# 9. Sharing systems





## Contents

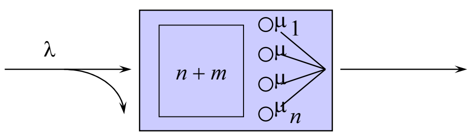






## Simple teletraffic model

- **Customers arrive** at rate $\lambda$ (customers per time unit)
  - $1/\lambda$ = average inter-arrival time
- Customers are **served** by $n$ parallel **servers**
- When busy, a server serves at rate $\mu$ (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

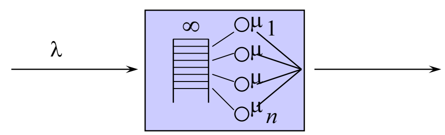







## Pure sharing system

- Finite number of servers ($n < \infty$), infinite number of service places ($n + m = \infty$), no waiting places
  - If there are at most $n$ customers in the system ($x \le n$), each customer has its own server. Otherwise ($x > n$), the total service rate ($n\mu$) is shared fairly among all customers.
  - Thus, the rate at which a customer is served equals $\min\{\mu, n\mu/x\}$
  - No customers are lost, and no one needs to wait before the service.
  - But the delay is the greater, the more there are customers in the system. Thus, delay is an interesing measure from the customer's point of view.



## Contents

- Refresher: Simple teletraffic model
- M/M/1-PS ($\infty$ customers, 1 server, $\infty$ customer places)
- M/M/$n$-PS ($\infty$ customers, $n$ servers, $\infty$ customer places)
- Application to flow level modelling of elastic data traffic
- M/M/1/$k$/$k$-PS ($k$ customers, 1 server, $k$ customer places)





## M/M/1-PS queue

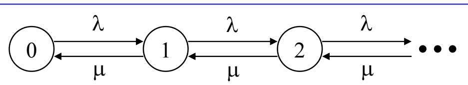
- Consider the following simple teletraffic model:
  - Infinite number of independent customers ($k = \infty$)
  - Interarrival times are IID and exponentially distributed with mean $1/\lambda$
    - so, customers arrive according to a Poisson process with intensity $\lambda$
  - One server ($n = 1$)
  - Service requirements are IID and exponentially distributed with mean $1/\mu$
  - Infinite number of customer places ($p = \infty$)
  - Queueing discipline: **PS**. All customers are served simultaneously in a fair way with equal shares of the service capacity $\mu$.
- Using Kendall's notation, this is an **M/M/1-PS queue**
- Notation:
  - $\rho = \lambda/\mu$ = 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 $i \to i+1$)
    - if $i > 0$, then, with prob. $i(\mu/i)h + o(h) = \mu h + o(h)$, a customer leaves the system (state transition $i \to i-1$)
- Process $X(t)$ is clearly a Markov process with state transition diagram




- Note that this is the same irreducible birth-death process with an infinite state space $S = \{0,1,2,...\}$ as for the M/M/1-FIFO queue.





## Equilibrium distribution (1)

- Local balance equations (LBE):

$$\pi_i \lambda = \pi_{i+1}\mu \qquad \text{(LBE)}$$
$$\Rightarrow \quad \pi_{i+1} = \frac{\lambda}{\mu}\pi_i = \rho\pi_i$$
$$\Rightarrow \quad \pi_i = \rho^i \pi_0, \quad i = 0,1,2,\ldots$$

- Normalizing condition (N):

$$\sum_{i=0}^{\infty} \pi_i = \pi_0 \sum_{i=0}^{\infty} \rho^i = 1 \qquad \text{(N)}$$
$$\Rightarrow \quad \pi_0 = \left(\sum_{i=0}^{\infty} \rho^i\right)^{-1} = \left(\frac{1}{1-\rho}\right)^{-1} = 1-\rho, \quad \text{if } \rho < 1$$

## Equilibrium distribution (2)

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

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

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

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

- **Remark**: Insensitivity with respect to service time distribution
  - The result for the PS discipline is **insensitive** to the service time distribution, that is: it is valid for **any** service time distribution with mean $1/\mu$
  - So, instead of the M/M/1-PS model, we can consider, as well, the more general M/G/1-PS model





## Mean delay

- Let $D$ denote the total time (delay) in the system of a (typical) customer
- Since the mean number of customers in the system, $E[X]$, is the same for all work-conserving queueing disciplines, also the mean delay is the same, by Little's result.
- Thus, we may apply the result derived for the FIFO discipline in Lect. 8:

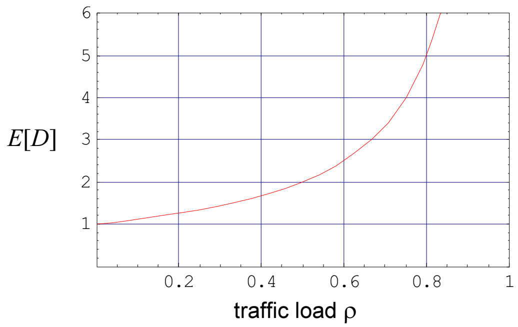
$$E[D] = \frac{1}{\mu} \cdot \frac{1}{1-\rho}$$





## Mean delay $E[D]$ vs. traffic load $\rho$

- Note that the time unit is the average service requirement $E[S]$





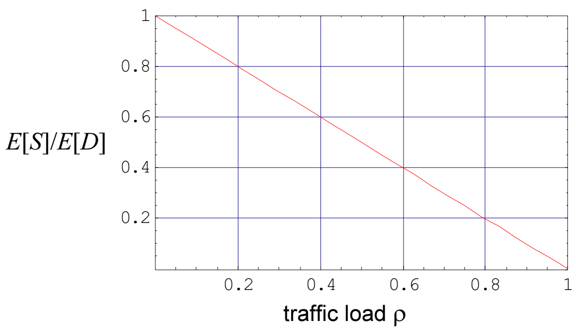



## Relative throughput

- A quality of service measure is the relative throughput $E[S]/E[D]$:

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

## Relative throughput $E[S]/E[D]$ vs. traffic load $\rho$








## Contents







## M/M/*n*-PS queue

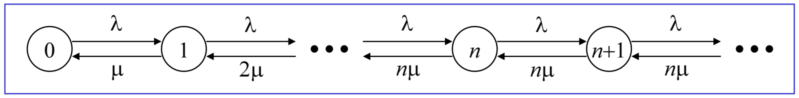
- Consider the following simple teletraffic model:
  - Infinite number of independent customers ($k = \infty$)
  - Interarrival times are IID and exponentially distributed with mean $1/\lambda$
    - so, customers arrive according to a Poisson process with intensity $\lambda$
  - Finite number of servers ($n < \infty$)
  - Service requirements are IID and exponentially distributed with mean $1/\mu$
  - Infinite number of customer places ($p = \infty$)
  - Queueing discipline: **PS**. If there are at most $n$ customers in the system ($i \le n$), each customer has its own server. Otherwise ($i > n$), the total service rate ($n\mu$) is shared fairly among all customers.
- Using Kendall's notation, this is an **M/M/*n*-PS queue**
- Notation:
  - $\rho = \lambda/(n\mu)$ = 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 $i \to i+1$)
    - if $i > 0$, then, with prob. $i \cdot \min\{\mu, n\mu/i\} \cdot h + o(h) = \min\{i,n\} \cdot \mu h + o(h)$, a customer leaves the system (state transition $i \to i-1$)
- Process $X(t)$ is clearly a Markov process with state transition diagram




- Note that this is the same irreducible birth-death process with an infinite state space $S = \{0,1,2,...\}$ as for the M/M/$n$-FIFO queue.

## Equilibrium distribution (1)

- Local balance equations (LBE) for $i < n$:

$$\pi_i \lambda = \pi_{i+1}(i+1)\mu \qquad \text{(LBE)}$$

$$\Rightarrow \quad \pi_{i+1} = \frac{\lambda}{(i+1)\mu}\pi_i = \frac{n\rho}{i+1}\pi_i$$

$$\Rightarrow \quad \pi_i = \frac{(n\rho)^i}{i!}\pi_0, \quad i = 0,1,\ldots,n$$

- Local balance equations (LBE) for $i \geq n$:

$$\pi_i \lambda = \pi_{i+1} n\mu \qquad \text{(LBE)}$$

$$\Rightarrow \quad \pi_{i+1} = \frac{\lambda}{n\mu}\pi_i = \rho\pi_i$$

$$\Rightarrow \quad \pi_i = \rho^{i-n}\pi_n = \rho^{i-n}\frac{(n\rho)^n}{n!}\pi_0 = \frac{n^n\rho^i}{n!}\pi_0, \quad i = n, n+1, \ldots$$





## Equilibrium distribution (2)

- Normalizing condition (N):

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

$$\Rightarrow \quad \pi_0 = \left(\sum_{i=0}^{n-1}\frac{(n\rho)^i}{i!} + \frac{(n\rho)^n}{n!}\sum_{i=n}^{\infty}\rho^{i-n}\right)^{-1}$$

$$= \left(\sum_{i=0}^{n-1}\frac{(n\rho)^i}{i!} + \frac{(n\rho)^n}{n!(1-\rho)}\right)^{-1} = \frac{1}{\alpha+\beta}, \quad \text{if } \rho < 1$$

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





## Equilibrium distribution (3)

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

$$\rho < 1 \quad \Rightarrow$$

$$P\{X = i\} = \pi_i = \begin{cases} \dfrac{(n\rho)^i}{i!}\cdot\dfrac{1}{\alpha+\beta}, & i = 0,1,\ldots,n \\ \dfrac{n^n\rho^i}{n!}\cdot\dfrac{1}{\alpha+\beta}, & i = n, n+1, \ldots \end{cases}$$

- **Remark**: Insensitivity with respect to service time distribution
  - The result for the PS discipline is **insensitive** to the service time distribution, that is: it is valid for **any** service time distribution with mean $1/\mu$
  - So, instead of the M/M/$n$-PS model, we can consider, as well, the more general M/G/$n$-PS model





## Mean delay

- Let $D$ denote the total time (delay) in the system of a (typical) customer
- Since the mean number of customers in the system, $E[X]$, is the same for all work-conserving queueing disciplines, also the mean delay is the same, by Little's result.
- Thus, we may apply the result derived for the FIFO discipline in Lect. 8:

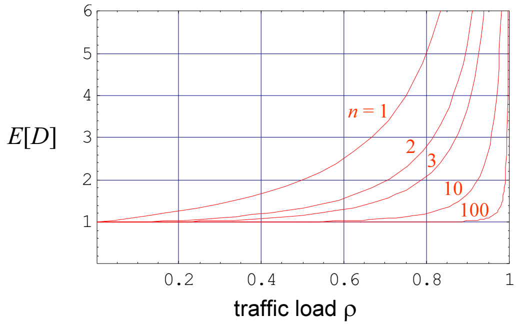
$$E[D] = \frac{1}{\mu}\cdot\left(\frac{p_W}{n(1-\rho)} + 1\right)$$

  - where $p_w$ refers to the probability

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

## Mean delay $E[D]$ vs. traffic load ρ

– Note that the time unit is the average service requirement $E[S]$








## Relative throughput

- A quality of service measure is the relative throughput $E[S]/E[D]$:

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

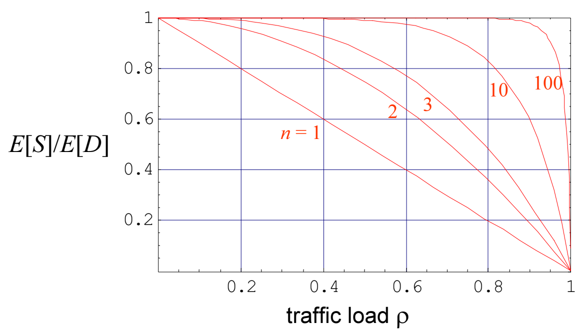
$$n=1: \ \frac{E[S]}{E[D]} = \frac{1-\rho}{p_W(1)+1-\rho} = 1-\rho$$

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





## Relative throughput $E[S]/E[D]$ vs. traffic load ρ








## Contents

- Refresher: Simple teletraffic model
- M/M/1-PS (∞ customers, 1 server, ∞ customer places)
- M/M/$n$-PS (∞ customers, $n$ servers, ∞ customer places)
- Application to flow level modelling of elastic data traffic
- M/M/1/$k$/$k$-PS ($k$ customers, 1 server, $k$ customer places)

## Application to flow level modelling of elastic data traffic

- M/G/$n$-PS model is applicable to flow level modelling of elastic data traffic
  - customer = TCP flow
  - $\lambda$ = flow arrival rate (flows per time unit)
  - $r$ = access link speed for a flow (data units per time unit)
  - $C = nr$ = speed of the shared link (data units per time unit)
  - $E[L]$ = average flow size (data units)
  - $E[S] = 1/\mu = E[L]/r$ = average flow transfer time with access link rate
  - $\rho = \lambda/(n\mu)$ = traffic load
- A quality of service measure is the throughput

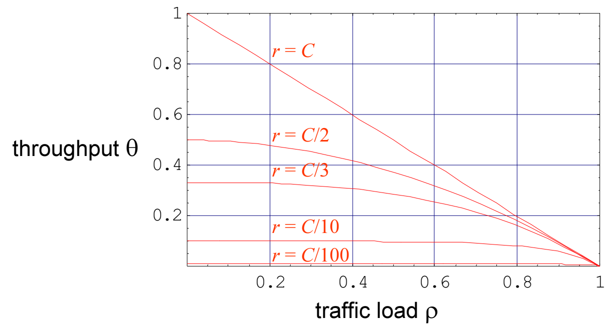
$$\theta = \frac{E[L]}{E[D]} = \frac{r \cdot E[S]}{E[D]} = \frac{r \cdot n(1-\rho)}{p_W(n)+n(1-\rho)} = C \cdot \frac{(1-\rho)}{p_W(n)+n(1-\rho)}$$





## Throughput $\theta$ vs. traffic load $\rho$

- Note that the rate unit is the link rate $C$








## Contents

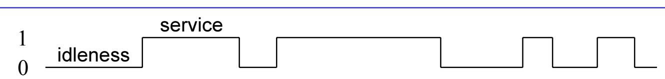






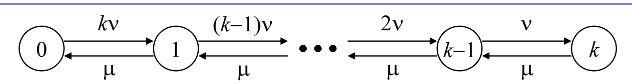
## M/M/1/$k$/$k$-PS queue

- Consider the following simple teletraffic model:
  - **Finite** number of independent customers ($k < \infty$)
    - **on-off type** customers (alternating between idleness and activity)
  - Idle times are IID and exponentially distributed with mean $1/\nu$
  - One server ($n = 1$)
  - Service requirements are IID and exponentially distributed with mean $1/\mu$
  - As many customer places as customers ($p = k$)
  - Queueing discipline: **PS**.
- Using Kendall's notation, this is an **M/M/1/$k$/$k$-PS queue**
- On-off type customer:



## 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 $(t, t+h]$:
    - if $i < k$, then, with prob. $(k-i)\nu h + o(h)$, an idle customer becomes active (state transition $i \to i+1$)
    - if $i > 0$, then, with prob. $i(\mu/i)h + o(h) = \mu + o(h)$, an active customer becomes idle (state transition $i \to i-1$)
- Process $X(t)$ is clearly a Markov process with state transition diagram




- Note that process $X(t)$ is an irreducible birth-death process with a finite state space $S = \{0,1,\ldots,k\}$





## Equilibrium distribution (1)

- Local balance equations (LBE):

$$\pi_i (k-i)\nu = \pi_{i+1}\mu \qquad \text{(LBE)}$$

$$\Rightarrow \quad \pi_i = \frac{\mu}{(k-i)\nu}\pi_{i+1}$$

$$\Rightarrow \quad \pi_i = \frac{1}{(k-i)!}\left(\frac{\mu}{\nu}\right)^{k-i}\pi_k, \quad i = 0,1,\ldots,k$$





## Equilibrium distribution (2)

- Normalizing condition (N):

$$\sum_{i=0}^{k}\pi_i = \pi_k \sum_{i=0}^{k}\frac{1}{(k-i)!}\left(\frac{\mu}{\nu}\right)^{k-i} = 1 \qquad \text{(N)}$$

$$\Rightarrow \quad \pi_k = \left(\sum_{i=0}^{k}\frac{1}{(k-i)!}\left(\frac{\mu}{\nu}\right)^{k-i}\right)^{-1}$$

$$\Rightarrow \quad \pi_i = \pi_k \cdot \frac{1}{(k-i)!}\left(\frac{\mu}{\nu}\right)^{k-i} = \frac{\frac{1}{(k-i)!}\left(\frac{\nu}{\mu}\right)^{i}}{\sum_{i'=0}^{k}\frac{1}{(k-i')!}\left(\frac{\nu}{\mu}\right)^{i'}}$$