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

Chapter 2.1

Little's Theorem and General Results

Performance Evaluation of the Internet of Things (IoT)

Module Course: Performance Evaluation of Distributed Systems

Prof. Tobias Hoßfeld, Summer Semester 2022

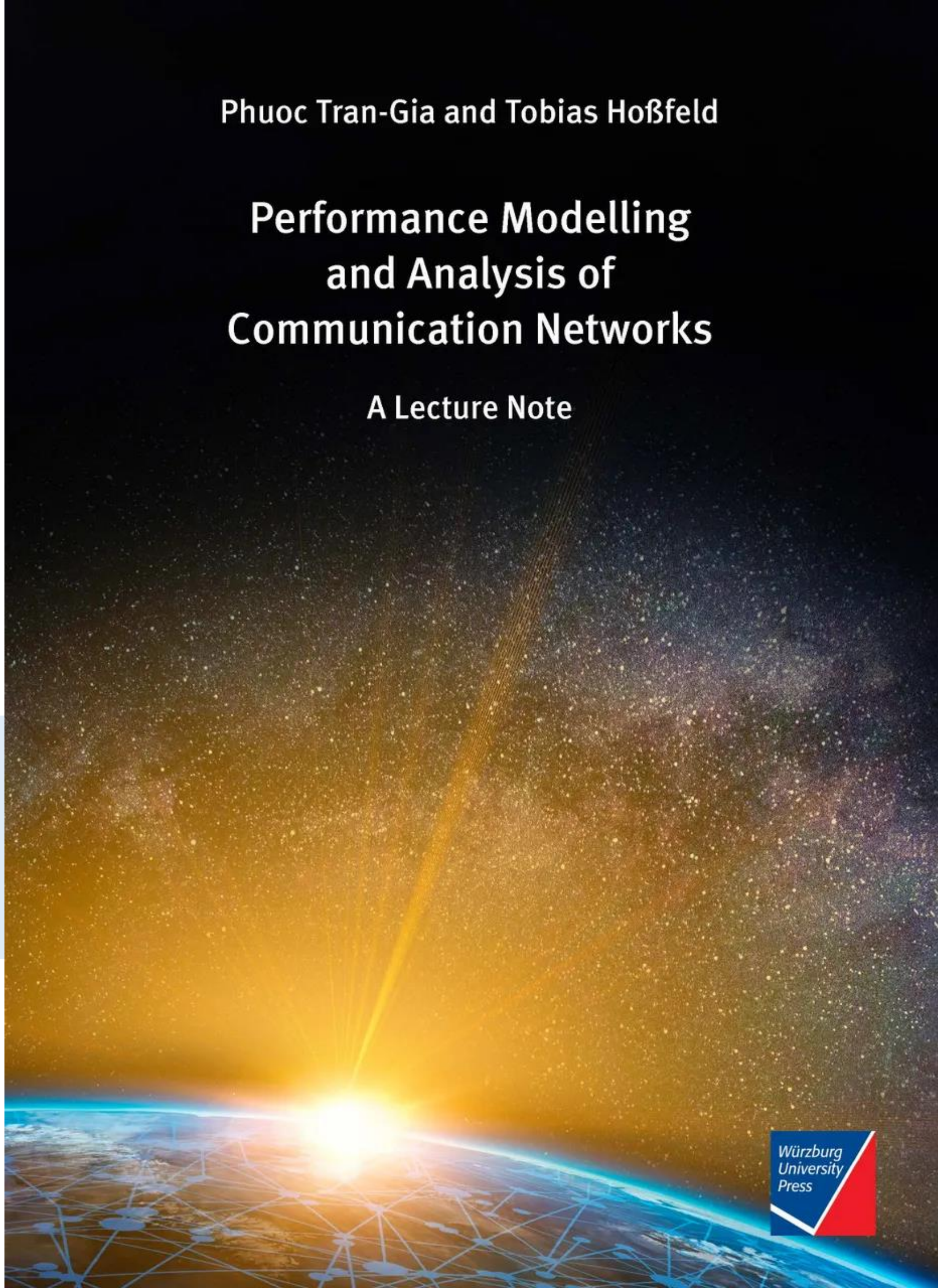
Disclaimer and Copyright Notice

Lecture slides, figures, and scripts are based on the open access text book "Performance Modeling and Analysis of Communication Networks". The book and scripts are licensed under the Creative Commons License Attribution-ShareAlike 4.0 International ([CC BY-SA 4.0](https://creativecommons.org/licenses/by-sa/4.0/)). If you remix, transform, or build upon the material, you must distribute your contributions under the same license as the original.

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/



Chapter 2

2 Fundamentals and Prerequisites

Little‘s Theorem

Little‘s law or Little‘s result

Operational Laws

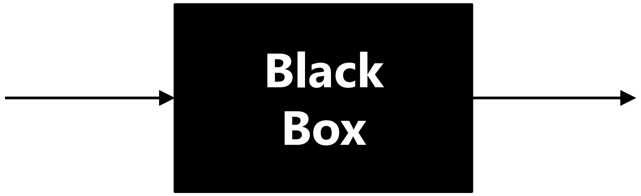
- System is considered as a black box
- Relation between quantities which do not require assumptions on the distribution of arrival times or service times
- “Operational” means directly measurable during operation
- Assumptions that can be verified during operation (in measurements).
  - For example: Is the number of arrivals equal to the number of completions?

Operational Quantities

- Quantities that can be measured directly in a finite observation period

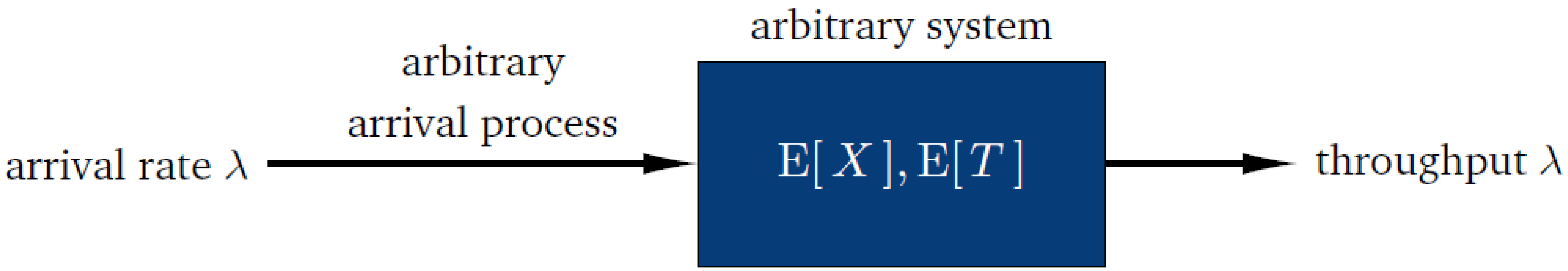
- t = observation period
- $A(t)$ = number of arrivals during (0,t)
- $D(t)$ = number of departures during (0,t)
- $B(t)$ = busy time during (0,t)

- λ_t = Arrival rate during (0,t)
- C_t = Throughput during (0,t)
- U_t = Utilization during (0,t)
- S_t = Mean service time during (0,t)



System

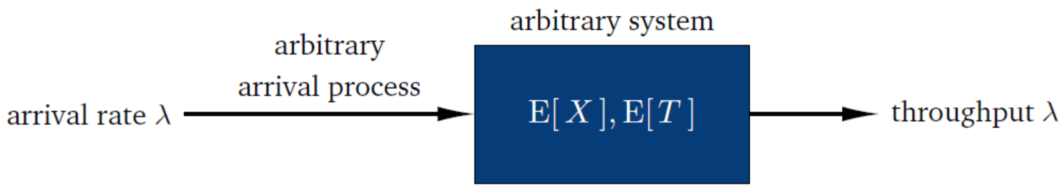
Little's Formula



$$\textbf{Little's Formula} \quad \lambda \cdot E[T] = E[X]$$

- λ mean arrival rate of jobs or customers
- $E[T]$ mean sojourn time of a job or a customer in the system
- $E[X]$ mean number of jobs or customers in the system

Assumptions of Little's Theorem



- General black box system
- No assumptions about
  - distribution of arrival or service process
  - operating discipline (FIFO, LIFO, …)
- Little's law is not valid when the system generates or destroys work
  - e.g. due to a failure in a router, some packets are not forwarded and are not departing from the system or may initiate additional (signaling) traffic within the system.
- Only required assumption
  - two of the quantities $\lambda, E[X], E[T]$ exist and are finite
  - third quantity follows
- Theorem can be applied to *any stable* system and to any *part* of the system.
- Most famous operational law: helpful for measurements and analysis!

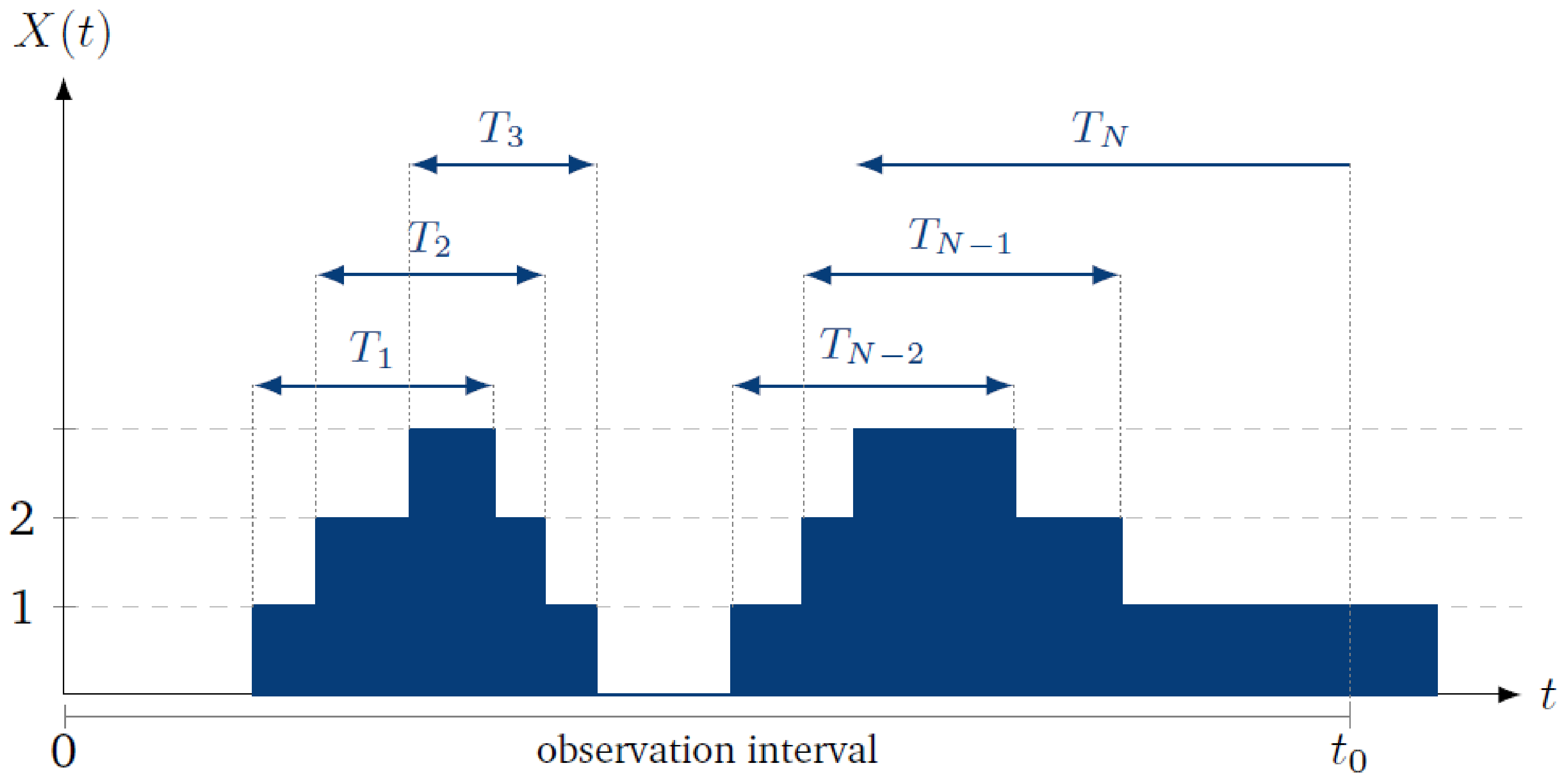
Simple Example of Little‘s Law

- Example 1: Waiting room at the doctor's
  - Arrival rate 10 patients/hour
  - Mean waiting time: 0.5 hours
  - Mean number of patients?

- Example 2: Waiting room at the doctor's
  - Arrival rate 10 patients/hour
  - Mean waiting time: 1.2 hours
  - Mean number of patients?

Proof of Little's Theorem

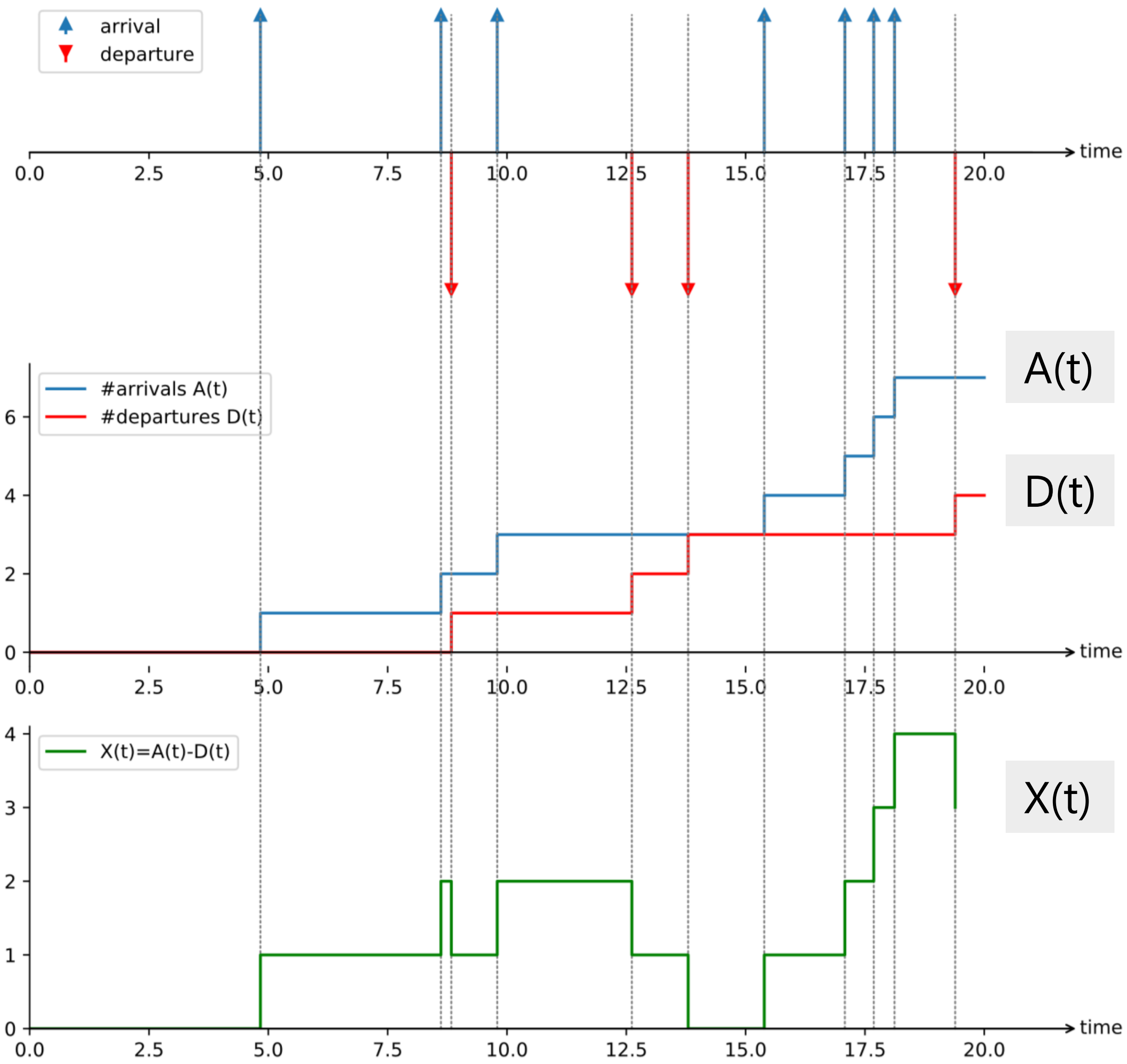
- During observation interval of length t_0: N arrivals of customers with sojourn times T_i.
- Number of customers at time t is $X(t)$.



lecture

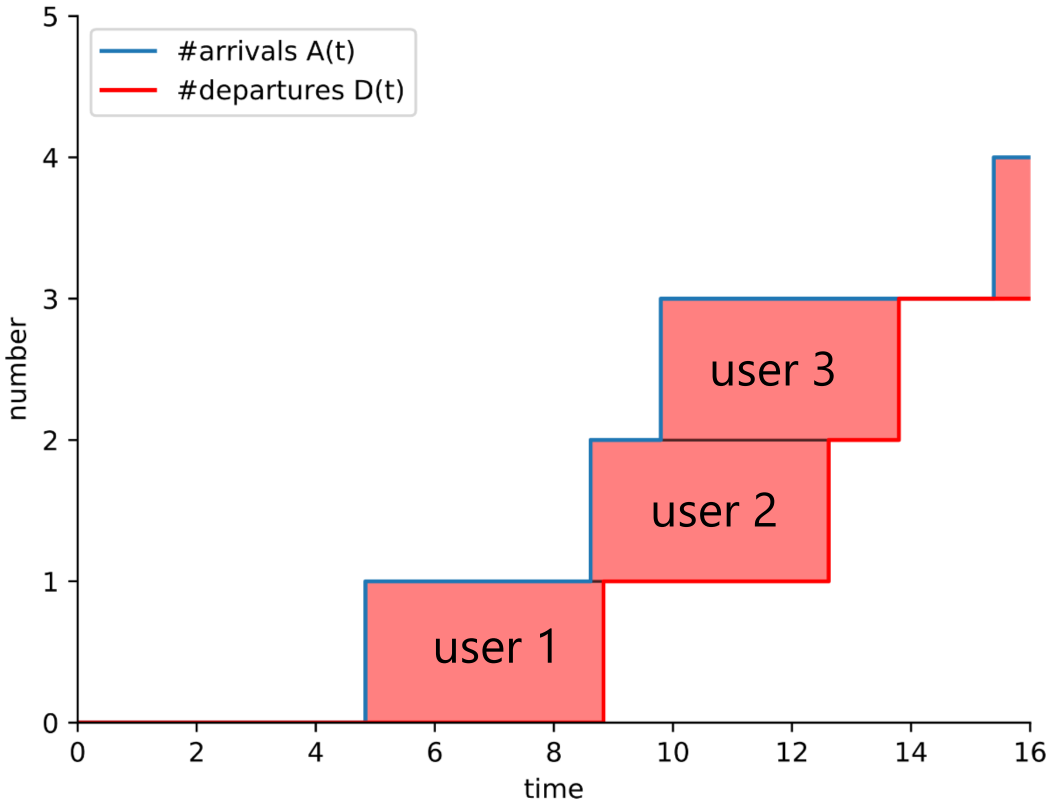
Whiteboard

Little's Law: Visualization (M/D/1)



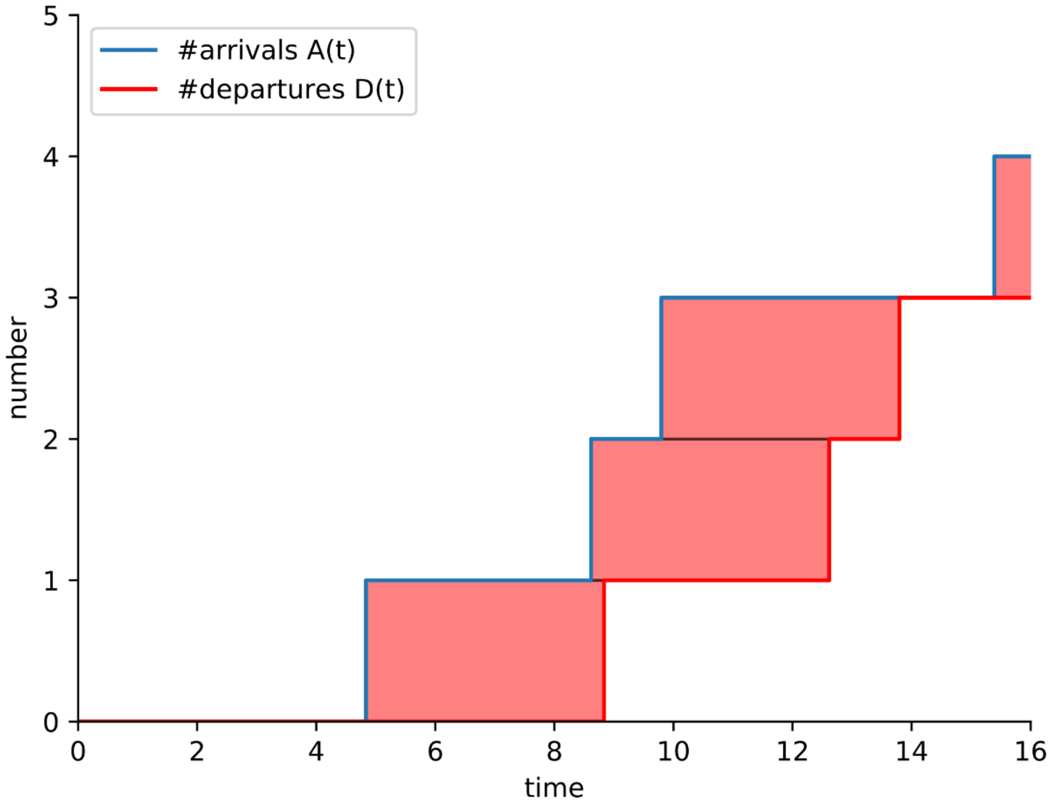
Accumulated Sojourn Time

- Area between $A(t)$ and $D(t)$: $\int_0^t X(\tau)d\tau$
- Accumulated sojourn time across all users



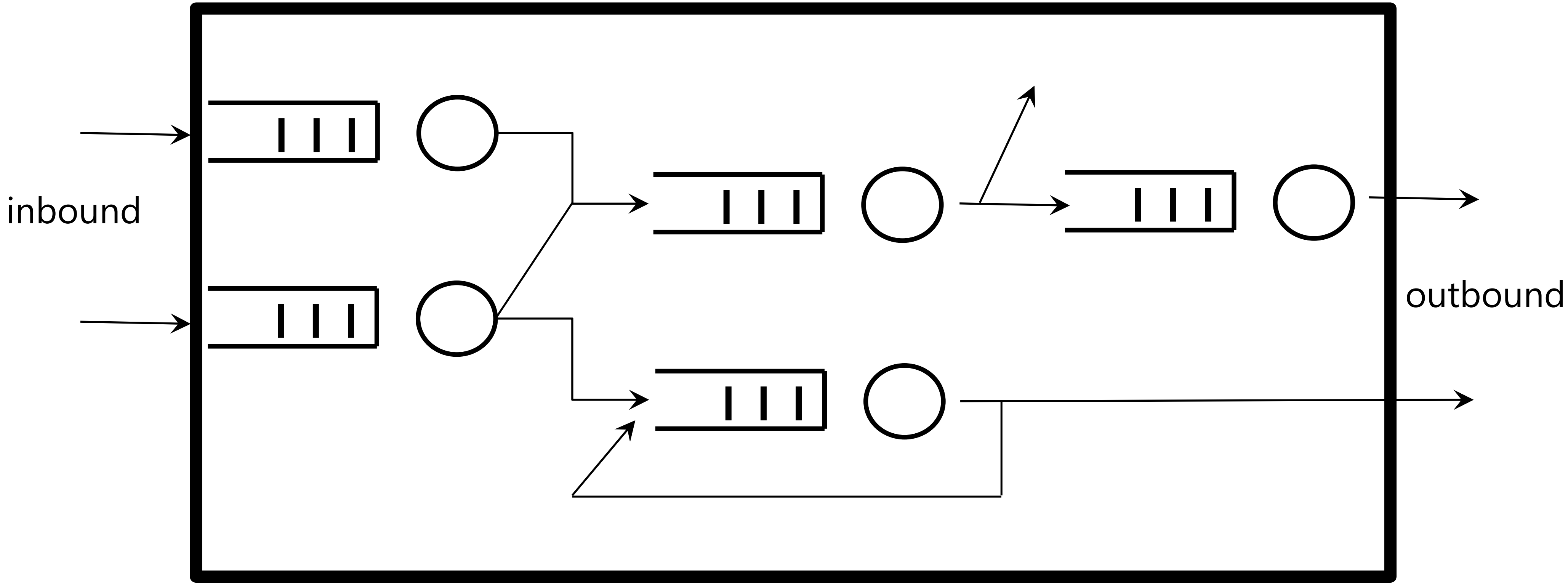
Queueing Discipline

- Do we observe the same mean waiting times for FIFO (first-in first-out) and LIFO (last-in first-out)?



Little‘s Law: Network of Queues

- Theorem can be applied to *any stable* system and to any *part* of the system.

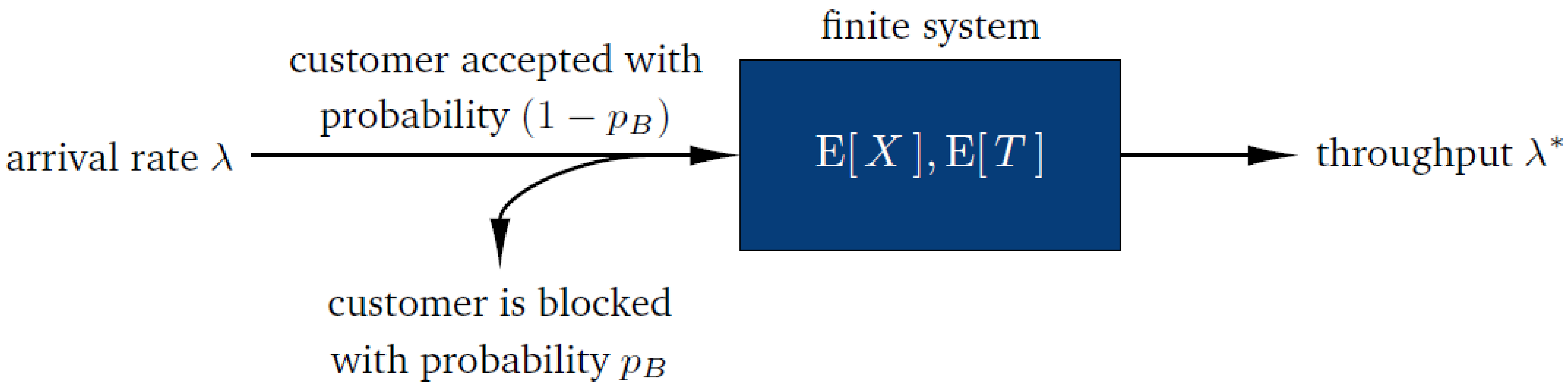


LITTLE'S THEOREM:
FINITE SYSTEM WITH BLOCKING

Loss system GI/GI/n-S

Finite System with Blocking

- Loss system GI/GI/n-S



lecture

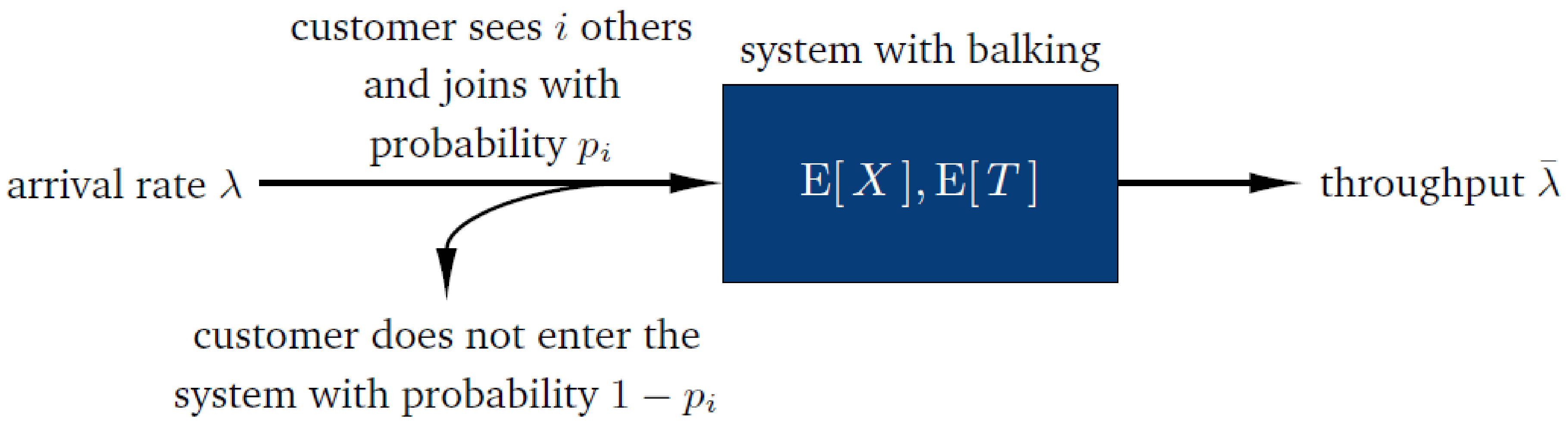
Whiteboard

LITTLE'S THEOREM: QUEUEING SYTEM WITH BALKING

Arriving customers refusing to enter the queue

Queueing System with Balking

- Customer arrives and sees the system in state i which means i other customers in system.
- With probability $1 - p_i$ the arriving customer refuses to enter the queue (balking)
- Example: waiting lines in a supermarket



lecture

Whiteboard

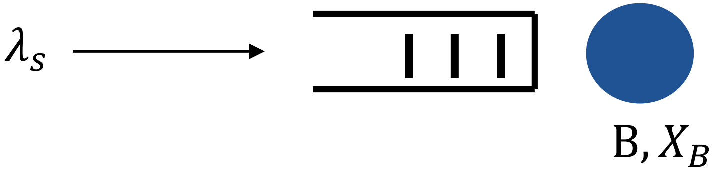
Utilization Law

Utilization of a server

The Utilization Law

- Applying Little's theorem to a server of a queueing system leads to the utilization law.

- Consider average number $E[X_B]$ of customers at a **single server** (i.e. not in waiting queue)
  - identical to **utilization** of the server: fraction of the time the server is busy
  - arrival rate (of accepted) customers at the server: λ_s
  - average service time of a customer: $\mathrm{E}[B]$



- Little's law yields the utilization law: $E[X_B] = \lambda_s \cdot E[B] < 1$

Delay System GI/GI/1-∞

- Consider the single server delay system with infinite waiting room: GI/GI/1-∞

- Utilization law leads to the probability that the system is empty

$$x(0) = P(X = 0) = 1 - E[X_B] = 1 - \lambda_s \cdot E[B]$$

- Example: IoT gateway
  - The measured throughput is 125 packets per second.
  - Each packet requires a forwarding time of 0.002 seconds.
  - What is the utilization?
  - What is the probability that the gateway is idle?

Serendipity…

A Note of Personal History (Little)

How did a sensible young PhD like me get involved in a crazy field like this? From 1957–1962, I taught operations research at the Case Institute of Technology in Cleveland (now Case Western Reserve University). I was asked to teach a course on queuing. OK. Initially I used my own notes, but when Morse (1958) came out, I used his book extensively. Queuing was taken by most of the OR graduate students and, indeed, one of these, Ron Wolff, went on to become a first class queuing theorist (Wolff 1989). One year we were at the point when we had done the basic Poisson-exponential queue and moved through multi-server queues, and some other general cases. I remarked, as many before and after me probably have (and Morse does), that the often reappearing formula $L = \lambda W$ seemed very general. In addition I gave the heuristic proof that is essentially Fig. 5.2 at the beginning of this chapter. After class I was talking to a number of students and one of them (Sid Hess) asked, "How hard would it be to prove it in general?" On the spur of the moment, I obligingly said, "I guess it shouldn't be too hard." Famous last words. Sid replied, "Then you should do it!"

The remark stuck in my mind and I started to think about the question from time to time. Clearly there was something fundamental going on, since, when you

Little, John DC, and Stephen C. Graves. "Little's law." *Building Intuition*. Springer US, 2008. 81-100.

General Results for Delay Systems

GI/GI/n delay systems

Notation

- Notation of random variables
  - interarrival time A
  - waiting time W
  - number of customers in queue X_W
  - service time B
  - number of customers in service X_B

- W, X_W
- B, X_B

- Notation of rates
  - arrival rate $\lambda = \frac{1}{E[A]}$ of customers
  - service rate $\mu = \frac{1}{E[B]}$ of a single server

- Notation of load
  - offered load $a = \lambda \cdot E[B]$
  - normalized offered load $\rho = \frac{a}{n} = \frac{\lambda}{n\mu}$

General Results for GI/GI/n Delay Systems

- **Sojourn time** or **response time:** $T = W + B$
- Number of customers in the system: $X = X_W + X_B$

- **Stability condition** for delay systems: $a = \lambda \cdot E[B] < n$ or $\rho = a/n < 1$
- *Note*: Finite buffer systems GI/GI/n-S are always stable due to blocking.

- Mean number of **busy servers**: $E[X_B] = \lambda \cdot E[B] = a$

- **Utilization**, i.e. fraction of time each server is busy: $\rho = a/n$

Loss Formula

GI/GI/n-S loss systems

Loss Formula for GI/GI/n-S Loss Systems

- Consider a single server in GI/GI/n-S system with ρ_s being the mean offered load of the server
- Mean arrival rate at the considered server is $\lambda_s = \lambda / n$
- It is: $\rho_s = \frac{\lambda \cdot E[B]}{n} = \lambda_s \cdot E[B]$

- Assuming that an arbitrarily chosen server is idle with probability ϕ_s

- Blocking probability that an arbitrary customer is blocked at that server

loss formula $p_B = 1 - \frac{1-\phi_s}{\rho_s}$

- GI/GI/1-∞ delay system?