Analysis of System Performance

IN2072

Chapter 4 – Analysis of Markov Systems (Part 2/2)

Dr. Alexander Klein
Prof. Dr.-Ing. Georg Carle

Chair for Network Architectures and Services
Department of Computer Science
Technische Universität München
http://www.net.in.tum.de

Markov Systems

- Content:
  - Waiting system M/M/n-∞
    - Erlang-C equation
    - Waiting probability
    - State probabilities
    - Multiplexing gain
    - Waiting time distribution
  - Loss system M/M/n-s (Finite number of sources)
    - State probabilities
    - Blocking probability
    - Engset equation

Analysis of System Performance

IN2072

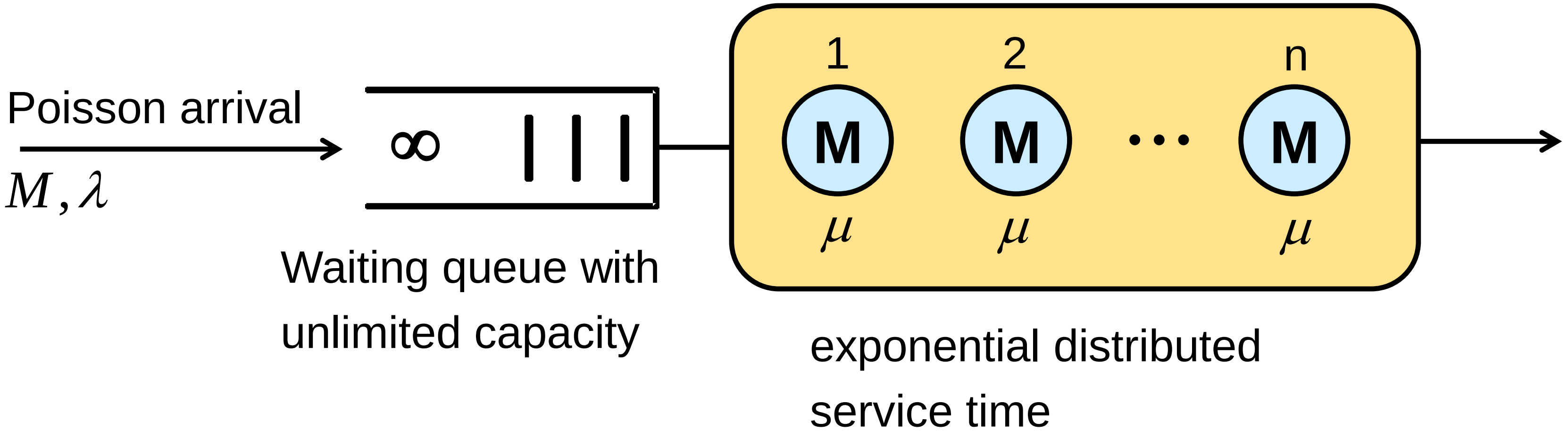
Chapter 4 – Analysis of

Markov Systems (Part 2/2)

M / M / n – Waiting system

M / M / n – Waiting system

- Model and parameter description:


Poisson arrival
M, λ
∞
Waiting queue with unlimited capacity
1
2
n
M
M
…
M
μ
μ
μ
exponential distributed service time

- Model and parameter description:
  - M / M / n – ∞ (No jobs are blocked!)
  - Arrival process is a Poisson process with an exponential distributed inter-arrival time A
  - Service time B is also exponential distributed
  - Jobs that arrive at a point in time when all service units are busy, are queued and served in FIFO order as soon as a free serving unit is available.

M / M / n – Waiting system

- Arrival process:

  Arrival rate λ

  Average number of arriving jobs per time unit.

$$A(t) = P(A \le t) = 1 - e^{-\lambda t}, \qquad E[A] = \frac{1}{\lambda}$$

- Service process:

  Service rate μ

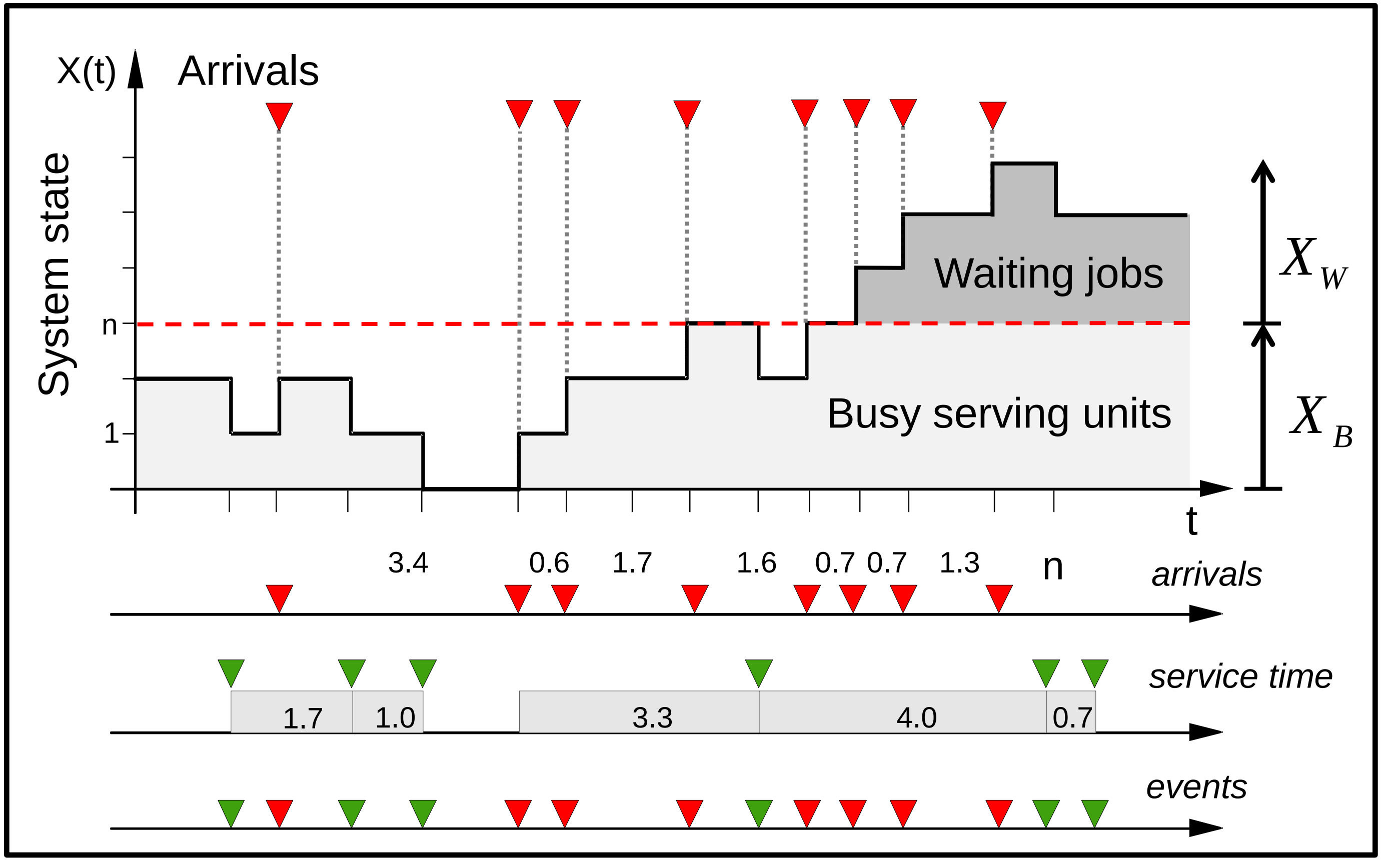
  Average number of service completions per time unit. (assuming a service unit only has two states – idle or busy)

$$B(t) = P(B \le t) = 1 - e^{-\mu t}, \qquad E[B] = \frac{1}{\mu}$$

- System:
  - Waiting system
  - Waiting queue with unlimited capacity
  - Queuing strategy – First In First Out (FIFO)

M / M / n – Waiting system

- State space:
  - Random variable $X(t)$ describes the number of (waiting and currently served) jobs in the system.
  - State process is state discrete and time continuous stochastic process



M / M / n – Waiting system

- State space:
  - Random variable $X_W(t)$ describes the number of waiting jobs in the system at the time of observation t.
  - Random variable $X_B(t)$ represents the number of served jobs at the time of observation t.

⇨ $X_W(t) = 0$ if $X_B(t) \leq n$

⇨ Number of jobs in the system: $X(t) = X_B(t) = X_W(t) = 0$

M / M / n – Waiting system

- Characteristics.
  - Utilization of service units: $a = \frac{\lambda}{\mu} = \lambda \cdot E[B]$

⇨ The utilization of the service units is identical with the offered load since no jobs are blocked.

  - Utilization of a single service unit: $\rho = \frac{a}{n} = \frac{\lambda}{n\mu}$

  - Stationary criteria: $a < n \quad \vee \quad \rho < 1$

⇨ The system becomes instable if the average number of arrivals is larger than the average number of served jobs since the waiting queue would steadily increase.

M / M / n – Waiting system

- Description:
  - State $X(t)$ is incremented if a new job arrives.
  - State $X(t)$ is decremented if a service is completed.

❗ Due to the memory-less characteristics of the arrival and the service process, the system is memory-less at any time of the process development.

- Transient phase:
  - The system starts in state $X(0)$ from which it develops through an instationary phase until it reaches a stationary state.
  - The state probabilities do not change any further as soon as the stationary state is reached which allows us to remove the time dependency of variables X_B, X_W and X.

- State probabilities:

$$\Rightarrow \quad x(i) = P(X(t) = i) = P(X = i), \qquad i = 0,1,\ldots,n$$

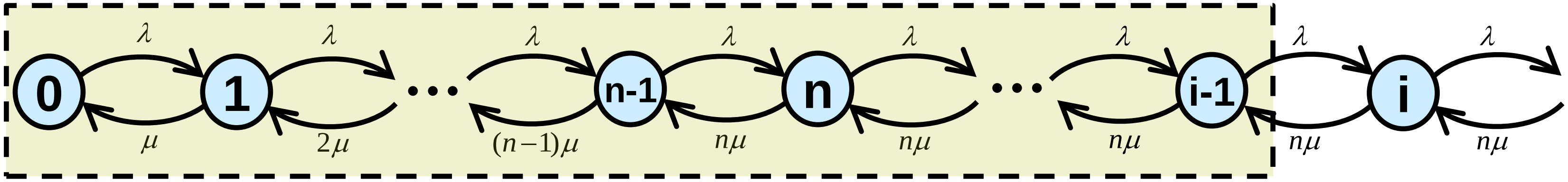
M / M / n – Waiting system

Arrival event:

- According to the definition of a Poisson process the transition from $[X=i]\rightarrow[X=i+1]$ occurs with rate λ if the system is in state $x(i), \quad i=0,1,\ldots,\infty.$

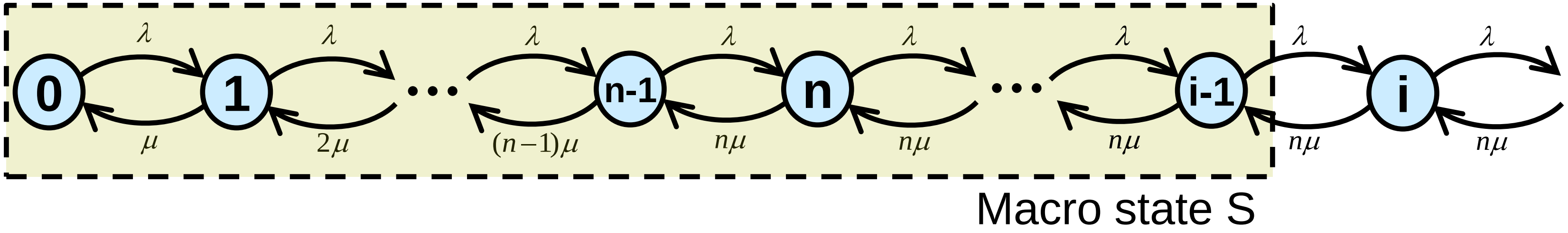
Service completion event:

- Sytem in state $x(i), \quad i\le n:$
  - i service units are busy / i jobs are currently served.
  - The transition from $[X=i]\rightarrow[X=i-1]$ occurs with rate $i\mu, \quad i=1,\ldots,n$
  - No jobs are waiting.
- System in state $x(i), \quad i>n:$
  - All n service units are busy.
  - i-n jobs are are waiting.
  - The transition from $[X=i]\rightarrow[X=i-1]$ occurs with rate $n\mu, \quad i=n+1,\ldots,\infty$



Macro state S

M / M / n – Waiting system



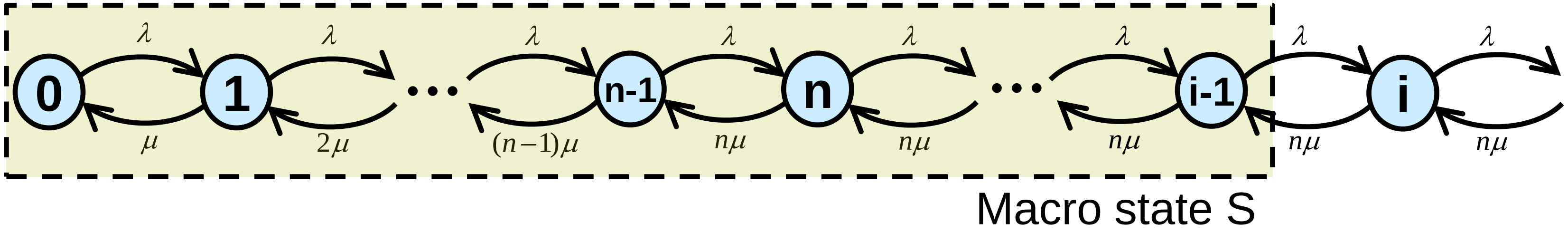
Equilibrium state of the system of equations of the **micro states S** is given by:

$$\lambda x(i-1) = i\mu x(i) \qquad i = 1,2,\ldots,n$$

$$\lambda x(i-1) = n\mu x(i) \qquad i = n+1, n+2,\ldots$$

$$\sum_{i=0}^{\infty} x(i) = 1$$

M / M / n – Waiting system



This system of equations can be resolved by succesive insertion of the micro states.

$$\Rightarrow \quad x(i) = \begin{cases} x(0)\cdot\dfrac{a^i}{i!} & i = 0,1,\ldots n \\ x(0)\cdot\dfrac{a^n}{n!}\left(\dfrac{a}{n}\right)^{i-n} = \underbrace{x(n)\rho^{i-n}}_{\text{Geometric tail}} & i > n \end{cases}$$

Geometric tail

M / M / n – Waiting system

- Idle state probability:

$$\Rightarrow \quad x(0)^{-1} = \sum_{k=0}^{n-1} \frac{a^k}{k!} + \frac{a^n}{n!} \sum_{k=0}^{\infty} \rho^k \qquad \rho = \frac{a}{n} = \frac{\lambda}{n\mu}$$

- Stationary condition:

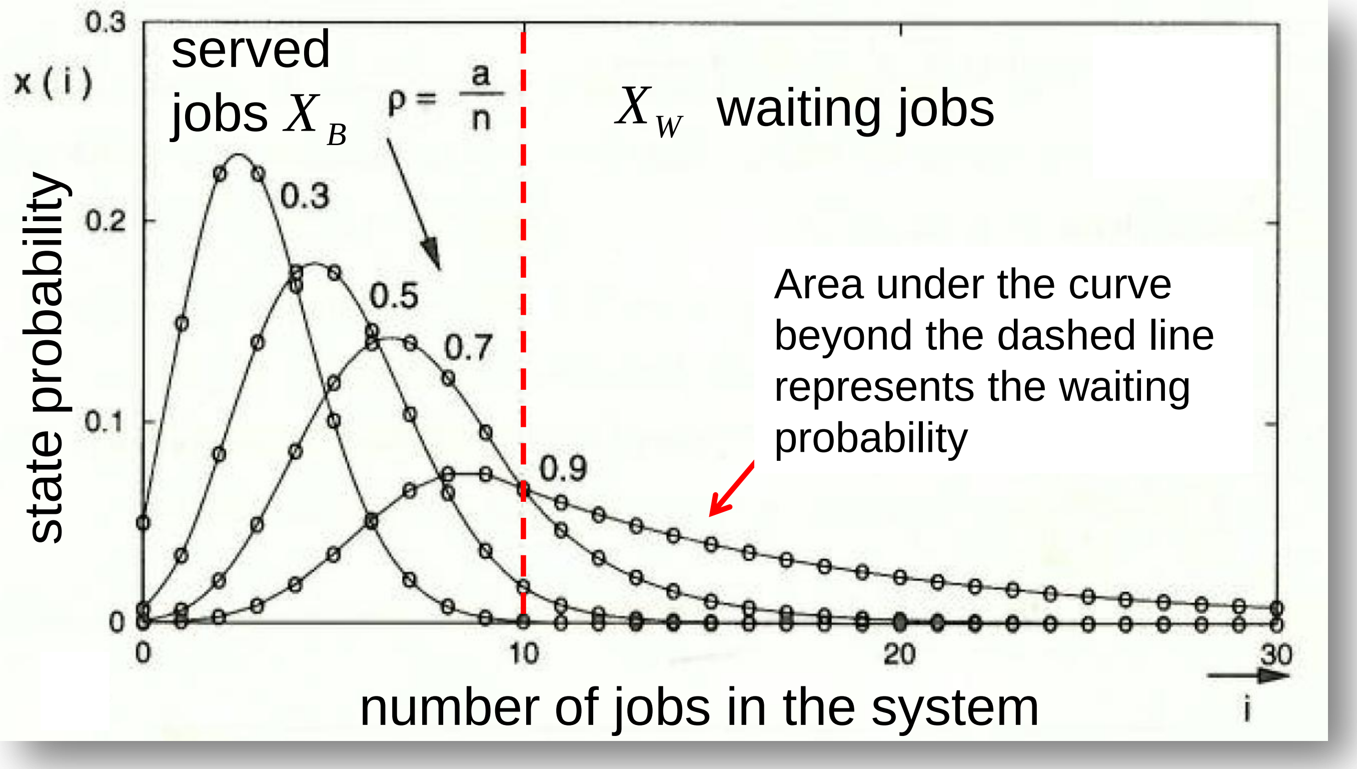
$$\Rightarrow \quad a < n \quad \rightarrow \quad \rho < 1$$

$$\Rightarrow \quad x(0)^{-1} = \sum_{k=0}^{n-1} \frac{a^k}{k!} + \frac{a^n}{n!} \cdot \frac{1}{1-\rho}$$

Picture taken from Tran-Gia, Einführung in die Leistungsbewertung und Verkehrstheorie

M / M / n – Waiting system

- State probability – M / M / 10 - ∞

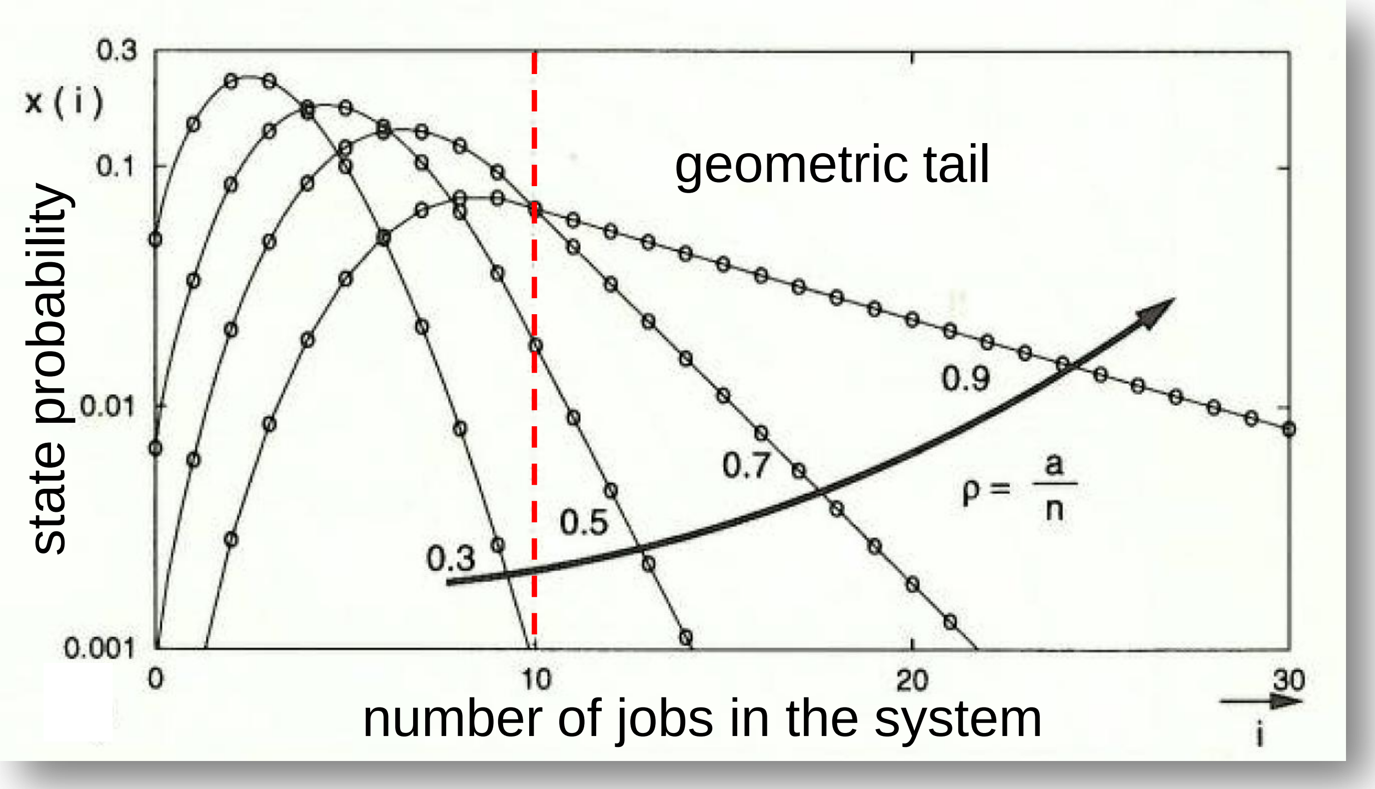


Area under the curve beyond the dashed line represents the waiting probability

Picture taken from Tran-Gia, Einführung in die Leistungsbewertung und Verkehrstheorie

M / M / n – Waiting system

- State probability – M / M / 10 - ∞



Picture taken from Tran-Gia, Einführung in die Leistungsbewertung und Verkehrstheorie

M / M / n – Waiting system

- Waiting probability:

  An arriving job has to wait if all n service units are busy at the time of arrival. Thus, the waiting probability is given by the sum of the state probabilities of the states $x(i), \quad i = n, n+1, \ldots, \infty.$

⇨ $$p_W = \sum_{i=n}^{\infty} x(i) = x(n) \sum_{i=0}^{\infty} \rho^i = x(n) \cdot \frac{1}{1-\rho}$$

❗ $\rho < 1$

With $x(n)$ from previous equation:

➡ $$p_W = \frac{\frac{a^n}{n!} \cdot \frac{1}{1-\rho}}{\sum_{i=0}^{n-1} \frac{a^i}{i!} + \frac{a^n}{n!} \cdot \frac{1}{1-\rho}}$$

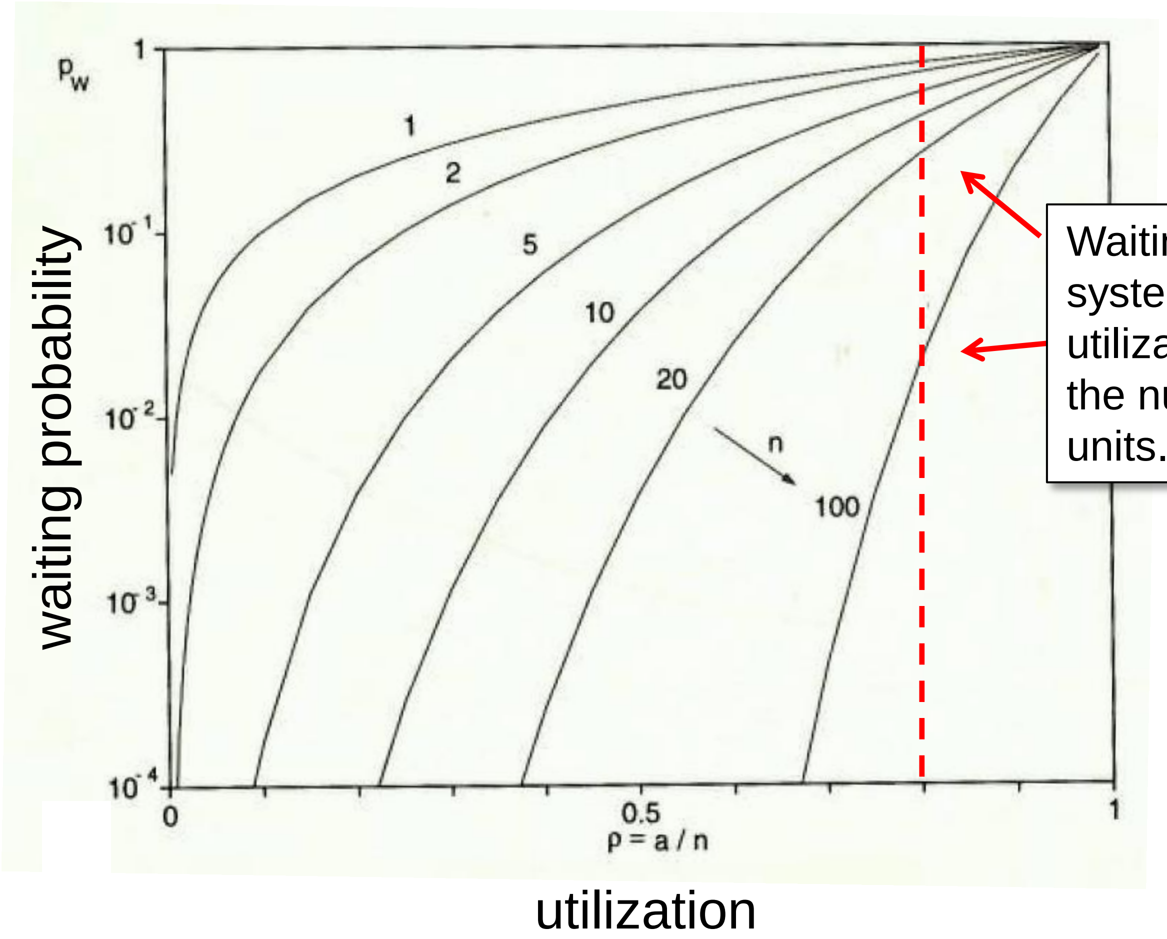
Erlang-C equation

(Erlang-Warteformel)

Picture taken from Tran-Gia, Einführung in die Leistungsbewertung und Verkehrstheorie

M / M / n – Waiting system

- Waiting probability – M / M / n - ∞



Waiting probability for a system with 80% utilization depending on the number of service units.

Picture taken from Tran-Gia, Einführung in die Leistungsbewertung und Verkehrstheorie

M / M / n – Waiting system

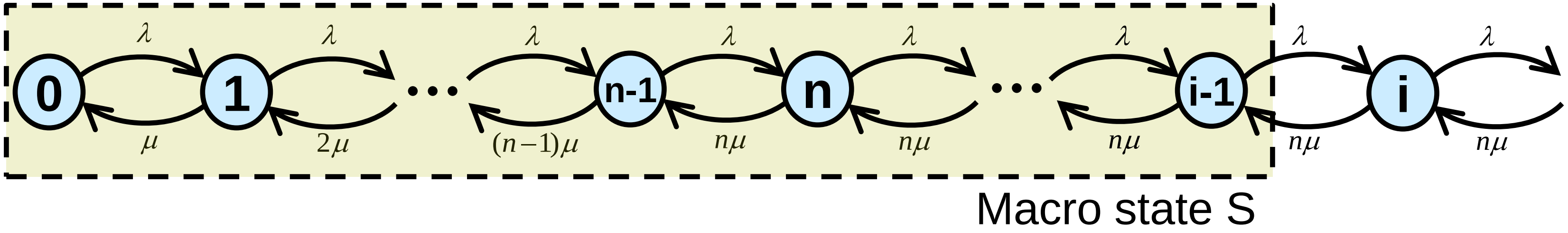
- Multiplexing gain:

⇨ The multiplexing gain converges for large values of n.

⇨ For constant utilization ρ, the waiting probability decreases if the number of service units is increased.

⇨ System design is always a trade-off between efficient use of available resources and their costs! (also true for waiting systems)

M / M / n – Waiting system



- Traffic load:

  Describes the average number of busy service units.

$$\Rightarrow \quad Y = E[X_B] = \sum_{i=0}^{n-1} i \cdot x(i) + n \sum_{i=n}^{\infty} x(i) = a = \frac{\lambda}{\mu}$$

M / M / n – Waiting system

- Average waiting queue length:

$$\Omega = E[X_W]$$

$$= \sum_{i=n}^{\infty} (i-n) \cdot x(i) = \sum_{i=n}^{\infty} (i-n) \cdot x(n) \cdot \rho^{i-n}$$

$$= x(n) \sum_{i=0}^{\infty} i \cdot \rho^{i} = x(n) \frac{\rho}{(1-\rho)^2} = x(0) \cdot \frac{a^n}{n!} \cdot \frac{\rho}{(1-\rho)^2}$$

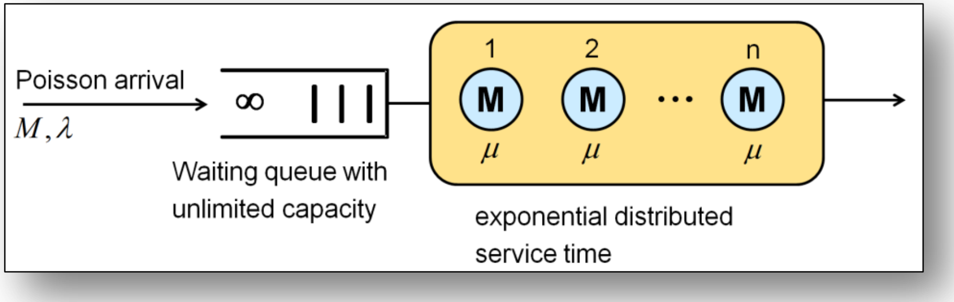
$$= p_W \cdot \frac{\rho}{1-\rho}$$

M / M / n – Waiting system

- Average waiting time:

For performance evaluation of systems it is necessary to distinguish between average waiting time of all jobs $E[W]$ and the average waiting time of waiting jobs $E[W_I]$.

! Depending on the utilization of the system and the number of service units, the average waiting time of waiting jobs can be much higher than the average waiting time of all jobs!

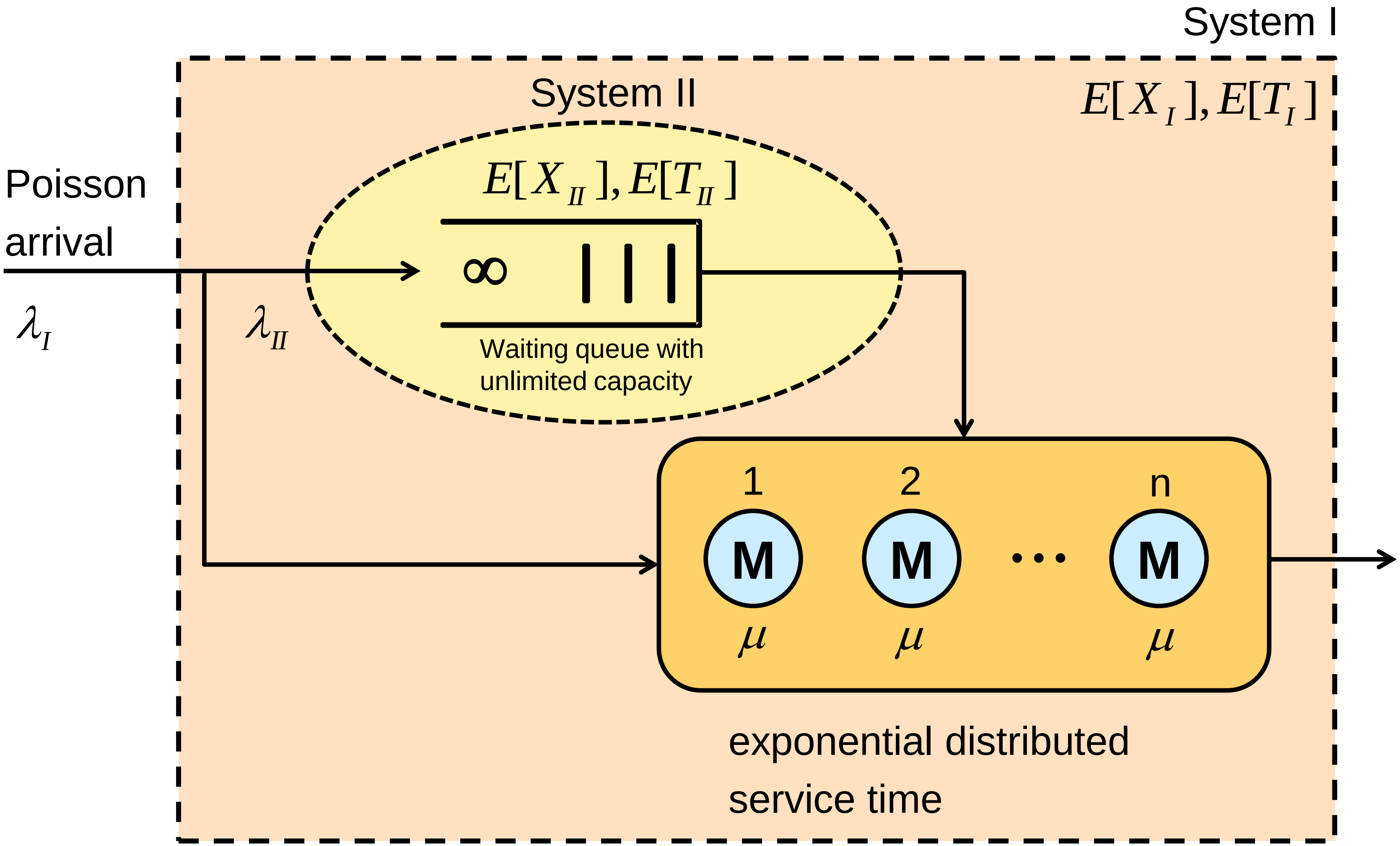

Poisson arrival
M, λ
∞
Waiting queue with unlimited capacity
1
2
n
M
M
M
μ
μ
μ
exponential distributed service time

Idea:

Divide the systems into two systems wheras one describes the system from the perspective of waiting jobs and another one that describes the system from the perspective of jobs that are immediately serverd.

M / M / n – Waiting system

- System in a System



M / M / n – Waiting system

- System I:

  (The total) M / M / n - ∞ Waiting system

  - Average arrival rate λ_I : total arrival rate $\lambda_I = \lambda$
  - Average number of jobs within the system $E[X_I]$
    (number of waiting + number of currently served jobs)

⇨ $$E[X_I] = E[X_W] + E[X_B] = \Omega + Y$$

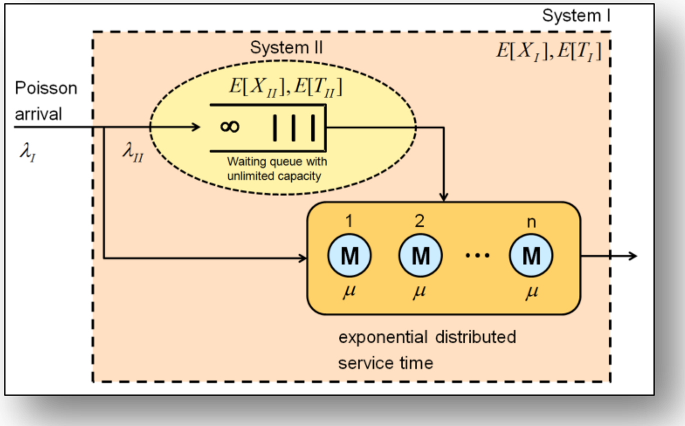
  - Average retention time within the system $E[T_I]$
    (average waiting time of all jobs + average service time)

⇨ $$E[T_I] = E[W] + E[B]$$

Little Theorem: $\lambda_I \cdot E[T_I] = E[X_I]$

➡ $$E[W] = \frac{\Omega}{Y}$$

M / M / n – Waiting system



- System II:
  (the inner waiting queue)

  - Average arrival rate λ_{II}
    (arrival rate of waiting jobs)

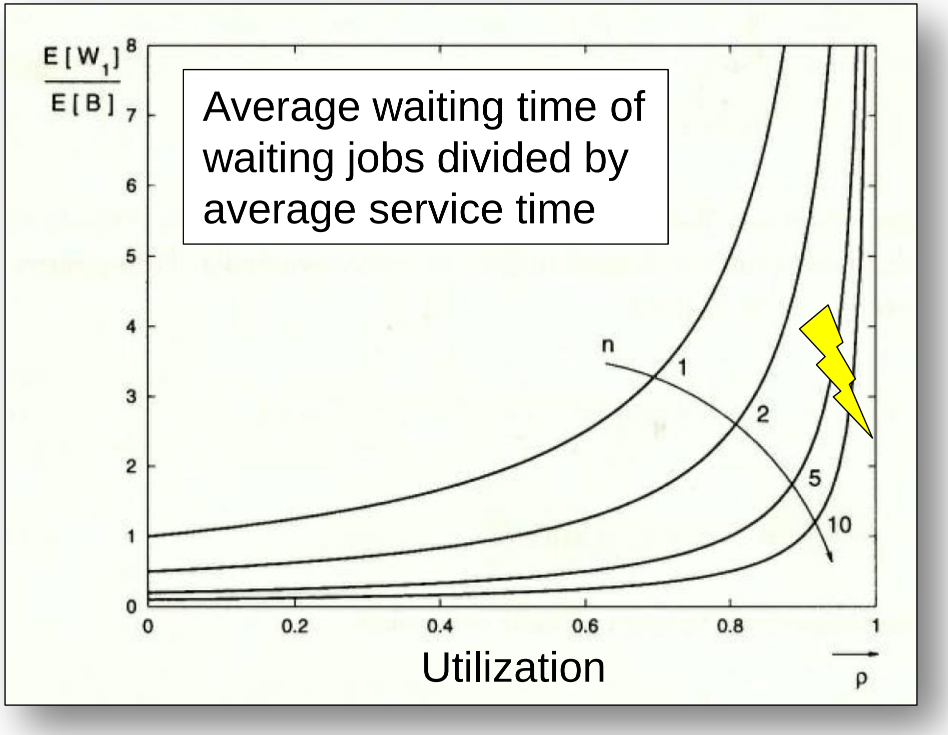
$$\Rightarrow \quad \lambda_{II} = \lambda \cdot p_W$$

  - Average number of jobs within the system $E[X_{II}] = \Omega$
    (number of waiting jobs)

  - Average retention time within the system $E[T_{II}]$
    (average waiting time of waiting jobs) $E[W_I]$

M / M / n – Waiting system

- Average waiting time of **waiting** jobs:



Average waiting time of waiting jobs divided by average service time

Utilization

- Multiplexing gain:
  - The waiting time of waiting jobs strongly increases for high utilizations.
  - A higher number of service units results in a much lower waiting time of waiting jobs.

Picture taken from Tran-Gia, Einführung in die Leistungsbewertung und Verkehrstheorie

M / M / n – Waiting system

- Average waiting time of **waiting** jobs:

Little Theorem $\lambda_{II} \cdot E[T_{II}] = E[X_{II}]$

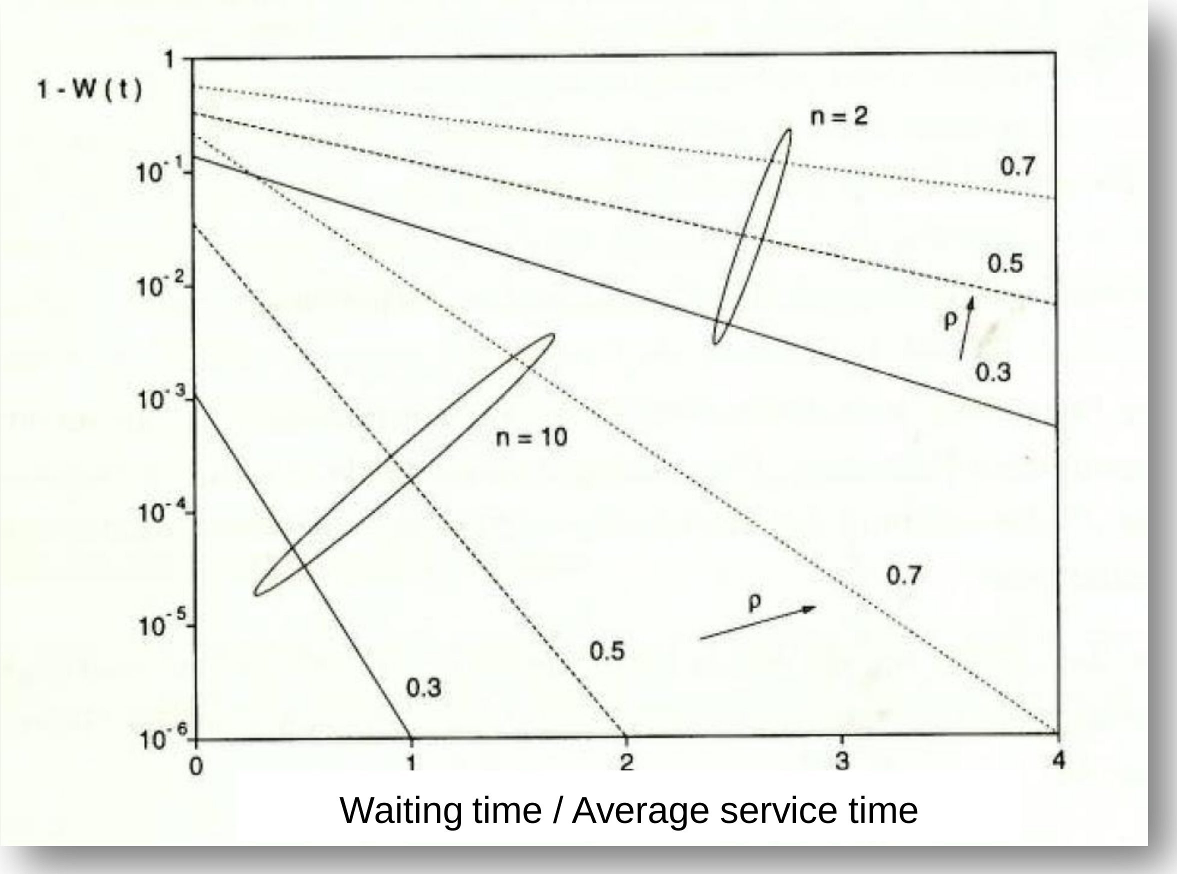
⇨ $$E[W_1] = \frac{\Omega}{\lambda \cdot p_W} = \frac{1}{\lambda} \cdot \frac{\rho}{1-\rho}$$

❗ Short bursts result in a high increase of waiting time for systems with high utilization.

❗ Communication systems should be dimensioned such that the average utilization is about 50% in order to be robust against temporary load variations.

M / M / n – Waiting system

- Complementary waiting time distribution



Picture taken from Tran-Gia, Einführung in die Leistungsbewertung und Verkehrstheorie

Questions

- Can you describe an M/M/n – waiting system?
- Derive the average waiting queue length.
- How does the utilization affect the waiting time of jobs?
- How does the state distribution of a M/M/n – waiting system change with higher utilization?
- How does the number of serving units affect the waiting time of a M/M/n – waiting system?
- What is the difference between waiting time of waiting jobs and waiting time of all jobs?

Analysis of System Performance

IN2072

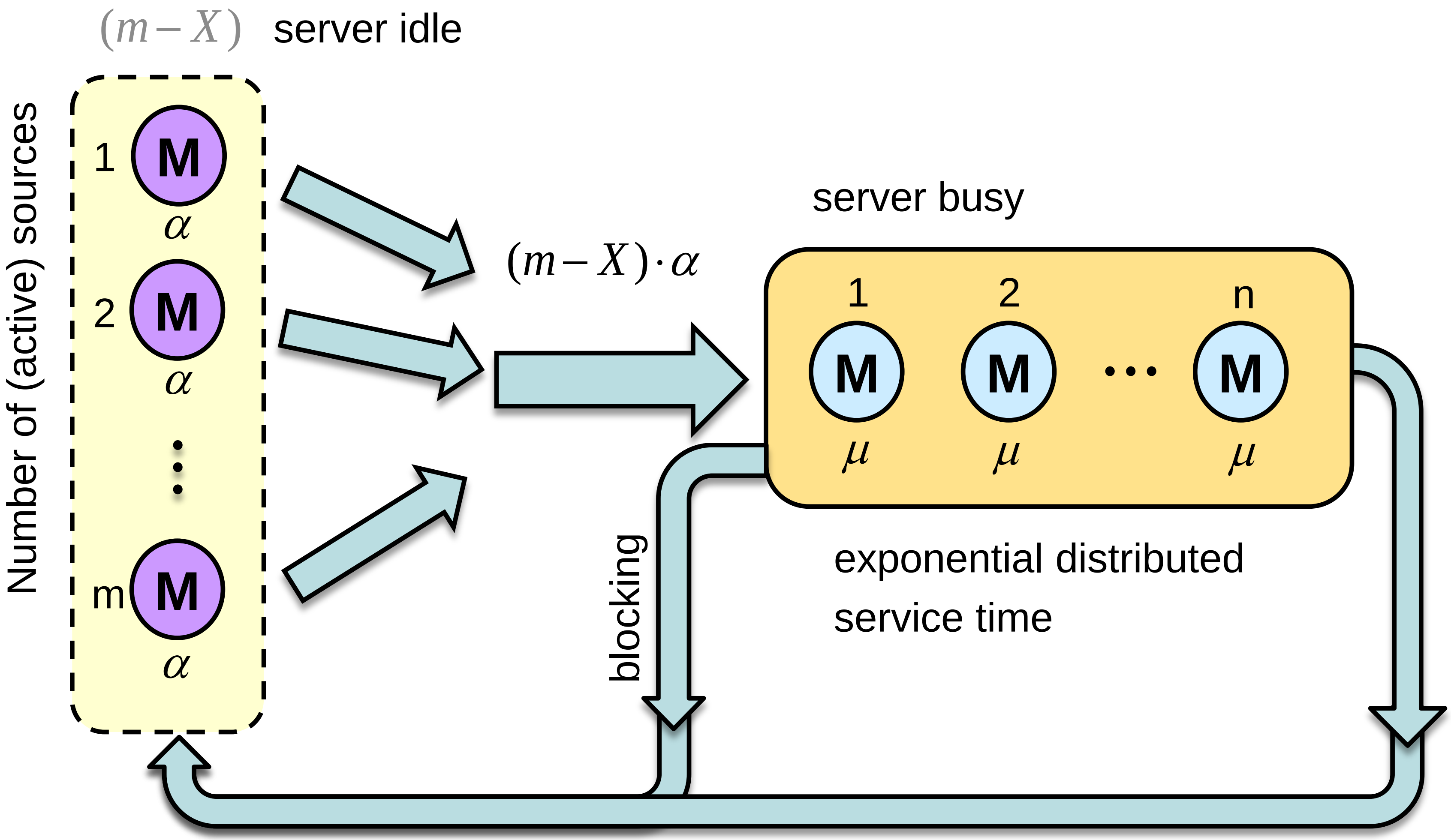
Chapter 4 – Analysis of

Markov Systems (Part 2/2)

M / M / n – Loss system with finite sources

M / M / n – Loss system **with finite sources**

- Model description:



M / M / n – Loss system **with finite sources**

- Arrival process:

  Arrival rate λ

  Average number of arriving jobs per time unit.

$$A(t) = P(A \le t) = 1 - e^{-\lambda t}, \qquad E[A] = \frac{1}{\lambda}$$

- Service process:

  Service rate μ

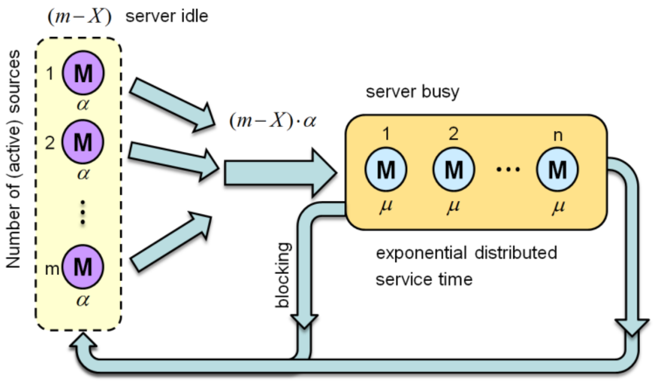
  Average number of service completions per time unit. (assuming a service unit with 100% utilization)

$$B(t) = P(B \le t) = 1 - e^{-\mu t}, \qquad E[B] = \frac{1}{\mu}$$

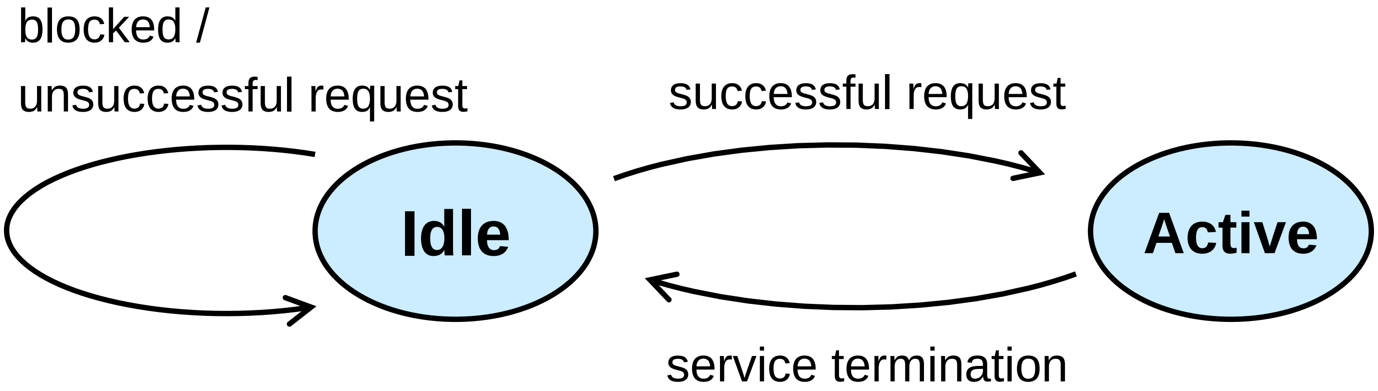
- System:
  - Loss system
  - No waiting queue
  - **Number of busy service units affects the future development of the system**

M / M / n – Loss system **with finite sources**

- Subscriber state diagram:



  - Active:
    - Subscriber is currently served.
    - Duration of the active period is identical with the busy period of the service unit.
  - Idle:
    - Subscriber remains in the idle state until its next service request.
    - Subscriber returns to the idle state after it has been serverd or is blocked.



blocked /
unsuccessful request

successful request

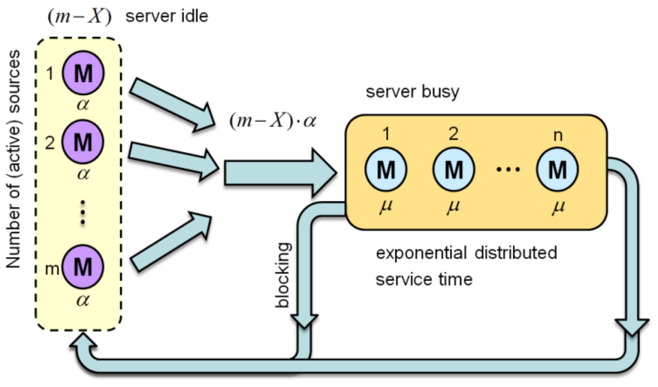
Idle

Active

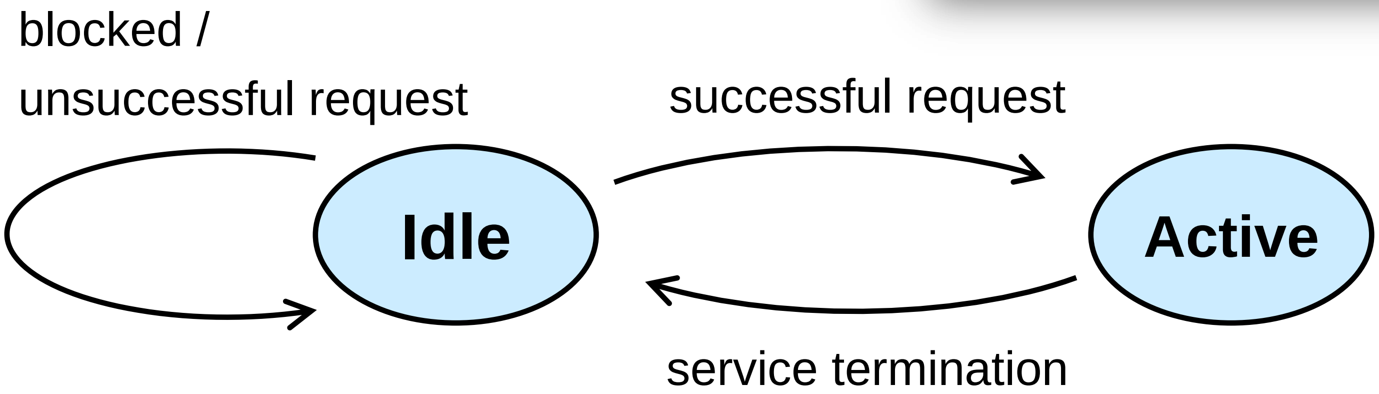
service termination

M / M / n – Loss system **with finite sources**

- Idle period:



$$I(t)=P(I\le t)=1-e^{-\alpha t}, \qquad E[I]=\frac{1}{\alpha}$$



blocked / unsuccessful request

successful request

Idle

Active

service termination

- Subscriber enters an idle phase after being served or blocked which is described by the idle distribution I(t).
- Subscriber is blocked if all servers are busy at the time of its arrival/request.

M / M / n – Loss system **with finite sources**

- Description:
  - State $X(t)$ is incremented if a job can be served by an idle service unit .
  - State $X(t)$ is decremented if a service is completed.
  - State $X(t)$ affects the arrival process.

Due to the memory-less characteristics of the arrival and the service process, the system is memory-less at any time of the process development.

- Transient phase:
  - The system starts in state $X(0)$ from which it develops through an instationary phase until it reaches a stationary state.
  - The state probabilities do not change any further as soon as the stationary state is reached.

- State probabilities: $x(i) = P(X(t) = i) = P(X = i), \qquad i = 0,1,\ldots,n$

- State probability vector: $X = \{x(0), x(1), \ldots, x(n)\}$

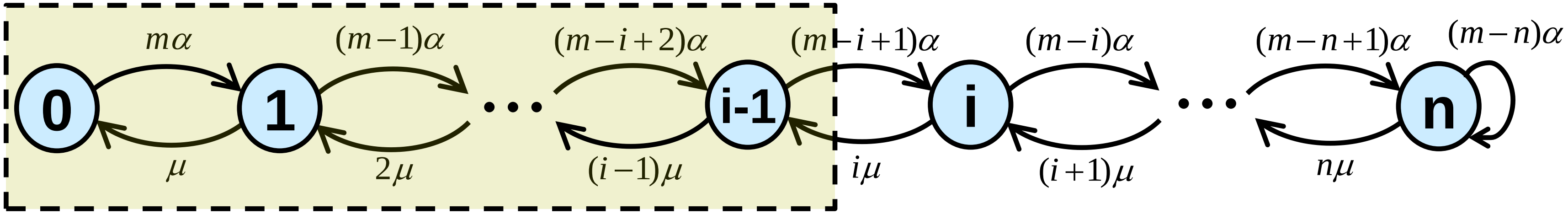
M / M / n – Loss system **with finite sources**

Arrival event:

- According to the definition of a Poisson process the transition from $[X=i]\rightarrow[X=i+1]$ occurs with rate $(m-i)\cdot\alpha$ if the system is in state $x(i),\quad i=0,1,\ldots,n-1$ since i subscribers are currently active.
- Otherwise the system is in state $x(n)$ which results in the blocking of the arriving job.

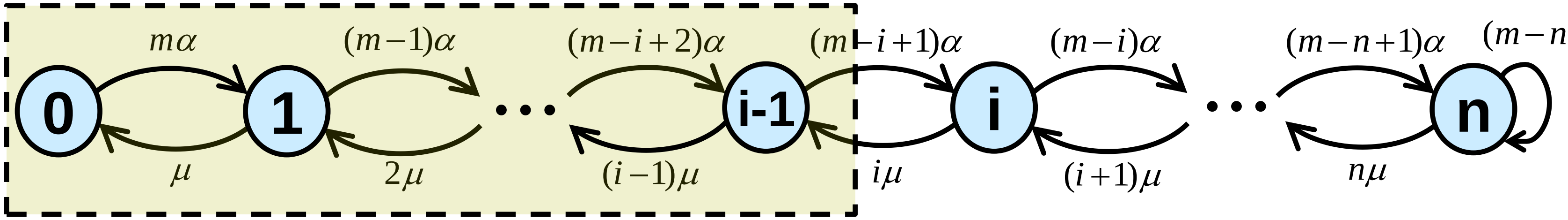
Service completion event:

- If the system is in state $x(i)$, i jobs are in the system.
- Thus, i service units are busy / i jobs are served.
- The transition from $[X=i]\rightarrow[X=i-1]$ occurs with rate $i\mu,\quad i=1,\ldots,n$ if one of the currently served jobs has finished.



Macro state S

M / M / n – Loss system **with finite sources**



Macro state S

Macro state S consists of micro states $\{X = 0,1,\ldots,i-1\}$.

⇨ $$(m-i+1)\alpha\cdot x(i-1) = i\cdot\mu\cdot x(i), \qquad i = 1,2,\ldots,n$$

⇨ $$\sum_{i=0}^{n} x(i) = 1$$

with $a^{*} = \dfrac{\alpha}{\mu}$ (load offered by a single subscriber)

➡ $$x(i) = \frac{\binom{m}{i}\cdot a^{*i}}{\sum_{k=0}^{n}\binom{m}{k}\cdot a^{*k}}, \qquad i = 0,1,2,\ldots,N$$ State probability

M / M / n – Loss system **with finite sources**

- Blocking probability:

  Subscribers are blocked if all service units are busy.

  Idea:

  - Describe the system from the perspective of a subscriber.
  - Exclude the subscriber from the system.
  - From the viewpoint of an idle subscriber S, the system can be regarded as a system with $(m-1)$ subscribers.

M / M / n – Loss system **with finite sources**

- Blocking probability:

$$x_A(i) = \frac{\binom{m-1}{i} \cdot a^{*i}}{\sum_{k=0}^{n} \binom{m-1}{k} \cdot a^{*k}}, \qquad i = 0,1,2,\ldots,N$$

State probability of a system with (m-1) subscribers.

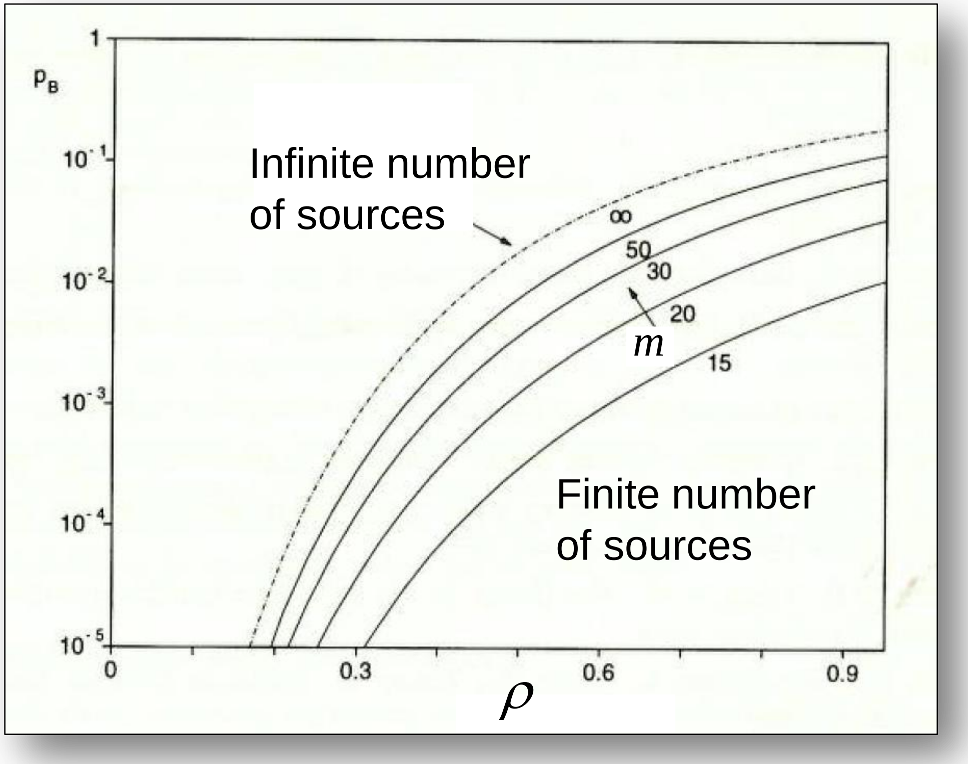
Subscriber is blocked if all service units are busy. → $[X_A = n]$

$$p_B = x_A(n) = \frac{\binom{m-1}{n} \cdot a^{*n}}{\sum_{k=0}^{n} \binom{m-1}{k} \cdot a^{*k}}$$

Blocking probability
(Engset equation)

M / M / n – Loss system **with finite sources**

- Engset equation:



Infinite sources: $\rho = \frac{a}{n} = \frac{\lambda}{n\mu}$

Finite sources: $\rho = \frac{a^* \cdot m}{n} = \frac{\lambda}{n\mu}$

Engset converges against Erlang with increasing number of sources.

Picture taken from Tran-Gia, Einführung in die Leistungsbewertung und Verkehrstheorie

M / M / n – Loss system **with finite sources**

- State and blocking probabilities:

  The derivation of the state and blocking probabilities discussed in this chapter is only valid for negative-exponential distributed service times. Howerver, it can be shown that they also hold for GI distributed service times as well.

Syski, R., Introduction to Congestion Theory in Telephone Systems, North-Jolland, Amsterdam 1985.

Questions

- Describe a M/M/n loss system with finite sources.
- What is the difference between the Engset and the Erlang formula?
- Describe the differences between a system with infinite and one with finite sources.
- What is the probability remaining in a macro state as which consists of micro states x(i) with i=0,1,2,…,i-1 ?
- How can you calculate the blocking probability of a M/M/n loss system with finite sources?