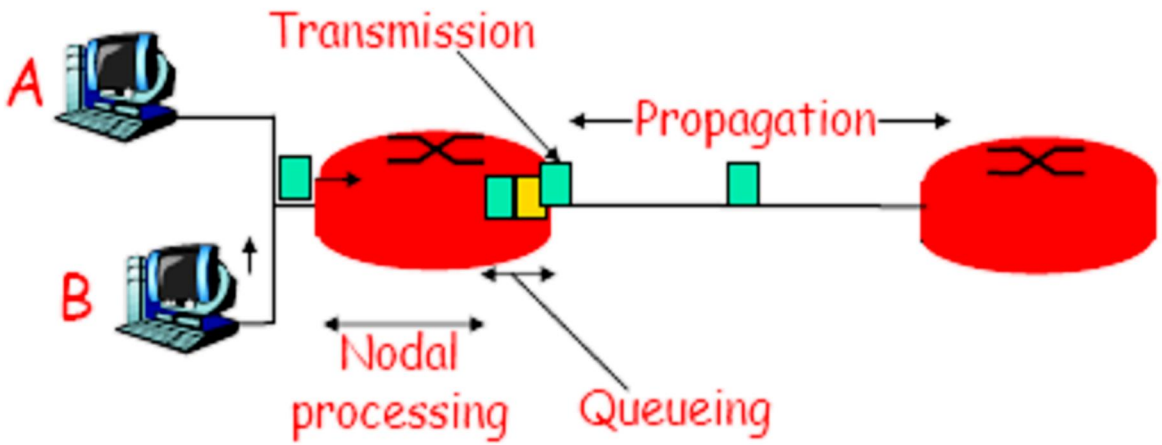
EECS 122: Introduction to Communication Networks

Unit 10: Queuing

Delay on the way – summary

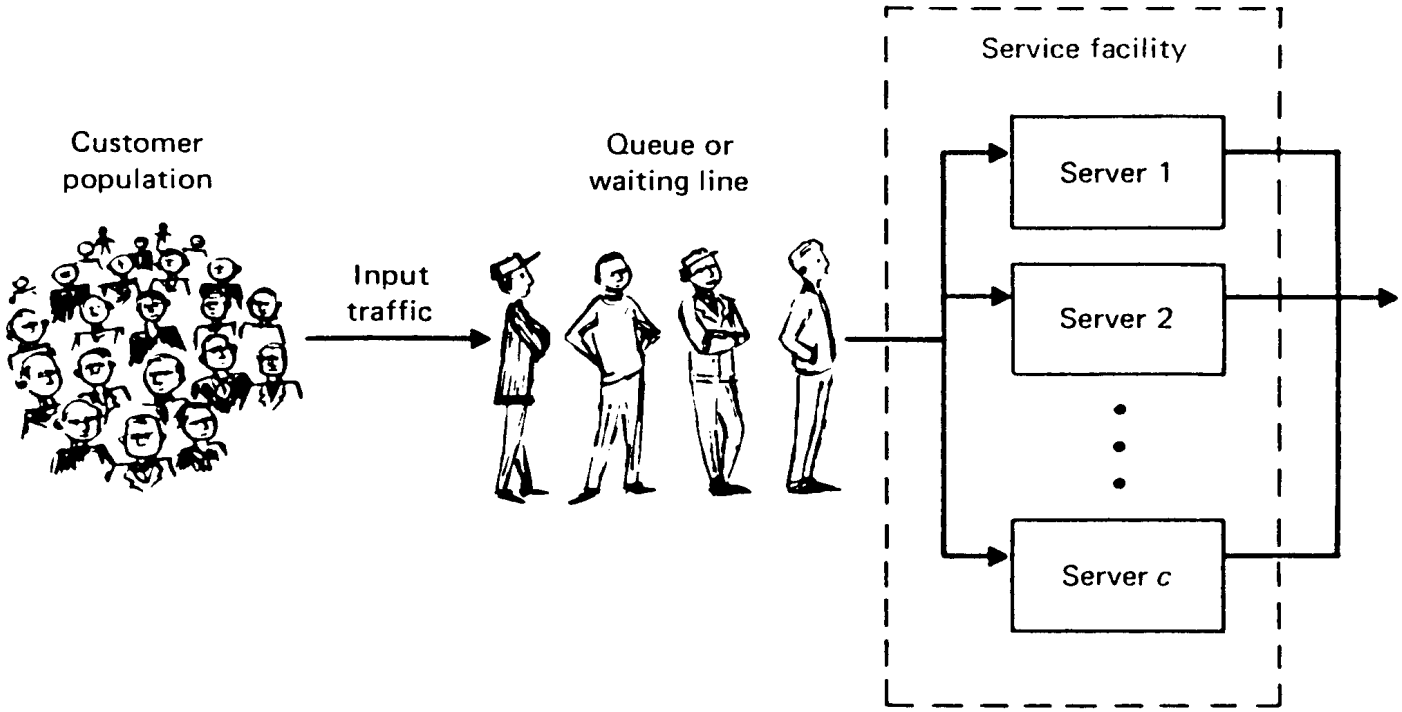
1. Nodal processing:
   - Check bit errors
   - Determine output
2. Queueing
   - Time waiting at output for trans.
   - Depends on congestion at router
3. Transmission delay:
   - R=link bandwidth (bps)
   - L=packet length (bits)
   - Time to send bits into link: L/R
4. Propagation delay:
   - d = length of physical link
   - s = propagation speed in medium
   - Propagation delay = d/s



Just to remind you the issue of queuing...

So what is queuing about?

[O Allen]



Whenever there is irregular demand for some Ăserviceñtaking a random time ï there unavoidably appear queues....
Only fully deterministic demand, and fully deterministic service (which never happens!) would eliminate the queuing effect...

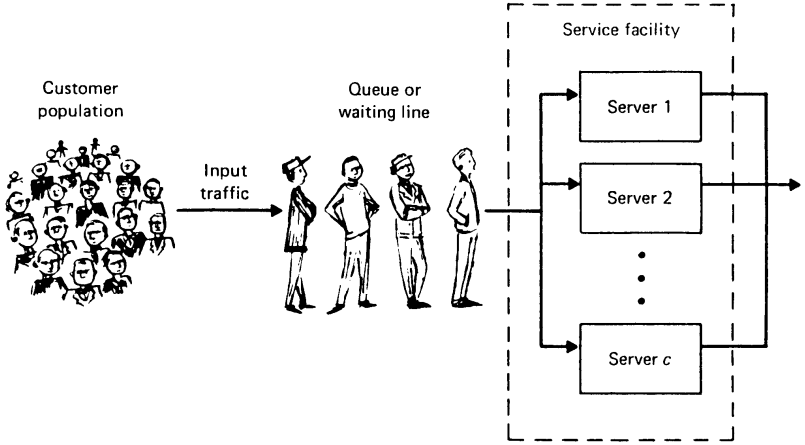
Queuing systems parameters:

- **Customers** (eg. ï packets!; phone calls!) arrive individually in discrete, randomly distributed time intervals according to the **inter-arrival time** distribution.
  - Note that if there is no possibility to form a queue (e.g. Buffering of packets), the customers experiencing all the servers busy might be just lost (busy line in telephone systems).
- The service can be provided on a number of parallel (for simplicity identical) **servers** (e.g. Transmission lines)
- Each customer has a randomly distributed **service time** (e.g. Packet transmission duration, resulting out of packet length; e.g. duration of a phone call).

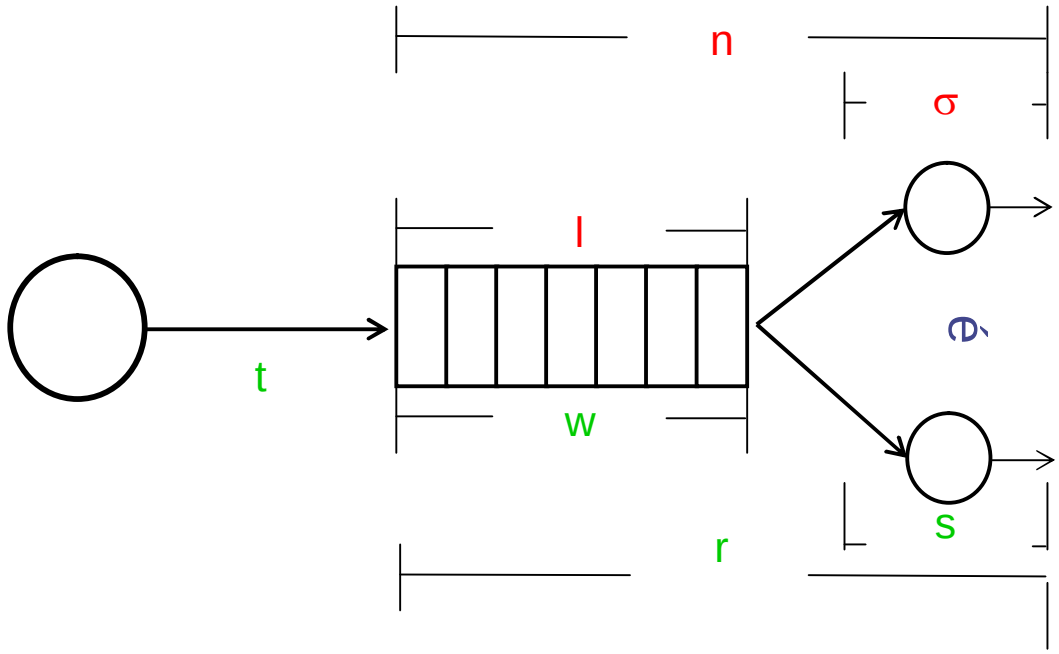
So , what would we like to know?

- There are values interesting for the customer:
  - The waiting time until the service begins
  - The complete time spend in the system (arrival to departure)
  - The probability of not obtaining the service ….
- There are values interesting to the service provider:
  - The utilization of the servers
    Comment: setting up the servers is usually an investment – but also a source of revenue ☺
  - The length of the queue (how to dimension buffers?)
- Basic challenge of the queuing theory: How to determine the distributions of the interesting values from the queuing system parameters…

Queuing Components & Random Variables



Customer Source Queue Server



Variable Descriptions

- Random variables related to Customer Numbers:
  - n : Total number of customers in the queuing system
  - l : Number of customers in the queue
  - σ : Number of customers currently served
- Random variables related to Time:
  - r : Total time required to pass the queuing system
  - w : Time spent in the queue
  - t : Inter-arrival time between two customers
  - s : Service demand of a customer
- m : number of parallel servers

Some useful abstractions

- An infinite amount of potential customers
  - Abstraction for a LARGE number of potential customers, or – in other words – for the case in which the length of the queue does not influence the arrival of new customers….
- An infinite queue length
  - Abstraction for the case of a "waiting capacity" being long enough
- The selected distributions of random variables
  - This is - for analytical studies – always a tradeoff between real, observed features and mathematical abstractions which make it possible to compute the interesting values.

Kendall Notation

- Imagine differences of queuing systems in practice:
  - Type of input traffic (deterministic vs. stochastic with a certain PDF)
  - Type of service time (det. vs. stochastic with a PDF)
  - Number of servers
  - Number of available slots in the queue
  - Queuing discipline
- In order to classify these different queuing systems, we use a notation introduced by Kendall:

A/B/X/Y/Z

  - A is the inter-arrival-time pattern
  - B the service time pattern
  - X the number of parallel servers
  - Y the restriction on system capacity (total amount of slots in the system)
  - Z the queue discipline

Kendall Notation for Queuing Systems

- Kendall Notation A/B/X/Y/Z
  - A is the inter-arrival-time distribution, B the service pattern
  - X the number of parallel service channels, Y the restriction on system capacity

Characteristics	Symbol	Explanation
Interarrival-time distribution (A)	M	Exponential
	D	Deterministic
	E_k	Erlang type *k* (*k = 1, 2, ...*)
	H_k	Hyperexponential type *k*
	PH	Phase type
	G	General
Service-time distribution (B)	M	Exponential
	D	Deterministic
	E_k	Erlang type *k* (*k = 1, 2, ...*)
	H_k	Hyperexponential type *k*
	PH	Phase type
	G	General
Number of parallel servers (X)	1,2, ...	
Restriction on system capacity (Y)	1,2, ...	
Queue discipline (Z)	FCFS	First come, first serve
	LCFS	Last come, first serve
	RSS	Random selection for service
	PR	Priority
	GD	General discipline
	PS	Processor sharing
	RR	Round Robin

Unlimited Buffer Case - Some Formula

Å**Arrival rate**: $\lambda = 1 / E[t]$

ï average rate of customers per unit time arriving to the queue

Å**Service rate**: $\mu = 1 / E[s]$

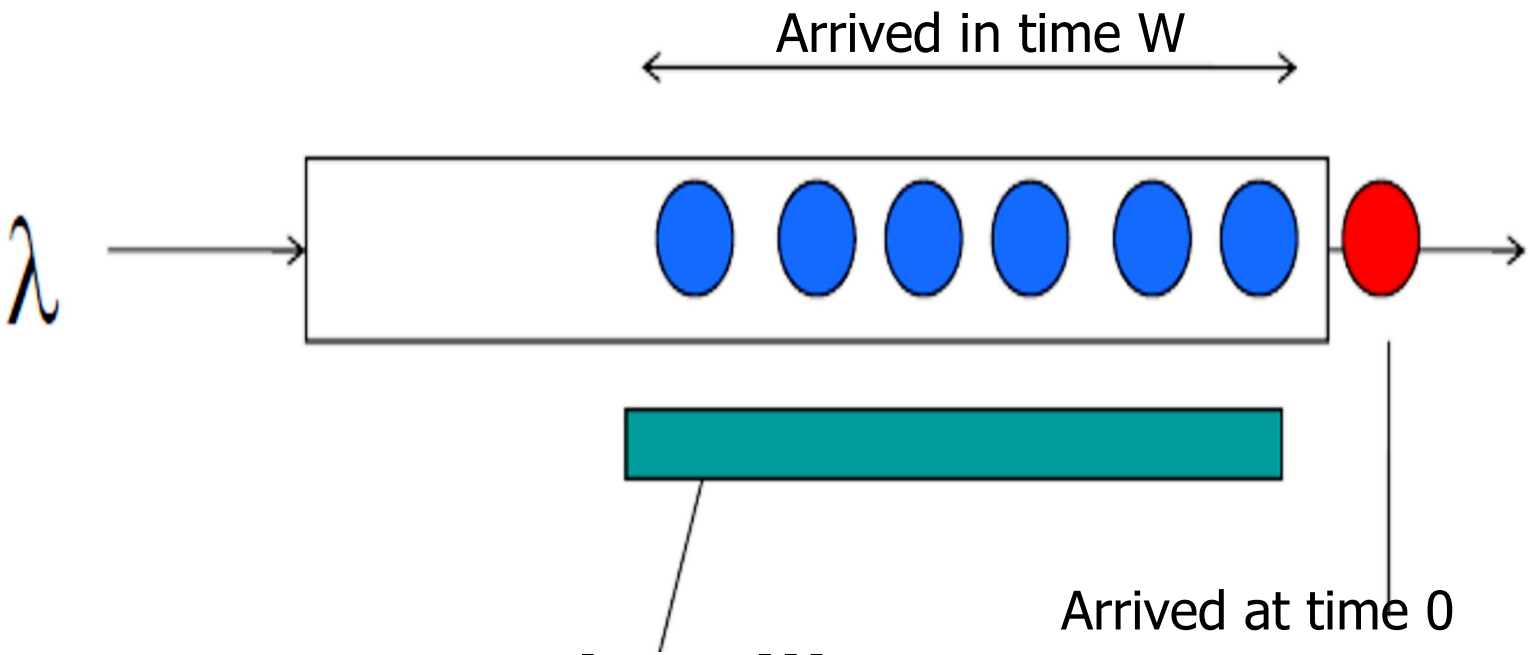
ï average service rate of the queuing system

- The **traffic intensity** $\boxed{\rho : \rho = \lambda/(\mu\ m)}$: (input to output relation)
  - Note that a system with unlimited buffer size DOES NOT NECESSARILY have to be stable, i.e. It could happen that the queue „tends to grow indefinitely".
  - Systems are stable only if $\boxed{\rho < 1\ !}$
- In the stable case, **server utilization** U_k: $\boxed{U_k = \lambda/(\mu\ m)}$
  and **throughput** $\boxed{\Lambda = \lambda.}$
  - If the system is stable the throughput Λ must equal the long-run rate at which customers arrive.
- Further in this considerations we will discuss only features of stable queuing systems.

The Little Theorem….

Relates the average queue lenght with the average waiting time

In systems without losses…. Under pretty general assumptions….



Average in the queue $\mathbf{L = \lambda W}$

or

Average in the SYSTEM $\mathbf{n = \lambda r}$

Usual simplifications …

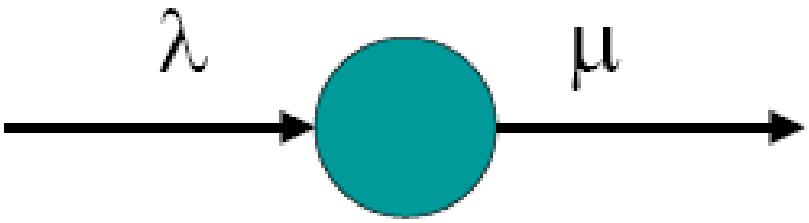
- Customers (e.g. – packets!; phone calls!) arrive (*individually*) in discrete, random, *independent, identically distributed (i. i. d.)* time intervals.
- The parallel servers are identical and operate independently
- Each customer has an random *i. i. d. service time* (e.g. packet transmission duration, resulting out of packet length; e.g. duration of a phone call).

➔ ***Italics*** emphasizes the simplifications

(which do not always hold !!!)

The basic queuing system

- Example: M/M/1 Queue



Arrivals are Poisson with rate λ

Service times are exponentially distributed with mean $1/\mu$

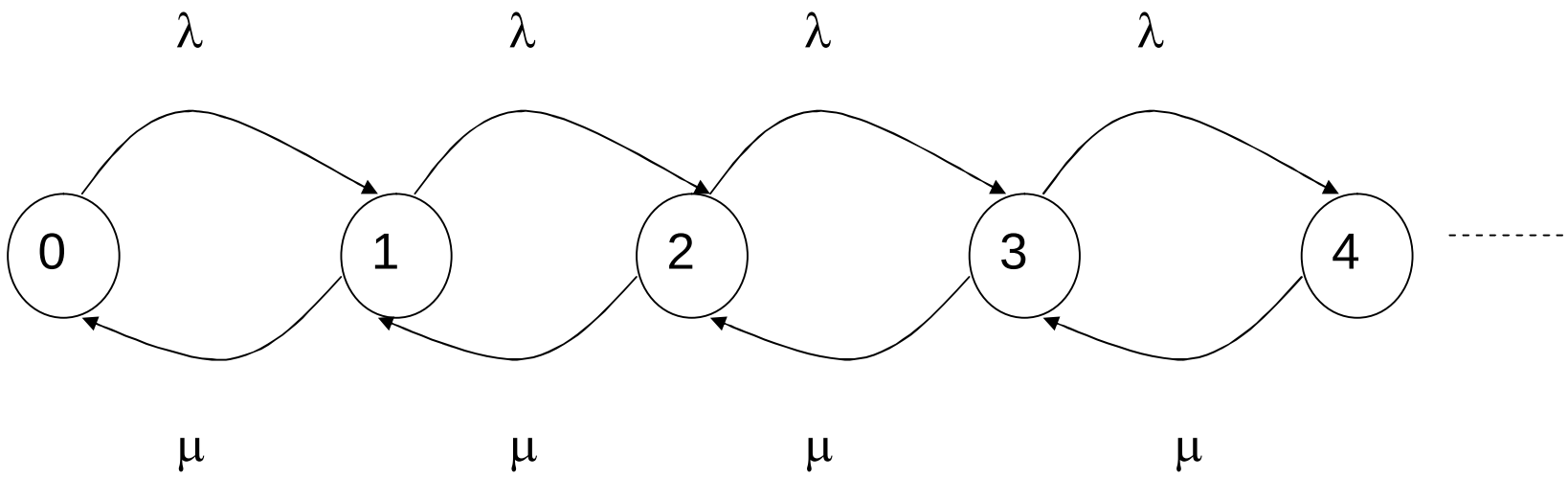
Average delay per packet

E(r) = $1/(\mu - \lambda) = (1/\mu)/(1 - \rho)$ where $\rho = \lambda/\mu$ = utilization

For instance, $1/\mu$ = 1ms and ρ = 80% => Q = 5ms

The "Flow-Balancing Approach"

- In the "rate diagram" given below, think of the following:



- Each circle representing a state (i.e., number of customer in the system) has an unknown probability π_j, j= 0, 1, 2, …
- This is called "birth and death process" with constant parameters

Flow balance equations…

$$\text{Balance equations}\begin{cases} \text{state } 0: & \pi_1\mu = \pi_0\lambda \\ \text{state } 0\text{-}1: & \pi_2\mu = \pi_1\lambda \\ \vdots & \\ \text{state } 0\text{-}1\text{-}\ldots\text{-}n: & \pi_{n+1}\mu = \pi_n\lambda \\ \vdots & \end{cases}$$

$$\text{Normalization equations: } \sum_{n=0}^{\infty} \pi_n = 1$$

- Per iterative elimination

$$\pi_j = \pi_0 c_j \quad (j = 1,2,\ldots) \qquad \text{where}$$

$$C_j = (\lambda_0\ \lambda_1\ \lambda_2 \text{é é}\ \ \lambda_{j-1})/(\mu_1\ \mu_2\ \mu_3 \text{é}\ ..\ \mu_j)$$

The result…

- If $\sum_{j=1}^{j=\infty} c_j$ is finite, we can solve the normalization for:

$$\pi_0 = \frac{1}{1 + \sum_{j=1}^{j=\infty} c_j}$$

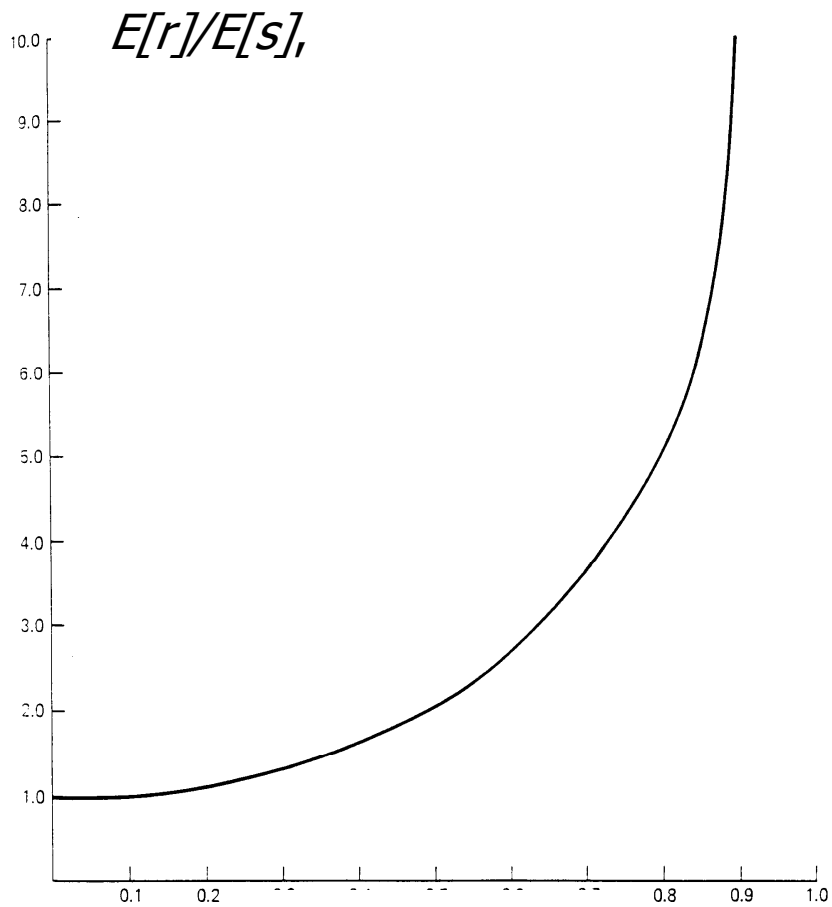
$$\pi_0 = 1 - \rho$$

$$\Rightarrow \pi_n = \rho^n \cdot (1 - \rho)$$

- It can be shown that if $\sum_{j=1}^{j=\infty} c_j$ is infinite, then no steady-state distribution exists.
- The most common reason for a steady-state failing to exist is that the arrival rate is at least as large as the maximum rate at which customers can be served.

Some observations of the behavior...

- The slope of the curve increases rapidly as ρ grows
- a small change in λ, assuming causes a huge change in *E[r]*
- With ρ approaching the value of 1 the system becomes UNSTABLE, i.e. the mean queue length tends to infinity!



λ* **E(S) =** ρ

Normalized average time in the system, *E[r]/E[s]*, for M/M/1 queuing system.

Some observations for the behavior... Detailed

- The **slope of the curve** increases rapidly as ρ grows beyond **ca. 0.8**, as

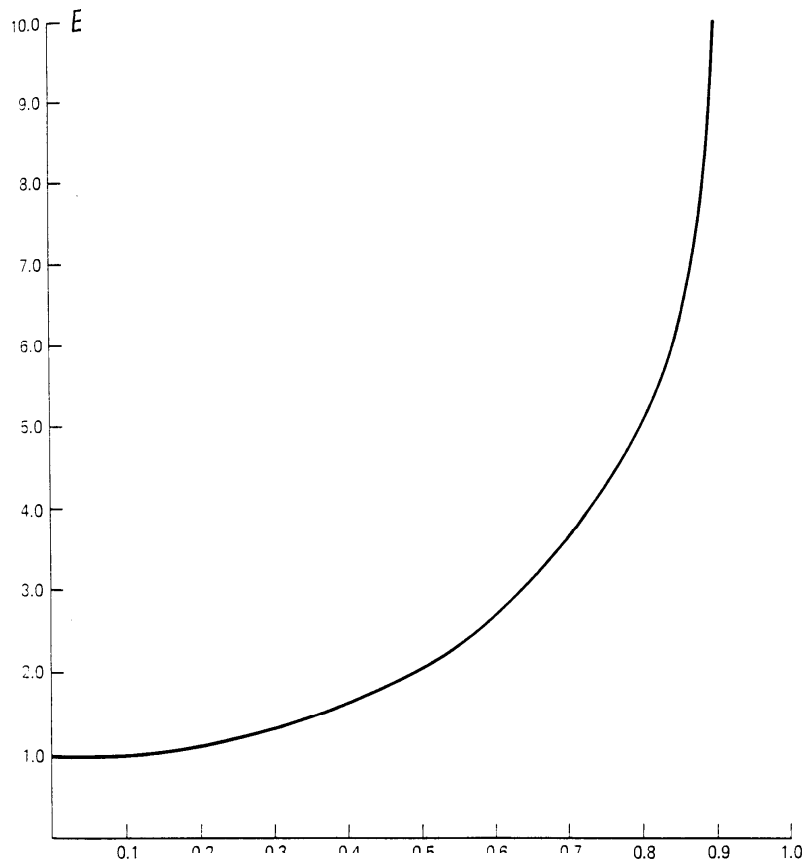
$$\frac{dE[r]}{d\rho} = E[S](1-\rho)^{-2} = E[S](1-E[S]\cdot\lambda)^{-2}$$

- A **small change** in ρ (due to a small change in $\Delta\lambda$, assuming E[s] is fixed) causes a change in E[r] given approximately by

$$\left(\frac{dE[r]}{d\rho}\right)\Delta\rho = \left(\frac{dE[r]}{d\rho}\right)E[S]\Delta\lambda = E[S]^2(1-\lambda E[S])^{-2}\Delta\lambda$$

- Thus, if ρ = 0.5, a change in $\Delta\lambda$ in λ will cause a change in E[r] of about **4***E[s]2$\Delta\lambda$, while, if ρ = 0.9, the change in E[r] will be about **100***E[s]2$\Delta\lambda$ (i.e. 25 times the size of the change that occurred for ρ = 0.5!).

E[r]/E[s]

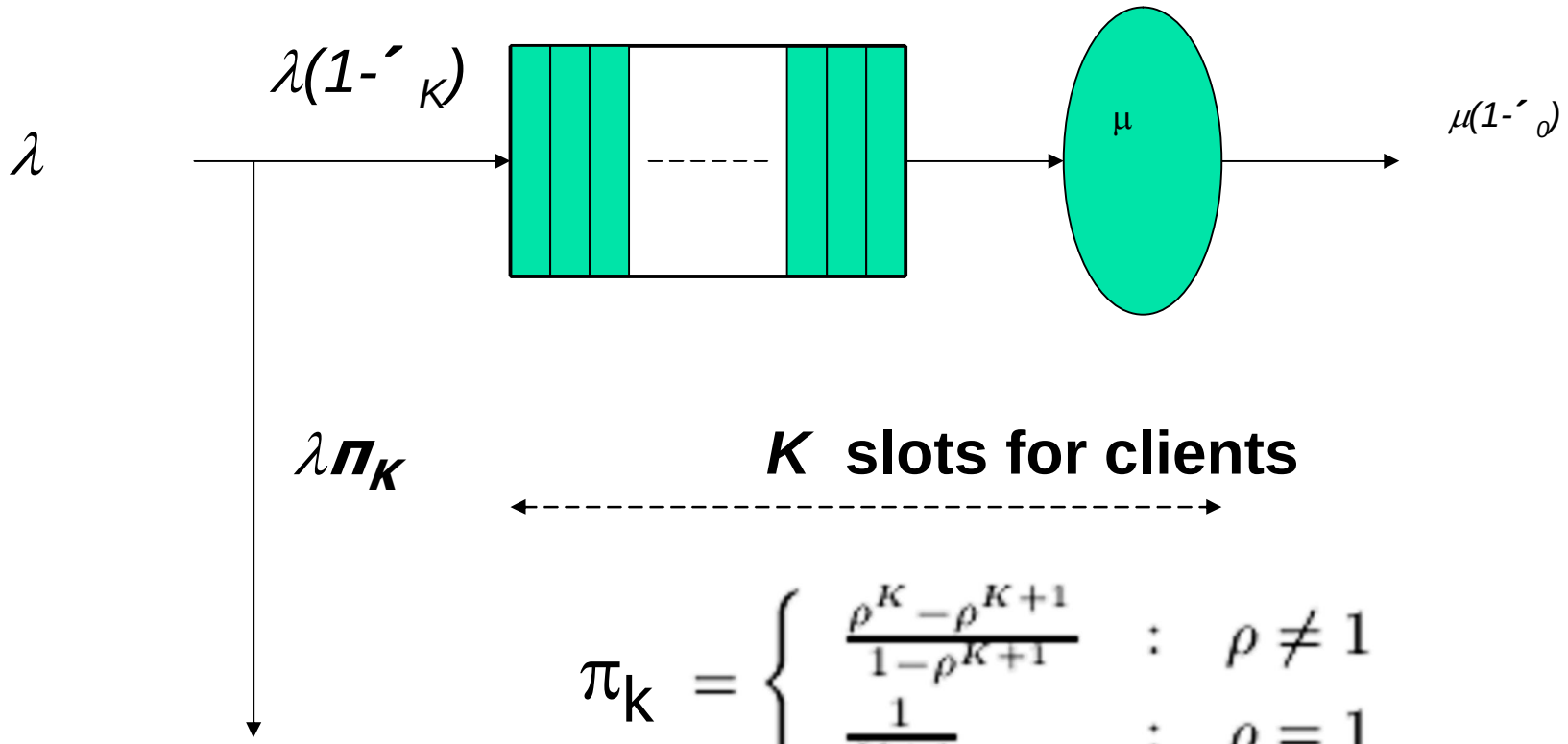


$\rho = \lambda E[s]$

Normalized average time in the system, E[r]/E[s], for M/M/1 queuing system.

M/M/1/K

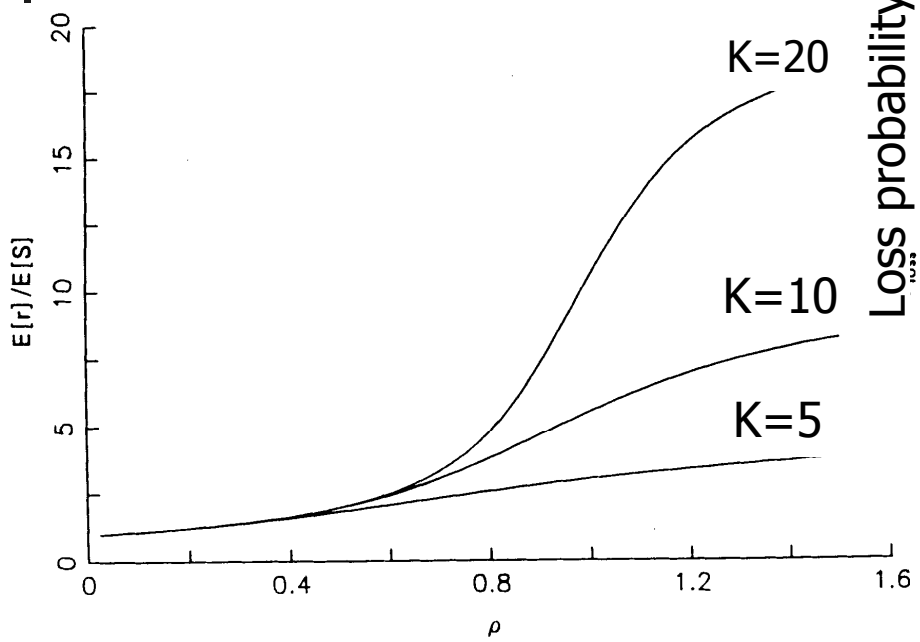
- What happens if we take a more realistic case of limited buffer capacity: all in all only N customers are allowed in the M/M/1/K queuing system

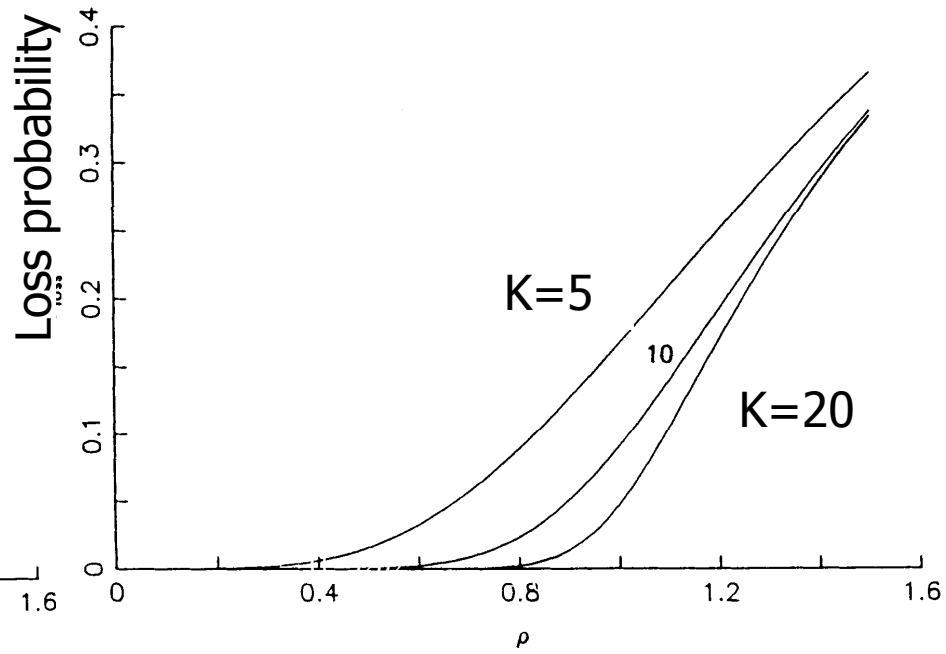


$$\pi_k = \begin{cases} \frac{\rho^K - \rho^{K+1}}{1-\rho^{K+1}} & : \quad \rho \neq 1 \\ \frac{1}{K+1} & : \quad \rho = 1 \end{cases}$$

Probability of customer loss

M/M/1/K

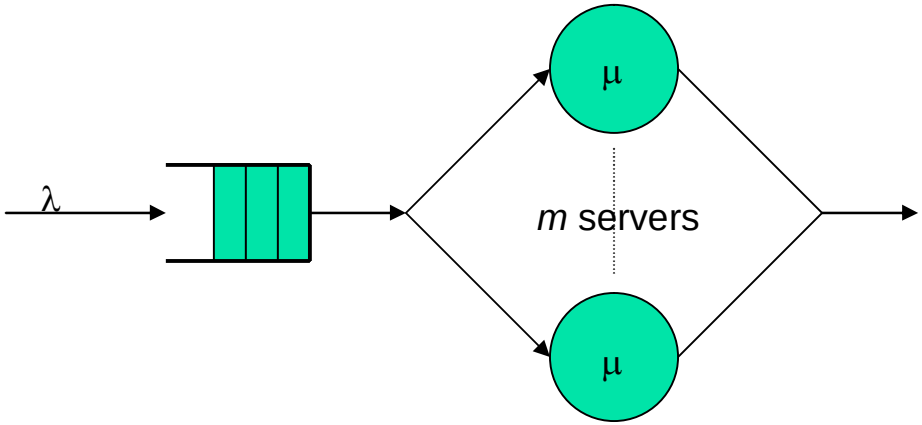




- This system is ALWAYS stable,
- The normalized mean delay tends asymptotically to K* E(S)
- Considering the normalized mean delay independently from the loss probability curves is not reasonable
- One typical question: How big should the buffer be, to assure loss probability smaller than some threshold....

M/M/m (1)

- The M/M/m queuing system



The case of multiple servers...

Comparison of Single Server and Multiserver Queues

a/ a service center having *a single server*

$$E[r(1, \lambda, \gamma)]$$

b/ a service center having *m servers*, each of which is *m times slower* than the server in a/

$$E[r(m, \lambda, \gamma/m)]$$

c/ a service center having *a single server* which is *m times slower* than the server in a/ and where the service center handles *(1/m)th* of the customers.

$$E[r(1, \lambda/m, \gamma/m)],$$

The following inequalities hold:

$$\boxed{E[r(1, \lambda, \gamma)] < E[r(m, \lambda, \gamma/m)] < E[r(1, \lambda/m, \gamma/m)],}$$

The inequalities are valid also for the M/G/m case, although the first one only for variability coefficients smaller than one.

Comparison of three systems – an example

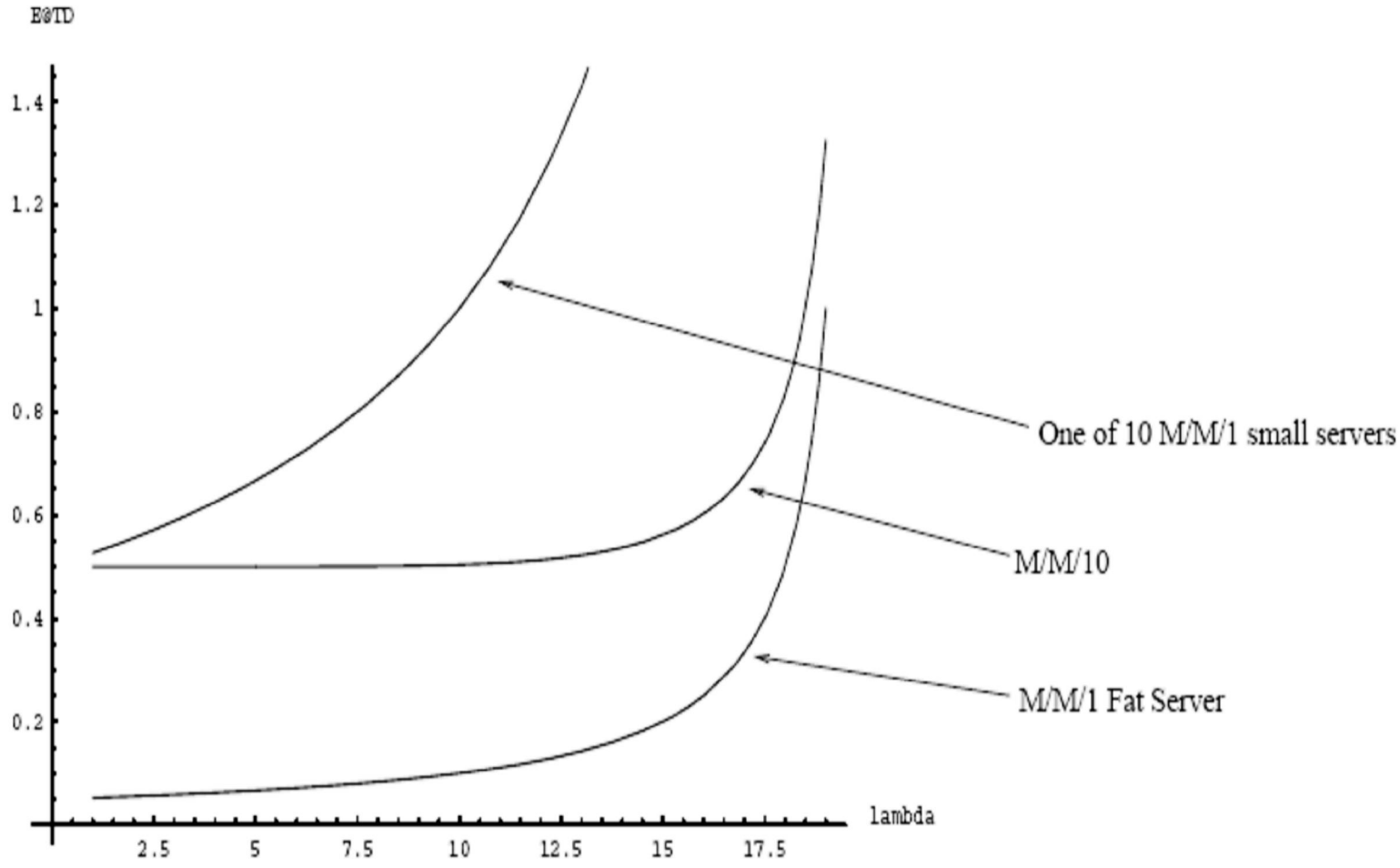
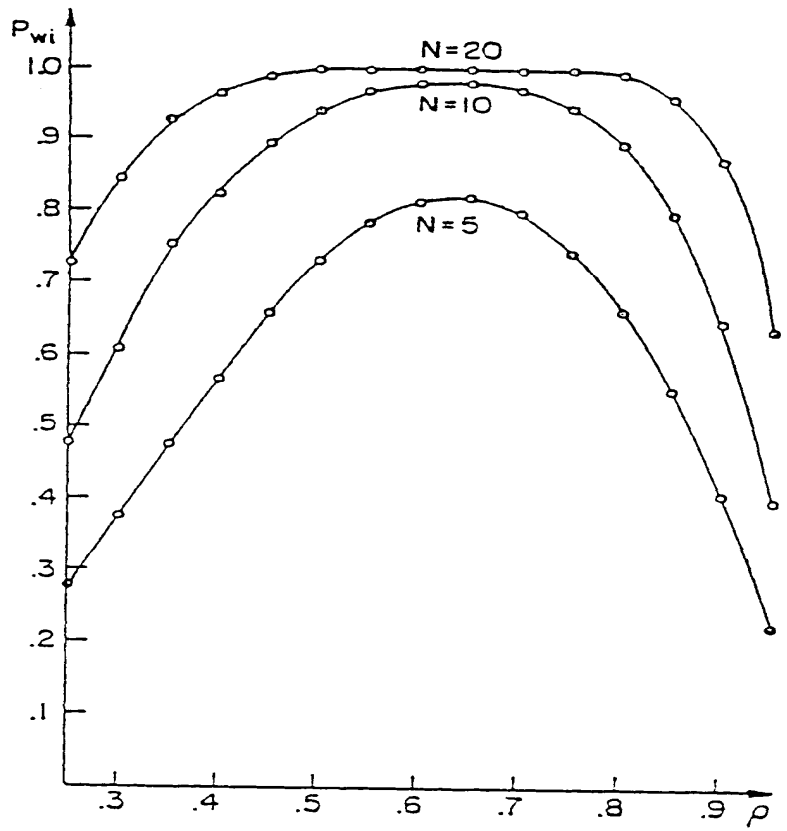


Figure 2.10: Mean Response Times for three different systems

Some explanations...the case of N separate queues

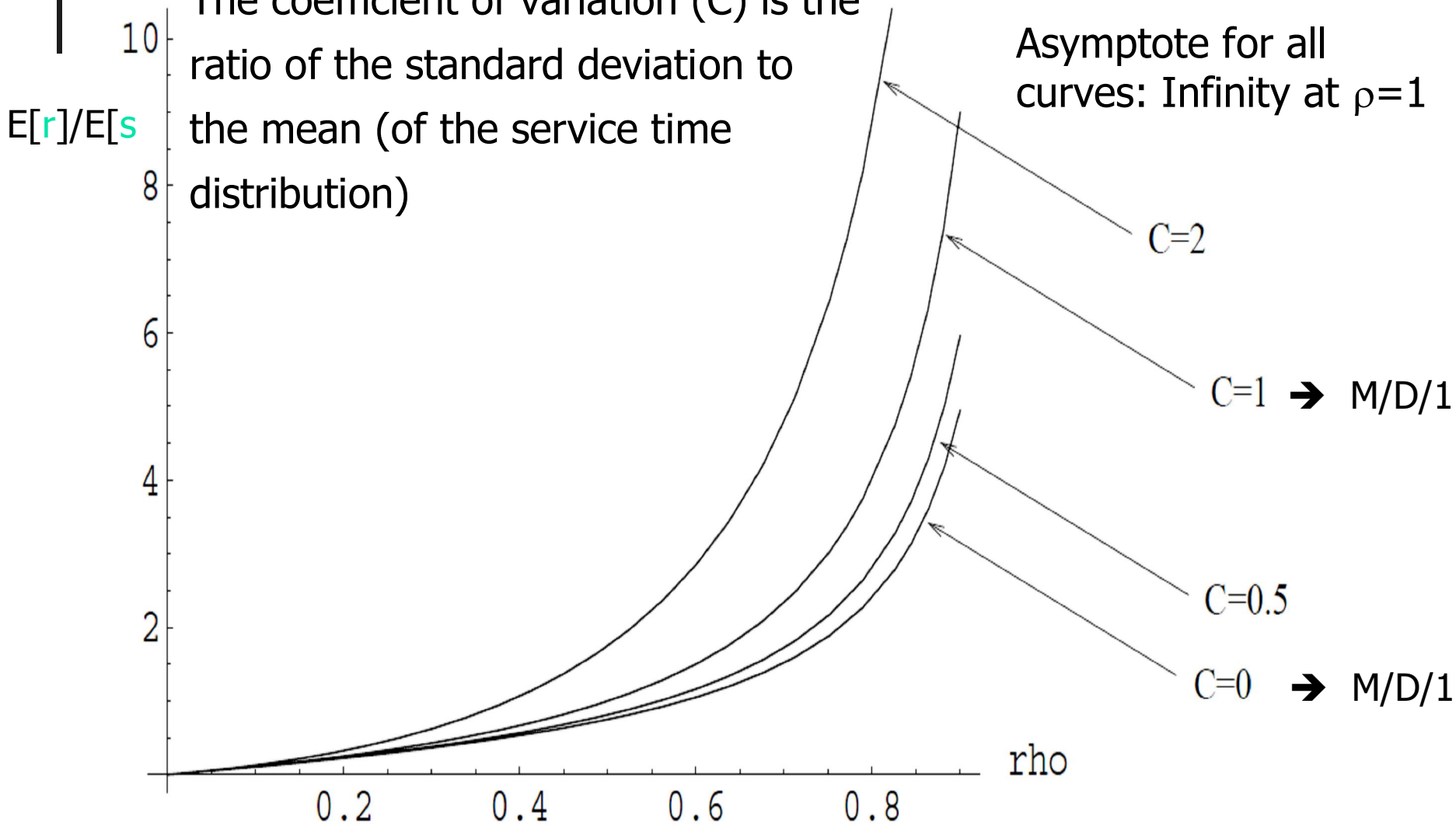


The probability of having at least one idle server in spite of at leat one non-empty queue in the system of N servers with splitted queues.

→ It is better to use a system with single queue and multiple servers than support a separate queue for each server..

M/G/1 Systems: General service time distribution

The coefficient of variation (C) is the ratio of the standard deviation to the mean (of the service time distribution)



Asymptote for all curves: Infinity at $\rho=1$

Telephony: multiple queues with no waiting

- Telephone traffic is the aggregate of telephone calls over a group of circuits or trunks with regard of the number and duration of calls.
- We measure traffic intensity (A) as : $A = \lambda t_m$, where
  - λ is the average arrival rate (ex., calls/hr)
  - t_m is the average holding time (ex., hrs)
- Units:
  - Erlangs (dimensionless) : calls-second per second
  - CCS : hundred (century) calls-second per hour
  - 1 erlang = (60)(60)/100 = 36 CCS
- Capacity of a single channel is one erlang
  - Interpretation: a telephone that is busy 10% of the time represents a load of 0.1 erlang on that particular line
- Example: 2 calls/hour with average holding time of 5 minutes
  - what is the traffic intensity in Erlangs? In CCS?

Erlang_B Formula.

Assumptions:

- **Arrivals from an infinite Poisson source (the inter-arrival times are exponentially distributed).**

  This corresponds to a situation when number of customers is much larger than the number of resources available to service them. Acceptable results if the number of customers is at least 10 times the total number of resources (N).

- **Calls which cannot be served are lost (and do NOT appear again)**

- **Probability of blockage at the switch due to congestion or "all trunks busy":**

$$E_B = \frac{\frac{A^n}{n!}}{\sum_{x=0}^{n} \frac{A^x}{x!}}$$

- **A is the mean of the offered traffic [Erlangs], n is the number of trunks**

Play with the calculator on http://owenduffy.net/traffic/erlangb.htm

Let us think in terms of a real network

- Example: A series of switches connected by links….
  - This can be modeled as a sequence of queuing systems
- But: are the service times independent?
  - Note: in a packet network the sending time of a packet, (i.e. the service time !) is - in reality **the same** in all queues (or differs by a constant factor, the inverse of the line speed)
- Kleinrock's independence assumption:
  - IGNOREE THIS FACT!

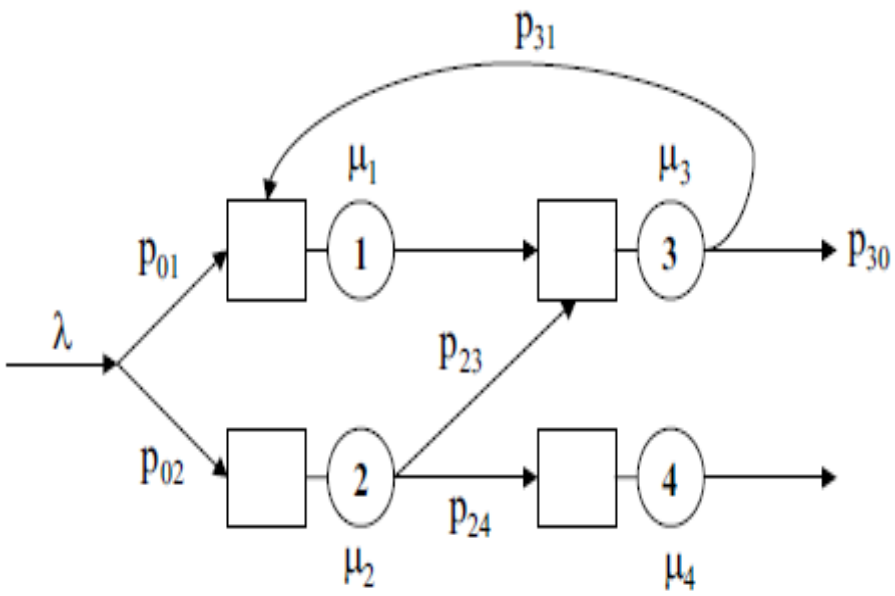
Kleinrock Independence Assumption

1. Interarrival times at various queues are independent
2. Service time of a given packet at the various queues are independent
   - Length of the packet is randomly selected each time it is transmitted over a network link
3. Service times and interarrival times: independent
4. Pretty good approximation when:
   - Poisson arrivals at entry points of the network
   - Packet transmission times "nearly" exponential
   - Several packet streams merged on each link
   - Densely connected network
   - Moderate to heavy traffic load

- But- even if the approximation is not good – it offers usually an UPPER BOUND!!! (anyway for a line of queues)

Jackson´s networks

- Jackson´s theorem: This is true for arbitrary "Open Queuing Networks" (i.e. with input and output) iff:
  - The arrivals are Poisson
  - The service times are exponential
  - The queues are infinite
  - Outputs of each node are "scattered with fixed probabilities among the outputs



Arrival rate to each queue can be computed as the sum of the proper incoming rates.