customers in the system, each of them served with equal rate \(\mu/i\)
  - see Lecture 9. Sharing systems

Contents

- Refresher: Simple teletraffic model
- Queueing discipline
  - M/M/1 (1 server, \(\infty\) waiting places)
  - Application to packet level modelling of data traffic
  - M/M/n (\(n\) servers, \(\infty\) waiting places)
8. Queueing systems

M/M/1 queue

- Consider the following simple teletraffic model:
  - Infinite number of independent customers ($k = \infty$)
  - Interarrival times are IID and exponentially distributed with mean $1/\lambda$
    - so, customers arrive according to a Poisson process with intensity $\lambda$
  - One server ($n = 1$)
  - Service times are IID and exponentially distributed with mean $1/\mu$
  - Infinite number of waiting places ($m = \infty$)
  - Default queueing discipline: FIFO
- Using Kendall’s notation, this is an $M/M/1$ queue
  - more precisely: $M/M/1$-FIFO queue
- Notation:
  - $\rho = \lambda/\mu = \text{traffic load}$

Related random variables

- $X$ = number of customers in the system at an arbitrary time
  = queue length in equilibrium
- $X^*$ = number of customers in the system at an (typical) arrival time
  = queue length seen by an arriving customer
- $W$ = waiting time of a (typical) customer
- $S$ = service time of a (typical) customer
- $D = W + S$ = total time in the system of a (typical) customer = delay

State transition diagram

- Let $X(t)$ denote the number of customers in the system at time $t$
  - Assume that $X(t) = i$ at some time $t$, and consider what happens during a short time interval $(t, t+h)$:
    - with prob. $\lambda h + o(h)$, a new customer arrives (state transition $i \to i+1$)
    - if $i > 0$, then, with prob. $\mu h + o(h)$, a customer leaves the system (state transition $i \to i-1$)
- Process $X(t)$ is clearly a Markov process with state transition diagram

Equilibrium distribution (1)

- Local balance equations (LBE):
  \[ \pi_i \lambda = \pi_{i+1} \mu \]  
  \[ \Rightarrow \pi_{i+1} = \frac{\lambda}{\mu} \pi_i = \rho \pi_i \]
  \[ \Rightarrow \pi_i = \rho^i \pi_0, \quad i = 0,1,2,\ldots \]
- Normalizing condition (N):
  \[
  \sum_{i=0}^{\infty} \pi_i = \pi_0 \sum_{i=0}^{\infty} \rho^i = 1
  \]
  \[ \Rightarrow \pi_0 = \left( \sum_{i=0}^{\infty} \rho^i \right)^{-1} = \left( \frac{1}{1-\rho} \right)^{-1} = 1 - \rho, \quad \text{if } \rho < 1 \]
8. Queueing systems

Equilibrium distribution (2)

- Thus, for a stable system ($\rho < 1$), the equilibrium distribution exists and is a geometric distribution:

  $$\rho < 1 \implies X \sim \text{Geom}(\rho)$$

  $$P\{X = i\} = \pi_i = (1 - \rho)\rho^i, \quad i = 0, 1, 2, \ldots$$

  $$E[X] = \frac{\rho}{1 - \rho}, \quad D^2[X] = \frac{\rho^2}{(1 - \rho)^2}$$

- Remark:
  - This result is valid for any work-conserving queueing discipline (FIFO, LIFO, PS, ...)
  - This result is not insensitive to the service time distribution for FIFO
    - Even the mean queue length $E[X]$ depends on the distribution
    - However, for any symmetric queueing discipline (such as LIFO or PS) the result is, indeed, insensitive to the service time distribution

Mean queue length $E[X]$ vs. traffic load $\rho$

Mean delay

- Let $D$ denote the total time (delay) in the system of a (typical) customer
  - Including both the waiting time $W$ and the service time $S$: $D = W + S$
- Little’s formula: $E[X] = \lambda E[D]$. Thus,

  $$E[D] = \frac{E[X]}{\lambda} = \frac{1}{\lambda} \cdot \frac{\rho}{1 - \rho} = \frac{1}{\mu} \cdot \frac{1}{1 - \rho} = \frac{1}{\mu - \lambda}$$

- Remark:
  - The mean delay is the same for all work-conserving queueing disciplines (FIFO, LIFO, PS, ...)
  - But the variance and other moments are different!
Mean waiting time

- Let $W$ denote the waiting time of a (typical) customer
- Since $W = D - S$, we have

$$E[W] = E[D] - E[S] = \frac{1}{\mu} \left( \frac{1}{1 - \rho} - 1 \right) = \frac{1}{\mu} \cdot \frac{\rho}{1 - \rho}$$

Waiting time distribution (1)

- Let $W$ denote the waiting time of a (typical) customer
- Let $X^*$ denote the number of customers in the system at the arrival time
- PASTA: $P\{X^* = i\} = P\{X = i\} = \pi_i$.
- Assume now, for a while, that $X^* = i$ – Service times $S_1, \ldots, S_i$ of the waiting customers are IID and ~ $\text{Exp}(\mu)$
- Due to the memoryless property of the exponential distribution, the remaining service time $S_i^*$ of the customer in service also follows $\text{Exp}(\mu)$-distribution (and is independent of everything else)
- Due to the FIFO queueing discipline, $W = S_1^* + S_2 + \ldots + S_i$
- Construct a Poisson (point) process $\tau_n$ by defining $\tau_1 = S_1^*$ and $\tau_n = S_1^* + S_2 + \ldots + S_n, n \geq 2$. Now (since $X^* = i$): $W > t \Leftrightarrow \tau_i > t$

Waiting time distribution (2)

- Since $W = 0 \Leftrightarrow X^* = 0$, we have

$$P\{W = 0\} = P\{X^* = 0\} = \pi_0 = 1 - \rho$$

$$P\{W > t\} = \sum_{i=1}^{\infty} P\{W > t \mid X^* = i\} P\{X^* = i\} = \sum_{i=1}^{\infty} P\{\tau_i > t\} \pi_i = \sum_{i=1}^{\infty} P\{\tau_i > t\} \pi_i \frac{(\mu \rho)^i}{i!} e^{-\mu \rho}$$

- Denote by $A(t)$ the Poisson (counter) process corresponding to $\tau_n$
  - It follows that: $\tau_i > t \Leftrightarrow A(t) \leq i - 1$
  - On the other hand, we know that $A(t) \sim \text{Poisson} (\mu t)$. Thus,

$$P\{\tau_i > t\} = P\{A(t) \leq i - 1\} = \sum_{j=0}^{i-1} \frac{(\mu t)^j}{j!} e^{-\mu t}$$

Waiting time distribution (3)

- By combining the previous formulas, we get

$$P\{W > t\} = \sum_{i=1}^{\infty} P\{\tau_i > t\} (1 - \rho) \rho^i = \sum_{i=1}^{\infty} \sum_{j=0}^{i-1} \frac{(\mu \rho)^j}{j!} e^{-\mu \rho} (1 - \rho) \rho^i$$

$$= \rho \sum_{j=0}^{\infty} \frac{(\mu \rho)^j}{j!} e^{-\mu \rho} (1 - \rho) \sum_{i=j+1}^{\infty} \rho^{i-(j+1)}$$

$$= \rho \sum_{j=0}^{\infty} \frac{(\mu \rho)^j}{j!} e^{-\mu \rho} = \rho \frac{e^{\mu \rho} - e^{-\mu \rho}}{\mu} = \rho e^{-\mu(1 - \rho) t}$$
8. Queueing systems

Waiting time distribution (4)

- Waiting time $W$ can thus be presented as a product $W = JD$ of two independent random variables $J \sim \text{Bernoulli}(\rho)$ and $D \sim \text{Exp}(\mu(1-\rho))$:

\[
P(W = 0) = P(J = 0) = 1 - \rho
\]
\[
P(W > t) = P(J = 1, D > t) = \rho \cdot e^{-\mu(1-\rho)t}, \ t > 0
\]
\[
E[W] = E[J]E[D] = \rho \cdot \frac{1}{\mu(1-\rho)} = \frac{\rho}{\mu} \frac{1}{1-\rho}
\]
\[
E[W^2] = P(J = 1)E[D^2] = \rho \cdot \frac{2}{\mu^2(1-\rho)^2} = \frac{2\rho}{\mu^2} \frac{1}{(1-\rho)^2}
\]
\[
D^2[W] = E[W^2] - E[W]^2 = \frac{1}{\mu^2} \cdot \frac{\rho(2-\rho)}{(1-\rho)^2}
\]

Application to packet level modelling of data traffic

- $\text{M}/\text{M}/1$ model may be applied (to some extent) to packet level modelling of data traffic
  - customer = IP packet
  - $\lambda$ = packet arrival rate (packets per time unit)
  - $1/\mu$ = average packet transmission time (aikayks.)
  - $\rho = \lambda/\mu$ = traffic load
- Quality of service is measured e.g. by the packet delay
  - $P_z$ = probability that a packet has to wait “too long”, i.e. longer than a given reference value $z$

\[
P_z = P(W > z) = \rho e^{-\mu(1-\rho)z}
\]

Multiplexing gain

- We determine load $\rho$ so that prob. $P_z < 1\%$ for $z = 1$ (time units)
- Multiplexing gain is described by the traffic load $\rho$ as a function of the service rate $\mu$
Contents

- Refresher: Simple teletraffic model
- Queueing discipline
- M/M/1 (1 server, ∞ waiting places)
- Application to packet level modelling of data traffic
- M/M/n (n servers, ∞ waiting places)

M/M/n queue

- Consider the following simple teletraffic model:
  - Infinite number of independent customers ($k = \infty$)
  - Interarrival times are IID and exponentially distributed with mean $1/\lambda$.
    - so, customers arrive according to a Poisson process with intensity $\lambda$
  - Finite number of servers ($n < \infty$)
  - Service times are IID and exponentially distributed with mean $1/\mu$
  - Infinite number of waiting places ($m = \infty$)
  - Default queueing discipline: FCFS
- Using Kendall’s notation, this is an M/M/n queue
  - more precisely: M/M/n-FCFS queue
- Notation:
  - $\rho = \lambda/(n\mu) = \text{traffic load}$

State transition diagram

- Let $X(t)$ denote the number of customers in the system at time $t$
  - Assume that $X(t) = i$ at some time $t$, and consider what happens during a short time interval $(t, t+h]$:
    - with prob. $\lambda h + o(h)$, a new customer arrives (state transition $i \rightarrow i+1$)
    - if $i > 0$, then, with prob. $\min(i,n)\cdot \mu h + o(h)$, a customer leaves the system (state transition $i \rightarrow i-1$)
- Process $X(t)$ is clearly a Markov process with state transition diagram

Equilibrium distribution (1)

- Local balance equations (LBE) for $i < n$:
  \[ \pi_i \lambda = \pi_{i+1} (i+1) \mu \] (LBE)
  \[ \Rightarrow \pi_{i+1} = \frac{\lambda}{(i+1)\mu} \pi_i = \frac{n \rho}{i+1} \pi_i \]
  \[ \Rightarrow \pi_i = \left(\frac{n \rho}{\mu}\right)^i \pi_0, \quad i = 0,1,\ldots,n \]
- Local balance equations (LBE) for $i \geq n$:
  \[ \pi_i \lambda = \pi_{i+1} n \mu \] (LBE)
  \[ \Rightarrow \pi_{i+1} = \frac{\lambda}{n\mu} \pi_i = \rho \pi_i \]
  \[ \Rightarrow \pi_i = \rho^{i-n} \pi_n = \rho^i \left(\frac{n \rho}{n!}\right)^n \pi_0 = \frac{n \rho^i}{n!} \pi_0, \quad i = n,n+1,\ldots \]
8. Queueing systems

### Equilibrium distribution (2)

- Normalizing condition (N):

\[
\sum_{i=0}^{\infty} \pi_i = \pi_0 \left( \sum_{i=0}^{n-1} \frac{(np)^i}{i!} + \sum_{i=n}^{\infty} \frac{n^n \rho^i}{n!} \right) = 1 \quad \text{(N)}
\]

\[
\Rightarrow \pi_0 = \left( \sum_{i=0}^{n-1} \frac{(np)^i}{i!} + \sum_{i=n}^{\infty} \frac{n^n \rho^i}{n!} \sum_{i=n}^{\infty} \rho^{i-n} \right)^{-1} = \frac{1}{\alpha + \beta}, \text{ if } \rho < 1
\]

Notation: \( \alpha = \sum_{i=0}^{n-1} \frac{(np)^i}{i!}, \beta = \frac{(np)^n}{n!} \)

### Equilibrium distribution (3)

- Thus, for a stable system (\( \rho < 1 \), that is: \( \lambda < n\mu \)), the equilibrium distribution exists and is as follows:

\[
\rho < 1 \Rightarrow P\{X = i\} = \pi_i = \frac{(np)^i}{i!} \cdot \frac{1}{\alpha + \beta}, \quad i = 0, 1, \ldots, n
\]

\[
P\{X = i\} = \frac{n^n \rho^i}{n!} \cdot \frac{1}{\alpha + \beta}, \quad i = n, n+1, \ldots
\]

\[
n = 1: \quad \alpha = 1, \quad \beta = \frac{\rho}{1-\rho}, \quad \pi_0 = \frac{1}{\alpha + \beta} = 1 - \rho
\]

\[
n = 2: \quad \alpha = 1 + 2\rho, \quad \beta = \frac{2\rho^2}{1-\rho}, \quad \pi_0 = \frac{1}{\alpha + \beta} = \frac{1-\rho}{1+\rho}
\]

8. Queueing systems

### Probability of waiting

- Let \( p_W \) denote the probability that an arriving customer has to wait
- Let \( X^* \) denote the number of customers in the system at an arrival time
- An arriving customer has to wait whenever all the servers are occupied at her arrival time. Thus,

\[
p_W = P\{X^* \geq n\}
\]

- PASTA: \( P\{X^* = i\} = P\{X = i\} = \pi_i \). Thus,

\[
p_W = P\{X^* \geq n\} = \sum_{i=n}^{\infty} \pi_i = \sum_{i=n}^{\infty} \pi_0 \cdot \frac{n^n \rho^i}{n!} = \pi_0 \cdot \frac{(np)^n}{n!} \cdot \frac{1}{\alpha + \beta}
\]

\[
\begin{align*}
n = 1: & \quad p_W = \rho \\
n = 2: & \quad p_W = \frac{2\rho^2}{1+\rho}
\end{align*}
\]

8. Queueing systems

### Mean number of waiting customers

- Let \( X_W \) denote the number of waiting customers in equilibrium
- Then

\[
E[X_W] = \sum_{i=n}^{\infty} (i-n)\pi_i = \pi_0 \cdot \frac{(np)^n}{n!} \sum_{i=n}^{\infty} (i-n) \cdot (1-\rho)^i-n
\]

\[
= p_W \cdot \frac{\rho}{1-\rho}
\]

\[
n = 1: \quad E[X_W] = p_W \cdot \frac{\rho}{1-\rho} = \frac{\rho^2}{1-\rho}
\]

\[
n = 2: \quad E[X_W] = p_W \cdot \frac{\rho^2}{1+\rho} \cdot \frac{\rho}{1-\rho} = \frac{2\rho^3}{1-\rho^2}
\]
8. Queueing systems

**Mean waiting time**

- Let $W$ denote the waiting time of a (typical) customer
- Little's formula: $E[X_W] = \lambda E[W]$. Thus,

$$E[W] = \frac{E[X_W]}{\lambda} = \frac{1}{\lambda} \cdot p_W \cdot \frac{\rho}{1-\rho} = \frac{1}{\mu} \cdot \frac{p_W}{n(1-\rho)} = p_W \cdot \frac{1}{n\mu-\lambda}$$

$n = 1$: $E[W] = \frac{1}{\mu} \cdot \frac{p_W}{1-\rho} = \frac{\rho}{1-\rho}$

$n = 2$: $E[W] = \frac{1}{\mu} \cdot \frac{p_W}{2(1-\rho)} = \frac{\rho^2}{1-\rho}$

**Mean delay**

- Let $D$ denote the total time (delay) in the system of a (typical) customer— including both the waiting time $W$ and the service time $S$: $D = W + S$
- Then,

$$E[D] = E[W] + E[S] = \frac{1}{\mu} \left( \frac{p_W}{n(1-\rho)} + 1 \right) = p_W \cdot \frac{1}{n\mu-\lambda} + \frac{1}{\mu}$$

$n = 1$: $E[D] = \frac{1}{\mu} \left( \frac{p_W}{1-\rho} + 1 \right) = \frac{1}{\mu} \left( \frac{\rho}{1-\rho} + 1 \right) = \frac{1}{\mu} \cdot \frac{1}{1-\rho}$

$n = 2$: $E[D] = \frac{1}{\mu} \cdot \frac{p_W}{2(1-\rho)} + \frac{1}{\mu} \left( \frac{\rho^2}{1-\rho^2} + 1 \right) = \frac{1}{\mu} \cdot \frac{1}{1-\rho^2}$

**Mean queue length**

- Let $X$ denote the number of customers in the system (queue length) in equilibrium
- Little's formula: $E[X] = \lambda E[D]$. Thus,

$$E[X] = \lambda \cdot E[D] = p_W \cdot \frac{\lambda}{n\mu-\lambda} + \frac{\lambda}{\mu} = p_W \cdot \frac{\rho}{1-\rho} + n\rho$$

$n = 1$: $E[X] = p_W \cdot \frac{\rho}{1-\rho} + \rho = \rho \cdot \frac{\rho}{1-\rho} + \rho = \frac{\rho}{1-\rho}$

$n = 2$: $E[X] = p_W \cdot \frac{\rho}{1-\rho} + 2\rho = \frac{2\rho^2}{1+\rho} \cdot \frac{\rho}{1-\rho} + 2\rho = \frac{2\rho}{1-\rho^2}$

**Waiting time distribution (1)**

- Let $W$ denote the waiting time of a (typical) customer
- Let $X*$ denote the number of customers in the system at the arrival time
- The customer has to wait only if $X* \geq n$. This happens with prob. $p_W$.
- Under the assumption that $X* = i \geq n$, the system, however, looks like an ordinary $M/M/1$ queue with arrival rate $\lambda$ and service rate $n\mu$.
  - Let $W'$ denote the waiting time of a (typical) customer in this $M/M/1$ queue
  - Let $X*'*$ denote the number of customers in the system at the arrival time
- It follows that

$$P\{W = 0\} = 1 - p_W$$

$$P\{W > t\} = P\{X* \geq n\} P\{W > t \mid X* \geq n\}$$

$$= p_W \cdot P\{W' > t \mid X*'* \geq 1\} = p_W \cdot e^{-n\mu(1-\rho)t}, \quad t > 0$$
Waiting time distribution (2)

- Waiting time $W$ can thus be presented as a product $W = JD'$ of two indep. random variables $J \sim \text{Bernoulli}(p_W)$ and $D' \sim \text{Exp}(n \mu (1-\rho))$:

$$P\{W = 0\} = P\{J = 0\} = 1 - p_W$$
$$P\{W > t\} = P\{J = 1, D' > t\} = p_W \cdot e^{-n \mu (1-\rho)t}, \ t > 0$$
$$E[W] = E[J]E[D'] = p_W \cdot \frac{1}{n \mu (1-\rho)} = \frac{1}{\mu} \cdot p_W \cdot \frac{n \mu (1-\rho)}{n \mu (1-\rho)}$$
$$E[W^2] = P\{J = 1\}E[D'^2] = p_W \cdot \frac{2}{n^2 \mu^2 (1-\rho)^2} = \frac{1}{\mu^2} \cdot p_W \cdot \frac{2}{n^2 (1-\rho)^2}$$
$$D^2[W] = E[W^2] - E[W]^2 = \frac{1}{\mu^2} \cdot p_W \cdot \frac{2 - p_W}{n^2 (1-\rho)^2}$$

Example (1)

- Printer problem
  - Consider the following two different configurations:
    - One rapid printer (IID printing times $\sim \text{Exp}(2\mu)$)
    - Two slower parallel printers (IID printing times $\sim \text{Exp}(\mu)$)
  - Criterion: minimize mean delay $E[D]$
    - One rapid printer ($\text{M/M/1}$ model with $\rho = \frac{\lambda}{2\mu}$):
      $$E[D_1] = \frac{1}{2\mu} \cdot \frac{1}{1-\rho}$$
    - Two slower printers ($\text{M/M/2}$ model with $\rho = \frac{\lambda}{2\mu}$):
      $$E[D_2] = \frac{1}{\mu} \cdot \frac{1}{1-\rho^2} = \frac{1}{2\mu} \cdot \frac{2}{(1-\rho)(1+\rho)} = E[D_1] \cdot \frac{2}{1+\rho} > E[D_1]$$

Example (2)

- Traffic load $\rho$
  - $E[D_1]/E[D_2]$
  - Traffic load $\rho$
  - $E[D_1] / E[D_2]$