1918
TALLINNA
TEHNIKAÜLIKOOL

IDU0010 ERP, CRM ja Datawarehouse süsteemid

Loeng 9

Järjekorrateooria

Enn Õunapuu

enn@cc.ttu.ee

Kava

- Järjekorrateooria
- Näited
- Järeldused
- Küsimused

Loengu eesmärk

Loengu eesmärgiks on anda alus mõningate tähtsate protsesside optimeerimiseks

Queuing theory definitions

- (Bose) “the basic phenomenon of queueing arises whenever a shared facility needs to be accessed for service by a large number of jobs or customers.”
- (Wolff) “The primary tool for studying these problems [of congestions] is known as queueing theory.”
- (Kleinrock) “We study the phenomena of standing, waiting, and serving, and we call this study Queueing Theory." "Any system in which arrivals place demands upon a finite capacity resource may be termed a queueing system.”
- (Mathworld) “The study of the waiting times, lengths, and other properties of queues.”

http://www2.uwindsor.ca/~hlynka/queue.html

Applications of Queuing Theory

- Telecommunications
- Traffic control
- Determining the sequence of computer operations
- Predicting computer performance
- Health services (eg. control of hospital bed assignments)
- Airport traffic, airline ticket sales
- Layout of manufacturing systems.
- In computers, jobs share many resources: CPU, disks, devices
- Only one can access at a time, and others must wait in queues
- Queuing theory helps determine time jobs spend in queue

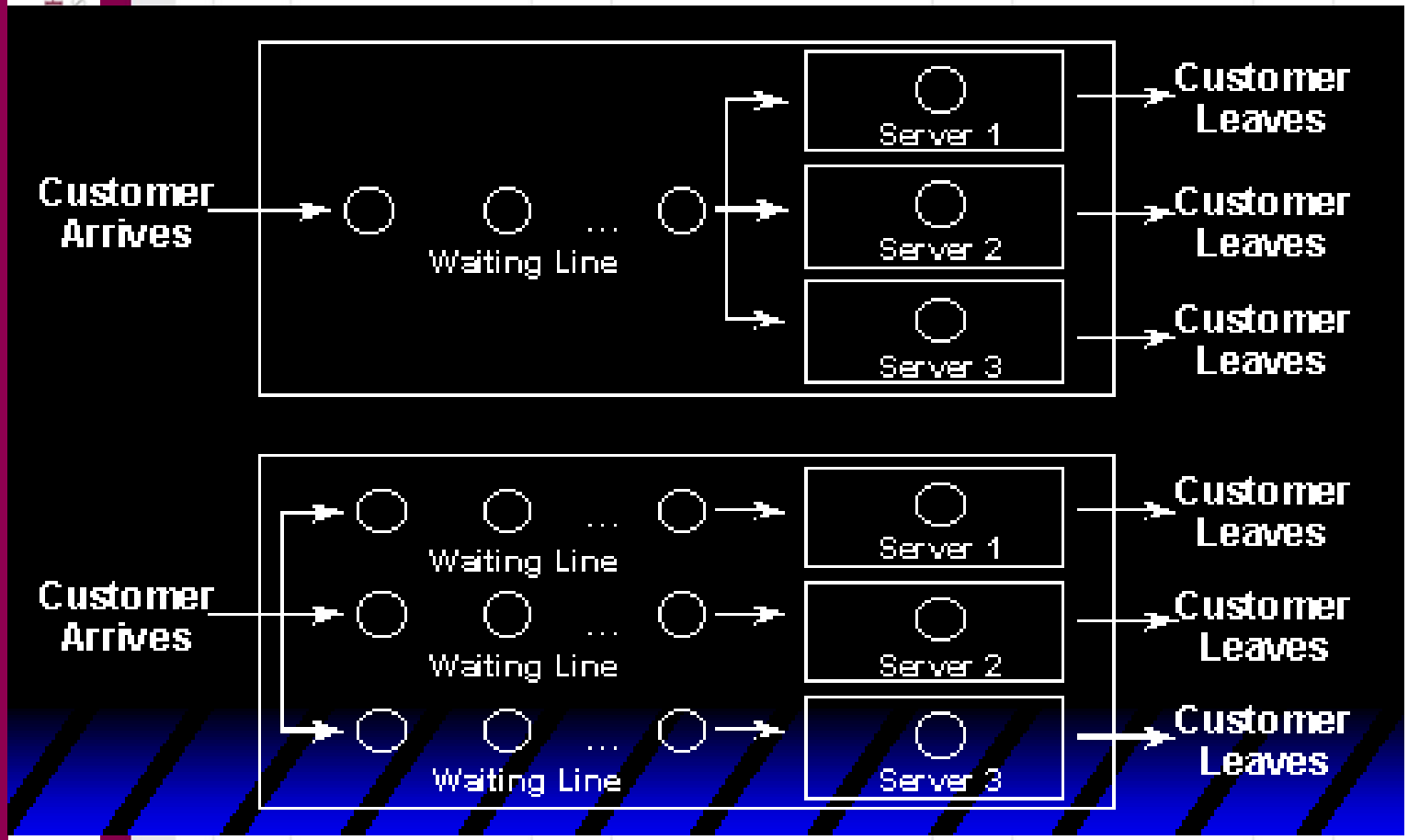
http://www2.uwindsor.ca/~hlynka/queue.html

Example application of queuing theory

- In many retail stores and banks
  - multiple line/multiple checkout system → a queuing system where customers wait for the next available cashier
  - We can prove using queuing theory that : throughput improves increases when queues are used instead of separate lines

http://www.andrews.edu/~calkins/math/webtexts/prod10.htm#0

Example application of queuing theory

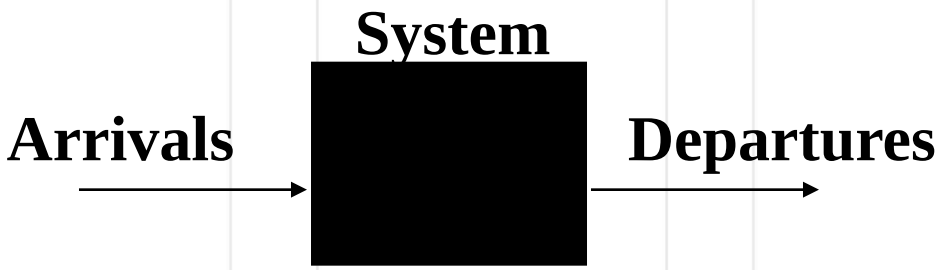


http://www.bsbpa.umkc.edu/classes/ashley/Chaptr14/sld006.htm

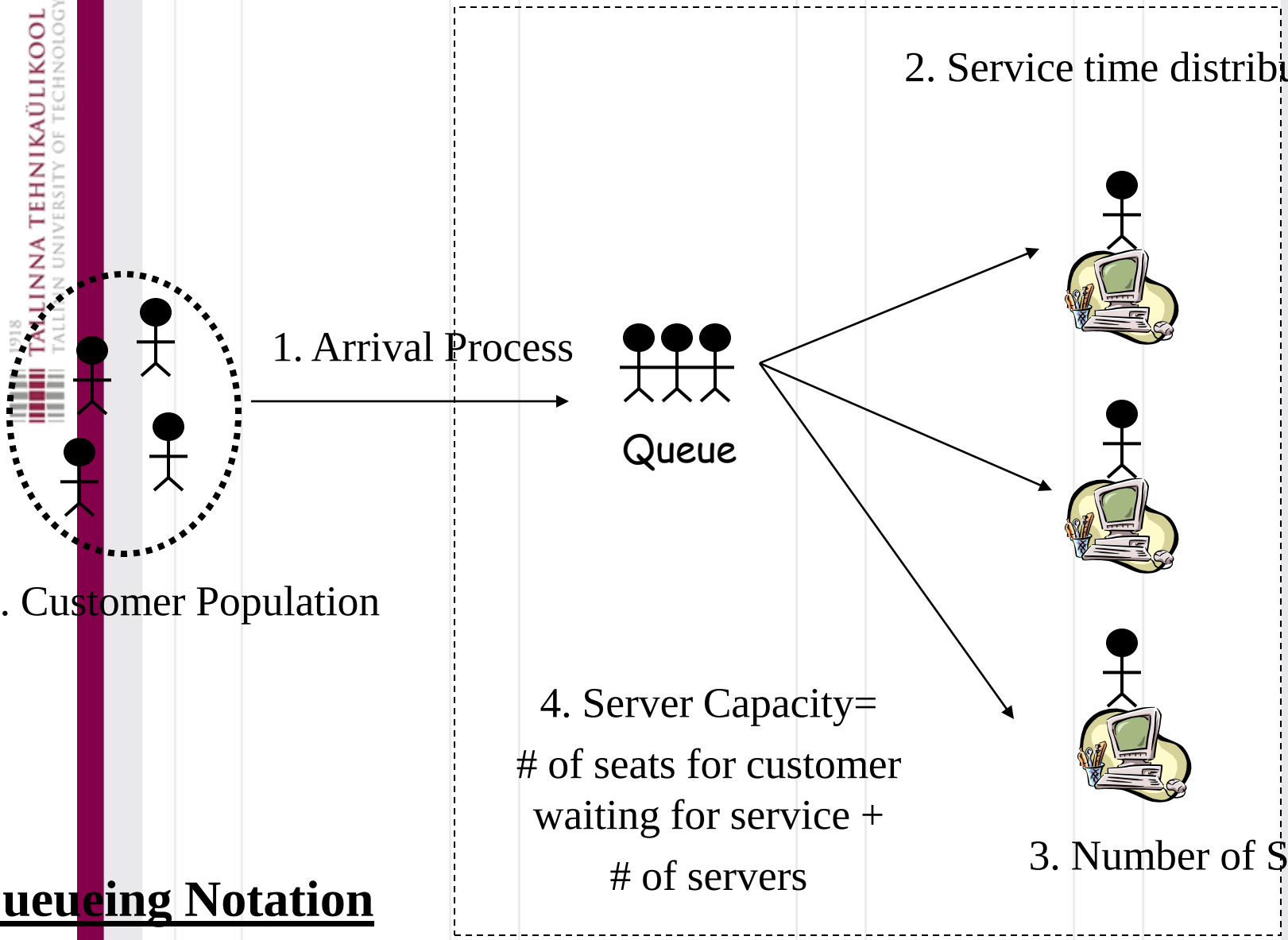
Queuing theory for studying networks

- View network as collections of queues
  - FIFO data-structures
- Queuing theory provides probabilistic analysis of these queues
- Examples:
  - Average length
  - Average waiting time
  - Probability queue is at a certain length
  - Probability a packet will be lost

Little's Law



- Little's Law:
  Mean number tasks in system = mean arrival rate x mean response time
  - Observed before, Little was first to prove
- Applies to any system in **equilibrium**, as long as nothing in black box is creating or destroying tasks



Queueing Notation

- Arrival process
  - Let's say customers arrive at $t_1, t_2, \ldots, t_j$
  - Random variables $\tau_j = t_j - t_{j-1}$ are inter arrival times.
  - Usually assume inter arrival times τ_j are independent, identically distributed (IID).
  - Most common arrival processes are Poisson arrivals: if inter arrival times are IID and exponentially distributed, then the arrival rate follows a Poisson distribution → *Poisson Process*
- Service time distribution
  - Amount of time each customer spends at the server
  - Again, usually assume IID
  - Most commonly used distribution is the exponential distribution.

Number of servers

If servers are not identical, divide them into groups of identical servers with separate queues for each group→ Each group is a queuing system

- System capacity

  the number of places for waiting customers + the number of servers
- Population size: The total number of potential customers who can ever come to the system.
- Service discipline: The order where the customers are served. (e.g., FCFS, LCFS, RR, etc)

Kendall notation

A/S/m/B/K/SD

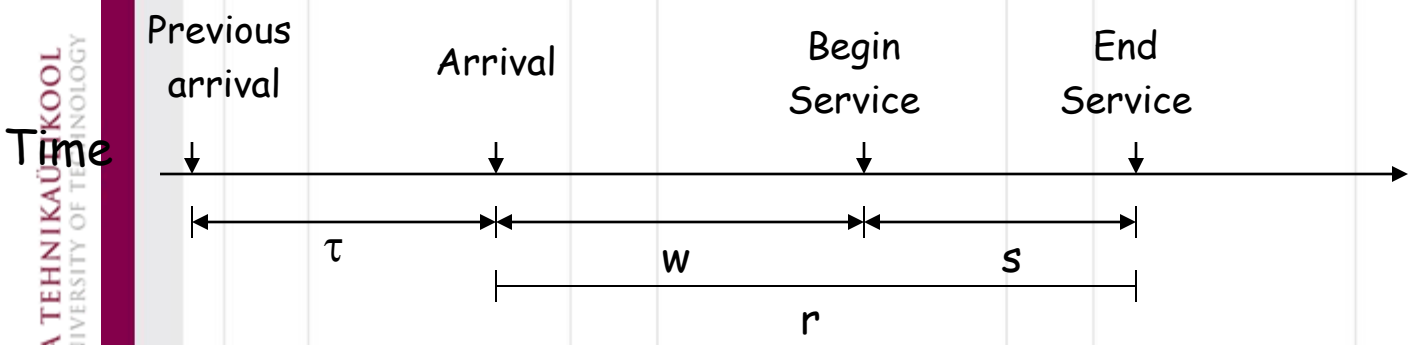
- A is Arrival time distribution
- S is Service time distribution
- m is number of servers
- B is number of buffers (system capacity)
- K is population size
- SD is service discipline

Kendall notation

- The distributions for interarrival time and service times are generally denoted by
  - M Exponential (M means “memoryless” in that the current arrival is independent of the past)
  - E_k Erlang with parameter k
  - H_k Hyperexponential with parameter k
  - D Deterministic (time are constant)
  - G General (the results are valid for all distributions)
- Assume only individual arrivals (no bulk arrivals)

Notation Example

- M/M/3/20/1500/FCFS – single queue system with:
  - Exponentially distributed arrivals
  - Exponentially distributed service times
  - Three servers
  - Capacity 20 (17 spaces for waiting customers)
  - Population is 1500 total
  - Service discipline is FCFS
- Often, assume infinite queue and infinite population and FCFS, so just → M/M/3



- τ = interarrival time
- λ = mean arrival rate
  - = $1/E[\tau]$
- s = service time per job
- μ = mean service rate per server
  - = $1/E[s]$ (total service rate for m servers is $m\mu$)

- n_q = number of jobs waiting to receive service.
- n_s = number of jobs receiving service
- n = number of jobs in system

$$n = n_q + n_s$$

- r = response time = w + s
- w = waiting time

Note that all of these variables are random variables except for λ and μ .

Rules for All Queues

Stability Condition

- If the number of jobs becomes infinite, the system is unstable. For stability, the mean arrival rate less than the mean service rate.

$$\lambda < m\mu$$

- Does not apply to finite buffer system or the finite population systems→ They are always stable.
  - Finite population: queue length is always finite.
  - Finite buffer system: arrivals are lost when the number of jobs in the system exceed the system capacity.

Rules for All Queues

Number in System vs. Number in Queue

- $n = n_q + n_s$
- $E[n] = E[n_q]+E[n_s]$
- Also, if the service rate of each server is independent of the number in queue
  - $Cov(n_q,n_s) = 0$
  - $Var[n] = Var[n_q]+Var[n_s]$

Rules for All Queues

Number vs. Time (Little's law)

- If jobs are not lost due to buffer overflow, the mean jobs is related to its mean response time as follows:

  mean number of jobs in system = arrival rate x mean response time

- Similarly

  mean jobs in queue = arrival rate x mean waiting time

- For finite buffers can use effective arrival rate, that is, the rate of jobs actually admitted to the system.

Rules for All Queues

Time in System vs. Time in Queue

- Time spent in system, response time, is the sum of waiting time and service time

$$r = w + s$$

- In particular:

$$E[r] = E[w] + E[s]$$

- If the service rate is independent of jobs in queue
  - Cov(w,s) = 0
  - Var[r] = Var[w] + Var[s]

Applying Little's Law

Example:

- A disk server satisfies an I/O request in average of 100 msec. I/O rate is about 100 requests/sec. What is the mean number of requests at the disk server?
- Mean number at server = arrival rate x response time
  = (100 requests/sec) x (0.1 sec)
  = 10 requests

Importance of the Queuing Theory

-Improve Customer Service, continuously.

-When a system gets congested, the service delay in the system increases.

- A good understanding of the relationship between congestion and delay is essential for designing effective congestion control for any system.
- Queuing Theory provides all the tools needed for this analysis.

Queuing Models

- Calculates the best number of servers to minimize costs.
- Different models for different situations (Like SimQuick, we noticed different measures for arrival and service times)
- Exponential
- Normal
- Constant
- Etc.

Queuing Models Calculate:

- Average number of customers in the system waiting and being served
- Average number of customers waiting in the line
- Average time a customer spends in the system waiting and being served
- Average time a customer spends waiting in the waiting line or queue.
- Probability no customers in the system
- Probability n customers in the system
- Utilization rate: The proportion of time the system is in use.

Assumptions

- Different for every system.
- Variable service times and arrival times are used to decide what model to use.
- Not a complex problem:
  - Queuing Theory is not intended for complex problems. We have seen this in class, where this are many decision points and paths to take. This can become tedious, confusing, time consuming, and ultimately useless.

Examples of Queuing Theory

- Outside customers (Commercial Service Systems) - Barber shop, bank teller, cafeteria line
- Transportation Systems - Airports, traffic lights
- Social Service Systems -Judicial System, healthcare
- Business or Industrial –Production lines

How the Queuing Theory is used in Supply Chain Management

- Supply Chain Management use simulations and mathematics to solve many problems.

- The Queuing Theory is an important tool used to model many supply chain problems. It is used to study situations in which customers (or orders placed by customers) form a line and wait to be served by a service or manufacturing facility. Clearly, long lines result in high response times and dissatisfied customers. The Queuing Theory may be used to determine the appropriate level of capacity required at manufacturing facilities and the staffing levels required at service facilities, over the nominal average capacity required to service expected demand without these surges.


Questions?