Julius-Maximilians-
UNIVERSITÄT
WÜRZBURG

Institute of Computer Science
Chair of Communication Networks
Prof. Dr. Tobias Hoßfeld

Chapter 4.2

Delay System M/M/n

Performance Evaluation of the Internet of Things (IoT)

Module Course: Performance Evaluation of Distributed Systems

Prof. Tobias Hoßfeld, Summer Semester 2022

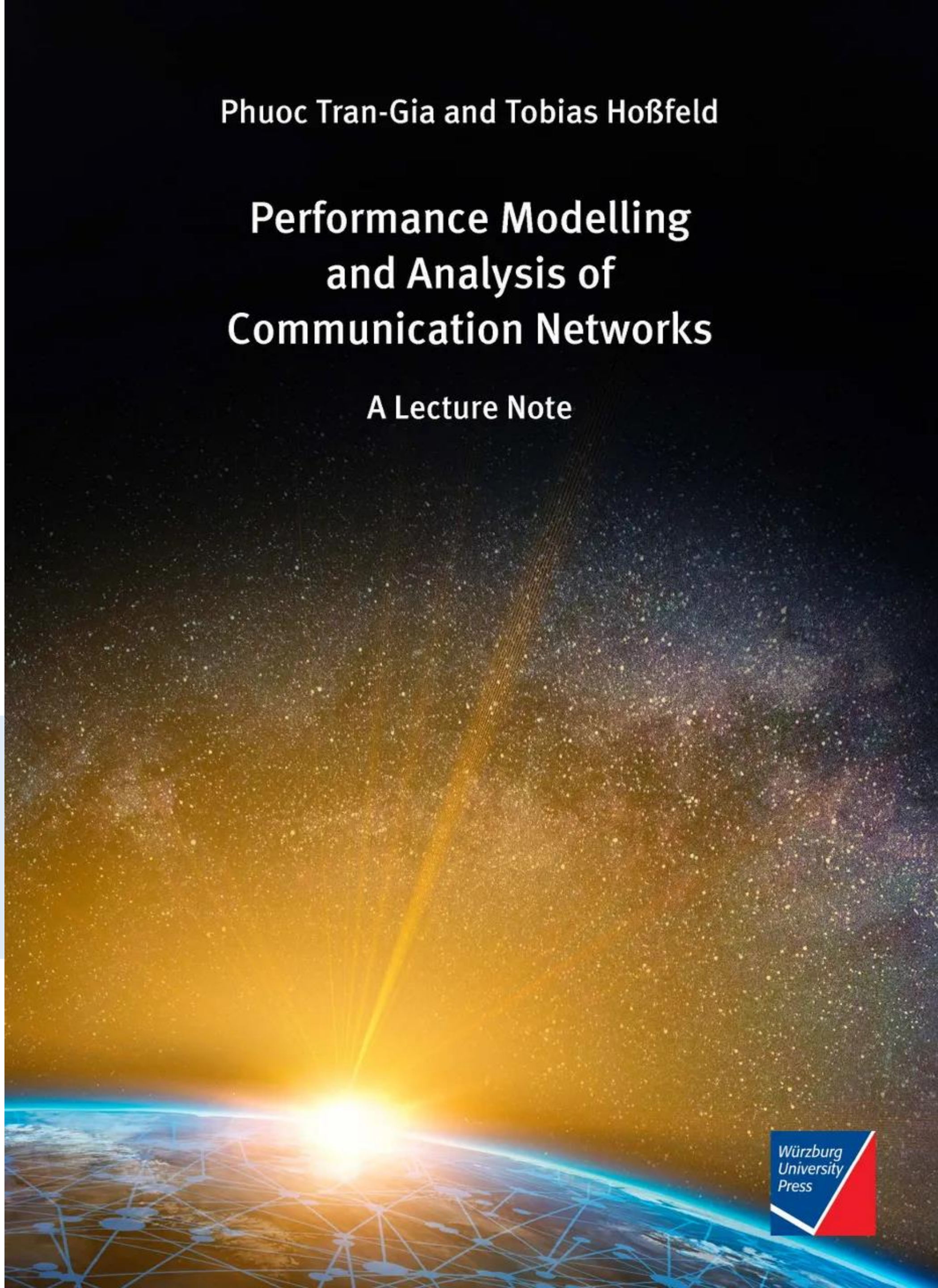
Disclaimer and Copyright Notice

The book must be cited and the disclaimer attached when using lectures slides or scripts.

Tran-Gia, P. & Hossfeld, T. (2021).
Performance Modeling and Analysis of Communication Networks - A Lecture Note. Würzburg University Press.
https://doi.org/10.25972/WUP-978-3-95826-153-2

Website to download book, exercises, slides and scripts:
https://modeling.systems/

Julius-Maximilians-
UNIVERSITÄT
WÜRZBURG



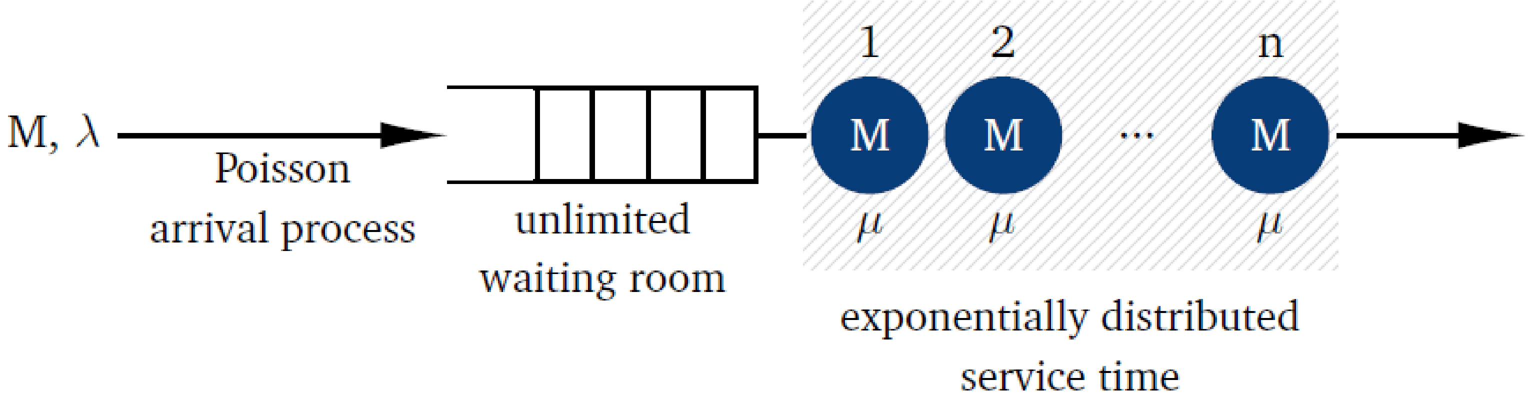
Chapter 4

4 Analysis of Markovian Systems

Model Structure and Parameters

Stability condition

Delay System M/M/n



- Interarrival time A with arrival rate λ
- Service time B with service rate μ
- Offered traffic $a = \frac{\lambda}{\mu}$ in pseudo-unit Erlang [Erl]

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

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

- Pure delay system
  - number of waiting places is assumed to be unlimited
  - arriving customer finding upon arrival all servers occupied will join the queue until a server becomes available

Utilization and Stability Condition

- **Offered traffic** is identical to mean number of occupied servers $E[X]$

$$a = \frac{\lambda}{\mu} = \lambda E[B]$$

- **Utilization** or occupancy of a single server is the mean offered traffic per server

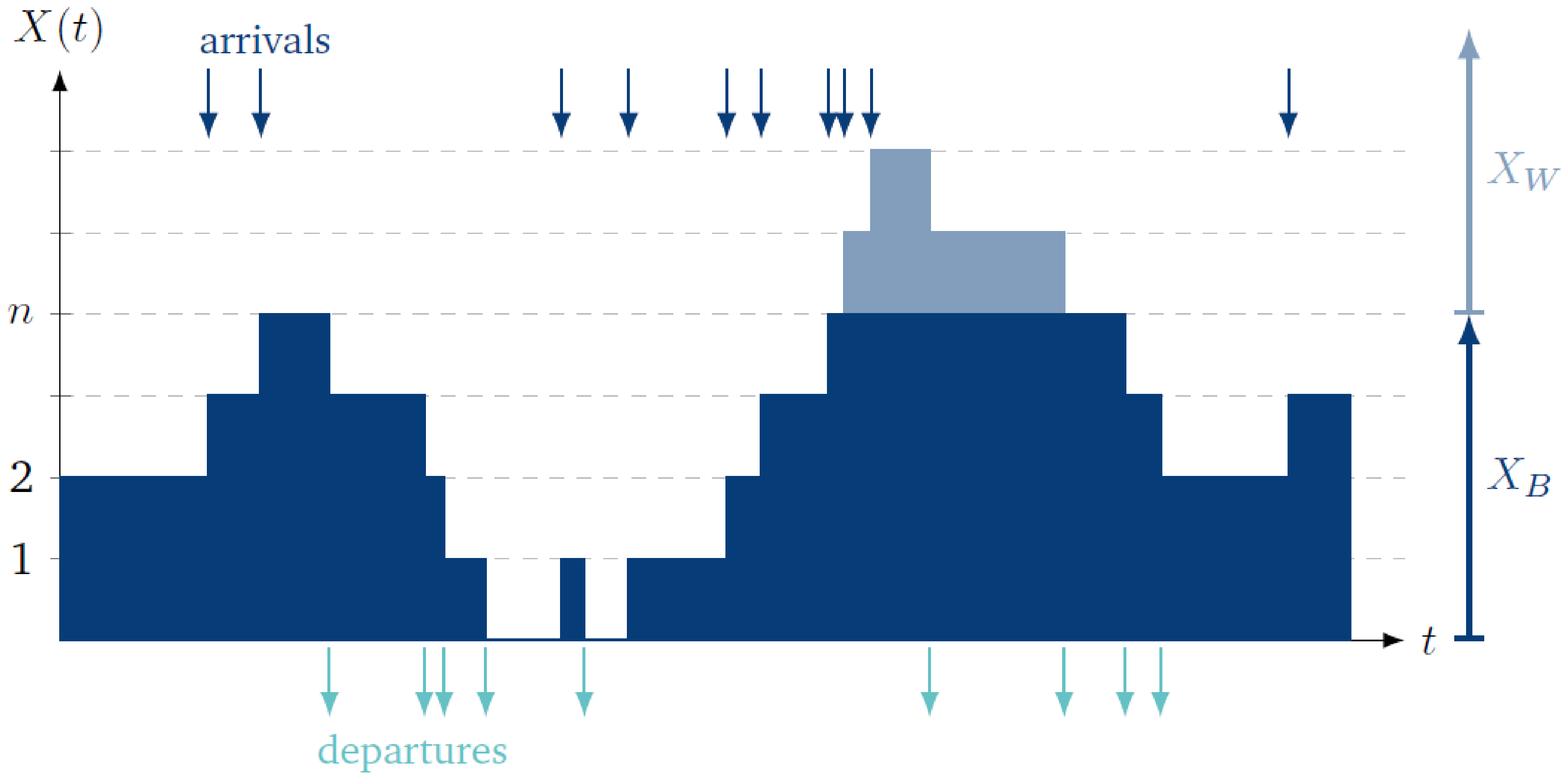
$$\rho = \frac{a}{n} = \frac{\lambda}{\mu n}$$

- **Stability condition**
  - during single service duration, $a = \lambda E[B]$ customers arrive on average
  - at most n customers can be served during single service duration: $\lambda E[B] < n$
  - Stable system requires $a < n$ or

$$\rho = \frac{a}{n} < 1$$

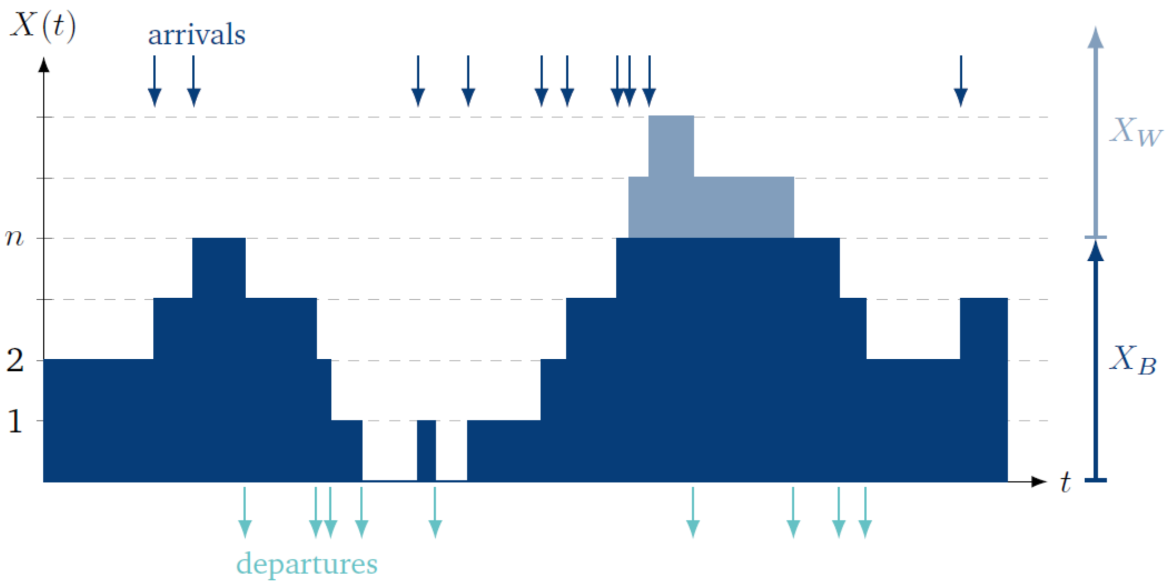
State Process and State Probabilities

State Process of M/M/n Delay System



State Process of M/M/n Delay System

- Complete description of stochastic process $X(t)$ at time t instead of $\{X_B(t), X_W(t)\}$



waiting customers $X_W(t)$

$X_W(t) = 0$ if $X_B(t) < n$

customers in service $X_B(t)$

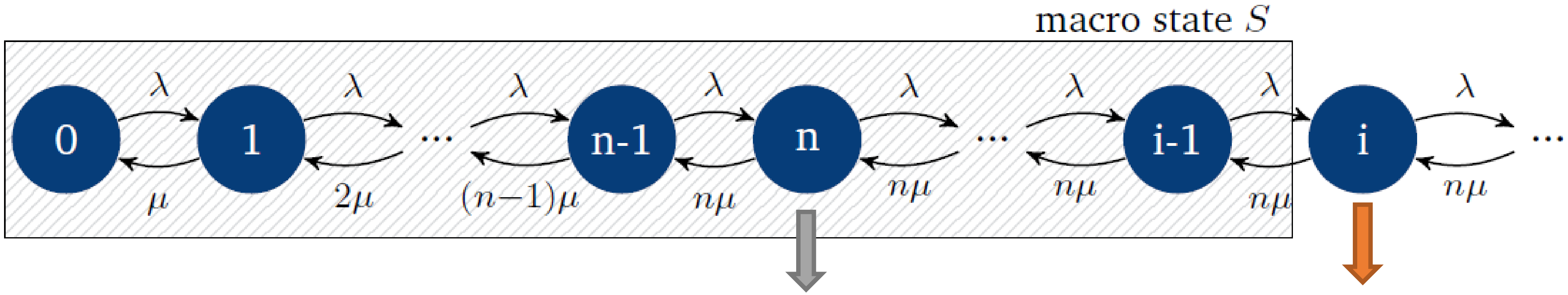
number of customers in system $X(t)$

$$X(t) = \begin{cases} X_B(t) & \text{for} \quad X_B(t) < n \quad (X_W(t) = 0) \\ X_B(t) + X_W(t) & \text{for} \quad X_B(t) = n \end{cases}$$

steady state probability

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

State Transition Diagram

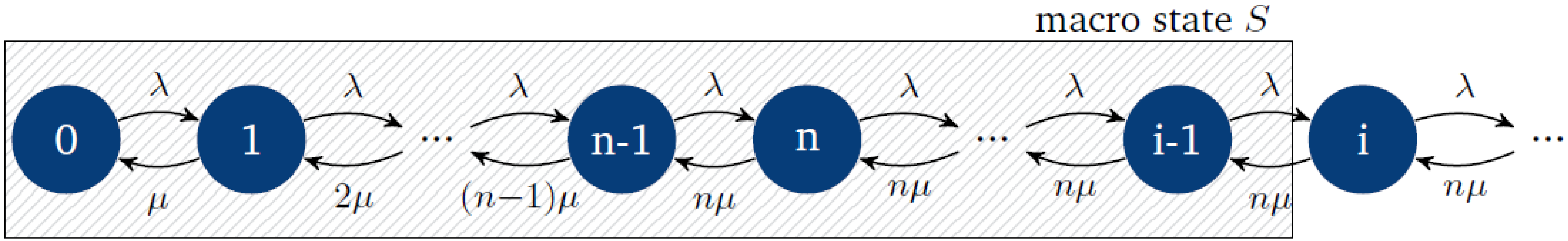


all servers occupied

(micro) state $[X = i]$
customers in the system

- Customer arrival: $[X\ =\ i] \rightarrow [X\ =\ i+1]$ with rate λ for $i = 0,1,\dots,\infty$ (Poisson process)
- Customer departure or service termination: $[X\ =\ i] \rightarrow [X\ =\ i-1]$ with rate $i\mu$
  - service time of one of the i customers ends ($i = 1,2,\dots,n$)
  - service rate for $X > n$: $n\mu$

Macro State Equations



- Differentiate macro states

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

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

- Normalization condition

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

Macro State Equations: Solution

Macro State Equations: Solution

- Successive iteration yields the solution

$$x(i) = \begin{cases} x(0)\dfrac{a^i}{i!} & i = 0,1,\ldots,n \\ x(0)\dfrac{a^n}{n!}\left(\dfrac{a}{n}\right)^{i-n} = x(n)\rho^{i-n} & i > n \end{cases}$$

- Probability of empty system using normalization condition

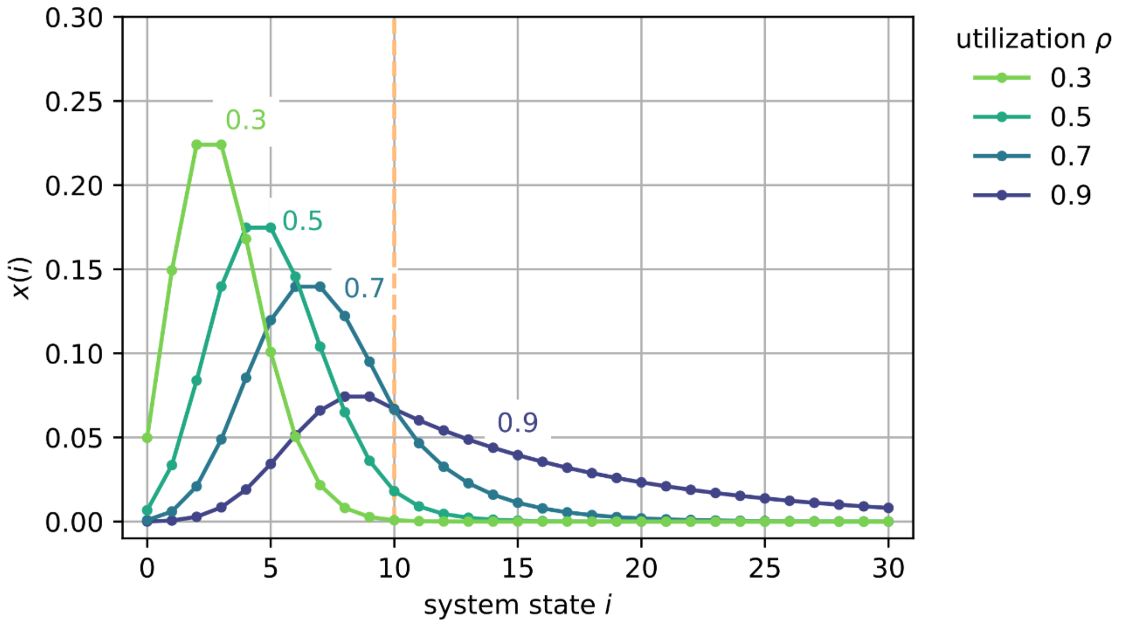
$$x(0) = \left(\sum_{k=0}^{n-1}\frac{a^k}{k!} + \frac{a^n}{n!}\sum_{k=0}^{\infty}\rho^k\right)^{-1}$$

- For stable systems $a < n$

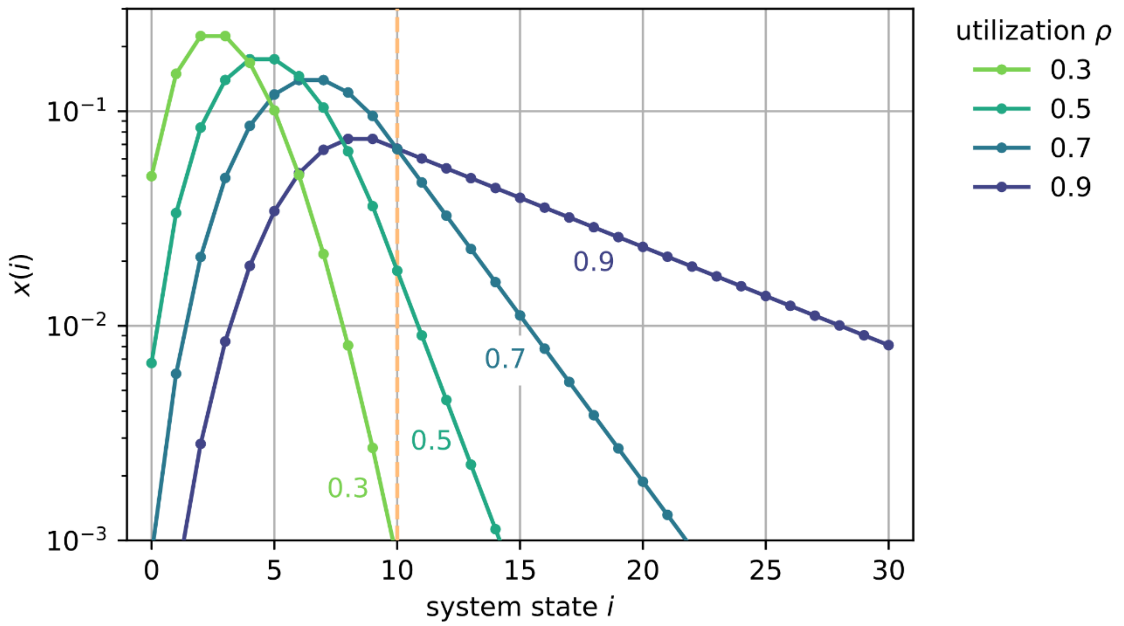
$$x(0) = \left(\sum_{k=0}^{n-1}\frac{a^k}{k!} + \frac{a^n}{n!}\cdot\frac{1}{1-\rho}\right)^{-1}$$

Steady State Distribution

- Example
  - M/M/n delay system for n=10



- Geometric tail for $i > n$
  - $x(i) = x(n) \cdot \rho^{i-n}$
  - logarithmic scale: linear for $i > n$



Julius-Maximilians-
UNIVERSITÄT
WÜRZBURG

Other System Characteristics

Erlang-C formula, waiting probability, mean waiting times

Waiting Probability

- Probability that an arriving customer sees all server busy

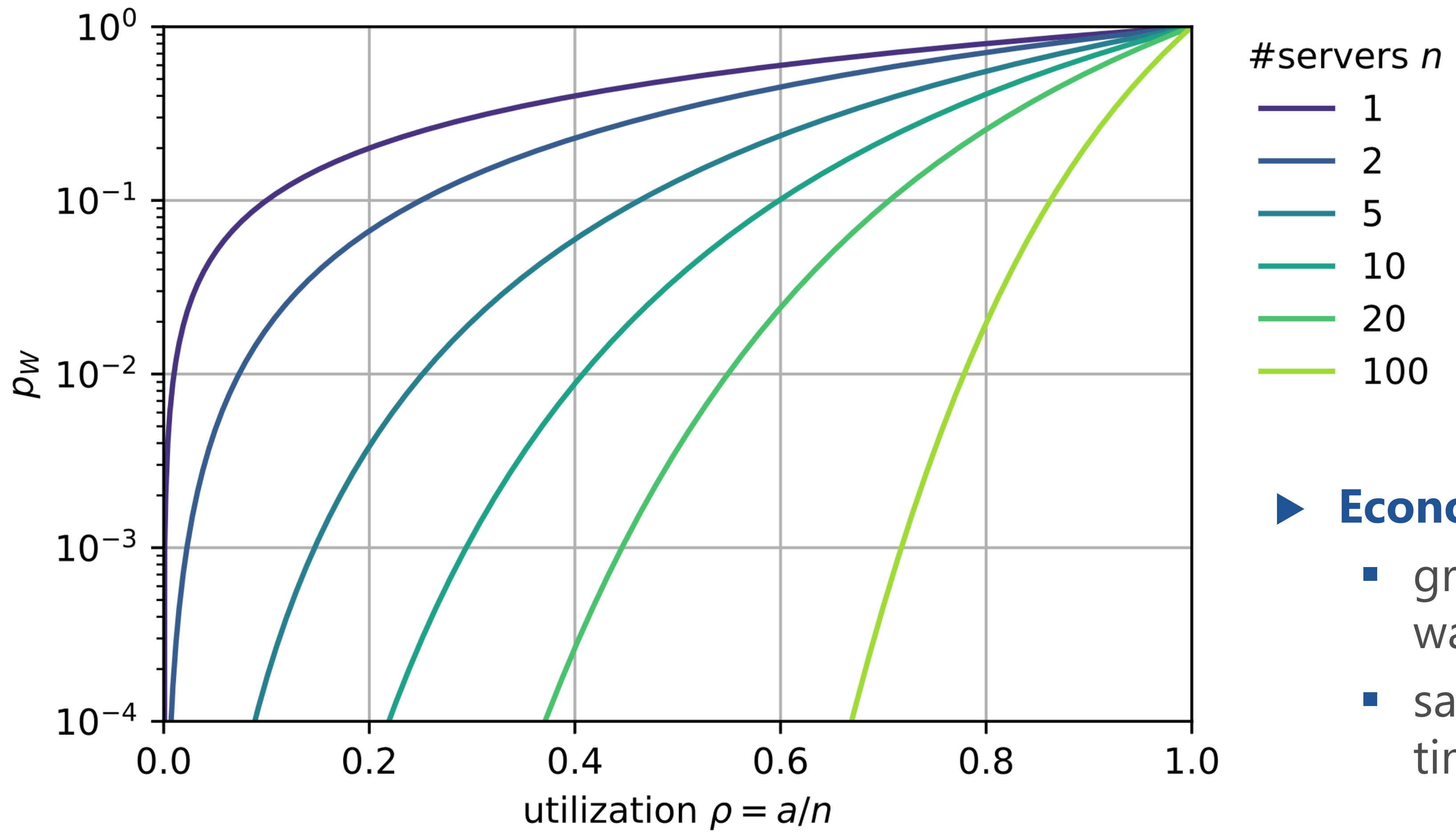
$$p_W = \sum_{i=n}^{\infty} x(i) = x(n)\sum_{i=0}^{\infty}\rho^i = x(n)\frac{1}{1-\rho}$$

- **Erlang-C formula**

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

- Example: M/M/1 delay system
  - $x(0) = 1-\rho$
  - $x(i) = (1-\rho)\cdot\rho^i$
  - $p_W = a = \rho$

Waiting Probability: Illustration



- **Economy of scale** in delay systems
  - grouping of servers least to lower waiting probabilities for same ρ
  - same result for mean waiting times (see later)

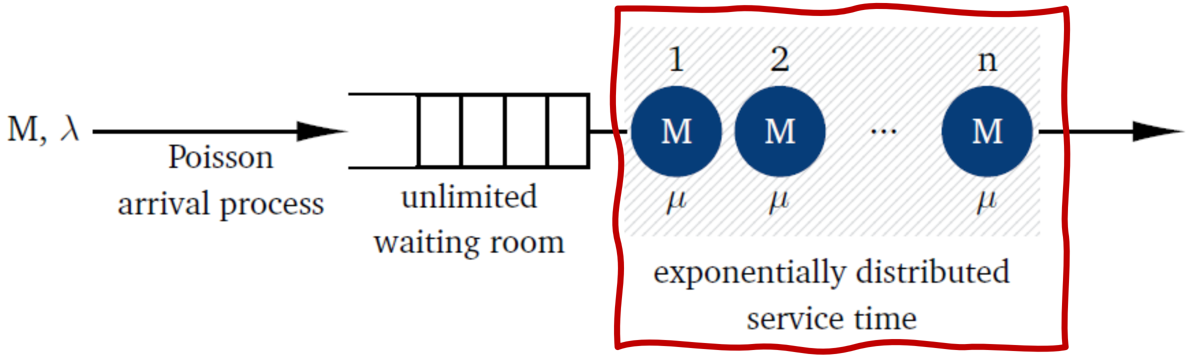
Carried Traffic

- Carried traffic Y is the mean number of occupied servers $E[X_B]$

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

- Using Little's theorem
  - mean arrival rate in the system: λ
  - mean sojourn time in the system: $E[B]$
  - mean number of customers in system: $E[X_B] = Y$

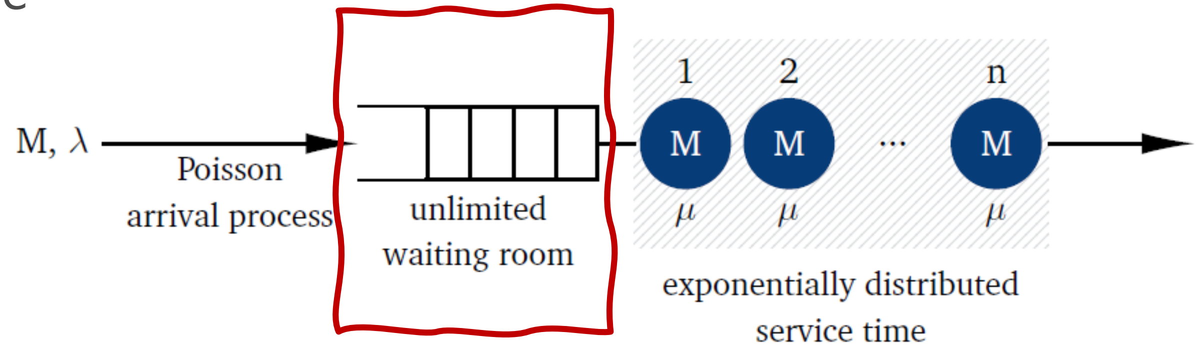
  - Little's law: $$Y = \frac{\lambda}{\mu} = a$$



Mean Queue Length

- Mean number of customers in the waiting space

$$\begin{aligned}\Omega &= E[X_W] \\ &= \sum_{i=n}^{\infty}(i-n)\cdot x(i) = \sum_{i=n}^{\infty}(i-n)\, x(n)\, \rho^{i-n} \\ &= x(n)\sum_{i=0}^{\infty} i\, \rho^i = x(n)\cdot\frac{\rho}{(1-\rho)^2} = x(0)\,\frac{a^n}{n!}\cdot\frac{\rho}{(1-\rho)^2}\end{aligned}$$

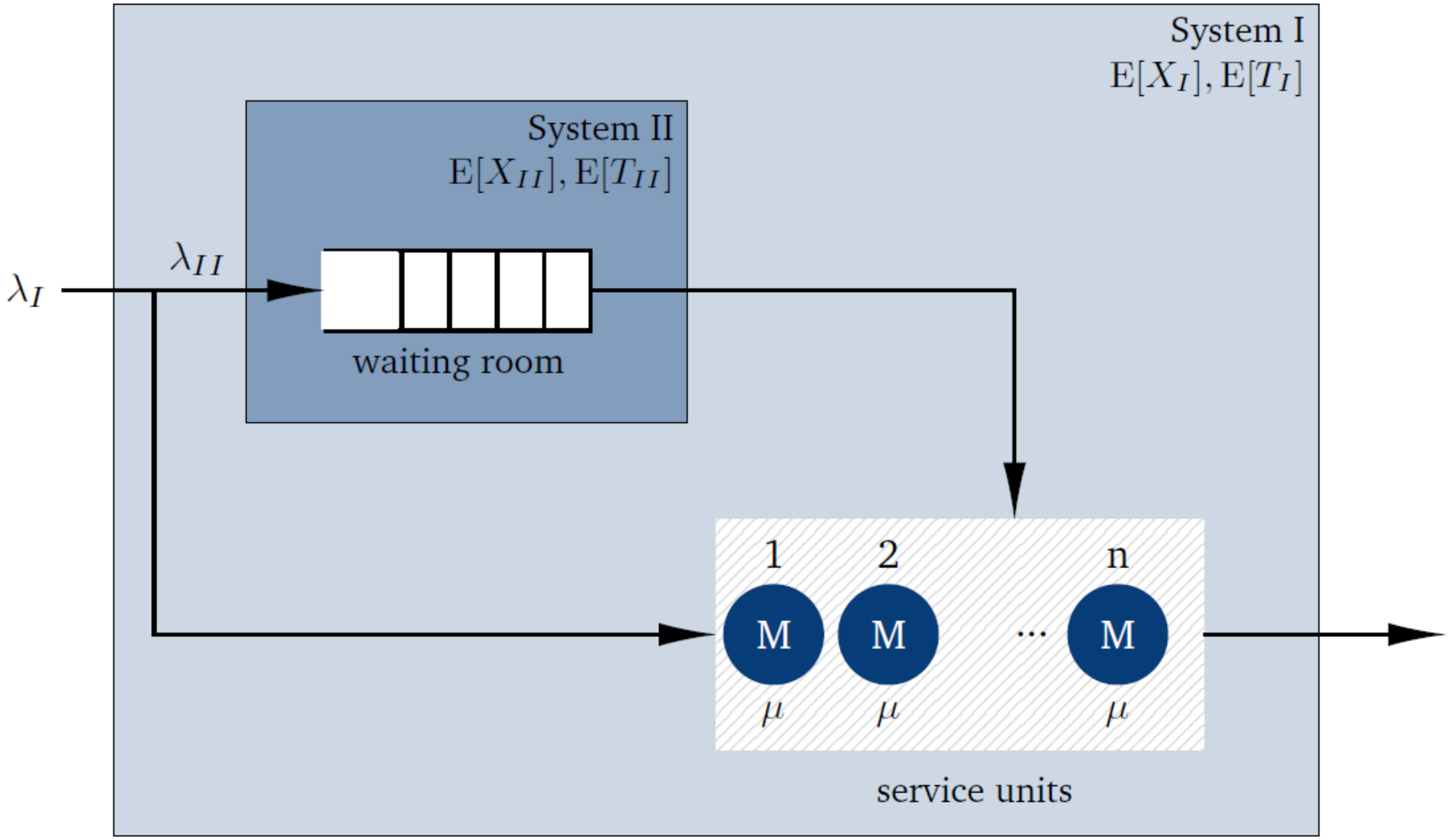


- Using waiting probability p_W

$$\Omega = \underbrace{\underbrace{x(0)\,\frac{a^n}{n!}}_{x(n)}\,\frac{1}{1\text{-}\rho}}_{p_W}\,\frac{\rho}{1\text{-}\rho} = p_W\,\frac{\rho}{1\text{-}\rho}$$

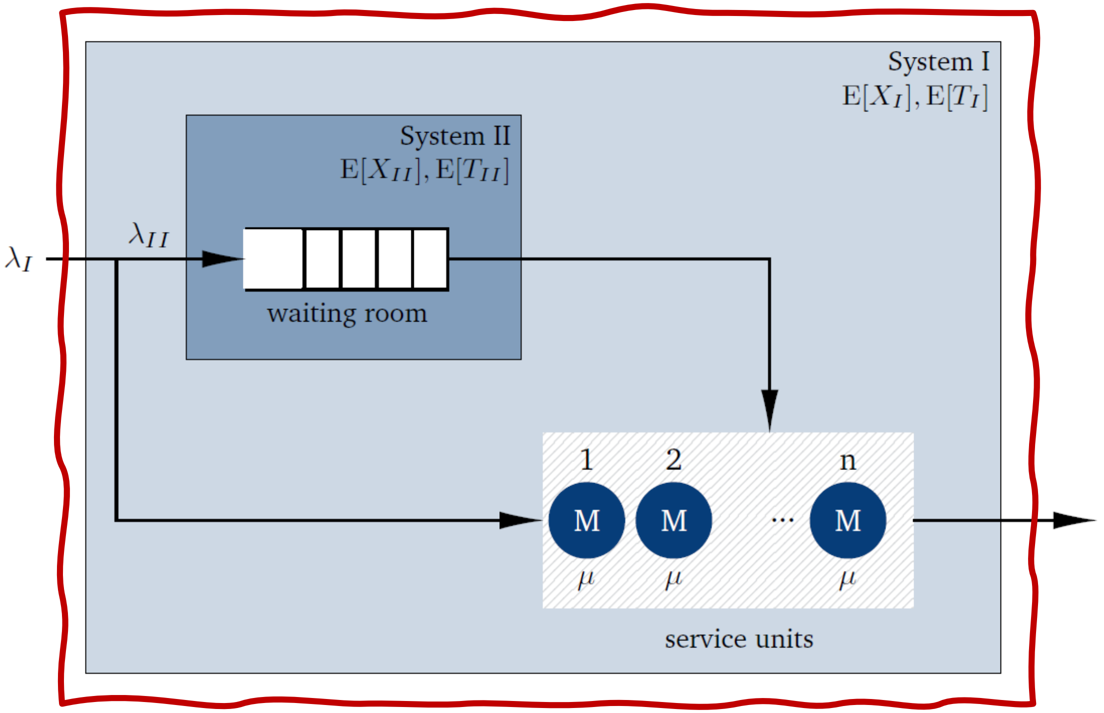
Mean Waiting Time

- Mean waiting times of all customers (system I) and of waiting customers (system II)



Mean Waiting Time of All Customers (System I)

- System I (all customers): entire M/M/n system
  - mean arrival rate λ_I
    $\lambda_I = \lambda$
  - mean number of customers in system $E[X_I]$
    $E[X_I] = E[X_W] + E[X_B] = \Omega + Y$
  - mean sojourn time in system $E[T_I]$
    $E[T_I] = E[W] + E[B]$

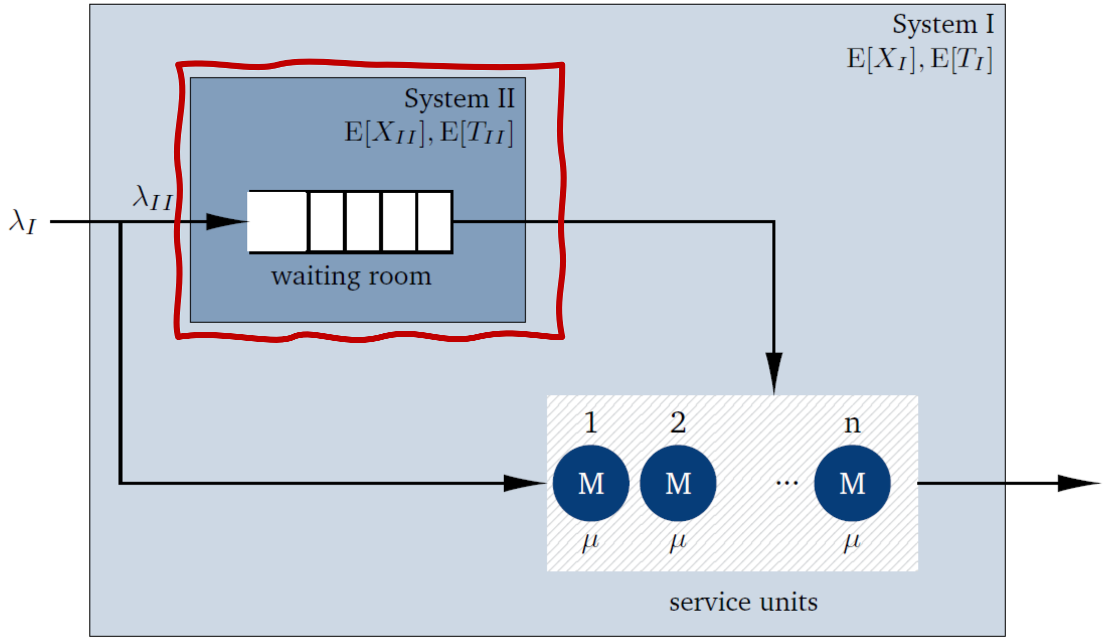


- **Little's law**

$$\lambda_I\, E[T_I] = E[X_I] \quad \Rightarrow \quad \lambda\, E[W] + \underbrace{\lambda\, E[B]}_{Y} = \Omega + Y \quad \Rightarrow \quad \boxed{E[W] = \frac{\Omega}{\lambda}}$$

Mean Waiting Time of Waiting Customers (System II)

- System II (only waiting customers): waiting queue of M/M/n system
  - mean arrival rate λ_{II}: arrival rate of waiting customers $\lambda_{II} = \lambda \cdot p_W$
  - mean number of customers in system $E[X_{II}]$: mean queue length $E[X_{II}] = \Omega = p_W \frac{\rho}{1-\rho}$
  - mean sojourn time in system $E[T_{II}]$: mean waiting time of waiting customers $E[T_{II}] = E[W_1]$

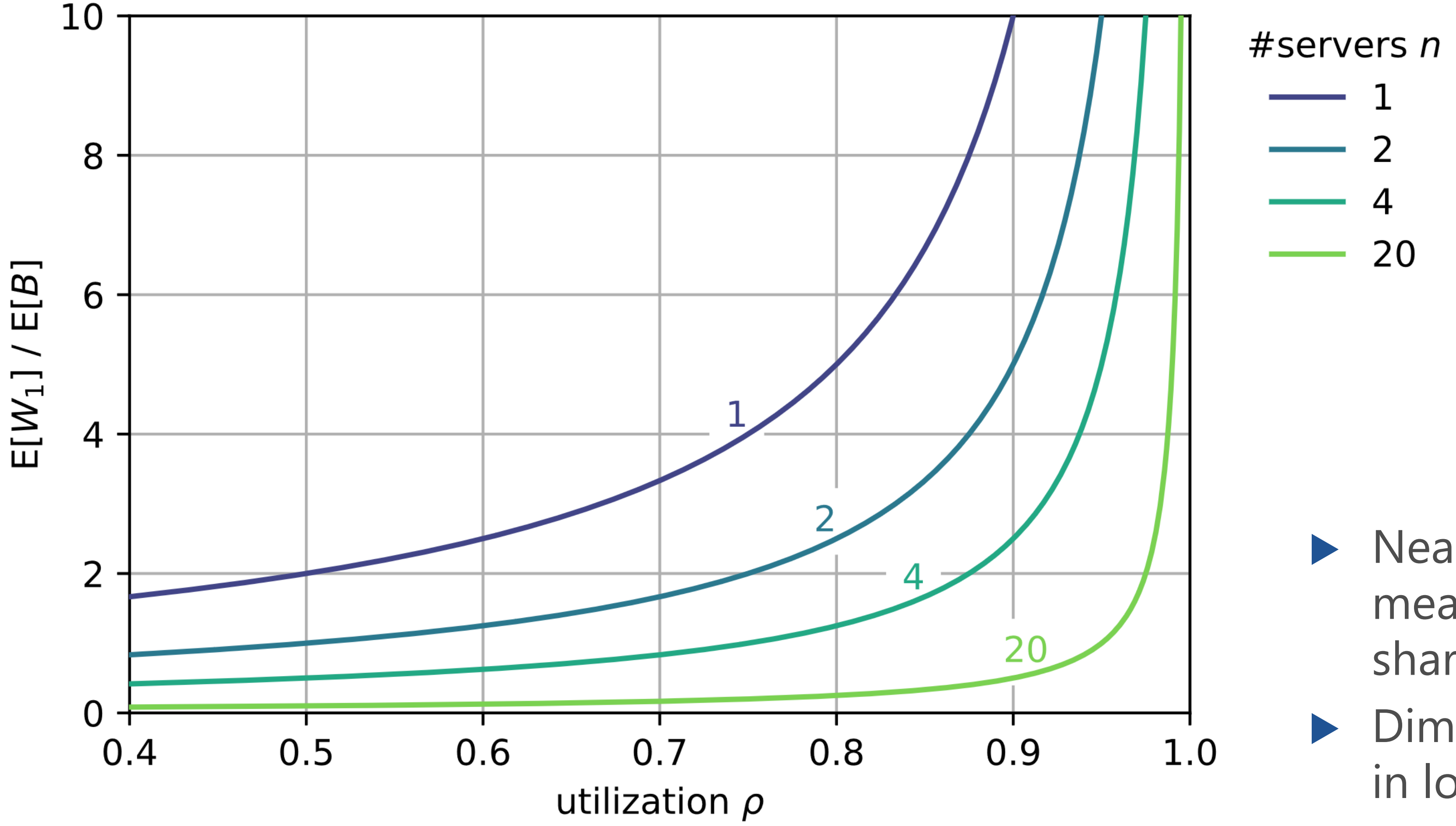


- **Little's law**

$$\lambda_{II} E[T_{II}] = E[X_{II}] \quad \Rightarrow \quad \boxed{E[W_1] = \frac{\Omega}{\lambda \cdot p_W} = \frac{1}{\lambda} \cdot \frac{\rho}{1-\rho}} \quad \Rightarrow \quad E[W_1] = \frac{E[W]}{p_W}$$

$$E[W] = \frac{\Omega}{\lambda}$$

Mean Waiting Time of Waiting Customers



- Near stability boundary $\rho = 1$, mean waiting time increases sharply
- Dimensioning the operation point in lower utilization range

DELAY DISTRIBUTION

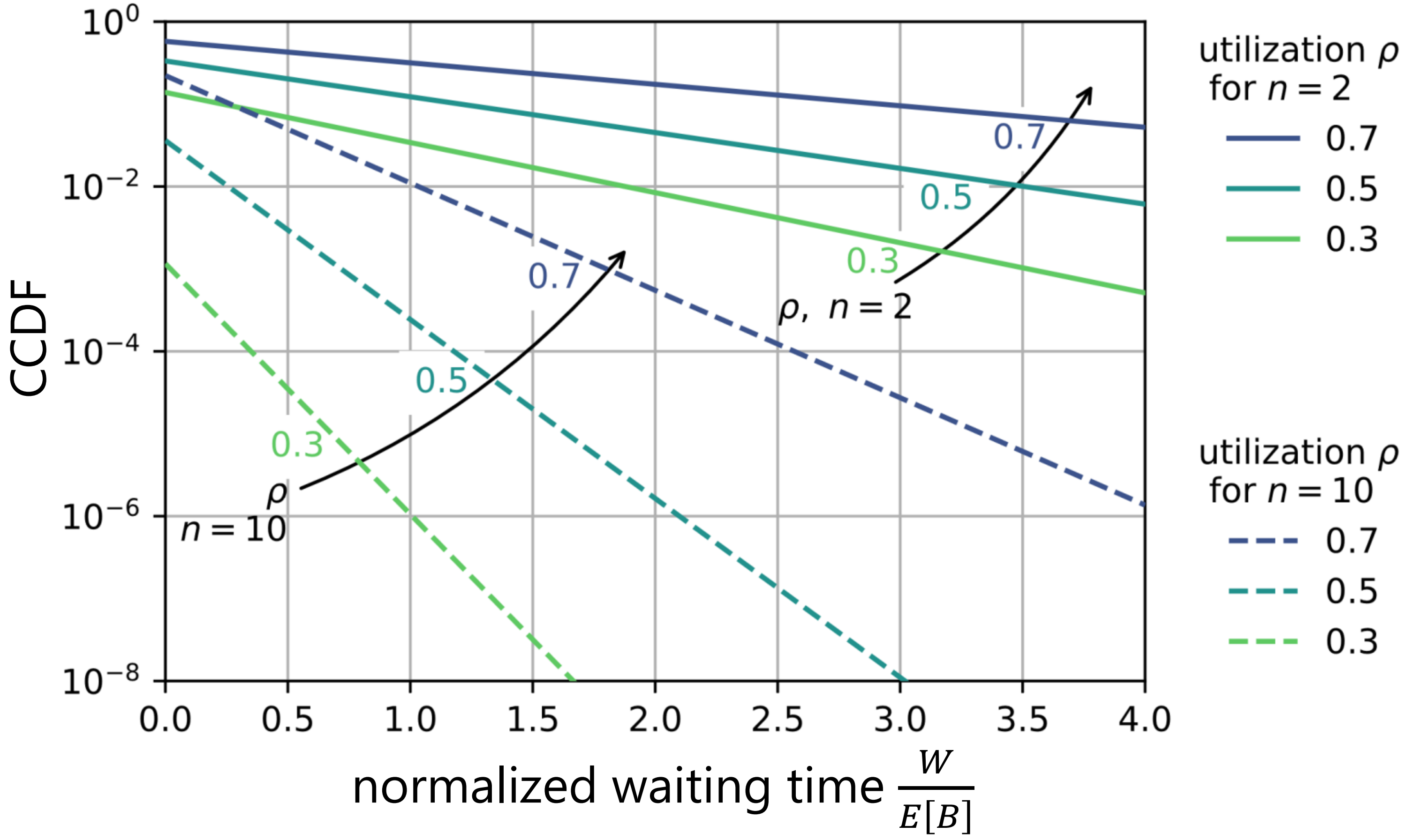
Waiting time distribution

Waiting Time Distribution for All Customers

- Random variable W is the waiting time for all customers

$$W(t) = 1 - p_W \cdot e^{-(1-\rho)n\mu t} = \begin{cases} 0 & t < 0, \\ 1 - p_W & t = 0, \\ 1 - p_W \cdot e^{-(1-\rho)n\mu t} & t > 0. \end{cases}$$

- Waiting probability is $1 - W(0) = p_W$



Waiting Time Distribution: Derivation

Waiting Time Distribution: Derivation (f.)

Lecture

EXAMPLES

Waiting time distribution, economy of scale

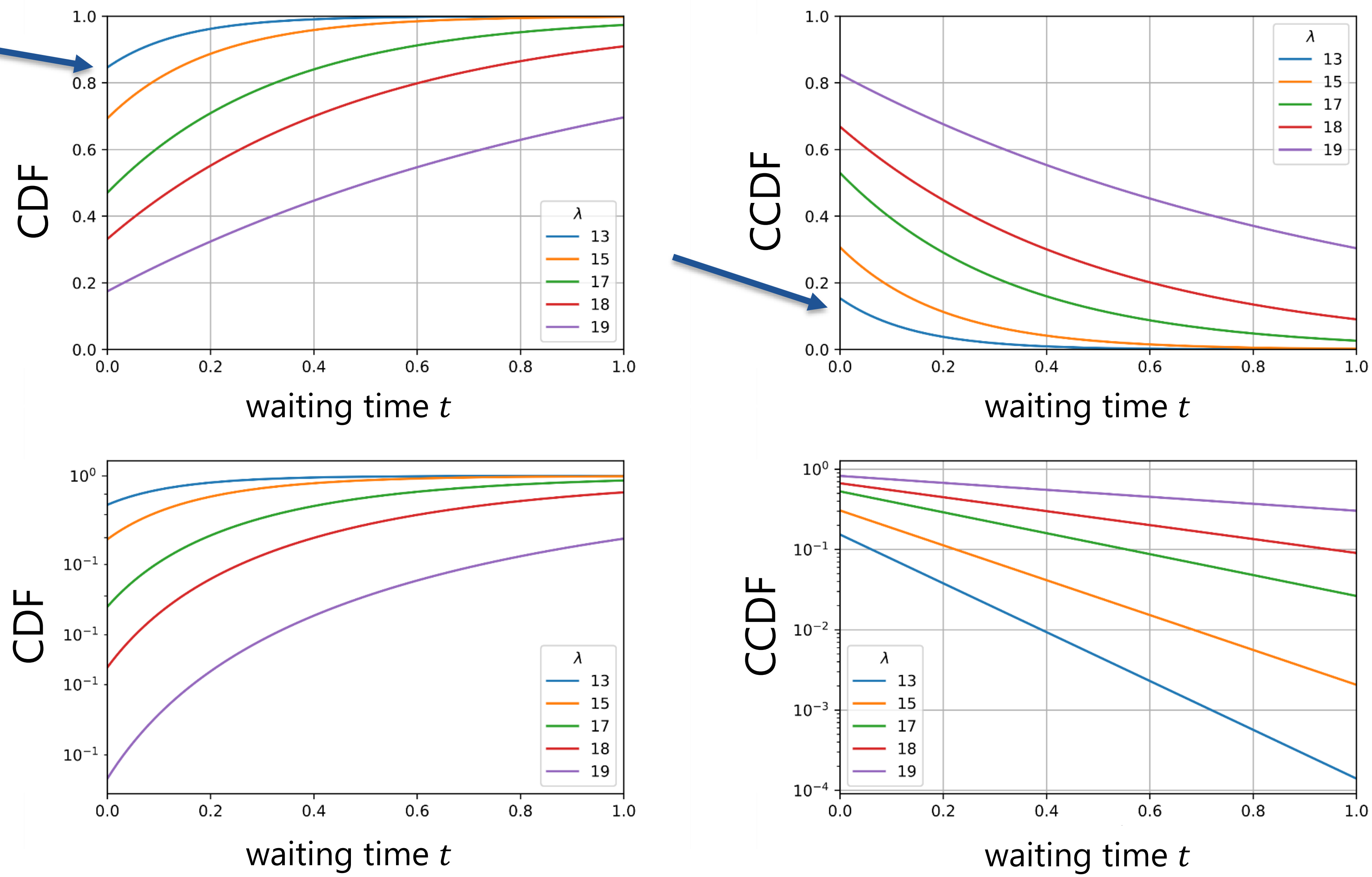
Example: Waiting Time Distributions M/M/n

$1 - p_W$

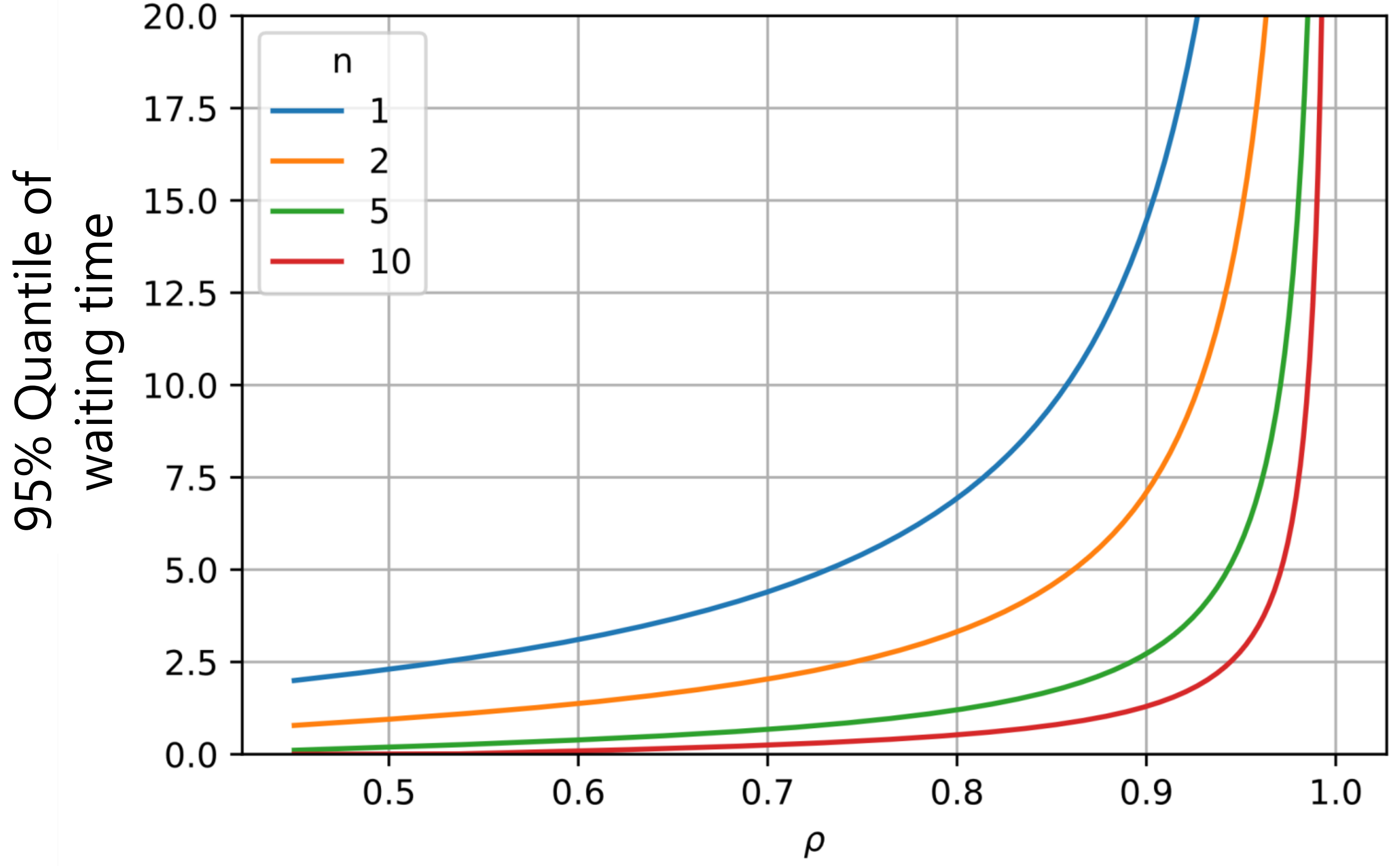
linear scale

parameter
$\mu = 2, n = 10$

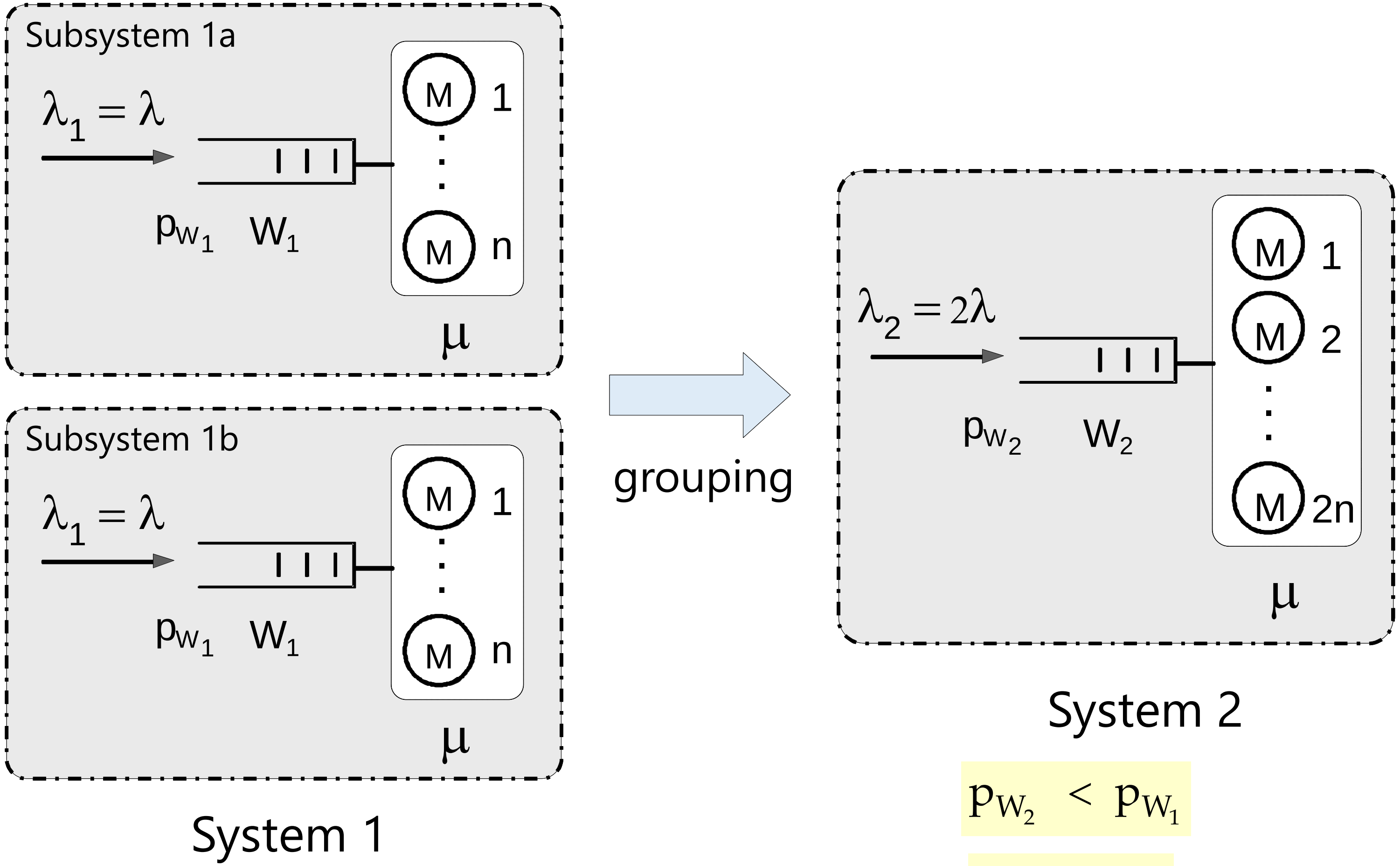
logarithmic scale



Quantiles of Waiting Times



Economy of Scale for Delay Systems



$p_{W_2} < p_{W_1}$

$W_2 < W_1$