Aalto University
School of Electrical Engineering

ELEC-E7450
Performance Analysis

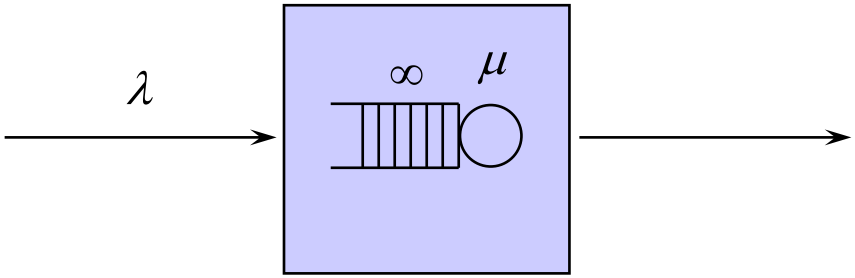
Single server queue M/G/1

Samuli Aalto
Department of Communications and Networking

Contents

M/G/1

- Customers arrive according to a Poisson process at rate λ
  - IID inter-arrival times
  - exponential inter-arrival time distribution with mean $1/\lambda$
- Customers are served by 1 server
  - IID service times S_i
  - general service time distribution with mean $E[S] = 1/\mu$
- There are ∞ customer places in the system



Service discipline

- **Definition**:
  Service discipline π determines the way the customers are served
  - It specifies whether the customers are served one-by-one or simultaneously
  - If the customers are served one-by-one, it specifies the order in which they are taken to service
  - If the customers are served simultaneously, it specifies how the service capacity is shared among them

- Service discipline is also called as queueing discipline, or scheduling discipline

- **Definition**:
  A service discipline is work-conserving if customers are served whenever the system is non-empty

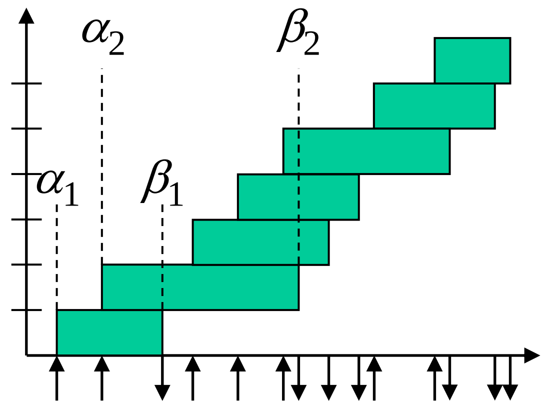
Work-conserving service disciplines

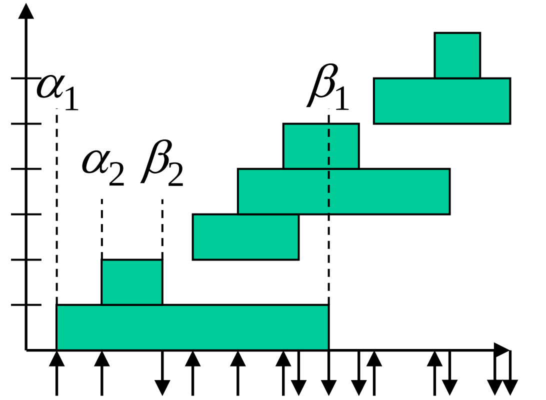
- First In First Out (FIFO)
  - customers are served one-by-one until completion
  - service in the arrival order ("ordinary queue")
  - the customer that arrived first is served with rate μ
  - also known as First Come First Served (FCFS)
- Processor Sharing (PS)
  - customers are served simultaneously
  - the service capacity is shared evenly among all customers ("fair queue")
  - when i customers are in the system, each of them is served with rate μ/i
  - ideal version of the Round Robin (RR) service discipline

Delay

- α_i = arrival time of customer i
- β_i^π = departure time of customer i
- T_i^π = delay (sojourn time) of customer i

$$T_i^\pi := \beta_i^\pi - \alpha_i$$





Queue length

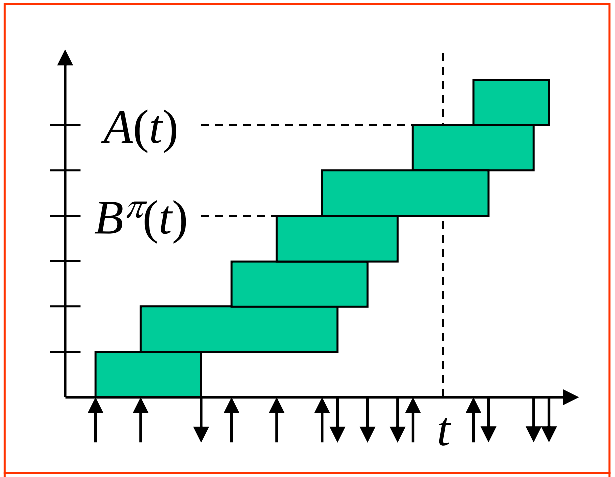
- $A(t)$ = number of arrivals until time t
  = arrival process
- $B^{\pi}(t)$ = number of departures until time t
  = departure process

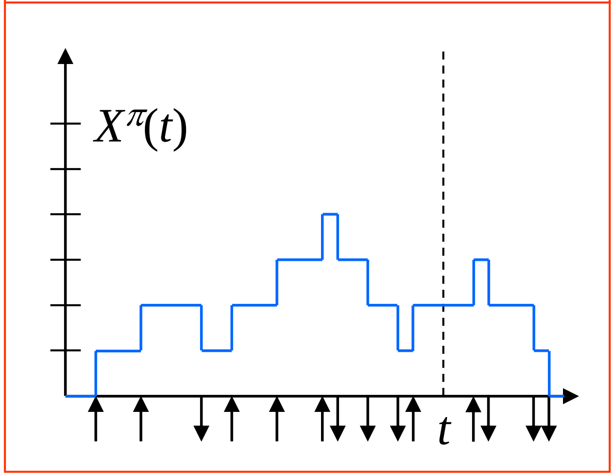
$$A(t) := \max\{i : \alpha_i \le t\}$$

$$B^{\pi}(t) := |\{i : \beta_i^{\pi} \le t\}|$$

- $X^{\pi}(t)$ = number of customers at time t
  = queue length process

$$X^{\pi}(t) := A(t) - B^{\pi}(t)$$





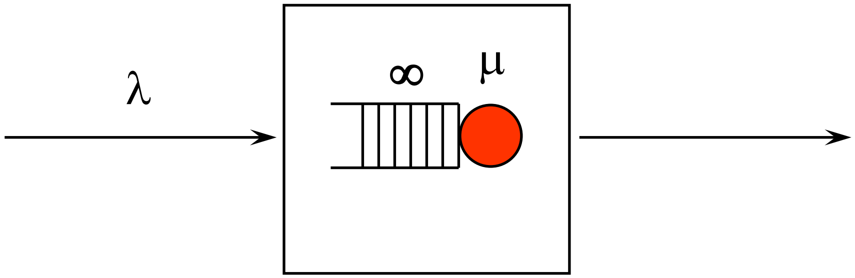
Traffic load in M/G/1

- **Definition**:
  Traffic load ρ is defined by

$$\rho := \frac{\lambda}{\mu}$$

- Applying Little's formula to the subsystem consisting of just the server:

$$P\{X>0\} = E[1_{\{X>0\}}] \overset{\text{Little}}{=} \lambda E[S] = \frac{\lambda}{\mu} = \rho, \quad P\{X=0\} = 1-\rho$$



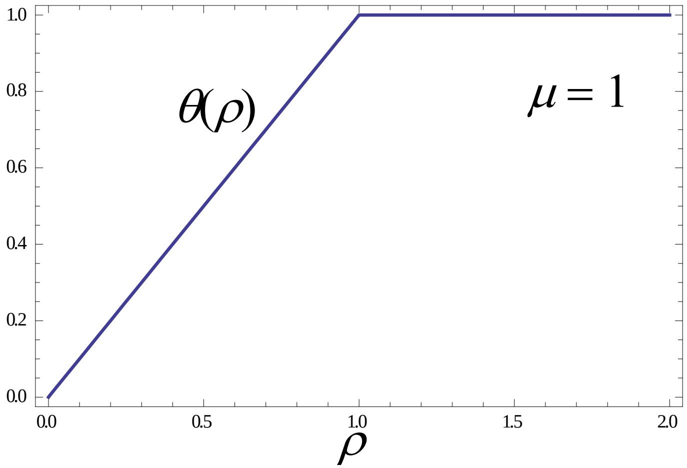
Throughput

- **Definition:**
  Throughput θ of service discipline π refers to the long-run average departure rate given by

$$\theta := \lim_{t\to\infty} \frac{1}{t} E[B^{\pi}(t)]$$

- **Proposition:**
  For any work-conserving service discipline π, the throughput is

$$\theta = \min\{\lambda, \mu\}$$



- **Corollary:**
  (i) If $\rho \le 1$, then $\theta = \lambda$
  (ii) If $\rho \ge 1$, then $\theta = \mu$

Stability

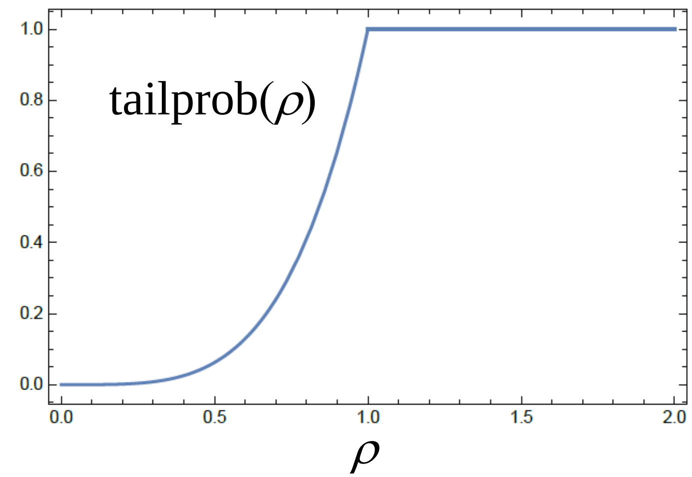
- **Definition:**
  Service discipline π is unstable if

$$\lim_{t\to\infty} P\{X^{\pi}(t) \geq n\} = 1 \quad \text{for all } n$$

Otherwise it is stable.

- **Proposition:**
  All service disciplines π are unstable if $\rho \geq 1$

- **Proposition:**
  All work-conserving service disciplines π are stable if $\rho < 1$



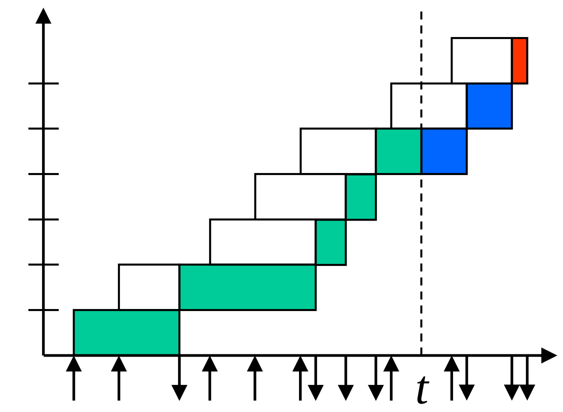
Unfinished work

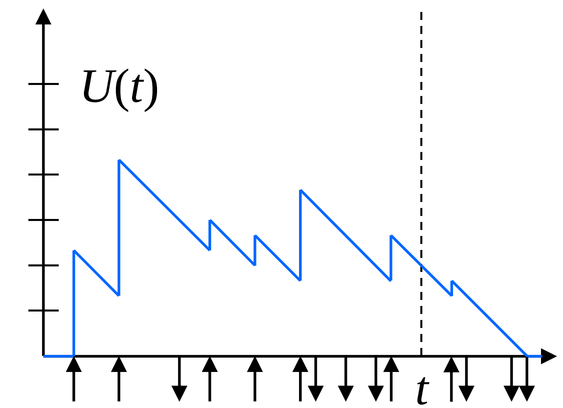
- $U(t)$ = sum of remaining service times of all customers in system at time t
  = total workload at time t
  = unfinished work process

$$U(t) := \sum_{i=1}^{A(t)} S_i - \int_0^t 1_{\{U(s)>0\}}\, ds$$

- **Proposition:**
  The unfinished work process $U(t)$ is the same for all work-conserving disciplines π. In addition,

$$\{X^{\pi}(t)=0\} = \{U(t)=0\}$$



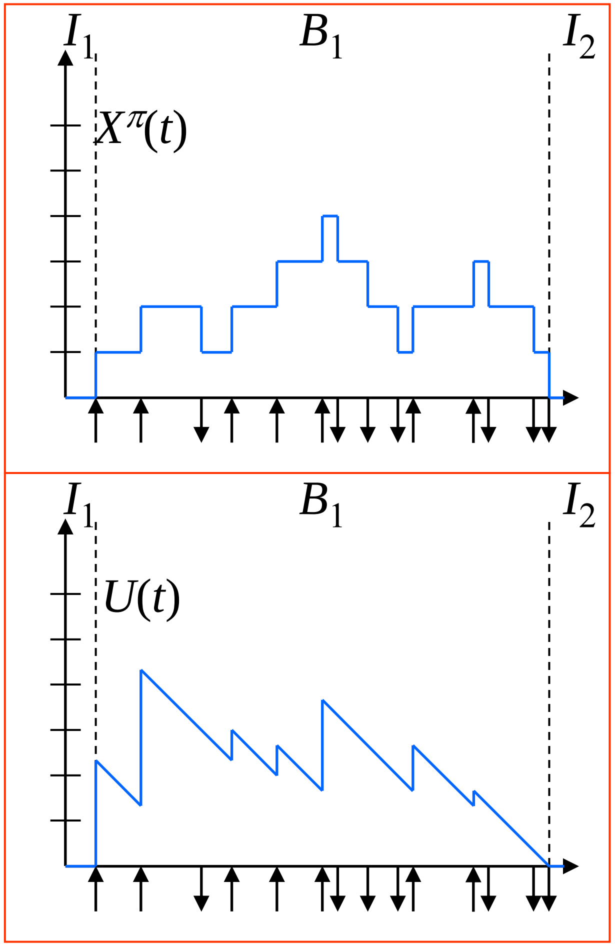


Busy and idle periods (1)

- **Definition:**
  The server is busy whenever the system is non-empty, and idle otherwise.
  A busy [idle] period is an unbroken interval during which the server is busy [idle].

- I_n = length of nth idle period
- B_n = length of nth busy period
- C_n = length of nth busy cycle
- N_n = number of customers served in the nth busy period

$$C_n = I_n + B_n$$

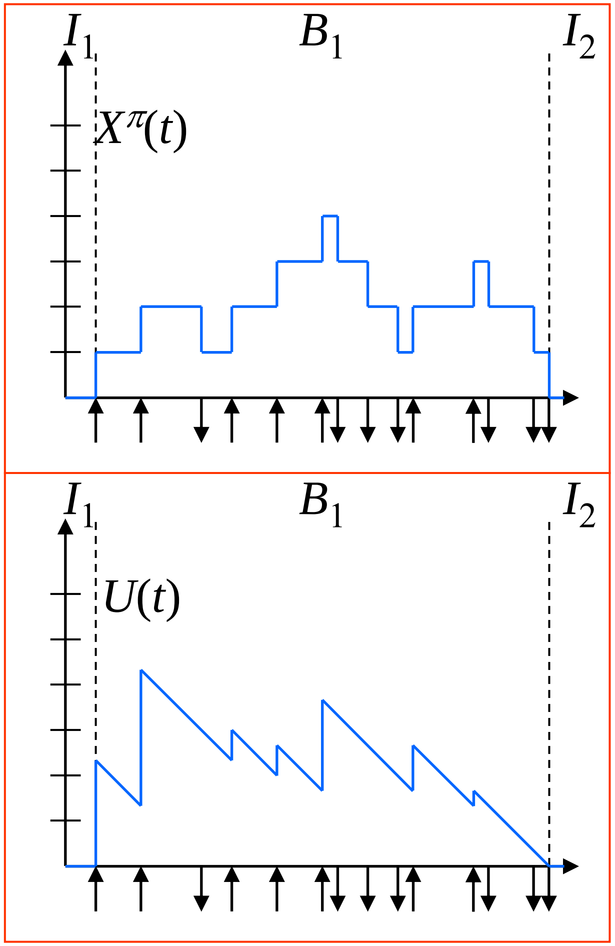


Busy and idle periods (2)

- Assume that $U(0) = 0$.

$$\gamma_1 := 0$$
$$I_1 := \inf\{t > 0 \mid U(t) > 0\} = \alpha_1$$
$$B_1 := \inf\{t > I_1 \mid U(t) = 0\} - I_1$$
$$C_1 := I_1 + B_1$$
$$N_1 := A(C_1)$$

$$\gamma_n := \gamma_{n-1} + C_{n-1}$$
$$I_n := \inf\{t > \gamma_n \mid U(t) > 0\} - \gamma_n$$
$$B_n := \inf\{t > \gamma_n + I_n \mid U(t) = 0\} - \gamma_n - I_n$$
$$C_n := I_n + B_n$$
$$N_n := A(\gamma_n + C_n) - A(\gamma_n)$$



Busy and idle periods (3)

- **Proposition:**
  Idle periods I_n, busy periods B_n, busy cycles C_n, and the number of customers N_n served in a busy period are the same for all work-conserving service disciplines π.

- **Proposition:**
  The busy cycles C_n constitute a renewal sequence (γ_n). In addition, $X^\pi(t)$ and $U(t)$ are regenerative processes with respect to the renewal sequence (γ_n) for all work-conserving service disciplines π.

- **Proposition:** Assume that $\rho < 1$.
  (i) Idle periods I_n are IID with mean

$$E[I] = 1/\lambda$$

  (ii) Busy periods B_n are IID with mean

$$E[B] = \frac{E[S]}{1-\rho}$$

  (iii) Busy cycles C_n are IID with mean

$$E[C] = \frac{1/\lambda}{1-\rho}$$

  (iv) Number of customers N_n served in a busy period are IID with mean

$$E[N] = \frac{1}{1-\rho}$$

Busy and idle periods (4)

- By previous propositions, the steady-state variables X^π and U,

$$P\{X^\pi \le x\} := \lim_{t\to\infty} P\{X^\pi(t) \le x\} = \frac{E[\int_0^C 1\{X^\pi(t)\le x\}dt]}{E[C]}$$

$$P\{U \le x\} := \lim_{t\to\infty} P\{U(t) \le x\} = \frac{E[\int_0^C 1\{U(t)\le x\}dt]}{E[C]}$$

are well-defined whenever the system is stable, $\rho < 1$.

- **Proposition:**
Assume that $\rho < 1$. For all work-conserving service disciplines π, we have

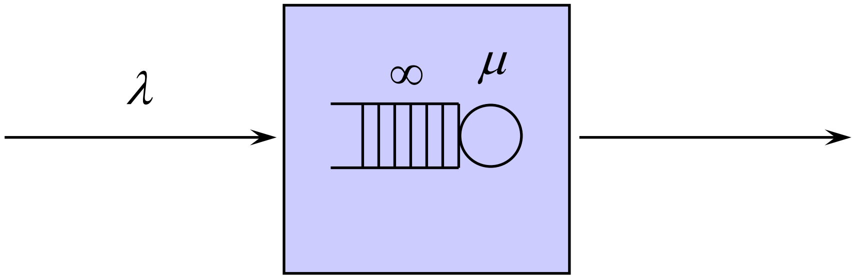
$$P\{X^\pi = 0\} = P\{U = 0\} = 1 - \rho$$

$$P\{X^\pi > 0\} = P\{U > 0\} = \rho$$

Contents

M/G/1-FIFO

- Customers arrive according to a Poisson process at rate λ
  - IID inter-arrival times
  - exponential inter-arrival time distribution with mean $1/\lambda$
- Customers are served by 1 server according to the FIFO service discipline
  - IID service times S_i
  - a general service time distribution with mean $E[S] = 1/\mu$
- There are ∞ customer places in the system



FIFO service discipline

- First In First Out (FIFO)
  - customers are served one-by-one until completion
  - service in the arrival order ("ordinary queue")
  - the customer that arrived first is served with rate μ
  - also known as First Come First Served (FCFS)

Waiting time

- In a FIFO system, the delay T_i of customer i consists of its waiting time W_i and service time S_i

$$T_i = W_i + S_i$$

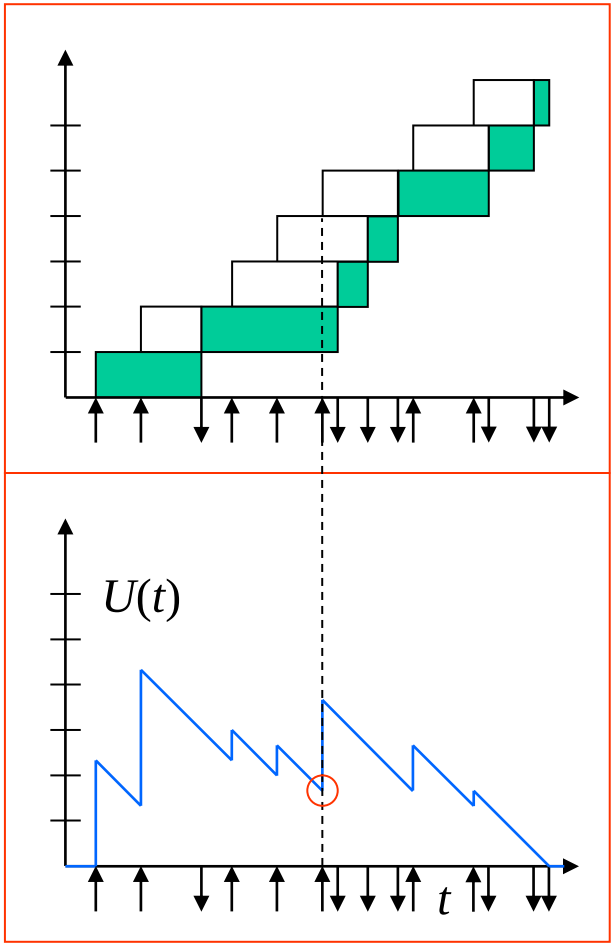
- Let Y_i^w denote the number of waiting customers that customer i sees upon its arrival,

$$Y_i^w = \max\{X(\alpha_i -) - 1, 0\}$$

- Now

$$W_i = U(\alpha_i -) = \sum_{j=1}^{Y_i^w} S_{i-j} + R(\alpha_i -)$$

where $R(t)$ denotes the remaining service time of the customer in service at time t (if any).



Remaining service time

- **Proposition:**
  Process $R(t)$ is regenerative with respect to the renewal sequence (γ_n).

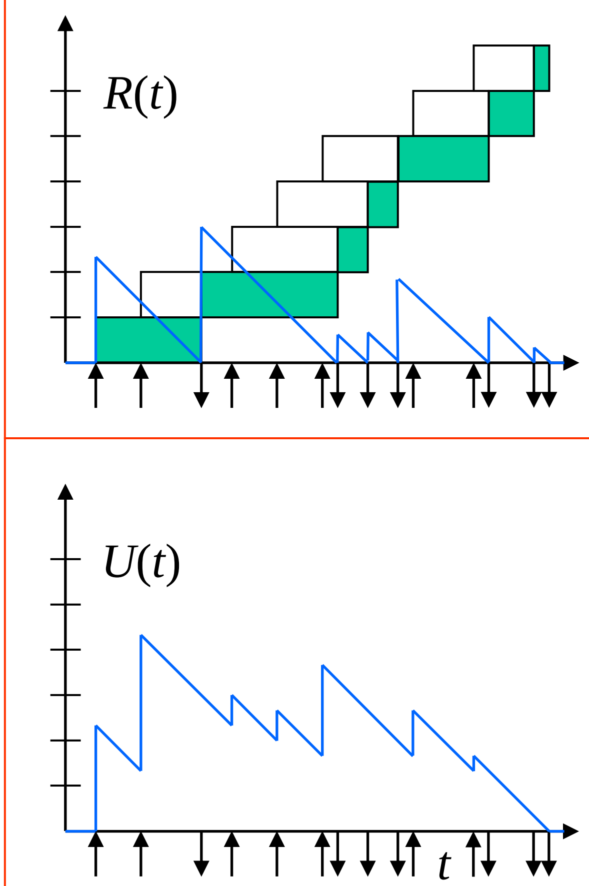
- Thus, the steady-state variable R,

$$P\{R \le x\} := \lim_{t \to \infty} P\{R(t) \le x\} = \frac{E[\int_0^C 1\{R(t) \le x\} dt]}{E[C]}$$

  is well-defined whenever $\rho < 1$.

- **Proposition:**
  Assume $\rho < 1$. Then

$$E[R] = \tfrac{\lambda}{2} E[S^2]$$



Pollaczek-Khinchin mean value formulas

- **Theorem:**
  Assume $\rho < 1$. For the M/G/1-FIFO queue, we have

$$E[W] = E[U] = \frac{\lambda E[S^2]}{2(1-\rho)}$$

$$E[T] = E[S] + \frac{\lambda E[S^2]}{2(1-\rho)}$$

$$E[X] = \rho + \frac{\lambda^2 E[S^2]}{2(1-\rho)}$$

- These are called the Pollaczek-Khinchin mean value formulas for M/G/1-FIFO

- Note that, if the mean service time $E[S]$ is kept fixed, then $E[W]$, $E[T]$, and $E[X]$ are increasing functions of the coefficient of variation $C[S]$ of the service time distribution,

$$C[S] := \sqrt{\frac{D^2[S]}{E[S]^2}} = \sqrt{\frac{E[S^2]}{E[S]^2} - 1}$$

$$\Rightarrow \quad E[S^2] = E[S]^2(1 + C^2[S])$$

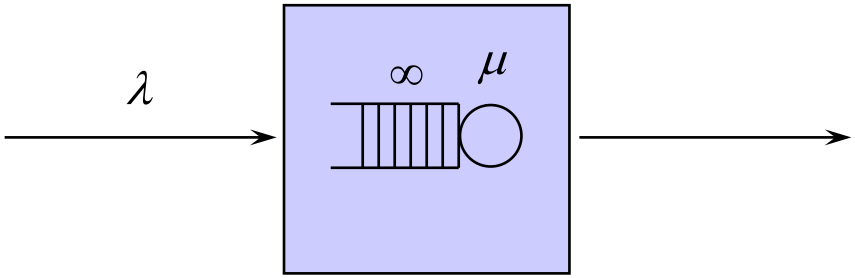
- **Examples:**
  - Erlang distribution: $C[S] < 1$
  - Exponential distribution: $C[S] = 1$
  - Hyperexponential distrib.: $C[S] > 1$
  - Pareto distribution: $C[S] > 1$

Contents

- M/G/1 with a work-conserving service discipline
- M/G/1-FIFO
- M/G/1-PS
- Performance comparison between FIFO and PS

M/G/1-PS

- Customers arrive according to a Poisson process at rate λ
  - IID inter-arrival times
  - exponential inter-arrival time distribution with mean $1/\lambda$
- Customers are served by 1 server according to the PS service discipline
  - IID service times S_i
  - a general service time distribution with mean $E[S] = 1/\mu$
- There are ∞ customer places in the system



PS service discipline

- Processor Sharing (PS)
  - customers are served simultaneously
  - the service capacity is shared evenly among all customers ("fair queue")
  - when i customers are in the system, each of them is served with rate μ/i
  - ideal version of the Round Robin (RR) service discipline

Exponential distribution

$$X \sim \mathrm{Exp}(\mu), \quad \mu > 0$$

$$S = (0, \infty)$$

$$F_X(x) := P\{X \le x\} = 1 - e^{-\mu x}$$

$$f_X(x) := \frac{d}{dx} F_X(x) = \mu e^{-\mu x}$$

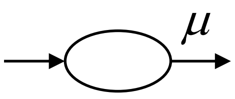
$$E[X] = \frac{1}{\mu}$$

$$E[X^2] = \frac{2}{\mu^2}$$

$$D^2[X] = \frac{1}{\mu^2}$$

$$D[X] = \frac{1}{\mu}$$

$$C[X] := \frac{D[X]}{E[X]} = 1$$



M/M/1-PS

- Let us first consider the M/M/1-PS queue
- So we assume that the service times obey the $\mathrm{Exp}(\mu)$ distribution
- In this case, the queue length process $X(t)$ is an irreducible (Markov) birth-death process with state space

$$S = \{0,1,2,\ldots\}$$

and transition rates

$$q(n,n+1) = \lambda$$

$$q(n+1,n) = (n+1)\frac{\mu}{n+1} = \mu$$

- **Proposition:**
  Assume $\rho < 1$. For the M/M/1-PS queue, the steady-state queue length distribution is

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

Erlang distribution

- IID exponential phases in a series

$$X \sim \mathrm{Erl}(K, K\mu), \quad \mu > 0$$

$$X = X_1 + \ldots + X_K, \quad X_k \sim \mathrm{Exp}(K\mu)$$

$$S = (0, \infty)$$

$$F_X(x) = 1 - \sum_{k=1}^{K} \frac{(K\mu x)^{K-k}}{(K-k)!} e^{-K\mu x}$$

$$f_X(x) = K\mu \frac{(K\mu x)^{K-1}}{(K-1)!} e^{-K\mu x}$$

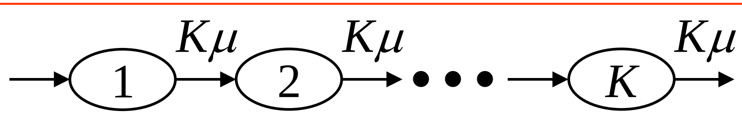
$$E[X] = \frac{1}{\mu}$$

$$E[X^2] = \frac{K+1}{K\mu^2}$$

$$D^2[X] = \frac{1}{K\mu^2}$$

$$D[X] = \frac{1}{\sqrt{K}\mu}$$

$$C[X] := \frac{D[X]}{E[X]} = \frac{1}{\sqrt{K}}$$



M/E$_K$/1-PS (1)

- Next we consider the M/E$_K$/1-PS queue
- Here we assume that the service times obey the $\mathrm{Erl}(K, K\mu)$ distribution with $K \geq 2$ exponential phases
- In this case, the queue length process $X(t)$ is no longer a Markov process, but we have to supplement the state description.
- To get a Markovian description of the system, we have to additionally keep track of the current phases of the customers.

- Let

$$N(t) = (N_1(t), \ldots, N_K(t))$$

where $N_k(t)$ refers to the total number of customers in phase k at time t

- Process $N(t)$ is an irreducible Markov process with state space

$$S = \{n = (n_1, \ldots, n_K) \mid n_k \in \{0,1,2,\ldots\}\}$$

and transition rates

$$q(n, n+e_1) = \lambda$$

$$q(n+e_k, n+e_{k+1}) = \frac{(n_k+1)K\mu}{n_1+\ldots+n_K+1}, \quad k < K$$

$$q(n+e_K, n) = \frac{(n_K+1)K\mu}{n_1+\ldots+n_K+1}$$

M/E$_K$/1-PS (2)

- **Proposition:**
  Assume $\rho < 1$. The steady-state distribution of process $N(t)$ is

$$P\{N = n\} = (1-\rho)(\rho / K)^{n_1+\ldots+n_K} \times \frac{(n_1+\ldots+n_K)!}{n_1!\ldots n_K!}, \quad n \in S$$

- Note that

$$X(t) = N_1(t) + \ldots + N_K(t)$$

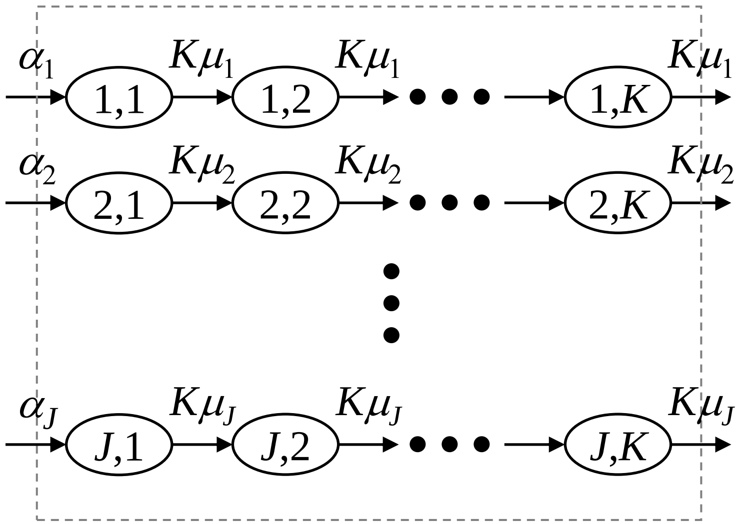
- **Corollary:**
  Assume $\rho < 1$. For the M/E$_K$/1-PS queue, the steady-state queue length distribution is

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

Phase-type distribution

- **Definition:**
  A phase-type (PH) distribution refers to the distribution of the absorption time in an absorbing finite-state Markov process

- **Examples:**
  - Exponential distribution
  - Erlang distribution
  - Hyperexponential distribution
  - Exponential distributions in series and/or parallel



M/PH/1-PS (1)

- Next we consider the M/PH/1-PS queue
- Here we assume (for simplicity) that the service times obey the phase-type distribution depicted in the previous slide with $JK \geq 2$ exponential phases
- In this case, the queue length process $X(t)$ is neither a Markov process, but we have to supplement the state description as before.
- Again, to get a Markovian description of the system, we have to additionally keep track of the current phases of the customers.

- Let

$$N(t) = (N_{1,1}(t), \ldots, N_{J,K}(t))$$

where $N_{j,k}(t)$ refers to the total number of customers in phase (j,k) at time t

- Process $N(t)$ is an irreducible Markov process with state space

$$S = \{n = (n_{1,1}, \ldots, n_{J,K}) \mid n_{j,k} \in \{0,1,2,\ldots\}\}$$

and transition rates determined from the underlying absorbing Markov process

- **Exercise:**
Determine the transition rates

M/PH/1-PS (2)

- **Proposition:**
  Assume $\rho < 1$. The steady-state distribution of process $N(t)$ is

$$P\{N = n\} = (1-\rho) \times \prod_{j=1}^{J} (\rho_j / K)^{n_{j,1}+\ldots+n_{j,K}} \times \frac{(n_{1,1}+\ldots+n_{J,K})!}{n_{1,1}!\ldots n_{J,K}!}, \quad n \in S$$

  where

$$\rho := \rho_1 + \ldots + \rho_J, \quad \rho_j := \frac{\lambda \alpha_j}{\mu_j}$$

- **Corollary:**
  Assume $\rho < 1$. For the M/PH/1-PS queue, the steady-state queue length distribution is

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

- Note that

$$X(t) = N_{1,1}(t) + \ldots + N_{J,K}(t)$$

Insensitive queue length distribution in M/G/1-PS

- The generalization of the previous result is based on the known fact that any service time distribution can be approximated (with an arbitrary precision) by a phase-type distribution.
- Since the queue length distribution remains the same for any service time distribution with the same mean $E[S]$, the steady-state queue length distribution of the PS service discipline is said to be insensitive to the service time distribution.
- Interestingly, the mean sojourn time $E[T]$ in the M/G/1-PS queue equals the mean busy period $E[B]$.

- **Theorem:**
  Assume $\rho < 1$. For the M/G/1-PS queue, the steady-state queue length distribution is

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

  with

$$E[X] = \frac{\rho}{1-\rho}$$

- **Corollary:**
  Assume $\rho < 1$. For the M/G/1-PS queue, the mean steady-state sojourn time is

$$E[T] = \frac{E[S]}{1-\rho}$$

Contents

Performance comparison between FIFO and PS

- **Proposition:**

  Assume $\rho < 1$. For the M/G/1 queue, we have

$$E[X^{\text{FIFO}}] < E[X^{\text{PS}}] \quad \Leftrightarrow \quad E[T^{\text{FIFO}}] < E[T^{\text{PS}}] \quad \Leftrightarrow \quad C[S] < 1$$

$$E[X^{\text{FIFO}}] = E[X^{\text{PS}}] \quad \Leftrightarrow \quad E[T^{\text{FIFO}}] = E[T^{\text{PS}}] \quad \Leftrightarrow \quad C[S] = 1$$

$$E[X^{\text{FIFO}}] > E[X^{\text{PS}}] \quad \Leftrightarrow \quad E[T^{\text{FIFO}}] > E[T^{\text{PS}}] \quad \Leftrightarrow \quad C[S] > 1$$

  where $C[S]$ refers to the coefficient of variation of the service time distribution

Summary

- M/G/1 with a work-conserving service discipline
  - traffic load, throughput, stability, unfinished work, busy period, busy cycle
- M/G/1-FIFO
  - FIFO, remaining service time, Pollaczek-Khinchin mean value formulas
- M/G/1-PS
  - PS, phase method, insensitivity
- Performance comparison between FIFO and PS
  - FIFO better than PS if the service time distribution less variable than exp