204528
QUEUEING THEORY AND APPLICATIONS IN NETWORKS

Assoc. Prof. Anan Phonphoem, Ph.D. (รศ.ดร. อนันต์ ผลเพิ่ม)
Computer Engineering Department, Kasetsart University

Outline

- Overview
- Queueing system
- Queueing process characteristics
- Notation
- Basic queueing system

Queue in real life situation

- Wait for buying lunch
- Wait for taking a ride in Disney World
- Wait for withdraw money from ATM
- Wait for a green light
- Wait for Bug 1113 to pick up our call
- Etc.

http://michael.toren.net/

Who like to wait?

- Customer does not
- Entrepreneur does not like it either
  - Cost more money
  - Cost more space for waiting
  - Customers loss
  - Unhappy customers

http://www.ac-nancy-metz.fr/enseign/anglais/Henry/transport.htm

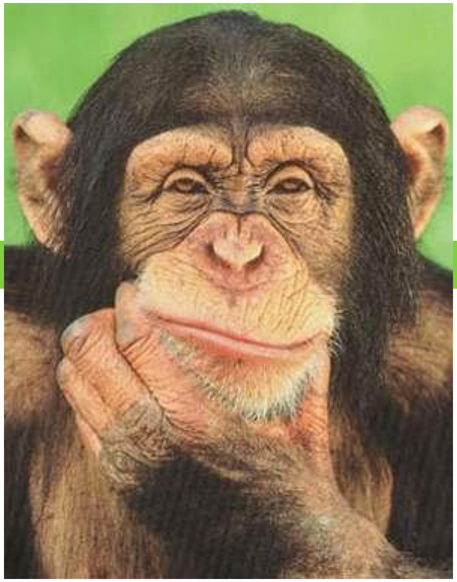
So, why waiting?

- Demand > Service availability
- Why service is not enough?
  - Not economics
  - No space
  - Unpredictable arrival
  - Slow servers
  - HOL (Head of line) blocking

Still Waiting …

- Interesting questions for customers?
  - How long do I need to wait?
  - How many people are now in the line?
  - When should I come to get serve faster?

Still Waiting …



http://gotoknow.org/file/lilygroup/thinkingshi.jpg

- Interesting questions for service provider?
  - How big is the waiting area?
  - How many customers leave?
  - Should we add some more tellers?
  - Should the system form 1 or many queues?
  - Should the system provide a fast lane?

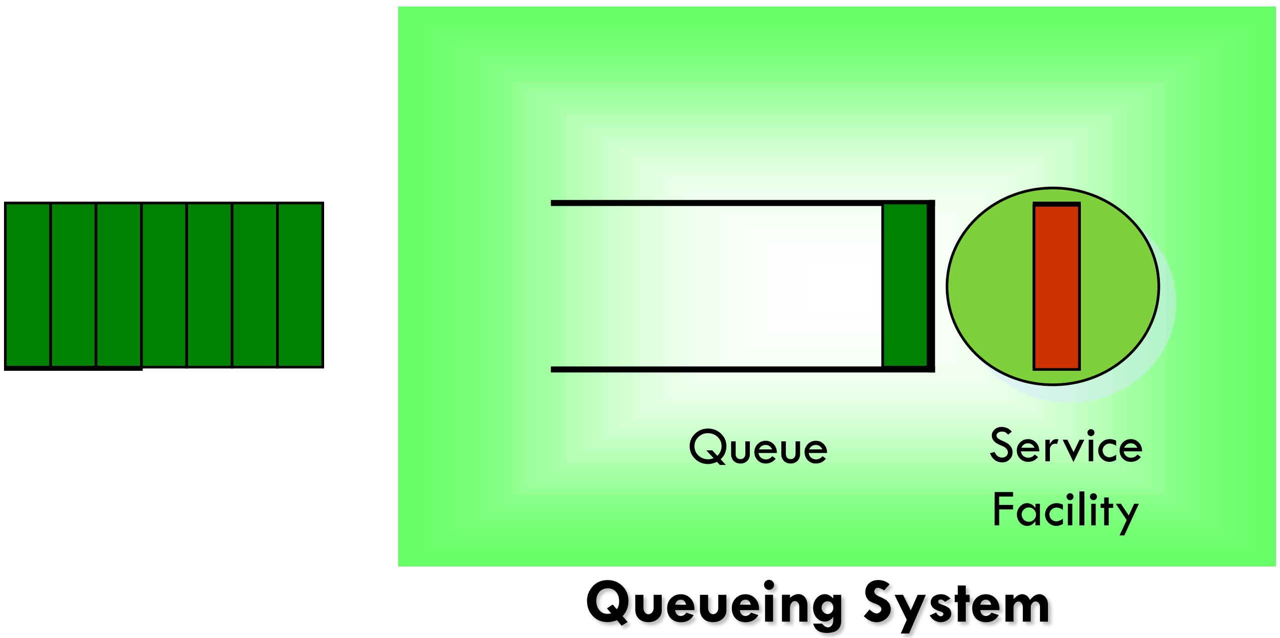
Here comes …Queueing Theory

- Describe the queue phenomena
  - Waiting and serving
- Model the system mathematically
- Try to answer those questions

Queueing System

- Arriving for service
- Waiting for service
- Getting serve
- Leaving the system

General queueing system



Queueing process characteristics

- Arrival pattern
- Service pattern
- Queue discipline
- System capacity
- Number of service channels
- Number of service stages

Arrival pattern

- Stochastic
  - Probability distribution
  - Single or batch arrival
- Behavior of customer
  - Patient customer
    - Wait forever
  - Impatient customer
    - Wait for a period and decide to leave
    - See the long line and decide not to join
    - Change the waiting line

Arrival pattern

- Is it time dependent?
  - Stationary arrival pattern
    (time independent – probability distribution)
  - Non-stationary arrival pattern

Service pattern

- Distribution for service time
- Single or batch (parallel machine) service
- Service process depends on number of customers waiting (state dependent)
- Very fast service → still have a line?
  - Depends also on the arrival
  - May assume mutually independent

Queue discipline

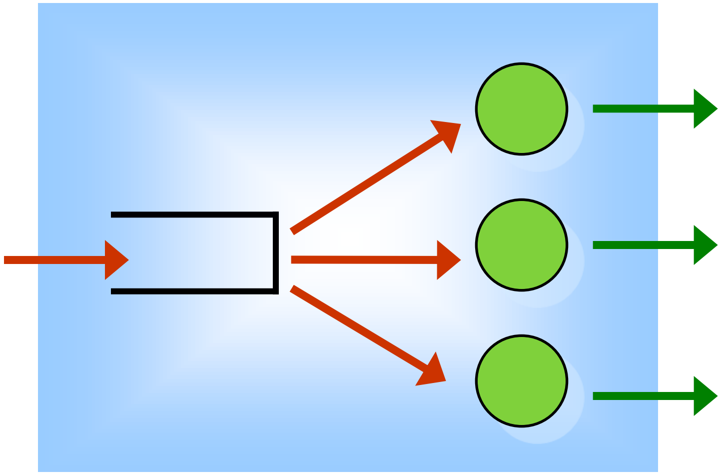
- Manner of customers to get serve
- First come, first serve
- Last come, first serve
- Random serve
- Priority serve
  - Preemptive
  - Nonpreemptive

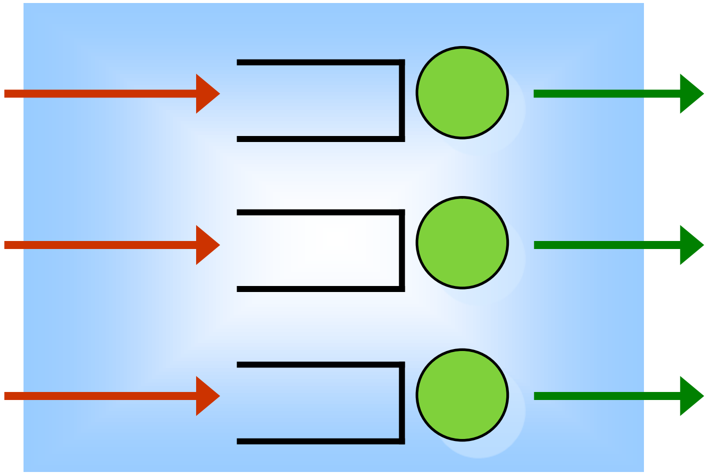
System capacity

- Finite capacity
  - Maximum system size
- Infinite capacity

Number of service channels

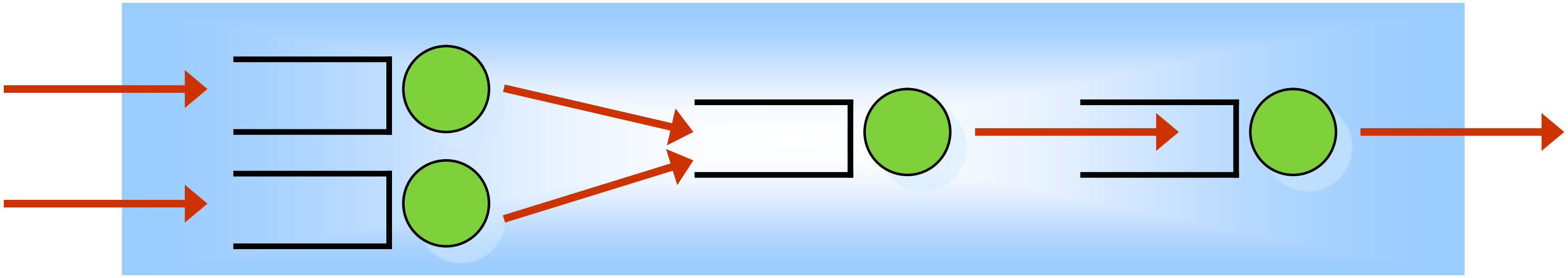
- Multiserver queueing system
  - Single line service
  - Multiple line service





Stages of service

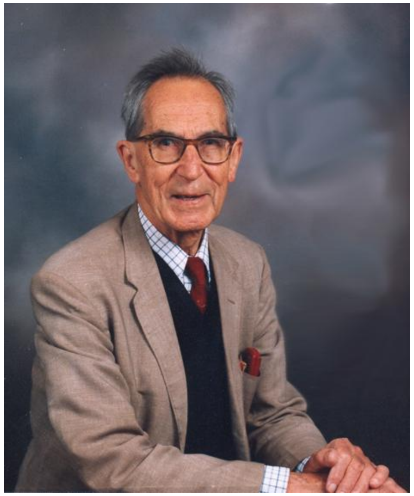
- Single stage
- Multiple stages
  - Without feedback (Entrance Exam)
  - With feedback (Manufacturing)



Queueing Notation

- Kendall's notation (1953)

  A / B / X / Y / Z



Prof. David George Kendall (1918-2007)
http://www.statslab.cam.ac.uk/kendall/index.html

A : Interarrival-time distribution

B : Service time distribution

X : # of parallel service channels

Y : System capacity

Z : Queue discipline

Queueing Notation A/B/X/Y/Z

Characteristics	Symbol	Explanation
A & **B** (Interarrival / Service Time)	M D E_k G	Exponential (Memory less) Deterministic Erlang General
X (# Servers)	1,2,…,∞	
Y (Capacity)	1,2,…,∞	
Z (Q discipline)	FCFS, PR	

Queueing Notation A/B/X/Y/Z

- $M/M/3/\infty/FCFS$
  - Exponential interarrival time
  - Exponential service time
  - 3 parallel servers
  - Unlimited space
  - First-come first-serve queue discipline

Queueing Notation A/B/X/Y/Z

- M/D/1
  - Exponential interarrival time
  - Deterministic service time
  - 1 server
  - (default) Unlimited space
  - (default) FCFS queue discipline

Queueing Notation A/B/X/Y/Z

- M/M/1
- M/M/c/k
- M/M/∞
- Ek/M/1
- M/G/1
- G/M/m
- G/G/1

Basic queueing system

- G/G/m
  - Interarrival time with distribution A(t)
  - Service time with distribution B(x)
  - m servers
- C_n: The n[th] customer enters system

Basic queueing system

- τ_n: arrival time for C_n
- t_n: Interarrival time ($\tau_n - \tau_{n-1}$)
- x_n: service time for C_n

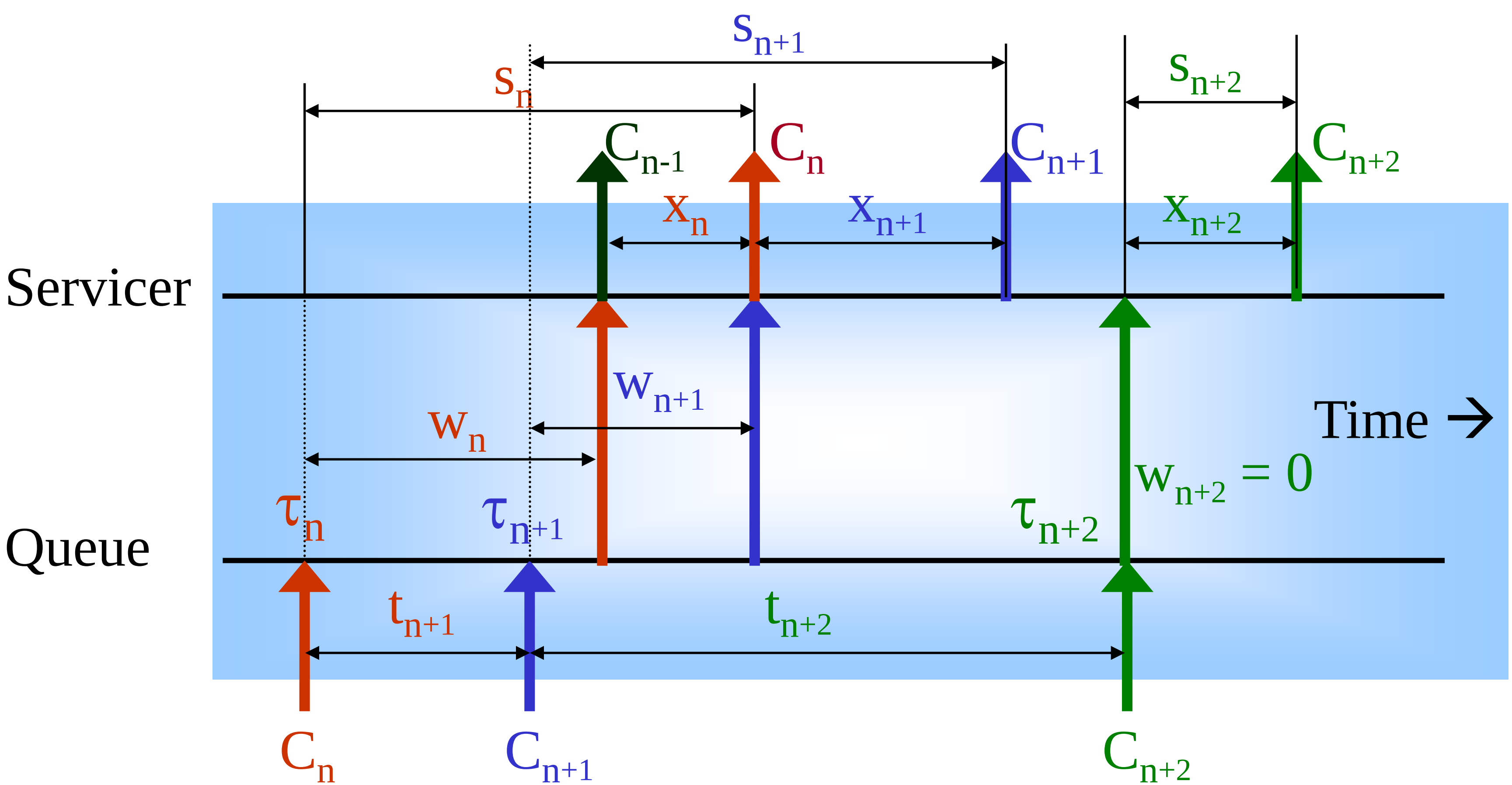
$$P[\ t_n \le t\ ] = A(t)$$

$$P[\ x_n \le x\ ] = B(x)$$

Basic queueing system

- w_n: waiting time in queue for C_n
- s_n: system time for C_n ➔ $(w_n + x_n)$
- λ : average arrival rate
- μ : average service rate
- $\tilde{t} = \lim_{n \to \infty} t_n = \dfrac{1}{\lambda}$
- $\tilde{x} = \lim_{x \to \infty} x_n = \dfrac{1}{\mu}$

Time diagram notation

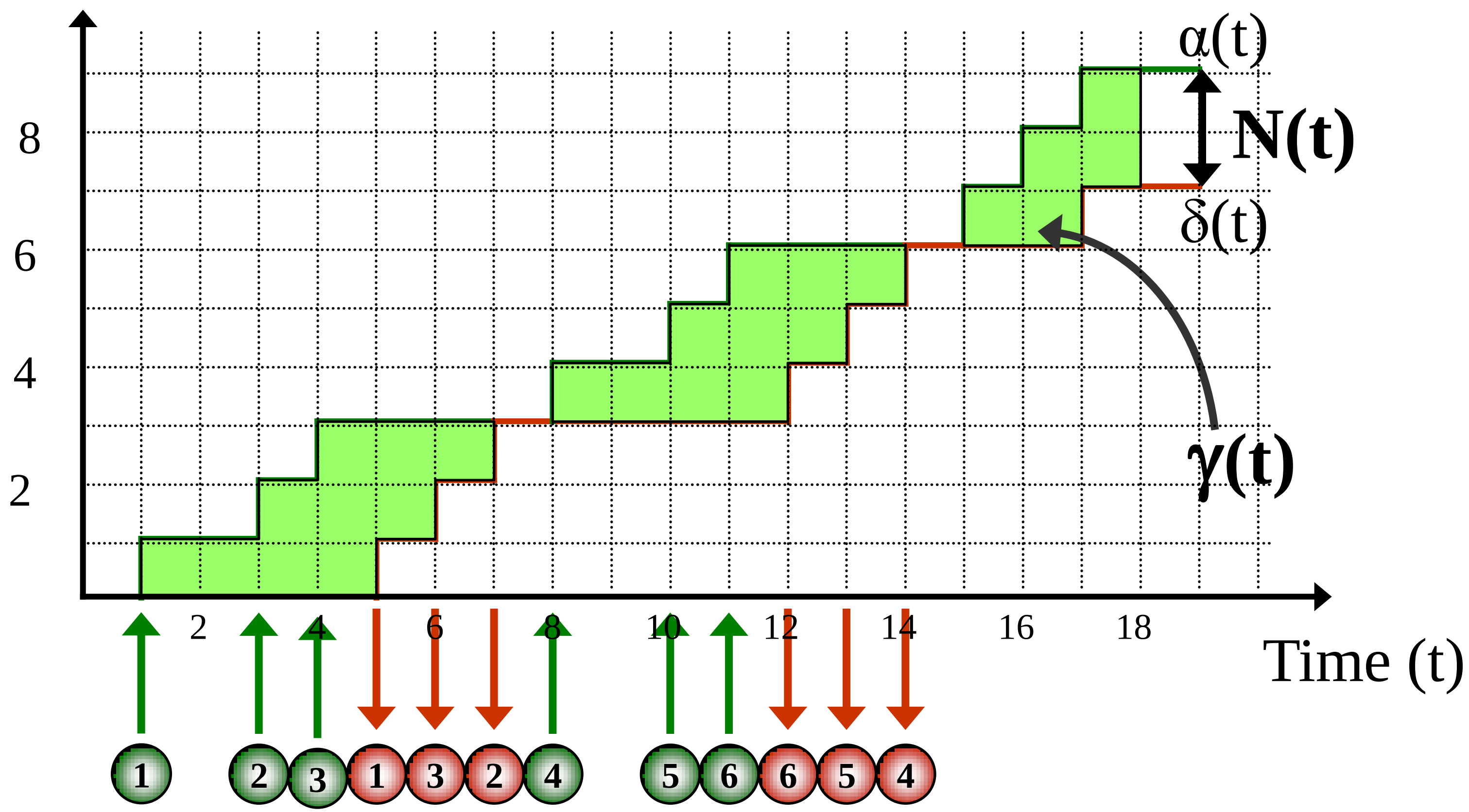


Basic queueing system

- N(t): # of customers in the system @time t
- U(t): Unfinished work @time t
  - U(t) = 0 ➔ System idle
  - U(t) > 0 ➔ System busy
- α(t): # of arrivals in (0,t)
- δ(t): # of departures in (0,t)

Basic queueing system

$N(t) = \alpha(t) - \delta(t)$



Basic queueing system

- λ_t : arrival rate
- $\lambda_t = \dfrac{\alpha(t)}{t}$ = # of arrival / time
- $\gamma(t)$: total time all customers spent in the system (customer-seconds)
- $T_t = \dfrac{\gamma(t)}{\alpha(t)}$ = system time / customer

Basic queueing system

- $\overline{N}_t = \dfrac{\gamma(t)}{t} =$ avg.# customers in system

$$= \frac{\alpha(t)}{t}\,\frac{\gamma(t)}{\alpha(t)}$$

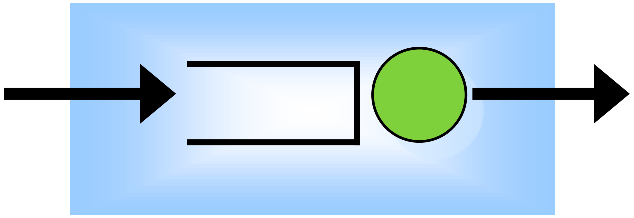
$$= \lambda_t T_t$$

- As $t \rightarrow \infty$
  - $\lim_{t\rightarrow\infty} \lambda_t \rightarrow \lambda$
  - $\lim_{t\rightarrow\infty} T_t \rightarrow T$

$$\boxed{\overline{N} = \lambda T}$$

Little's Result

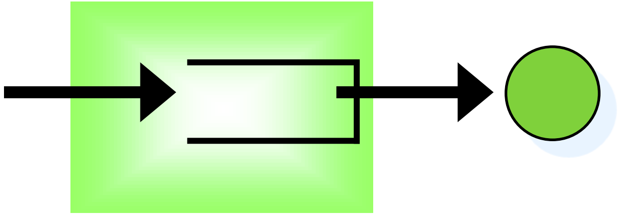
$$\overline{N} = \lambda T$$



"The *average number of customers* in a queueing system is equal to the arrival rate of customers to that system, times the average time spent in the system"

Little's Result

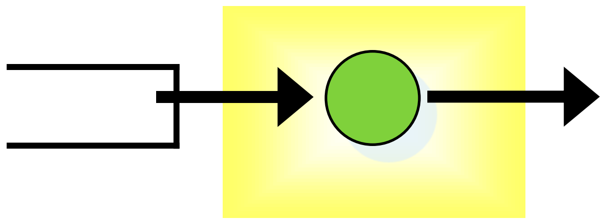
$$\overline{N}_q = \lambda W$$



- $\overline{N}_q$: avg.# of customers in queue
- λ : arrival rate
- W : avg. time spent in the queue

Little's Result

$$\overline{N}_s = \lambda\overline{x}$$



- $\overline{N}_s$: avg.# customers in service fac.
- λ : arrival rate
- $\overline{x}$: avg. time spent in the service fac.

Basic queueing system

- $T = W + \overline{x}$
- ρ : Utilization factor

  : rate of work / rate of max. capacity
- $\rho = \lambda \overline{x}$; for a single server
- $\rho = \dfrac{\lambda \overline{x}}{m}$; for m servers
- for G/G/1 to be stable: $0 \le \rho < 1$