

204528

**QUEUEING THEORY AND
APPLICATIONS IN NETWORKS**

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

Outline

2

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

Queue in real life situation

3

- 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?

4

- 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/enseignement/anglais/Henry/transport.htm>

So, why waiting?

5

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

Still Waiting ...

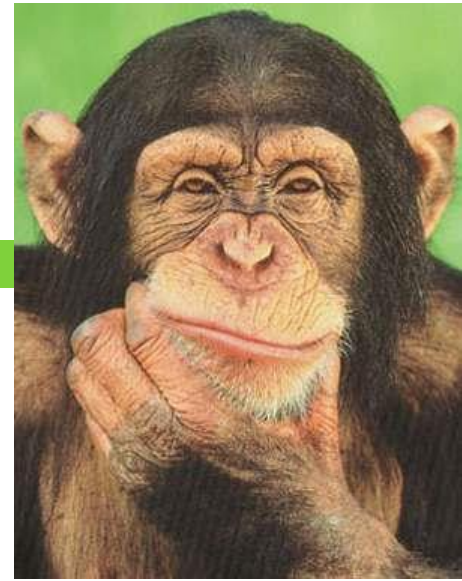
6

- 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 ...

7

- 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?



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

Here comes ...Queueing Theory

8

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

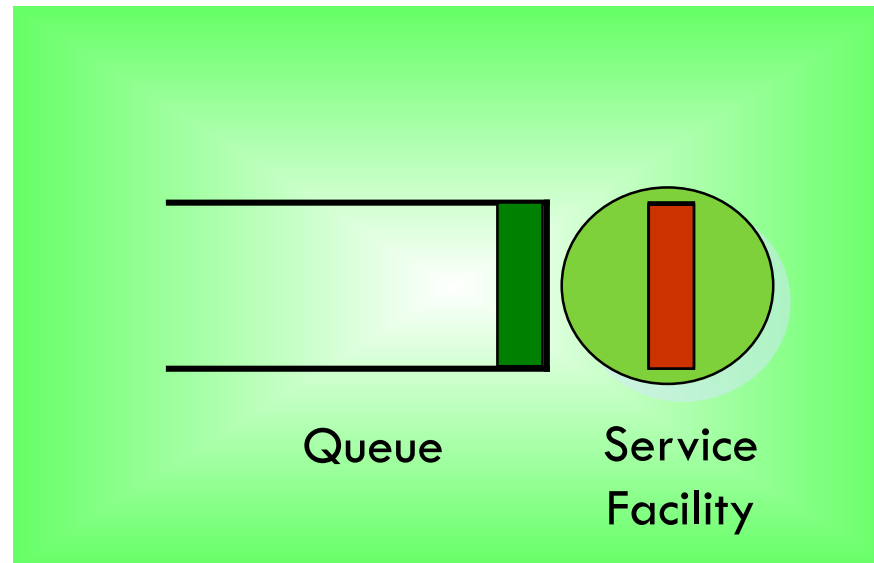
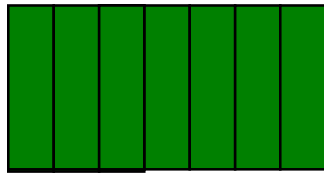
Queueing System

9

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

General queueing system

10



Queueing System

Queueing process characteristics

11

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

Arrival pattern

12

- 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

13

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

Service pattern

14

- 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

15

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

System capacity

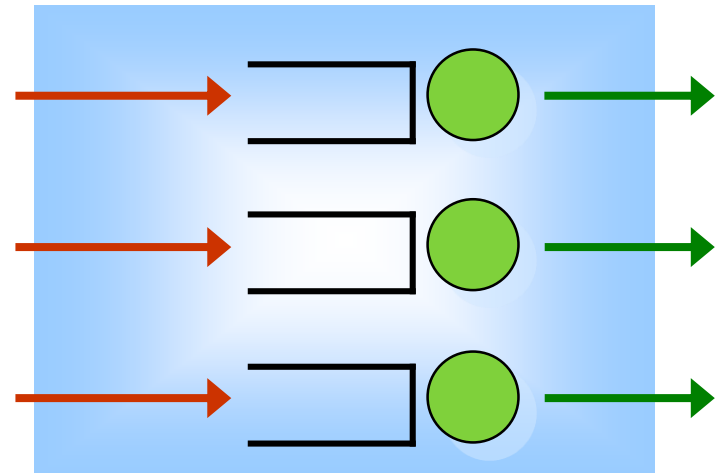
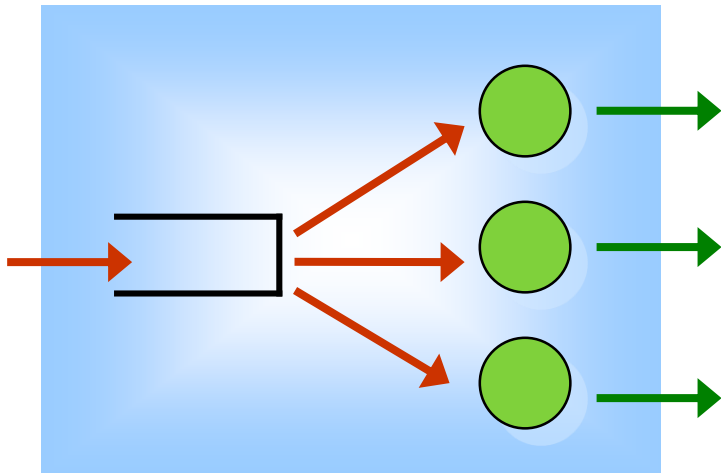
16

- Finite capacity
 - Maximum system size
- Infinite capacity

Number of service channels

17

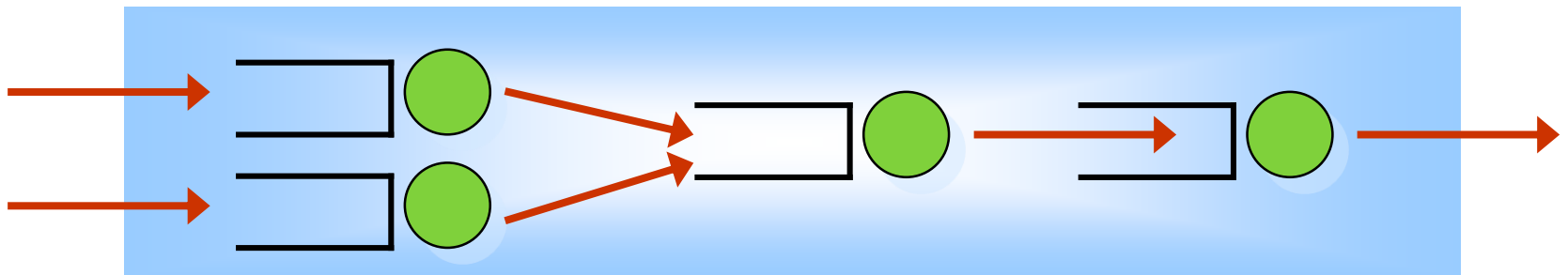
- Multiserver queueing system
 - Single line service
 - Multiple line service



Stages of service

18

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



Queueing Notation

19

- Kendall's notation (1953)

A / B / X / Y / Z

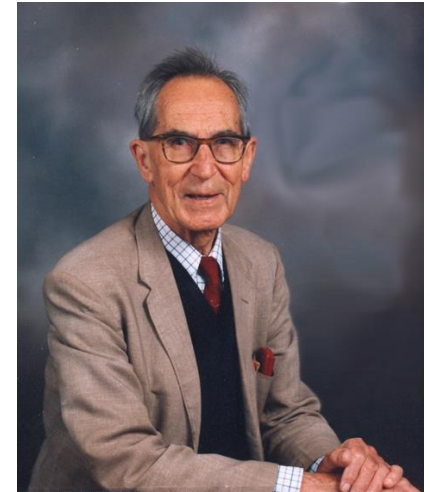
A : Interarrival-time distribution

B : Service time distribution

X : # of parallel service channels

Y : System capacity

Z : Queue discipline



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

Queueing Notation $A/B/X/Y/Z$

20

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

Queueing Notation $A/B/X/Y/Z$

21

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

Queueing Notation $A/B/X/Y/Z$

22

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

Queueing Notation $A/B/X/Y/Z$

23

- $M/M/1$
- $M/M/c/k$
- $M/M/\infty$
- $E_k/M/1$
- $M/G/1$
- $G/M/m$
- $G/G/1$

Basic queueing system

24

- $G/G/m$
 - 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

25

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

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

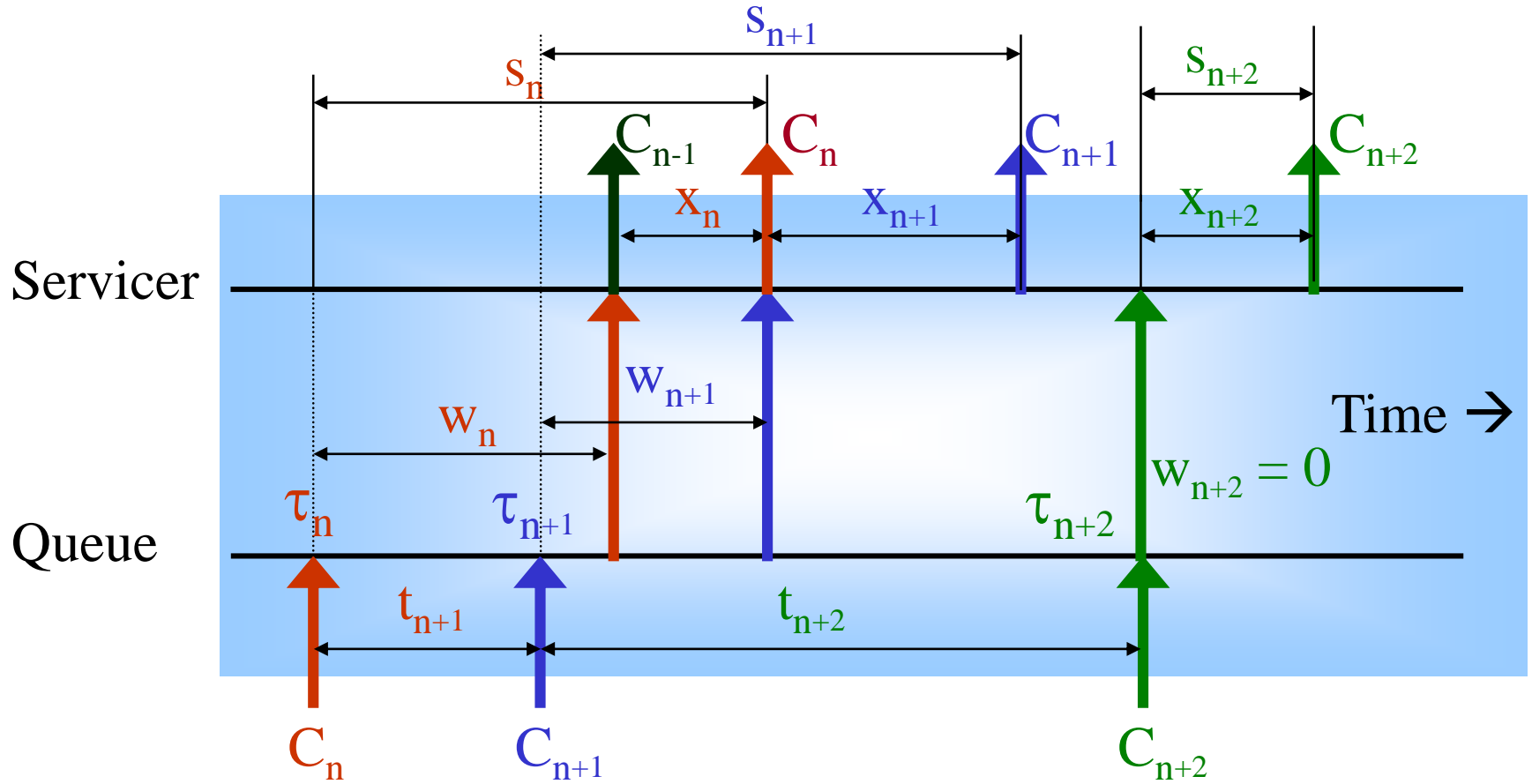
$$P[x_n \leq x] = B(x)$$

Basic queueing system

26

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

Time diagram notation



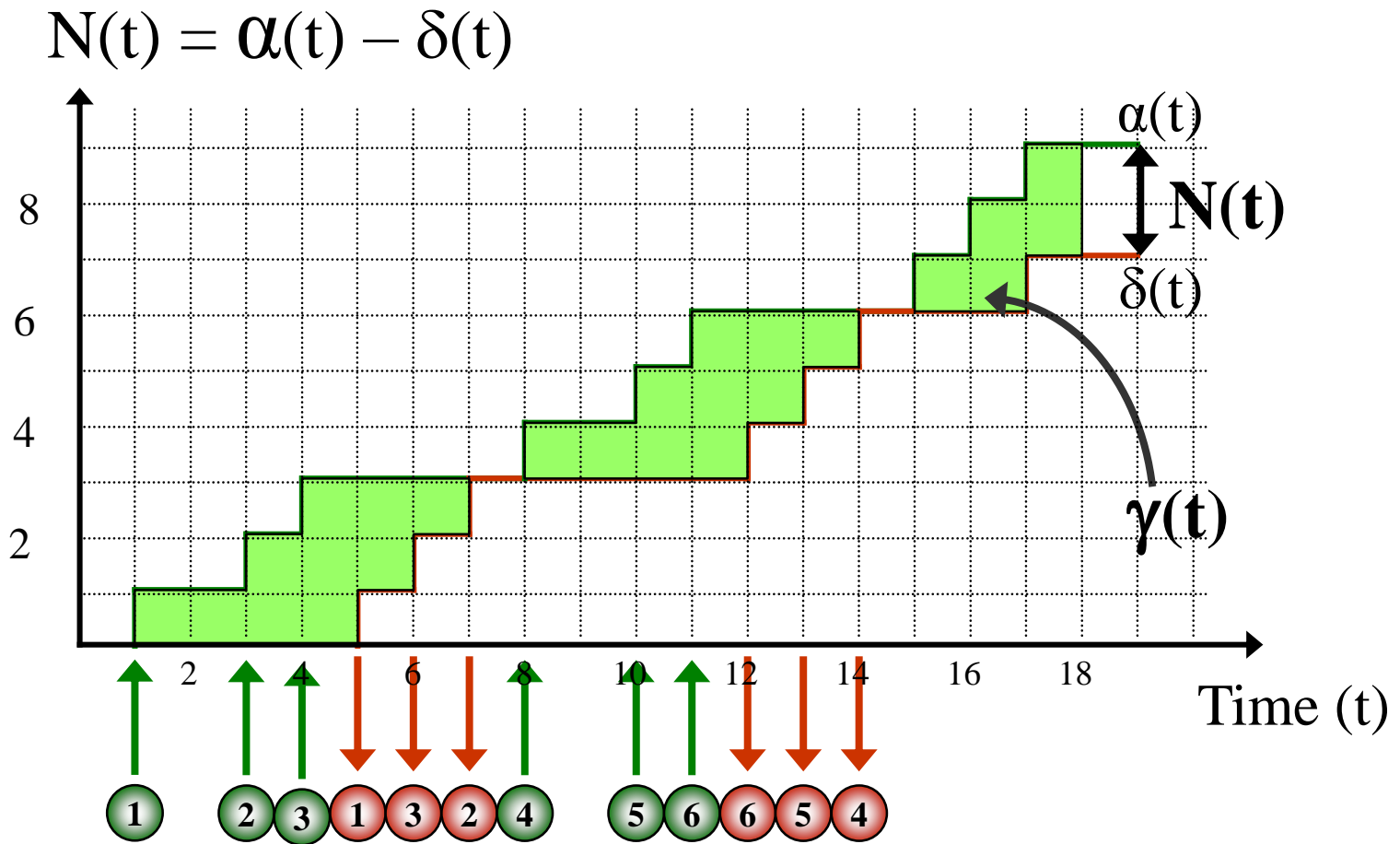
Basic queueing system

28

- $N(t)$: # of customers in the system @time t
- $U(t)$: Unfinished work @time t
 - $U(t) = 0 \quad \rightarrow$ System idle
 - $U(t) > 0 \quad \rightarrow$ System busy
- $\alpha(t)$: # of arrivals in $(0,t)$
- $\delta(t)$: # of departures in $(0,t)$

Basic queueing system

29



Basic queueing system

30

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

Basic queueing system

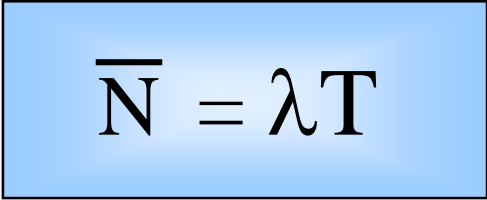
31

- $\bar{N}_t = \frac{\gamma(t)}{t} = \text{avg. \# customers in system}$

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

$$= \lambda_t T_t$$

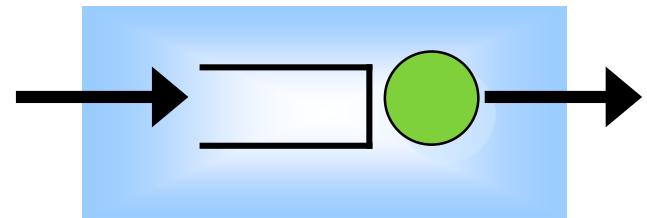
- As $t \rightarrow \infty$
 - $\lim_{t \rightarrow \infty} \lambda_t \rightarrow \lambda$
 - $\lim_{t \rightarrow \infty} T_t \rightarrow T$


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

Little's Result

32

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

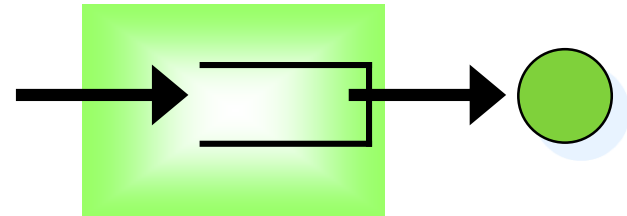


“The *average number of customers* in a queueing system is equal to the **arrival rate of customers** to that system, times the **average time spent in the system**”

Little's Result

33

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

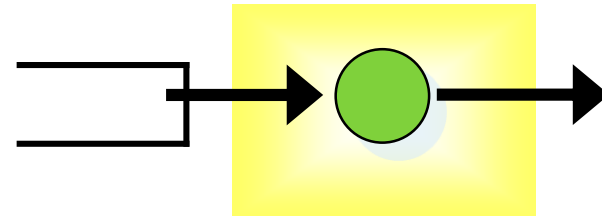


- \bar{N}_q : avg.# of customers in queue
- λ : arrival rate
- W : avg. time spent in the queue

Little's Result

34

$$\bar{N}_s = \lambda \bar{X}$$



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

Basic queueing system

35

- $T = W + \bar{X}$
- ρ : Utilization factor
: rate of work / rate of max. capacity
- $\rho = \lambda \bar{X}$; for a single server
- $\rho = \frac{\lambda \bar{X}}{m}$; for m servers
- for G/G/1 to be stable: $0 \leq \rho < 1$