EP2200 Queueing theory and teletraffic systems

Lecture 8

Semi-Markovian systems
The Method of Stages

Viktoria Fodor
KTH EES/LCN

Semi-Markovian system

- Advantages with M/M/*
  - The interarrival time and the service time distribution is memoryless
  - The state can be defined by the number of customers in the system – since the time of the next state transition does not depend on the time already spent in the state
- Applicability for real systems
  - The arrival process is often Poisson (large number of potential customers)
  - The service process is often **not** memoryless
    - E.g., packet size distribution on the Internet, file size distribution
    - The future of the system depends on the elapsed service time

Semi-Markovian system and Method of Stages

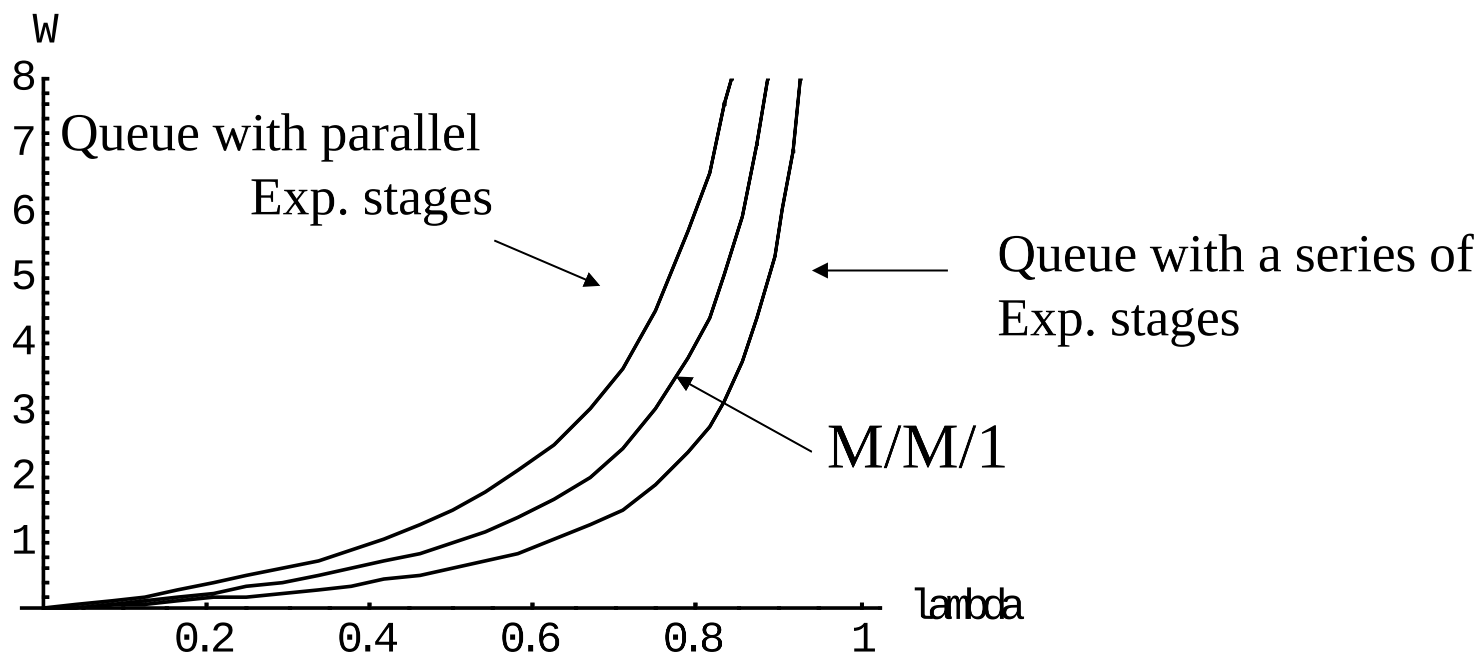
- Ways to handle the non-exponential service time

  1. Look at distributions consisting of several exponentially distributed stages in series or in parallel - Method of stages (this lecture)

  2. Describe the system only at specific points of time (e.g., end of service), where the Markovian property is satisfied – Semi-Markovian systems (semi half)

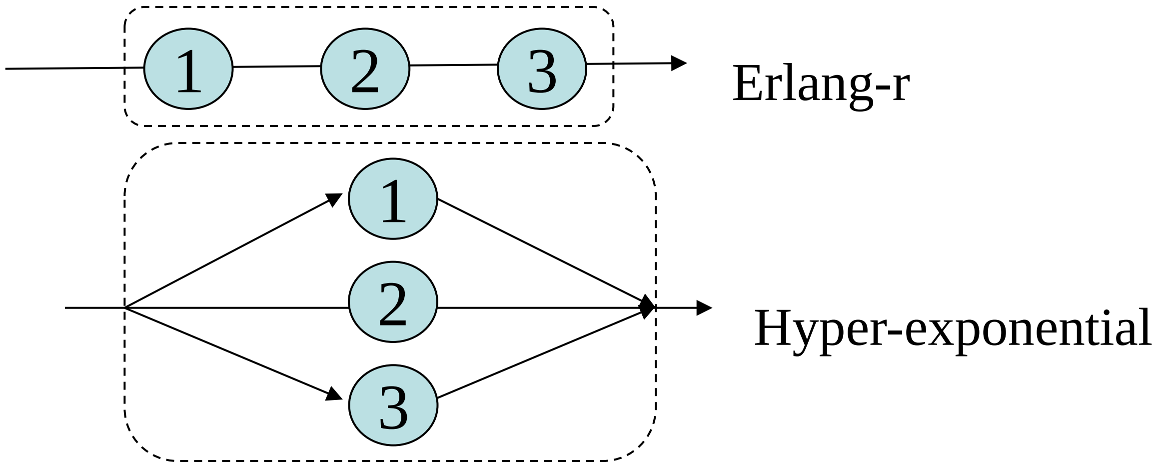
Method of Stages

- Use distributions that are composed of Exponential distributions

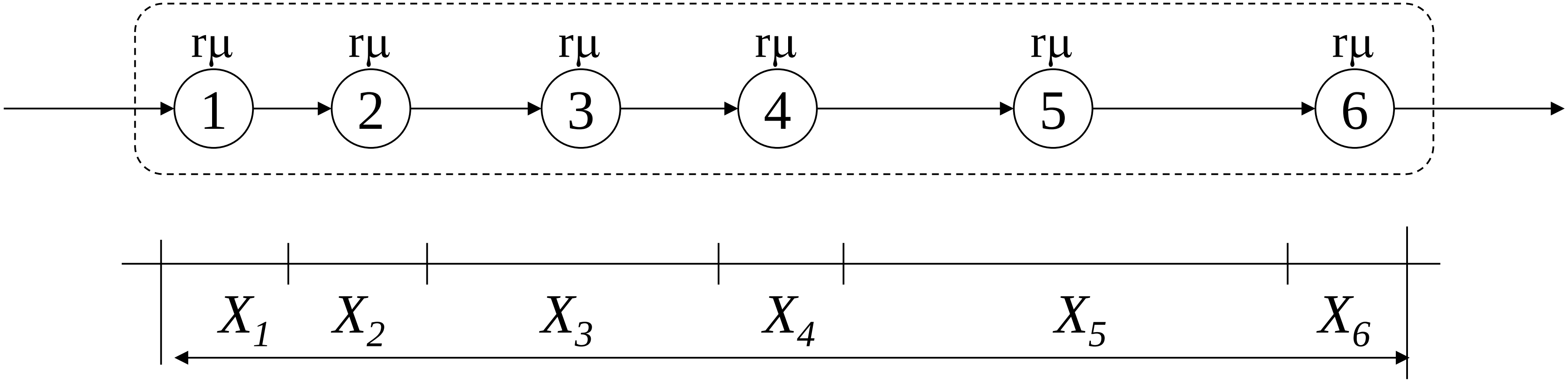


Method of Stages

- Each service stage is Exponential
- Series of stages: the customer has to finish r service stages before the next customer can enter the server $\rightarrow$ Erlang-r service time distribution
- Parallel (or alternative) stages: the customer selects one server randomly, but only one customer can be in the service unit $\rightarrow$ Hyper-exponential service time distribution (linear combination of Exp. distributions)



Erlang-r server (E_r)



X is E_6 distributed, E[X]=r/(rμ)=1/μ

- Goal: service time with average $E[X] = \bar{x} = {}^{1}/_{\mu}$
- Since $E[X] = \sum E[X_i]$ if $X = \sum X_i$, we select:
  - X_i a stochastic variable with Exponential distribution $b(x_i) = r\mu e^{r\mu x_i}$
  - $X = \sum_{i=1}^{r} X_i$, X_i, X_j independent, identically distributed
  - *That is, X* is Erlang-r distributed

Erlang-r server (E_r)

- For each exponential stage:

$$b(x_i) = r\mu e^{-r\mu x_i}$$

$$\left.\begin{aligned} E[X_i] &= \frac{1}{r\mu} \\ V[X_i] &= \left(\frac{1}{r\mu}\right)^2 \end{aligned}\right\} \quad C_{x_i}^2 \stackrel{\Delta}{=} \frac{V[X_i]}{E[X_i]^2} = 1$$

(coefficient of variation)

- For the service time:

$$X = X_1 + X_2 + \ldots X_r, \quad X_i, X_j \text{ i.i.d.}$$

$$b(x) = \frac{(r\mu)^r x^{r-1}}{(r-1)!} e^{-r\mu x} \quad (Erlang - r)$$

$$\left.\begin{aligned} E[X] &= rE[X_i] = \frac{1}{\mu} \\ V[X] &= rV[X_i] = \frac{1}{r\mu^2} \end{aligned}\right\} \quad C_x^2 = \frac{1}{r} \quad < 1$$

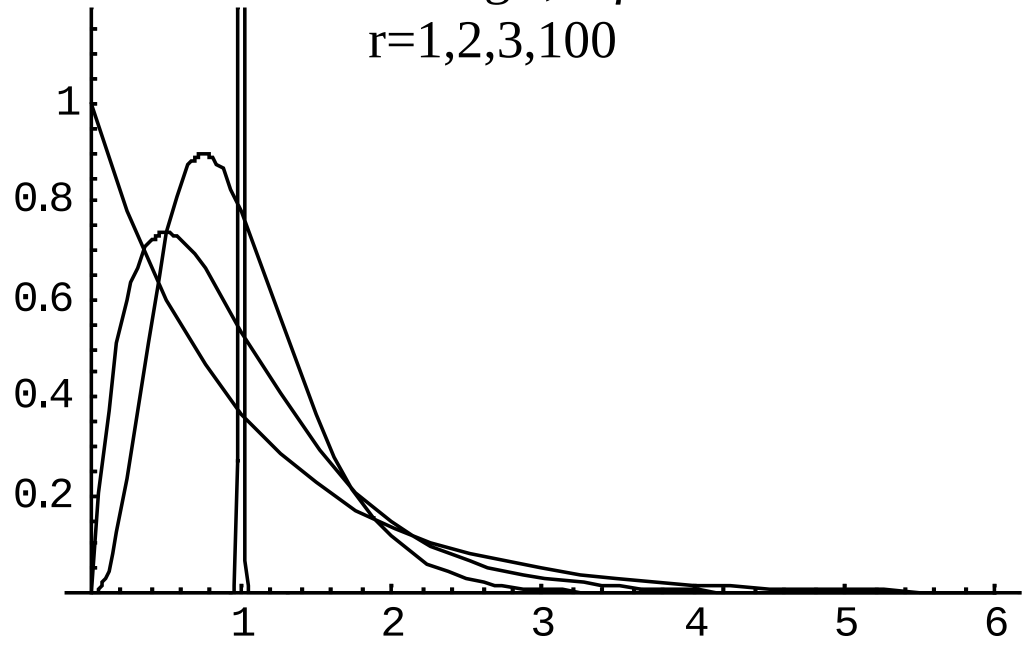
Erlang-r server (E_r)

$$X = X_1 + X_2 + \ldots X_r$$

$$L(b(x)) = \left(\frac{r\mu}{s + r\mu}\right)^r, \quad b(x) = \frac{(r\mu)^r x^{r-1}}{(r-1)!} e^{-r\mu x}$$

$$\left.\begin{aligned} E[X] &= rE[X_i] = \frac{1}{\mu} \\ V[X] &= rV[X_i] = \frac{1}{r\mu^2} \end{aligned}\right\} \quad C_x^2 = \frac{1}{r} < 1$$

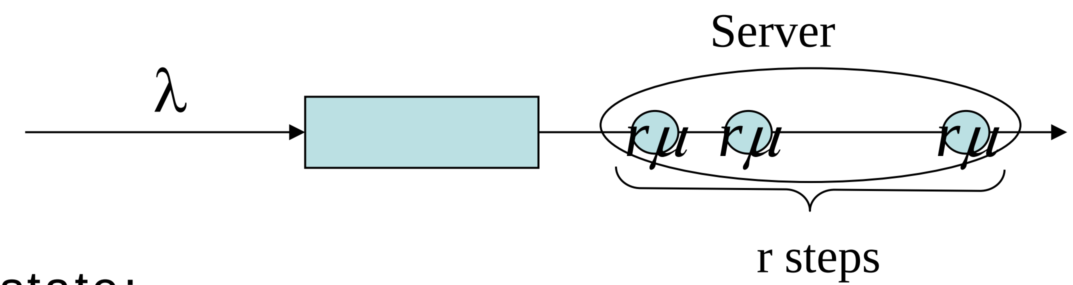
Erlang-r, 1/μ=1
r=1,2,3,100



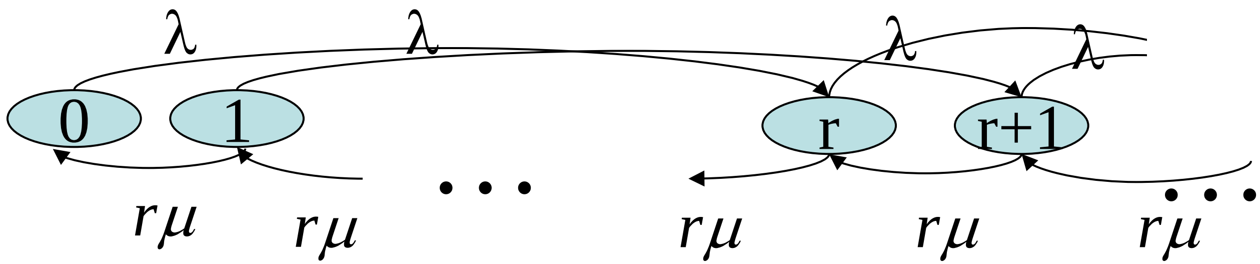
- As r→∞, V[X]→0, which means deterministic service time!

The $M/E_r/1$- queue

- If the system to be modeled has serial service or the service distribution has $C_x^2 < 1$ – approximate with Erlang-r

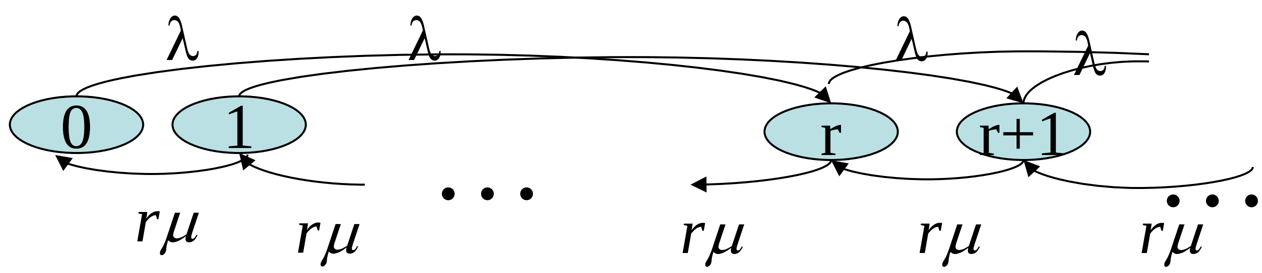


- System state:
  - {number of remaining service stages, number of customers}, or
  - number of remaining service stages + r*number of waiting customers
- The system can be modeled as a Markov chain



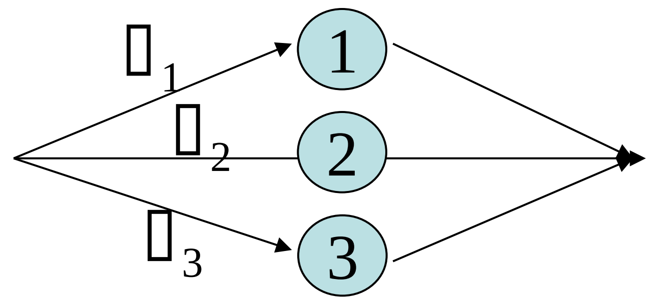
The $M/E_r/1$- queue

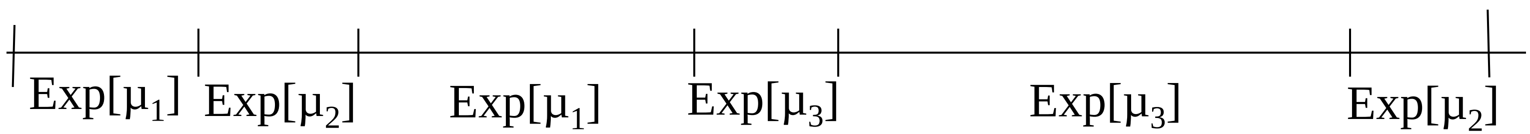
- System state:
  - number of remaining service stages + r*number of waiting customers
- Number of customers in the system in state i: $N_i = \lceil i/r \rceil$
- State probability distribution with z-transforms (Kleinrock p.127-128)
  - (not exam material)
- But, the followings hold:
  - PASTA
  - Little: $N_s=\lambda x=\lambda/\mu$=Utilization
  - For r=1: M/M/1, for r=∞: M/D/1
  - You will have to be able to calculate state probabilities and performance measures for limited buffer systems (e.g., *$M/E_2/1/3$*)!
  - Average performance for $M/E_r/1$ with general forms of M/G/1



Hyper-exponential server (H_r)

- r exponential servers with different μ_i-s
- Server *i* is chosen with the probability α_i
  - E.g., different types of packets intermixed
  - service time distribution is the linear combination (mixture) of Exp distributions





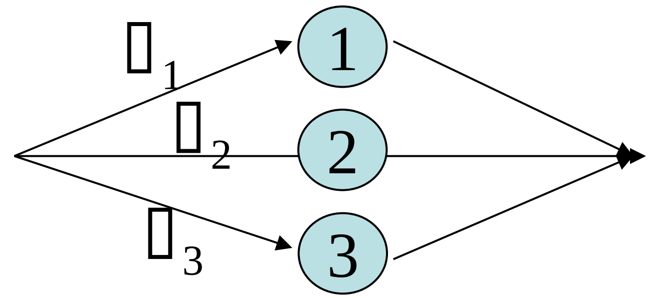
a possible sequence of service of 6 customers

$$b(x_i) = \mu_i e^{-\mu_i x}$$

$$b(x) = \alpha_1 \mu_1 e^{-\mu_1 x} + \ldots + \alpha_R \mu_R e^{-\mu_R x}, \quad \sum \alpha_i = 1$$

The hyper-exponential server (H_r)

- r exponential servers with different μ-s $\quad B(x_i) = 1 - e^{-\mu_i x}$
- *Server i* is chosen with the probability α_i



$$b(x) = \alpha_1 \mu_1 e^{-\mu_1 x} + \ldots + \alpha_R \mu_R e^{-\mu_R x}, \quad \sum \alpha_i = 1$$

$$L(b(x)) = \sum_{i=1}^{r} \alpha_i \frac{\mu_i}{s + \mu_i}$$

$$E[X] = \sum_i \frac{\alpha_i}{\mu_i}$$

$$E[X^2] = \sum_i \alpha_i \frac{2}{\mu_i^2}$$

$$V[X] = E[X^2] - E[X]^2$$

$$C_{x_i}^2 = \frac{V[X_i]}{E[X_i]^2} = \frac{E[X_i^2] - E[X_i]^2}{E[X_i]^2} = \frac{E[X_i^2]}{E[X_i]^2} - 1$$

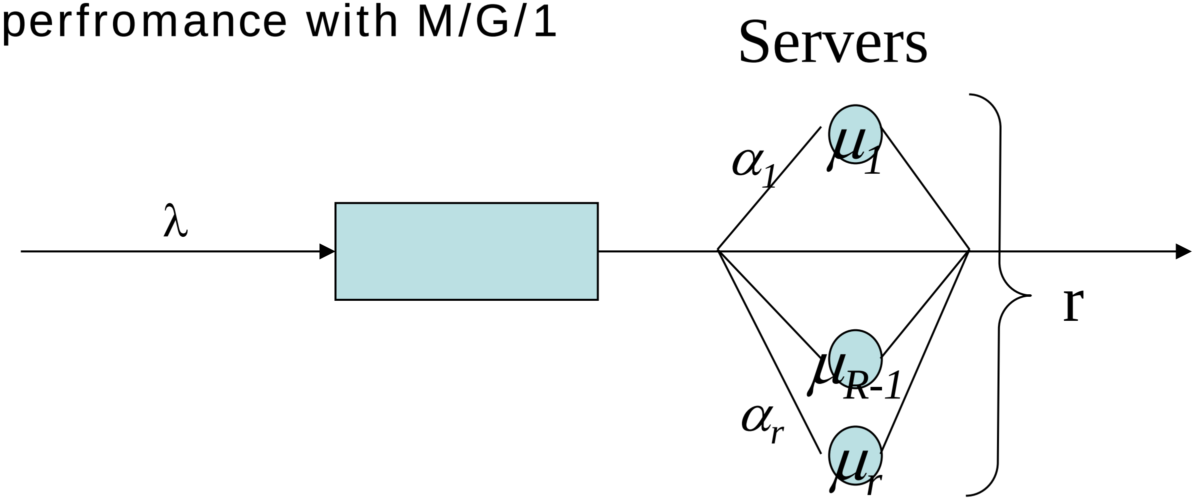
$$C_x^2 = \frac{E[X^2]}{E[X]^2} - 1 \quad \geq 1$$

- *For given coefficient of variation 2R-1* free parameters in total
  - *R-1* of α_i and *R* of μ_i

The M/H_r/1 queue

- If there are different service needs randomy intermixed
  - E.g., packet size distribution

  or if the service time distribution has $C_x^2>1$ – approximate with H_r
- The state represents the number of customers in the system and the actual server used (only one server used at a time!)
  - complicated Markov-chain (see notes from class)
  - you have to be able to handle it for limited buffer systems
  - Little, PASTA holds
  - Average perfromance with M/G/1



The M/H_r/1 queue

- Example problem: Packets of two types arrive to a multiplexer intermixed. The total arrival intensity is λ.

  Packet of type 1 arrives with probability α_1, its transmission time is exponential with parameter μ_1.

  Packet of type 2 arrives with probability α_2, its transmission time is exponential with parameter μ_2.

  There is no buffer.
- Give:
  - Kendall, Markov-chain
  - Balance equations (no need to calculate the sate probabilities...)
  - P(packet type 1 under transmission)
  - P(packet blocked)
  - Utilization

Method of stages for the arrival process

- Non-exponential inter-arrival times can be modeled similarly
- E.g., round-robin customer spreading: $E_r/M/1$

Semi-markovian system
Method of stages - Summary

- Ways to handle the non-exponential service / inter-arrival time
  - Method of stages: look at distributions consisting of several exponentially distributed stages in series or in parallel
  - Describe the system in specific points of time (end of service) – M/G/1, embedded Markov-chains

- Erlang-r service / inter-arrival times
  - series of stages in the real system, or
  - has distribution with $C_x^2 < 1$
  - can be modeled with Markov-chain
    state: r*number of customers + number of stages left from service
- Hyper-exponential service /inter-arrival times
  - parallel stages in the real system
  - has distribution with $C_x^2 > 1$
  - can be modeled with Markov chain
    state has 2 parameters: number of customers and server used