Chair of Future Communication
Prof. Dr. K. Tutschku
Department of Distributed and Multimedia Systems
Faculty of Computer Science

050069 VO Netzwerktechnologie für Multimedia Anwendungen

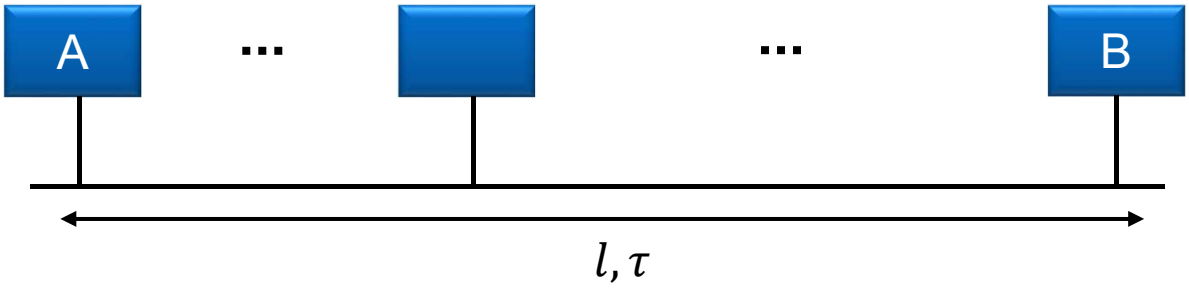
Lecture 2: Foundations of Numerical Performance Analysis and Queuing Theory

Prof. K. Tutschku (kurt.tutschku@univie.ac.at)

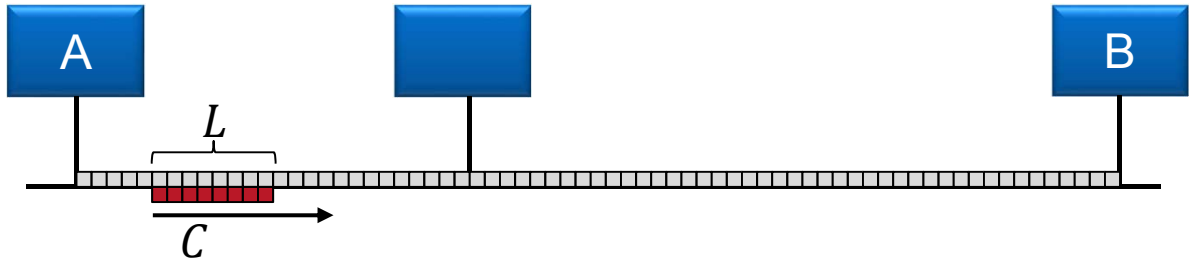
Endowed by

Bachelor Informatik (Medieninformatik)
WS 2010/11

Transmission Path



- Transmission path parameters
  - bus length — l in m (e.g. 5000m)
  - bus capacity — C in bps (bit per s, e.g. 64000 bps)
  - propagation speed — v in m/s (common value 2*10^8 m/s = 2/3c, c speed of light)
  - propagation delay — τ in s
- How are propagation delay, bus length, and propagation speed related to each other?



- bus capacity C in bps (bit per s, e.g. 64000 bps)
- logical bus length (in bit) $a_b = \frac{lC}{v}$

How many bits fit on the bus?

- packet length L in bit per packet (e.g. 1kbit)
- logical bus length (in packets) $a_N = \frac{lC}{Lv}$

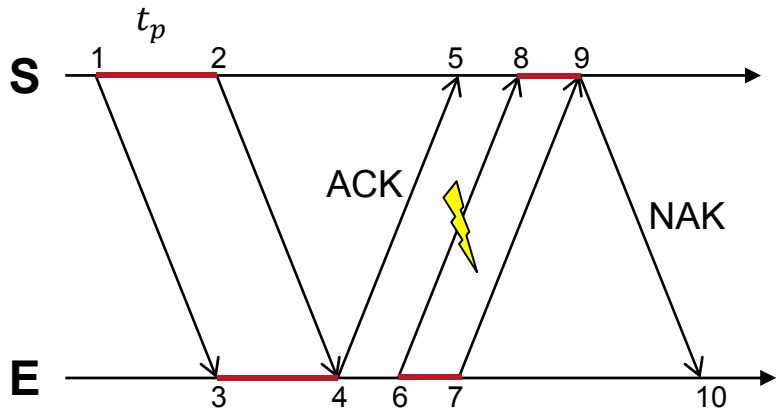
How many packets fit on the bus?

1000 or 1024, that's the question?

10^{-9}	10^{-6}	10^{-3}	10^{3}	10^{6}	10^{9}	10^{12}	10^{15}	10^{18}
nano	micro	milli	kilo	mega	giga	tera	peta	exa
			2^{10}	2^{20}	2^{30}	2^{40}	2^{50}	2^{60}
n	μ	m	k	M	G	T	P	E

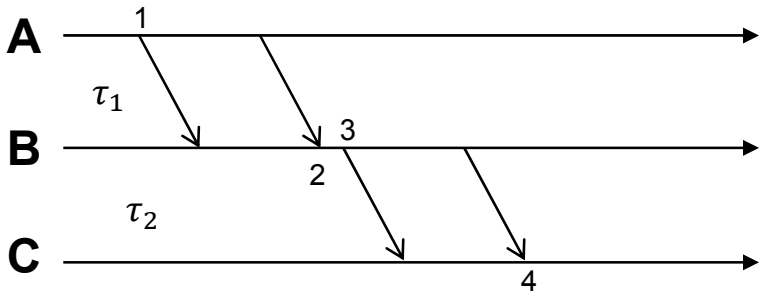
- **Decimal units (**factor 1000)
  - E.g.: storage systems (hard disk sizes)
- **Binary units** (factor 1024)
  - 1 Kbyte = 1024 Byte, 1 Mbyte = 1024 Kbyte, etc.
  - Alternatively special binary prefixes:
    - kibibyte (KiB), mebibyte (MiB, gibibyte (GiB), …
  - E.g.: transmission technology (data rates, packet sizes)

Path-Time Diagram



1. S sends first bit of packet 1
2. S sends last bit of packet 1 (2 takes place t_p after 1)
3. E receives first bit of packet 1 (3 takes place τ after 1)
4. E receives last bit of packet 1 and sends ACK (= ok) to S
5. S receives ACK
6. E sends first bit of packet 2
7. E sends last bit of packet 2
8. S receives first bit
9. S receives last bit and sends NAK (= failure) back because of an error
10. E receives NAK

Path-Time Diagram



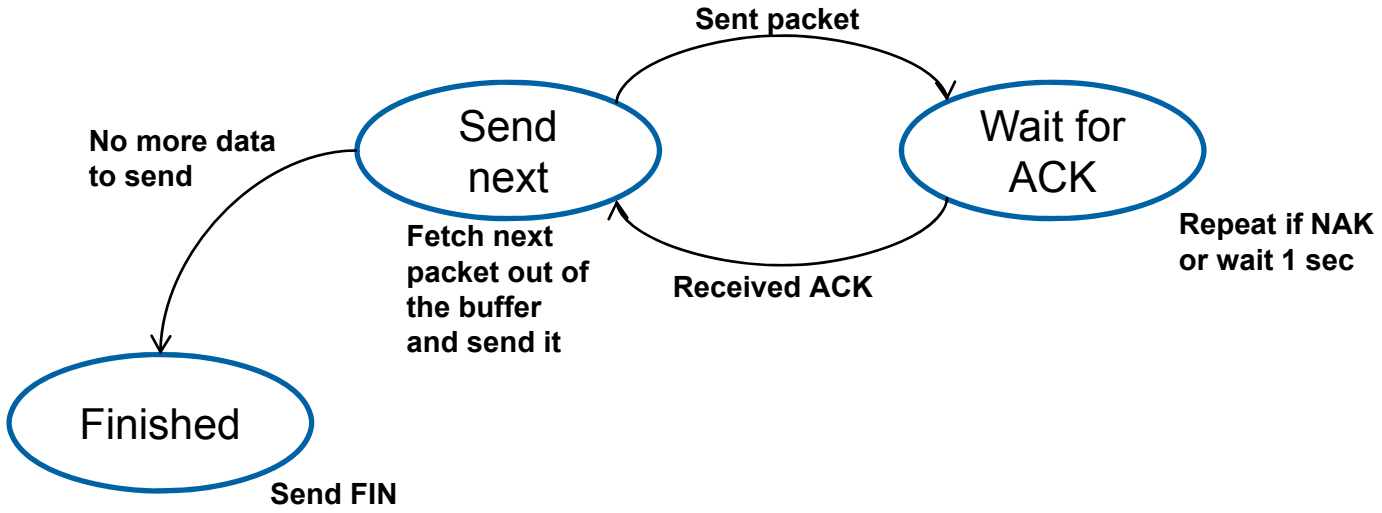
1. A sends first bit of the packet to B
2. B receives last bit of the packet
3. B sends first bit of the packet to C
4. C receives last bit of the packet

State Machine of a Protocol Instance

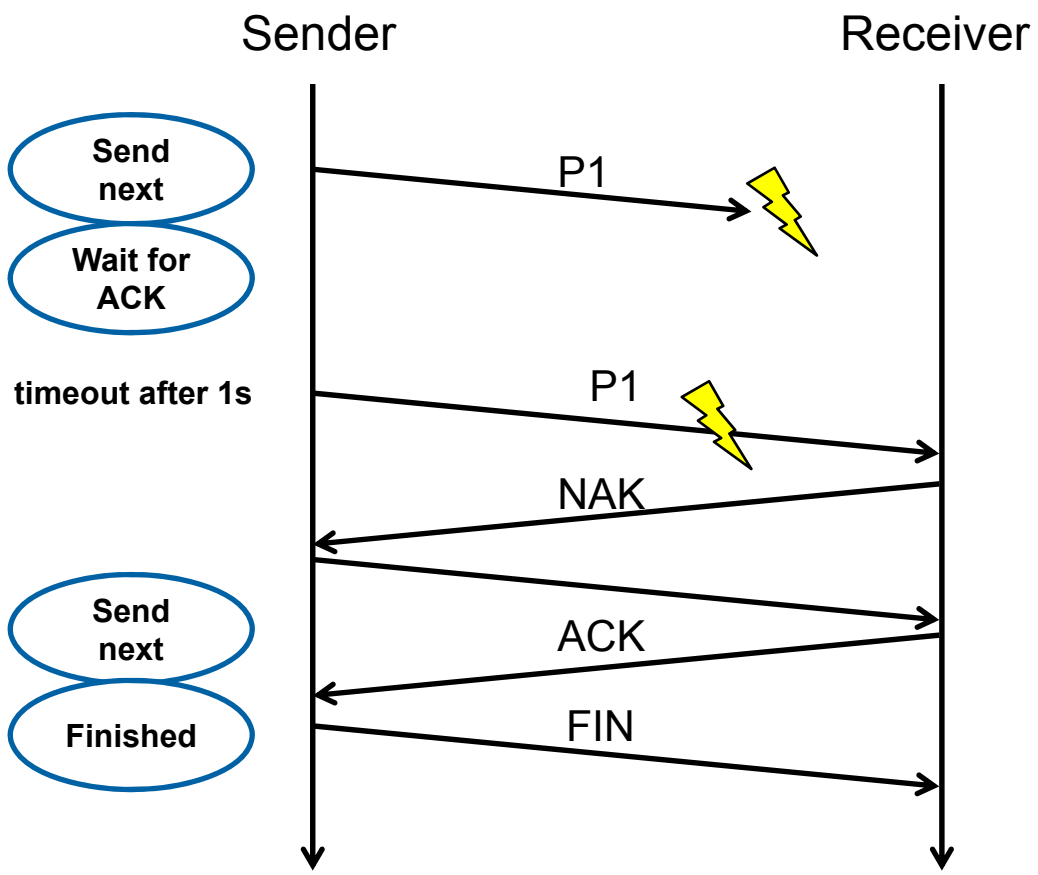
- **Goal:** retransmission in case of a transmission failure

Simple Protocol:

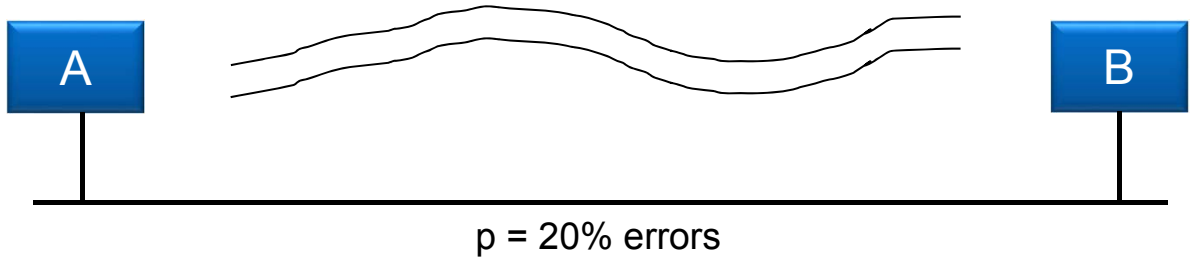
- **Receiver** is always in state “receive”, replies on a erroneous transmission with “NAK”, otherwise an “ACK” is send and the data is delivered to the next layer
- **Sender**



Path-Time Diagram v2



Transmission Error I



- Transmission with packet failure
  - First packet correct: $P(\text{“}success\text{”}) = 1 - p$
  - First packet erroneous: $P(\text{“}error\text{”}) = p$
  - First packet correct, second packet erroneous:
    - $P(\text{“}error\text{”}) * P(\text{“}success\text{”}) = p * (1-p)$
  - First and second packet erroneous:
    - $P(\text{“}error\text{”}) * P(\text{“}error\text{”}) = p * p$
- **In general**: $\boldsymbol{P(\text{“}i\ attempts\text{”}) = p^{i-1} * (1-p)}$

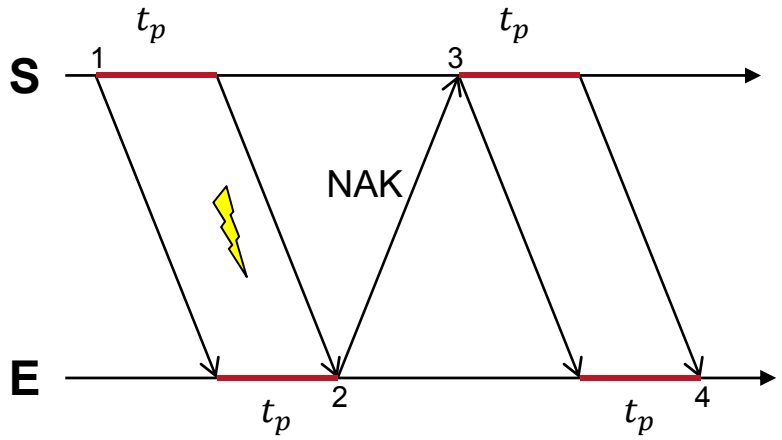
Transmission Error II

- **Remark:** This corresponds to the geometric distribution.

$$E[attempts] = 1 + \frac{p}{1-p} = \frac{1}{1-p}$$

- **Important:** This is based on the assumption, that the packet error probability is equal for each transmission.
  Often this is not the case. For example, due to external disruption many consecutive errors occur.

Virtual Transmission Time I



- How much time takes a transmission?
- Transmission time T = Time from 1 to 2
  - But then packet is not yet received by E

- **Virtual transmission time T_v**
  - Time until the packet is correctly received at the destination.
  - In this case: Time from 1 to 4

Virtual Transmission Time II

- E.g.: transmission time and virtual transmission time
- Transmission error as previously, error probability is p and independent

$$T = \tau + t_p$$

$$T_v = (i-1)T_{error} + T = \underbrace{\frac{p}{1-p}}_{E[attempts]-1}\underbrace{\left(2\tau + t_p\right)}_{T_{Error}} + \underbrace{\tau + t_p}_{T}$$

Statistics

- Properties of distributions or measured values
  - Mean $m = \frac{1}{n}\sum_{i=1}^{n} m_i$
  - Median
    - Let $m_1, m_2, \ldots, m_n$ be sorted, then $m_{(n+1)/2}$ is the median
    - If n is even, then different conventions are present, e.g. $m_{n/2}$ or $(m_{n/2} + m_{n/2+1})/2$
  - Variance $\sigma^2 = \frac{1}{n-1}\sum_{i=1}^{n}(m_i - m)^2$
  - Standard deviation $\sigma = \sqrt{\sigma^2}$
  - Variation (coefficient of variation) $c = \frac{\sigma}{m}$
    - standardized representation
    - c=1 means that the variation is as large as the mean
- Specification of mean and standard deviation or variation is common

Statistics

- **Cumulative Distribution Function**
  - $A(-\infty) = 0$ and $A(+\infty) = 1$
  - $A(x) \leq A(y)$, if $x \leq y$
  - $P(Value\ between\ x\ and\ y) = |A(y) - A(x)|$
- **Quantile**
  - x% of the values are on average smaller than the x%-quantile can be directly taken from the inverted distribution function.
  - E.g.: The median is the 50%-quantile
  - 10%-quantile and 90%-quantile are more useful than the minimum or the maximum
    - Easy to measure with a good accuracy
    - Minimum and maximum can be infinite

Recapitulation – Discrete Distributions

- **Geometric Distribution**
  - Success probability $1 - q$
  - Number of unsuccessful retries until success

$$P(i\ unsuccessful\ retries) = x(i) = q^i(1-q)$$

$$E[X] = \frac{q}{1-q}$$

- **Binomial Distribution**
  - Success probability $1 - q$
  - Fixed number of trials N
  - Number of successes

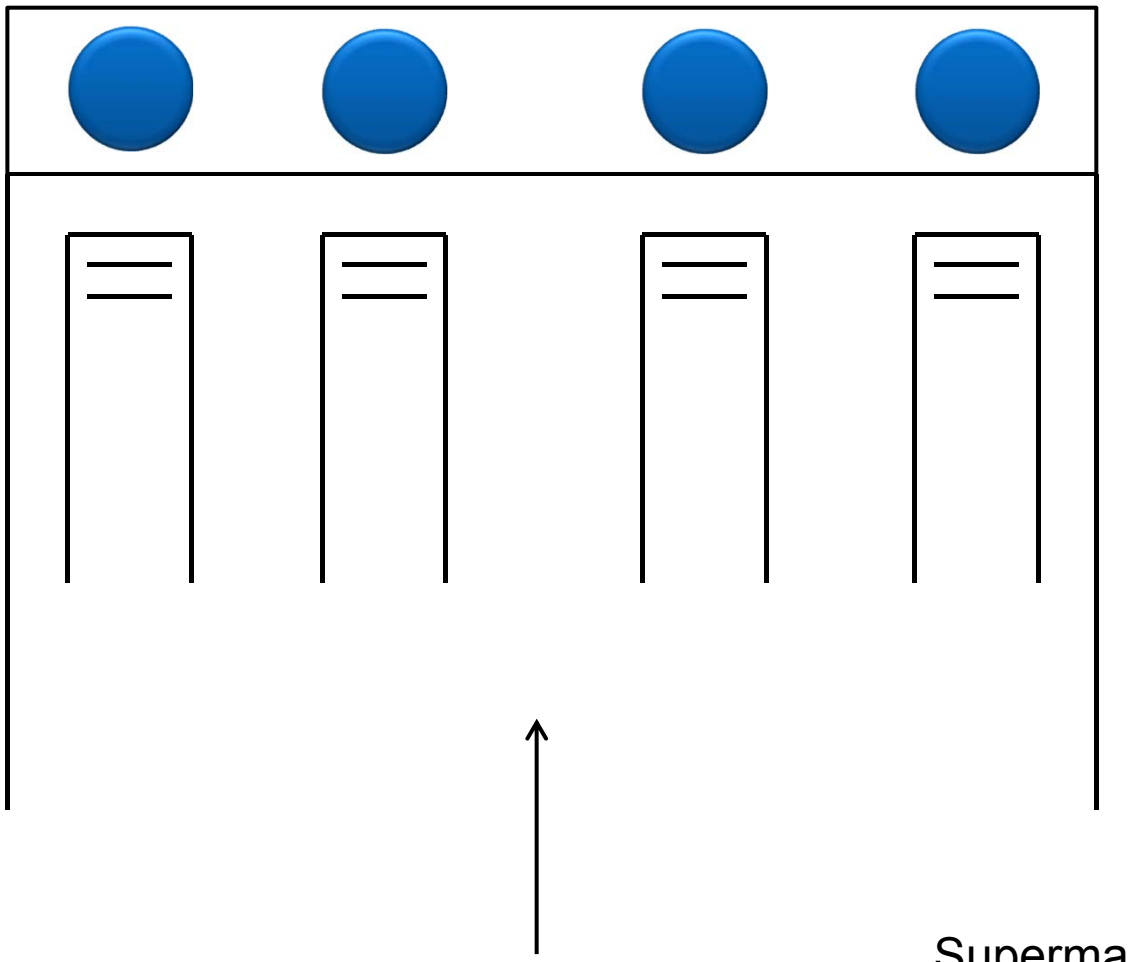
$$P(i\ successes) = x(i) = \binom{N}{i}(1-q)^i q^{N-i}$$

$$E[X] = N(1-q)$$

universität wien

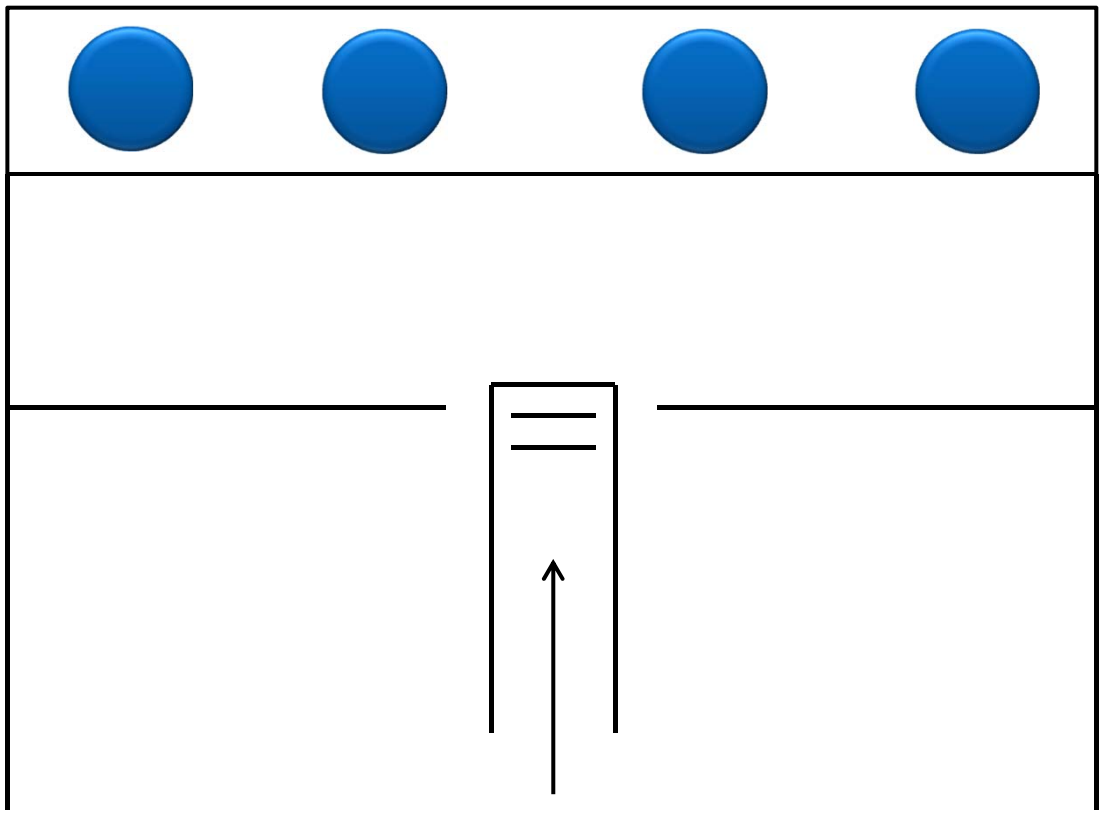
Queuing Theory

You know that?



Supermarket?

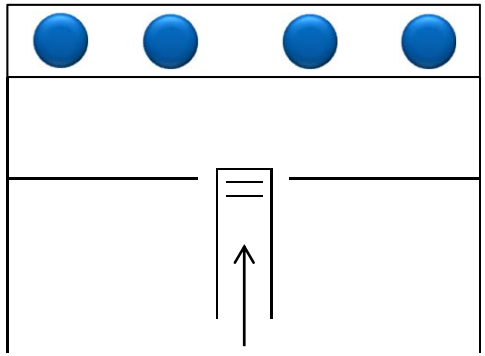
You know that?

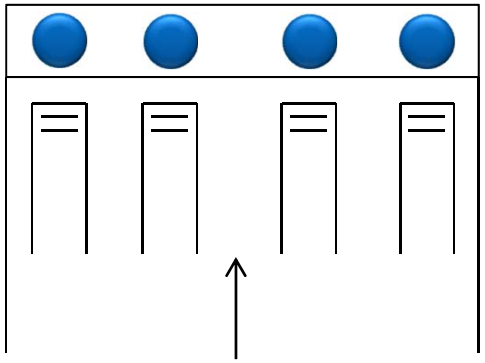


post office?

Which is better?

- What is the definition of “better”?





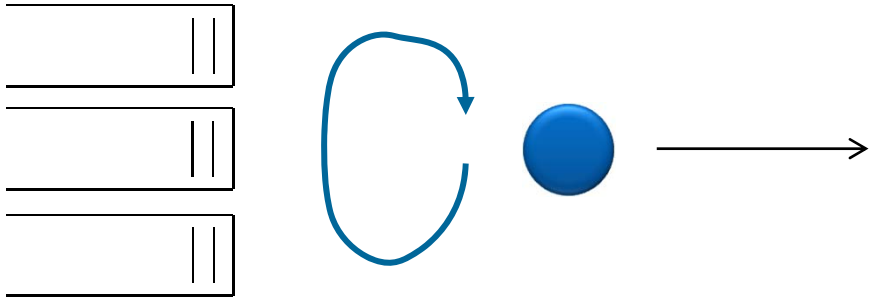
- Throughput?
- Average waiting time?
- Variance of the waiting time?
  - How unequal are equal customers are treated?
- Maximum waiting time?

Queuing Theory

- Queuing theory focus on the behavior of systems
- Systems are composed of:
  - Customers respectively request that enter the system and leave it after processing
  - Processing units or servers to do a certain task for the customer
  - Transport system to control the user movement
  - …
- Measurement parameters:
  - processing time of a request
  - utilization of the processing units
  - …

Queuing and Quality of Service in Networks

- Router in the Internet (schematically)



- Scheduling
  - FIFO (First In First Out, First Come First Serve): customers are processed in the order of arrival (1 queue)
  - Round-Robin: different queues are processed alternately (multiple queues)
  - There are approaches (e.g. weighted fair queuing) to priorities certain packets in order to fit a quality of service level (throughput, latency, deadlines, …)

Modeling of the Stochastic System Behavior

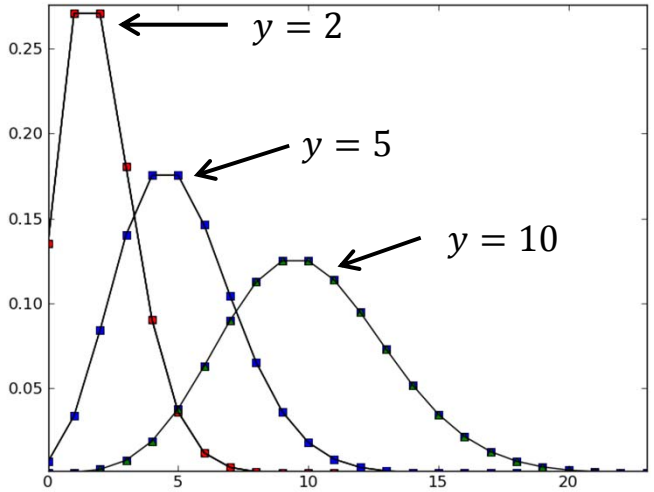
- Queuing theory
- Theory of stochastic processes

- Stochastic elements
- Request arrivals
- Request processing

- Behavior modeling
  - with distribution functions in the simple case
  - important distributions: Poisson-distribution and negative exponential distribution (memoryless, Markov-process, the next state depends only on the current and not on the previous states)
    - In the discrete case the binomial and geometric distributions are also memoryless

Poisson Distribution

- Number of events in a fixed continuous time interval
- Event is equiprobable at every point of time (memoryless)
- Distribution function $x(i) = \frac{y^i}{i!} e^{-y}$
- Mean $\bar{x} = E[x] = y$
- Variation $c = \sqrt{\frac{1}{y}}$

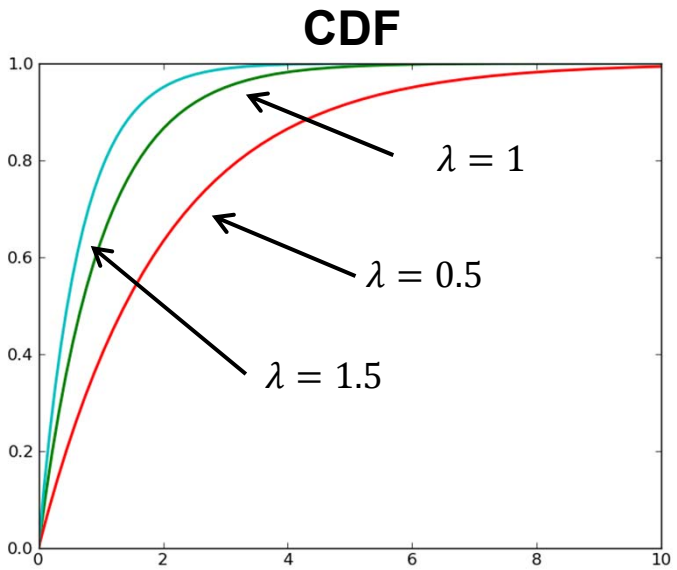


- Info: Distribution for infrequent events (approximates binomial distribution for large n and small success probabilities (1-q)), life time distribution
- E.g.: calls per hour

Exponential Distribution (M)

- Continuous time between two events
- Event is equiprobable at every point of time (memoryless)
- Distribution $A(t) = 1 - e^{-\lambda t}, t \geq 0$
- Probability density function $a(t) = \lambda e^{-\lambda t}, t \geq 0$
- Mean $E[X] = \frac{1}{\lambda}$
- Variation $c = 1$

- Info: Distribution for infrequent events, life time distribution
- E.g.: Time between two customer arrivals in a supermarket



Coherence of POIS and M

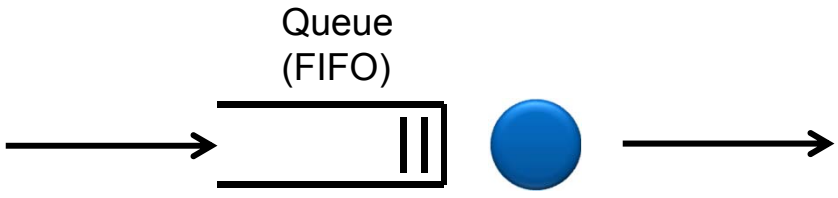
- Markov-System

Poisson (POIS)	Exponential (M)
Number of events in a fixed time interval T	Time between two events

$$\frac{y}{T} = \frac{E[Pois]}{T} = \frac{1}{E[M]} = \lambda$$

Fundamental Terms in Queuing Theory

- E.g.: single channel model



Arrivals
Requests
Customers
Clients
Packets

Server

Arrival Process

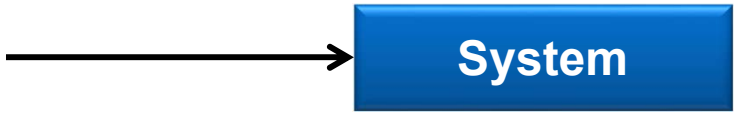
Serving Time Distribution

Processes

- The term “process” describes a procedure. This means, a sequence of values, which in the case of a stochastic process are of stochastically nature.
- This can be represented by the following example: A process is started on the computer and after a while it finishes. At the ending of a process an event takes place. This can either be a completion of a request or the end of the waiting period for a new request.
- The distribution specifies the time until the next event.

- E.g.:
  - A deterministic process with a mean of 5 s always takes 5 s
  - A geometric distributed process accomplishes a Bernoulli experiment at each time interval with a success probability of 1-q. With this experiment it is stated if the process ends or not. The end of the process is conditioned by this single experiment. Thus, the process is memoryless. In the continuous case the negative-exponential distribution is memoryless.

Arrival Process



- An arrival process describes the behavior of the event arrivals. The procedure is the waiting for the next arrival. At the end of a process an event arrival process occurs.
- Acts of peoples daily life occur very infrequent, but many people perform the same things. Thus, arrival processes are often negative-exponential (M) distributed
  - E.g.: user makes a phone call, user connects to the Internet, customer enters the supermarket, …
  - E.g.: Regarding a PCM transmission line (e.g. telephone) packets arrive uniformly. Thus, the arrival process is deterministically distributed.

Service Process

- A service process describes the duration to serve a request on a server. This equals the total service time.
- E.g.: If packets with a fixed size are copied to time slices on a transmission line by a multiplexer, the serving time is deterministic. Nevertheless, the total time in system (completion time) is not necessarily deterministic. Different packets could suffer different waiting times if the arrival process is stochastic.

Queues

- Queues are an important part in system investigations. Generally, not every request can be served at the arrival time. This raises the question, how these packets are treated.
- Queues are specified by size and scheduling mechanism. A basic scheduling is FIFO (first in first out), it serves the requests in the order of their arrival.

- A full queue leads to request rejection.
- If no size is specified for a queue, it is assumed that it is infinite.

Little's & Jackson's Theorems

- Little's Theorem
  - Every system with arbitrary arrival process and arrival rate λ fulfills the following term: $\lambda E[T] = E[X]$
    - $E[X]$: mean number of requests in the system
    - $E[T]$: mean serving time of the system
- Jackson's Theorem (informal)
  - In a system with memoryless arrivals and serving rates each server can be modeled with an independent Markov system.

universität wien

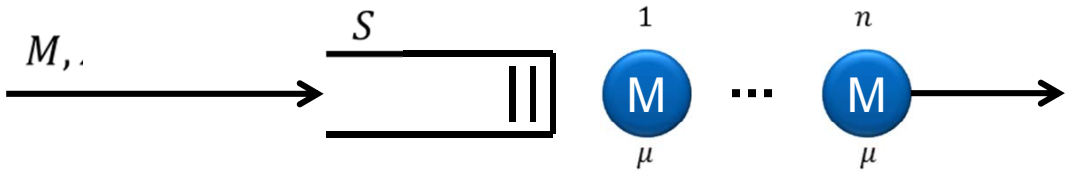
Chair of Future Communication

endowed by

A1 TELEKOM AUSTRIA

Markov Systems

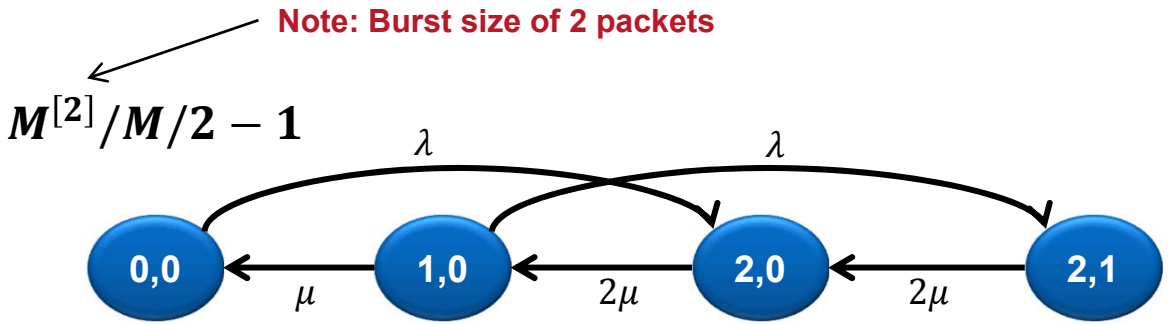
Markov Process



- In Markov systems the arrival process and the serving process are memoryless. Thus, each process has the Markov character. The arrival process is a Poisson process (inter arrival time is negative exponential distributed) and the serving time is negative exponential distributed.
- These systems are relatively easy to analyze. Many common systems nearly fit the Markov characteristic.
- E.g.: phone calls in a telephone network

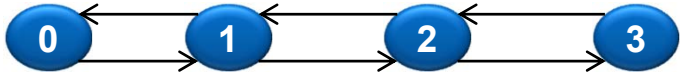
System States

- The state of a memoryless system can be described by the number of elements and their location within the system.
  - The distribution for a state change of each element is identical at each time. (memoryless)
- E.g.: Nodes are described by used servers and elements in the queue

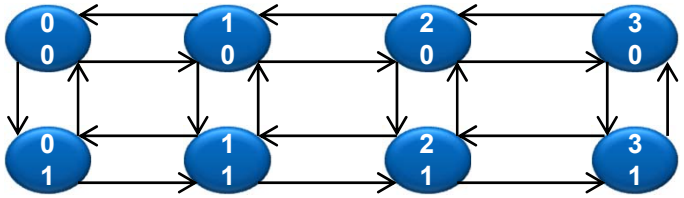


Birth-Death Process

- Def.: (Birth-Death Process)
  - Birth-death processes are a special case of a Markov process. There are only state changes between adjacent states.
- E.g.:

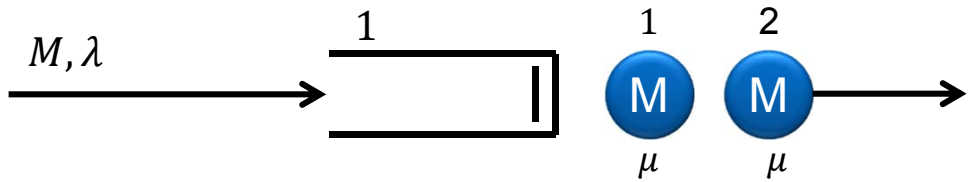


- Also multi dimensional birth-death processes are possible

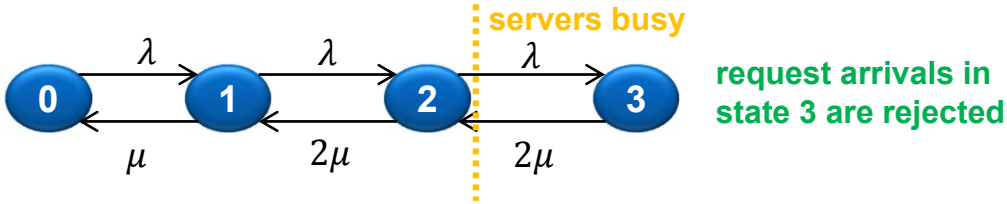


Birth-Death Process M/M/2-1 (or M/M/2/3)

- E.g.: Analysis of a birth-death process M/M/2-1 (M/M/2/3)
  - System with 2 servers and a queue length of 1



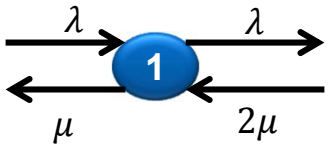
  - State space:



  - State number equals the number of request in the system
  - x(0): empty system x(1): single server busy
  - x(2): both servers busy (double serving rate)
  - x(3): both servers busy (double serving rate) and full queue

M/M/2-1 - Calculation

- Probability of a specific state
- Idea: The receiving rate has to be equal to the leaving rate in the stationary case.



- Equation system for the state probabilities

$$\lambda x(0) = \mu x(1)$$
$$(\lambda + \mu)x(1) = \lambda x(0) + 2\mu x(2)$$
$$(\lambda + 2\mu)x(2) = \lambda x(1) + 2\mu x(3)$$
$$\sum_{i=0}^{3} x(i) = 1$$
$$x(1) = \frac{\lambda}{\mu} x(0)$$

M/M/2-1 - Calculation

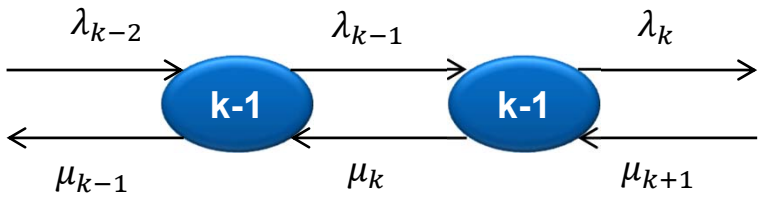
$$x(2) = -\frac{\lambda}{2\mu}x(0) + \frac{\lambda+\mu}{2\mu}x(1) = x(0)\frac{-\lambda + \frac{\lambda^2}{\mu} + \lambda}{2\mu} = x(0)\frac{\lambda^2}{2\mu^2}$$

$$x(3) = \frac{\lambda+2\mu}{2\mu}x(2) \pm \frac{\lambda}{2\mu}x(1) = \left(\frac{\lambda+2\mu}{2\mu}\frac{\lambda^2}{2\mu^2} + \frac{-\lambda}{2\mu}\frac{\lambda}{\mu}\right)x(0) = \frac{\lambda^3}{4\mu^3}x(0)$$

$$\sum_{i=0}^{3} x(i) = 1 = x(0)\left(1 + \frac{\lambda}{\mu} + \frac{\lambda^2}{2\mu^2} + \frac{\lambda^3}{4\mu^3}\right) \Rightarrow x(0) = \frac{1}{1 + \frac{\lambda}{\mu} + \frac{\lambda^2}{2\mu^2} + \frac{\lambda^3}{4\mu^3}}$$

General Solution of Birth-Death Processes

- General solution for birth-death processes with n states:
  - let λ_k, μ_k be the rates to leave state k (see state diagram)



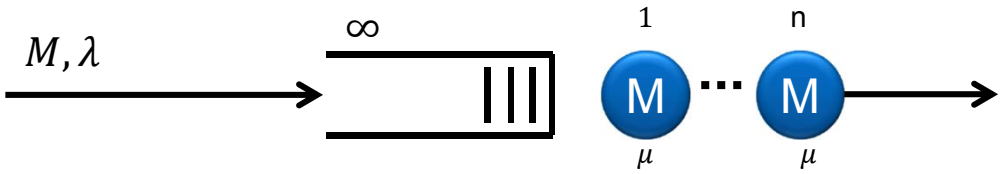
$$x(i) = x(0)\frac{\prod_{k=0}^{i-1}\lambda_k}{\prod_{k=1}^{i}\mu_k}$$

$$x(0) = \frac{1}{1 + \sum_{i=1}^{n-1}\frac{\prod_{k=0}^{i-1}\lambda_k}{\prod_{k=1}^{i}\mu_k}}$$

Results for M/M/n-Systems

- The section closes with the presentation of some classical M/M/n-systems:
  - M/M/n (M/M/n-waiting system)
  - M/M/n-0 (M/M/n-loss system)
  - M/GI/n-0

Waiting System M/M/n



- Offered load $a = \frac{\lambda}{\mu} = \lambda E[B]$
  - load is often given in the pseudo unit Erlang, a=1 equals 1 Erlang
- Traffic intensity = average number of occupied servers
  - $Y = a$
- Utilization of a single server $\rho = \frac{a}{n} = \frac{\lambda}{\mu n}$

Waiting System M/M/n

- Probability of a state

$$\frac{1}{x(0)} = \sum_{i=1}^{n-1} \frac{a^i}{i!} + \frac{a^n}{n!}\frac{1}{1-\rho}$$

$$x(i) = \begin{cases} x(0)\frac{a^i}{i!}, i = 0,1,\dots,n \\ x(0)\frac{a^n}{n!}\left(\frac{a}{n}\right)^{i-n}, i > n \end{cases}$$

- Waiting probability

$$p_w = \frac{\frac{a^n}{n!}\frac{1}{1-\rho}}{\sum_{i=0}^{n-i}\frac{a^i}{i!}+\frac{a^n}{n!}\frac{1}{1-\rho}}$$

(Erlang-Waiting-Formula)

universität wien

Chair of Future Communication

endowed by

- Average queue length

$$\Omega = p_w \frac{\rho}{1-\rho}$$

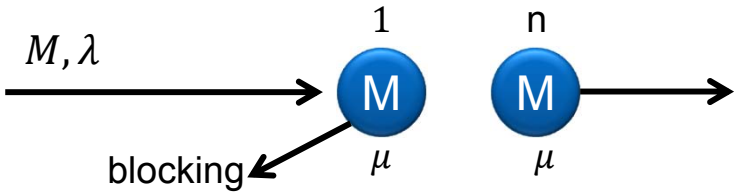
- Average waiting time for all request

$$E[W] = \frac{\Omega}{\lambda}$$

- Average waiting time for the waiting requests

$$E[W_1] = \frac{\Omega}{\lambda p_w} = \frac{1}{\lambda} \frac{\rho}{1-\rho}$$

Blocking System M/M/n-0 (M/M/n/n)



- M/M/n-0 is a birth-death process. The state probabilities can be calculated accordingly.
- Offered load

$$a = \frac{\lambda}{\mu}$$

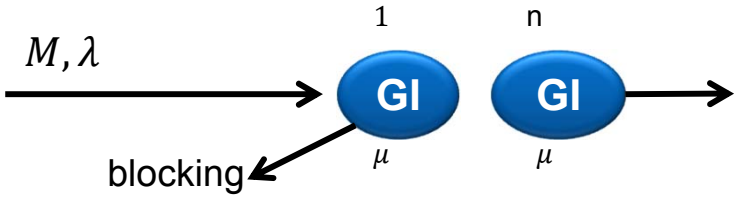
- Blocking probability

$$p_b = x(n) = \frac{\frac{a^n}{n!}}{\sum_{i=0}^{n} \frac{a^i}{i!}} \text{ (Erlang-Blocking-Formula)}$$

- Traffic intensity = average number of occupied servers

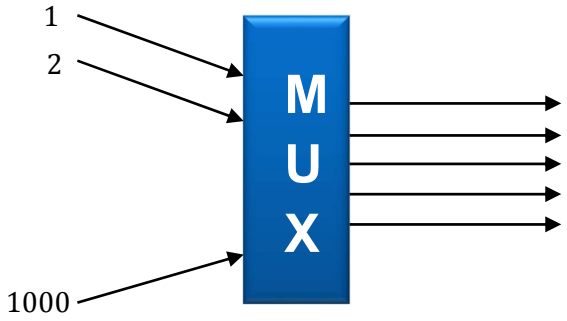
$$Y = a(1 - p_b) = \frac{\lambda}{\mu}(1 - p_b)$$

Blocking System M/GI/n-0 (M/GI/n/n)



- The blocking system (M/GI/n-0) equals the loss system M/M/n-0. Thus, the values for blocking probability, state probability and offered load can be adopted.

Example



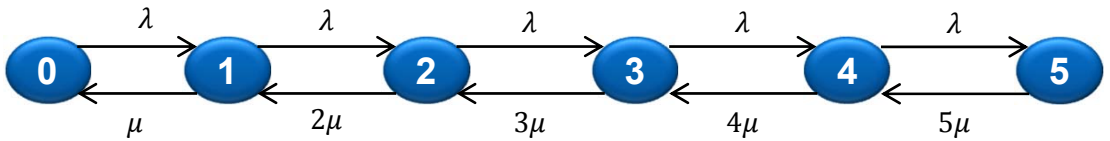
A multiplexer distributes requests from 1000 nodes to 5 outgoing channels with a capacity of 200 Mbps for each of them. Each Poisson distributed incoming stream utilizes the system with 1500000 messages per second. The size of the different messages is independent and its distribution is unknown with an average message size of 4 kBit. Messages that can not be transmitted directly are dropped. This equals a M/GI/n-0 model.

a) Calculate the blocking probability

b) How many outgoing channels are required to keep the blocking probability below 5%?

- System
  - M/GI/n-0
  - Arrival rate $\lambda = 150000 = 1.5 * 10^5 \frac{1}{s}$
  - Average length $l = 4000bit$
  - Bandwidth per channel $C = 200Mbps = 2 * 10^8 bps$
  - Serving rate $\mu = \frac{C}{l} = \frac{2*10^8 bit/s}{4*10^3 bit} = 5 * 10^4 \frac{1}{s} = 50000 \frac{1}{s}$
  - Offered load $a = \frac{\lambda}{\mu} = \frac{150000}{50000} = 3Erlang$
  - Channels $n = 5$

- a) It is given that M/M/n-0 = M/GI/n-0
  - Arrival rate $\lambda = 1.5 * 10^5 \frac{1}{s}$
  - Serving rate $\mu = \frac{c}{l} = 50000 \frac{1}{s}$
  - Channels $n = 5$



Solution using the stationary case with a linear equation system

$$\lambda x(0) = \mu x(1) \qquad (\lambda + 2\mu)x(2) = \lambda x(1) + 3\mu x(3) \qquad (\lambda + 4\mu)x(4) = \lambda x(3) + 5\mu x(5)$$
$$(\lambda + \mu)x(1) = \lambda x(0) + 2\mu x(2) \qquad (\lambda + 3\mu)x(3) = \lambda x(2) + 4\mu x(4) \qquad x(0) + x(1) + \cdots + x(5) = 1$$

→ we are looking for x(5), which is the state probability of state 5

- Alternative: Solution using known formula
  - Erlang-Blocking-Formula

$$p_b = \frac{\frac{a^n}{n!}}{\sum_{i=0}^{n} \frac{a^i}{i!}}$$

$$n = 1: \quad p_b = \frac{\frac{3}{1}}{1 + \frac{3}{1}} = 0.75$$

$$n = 2: \quad p_b = \frac{\frac{9}{2}}{1 + 3 + 4.5} = 0.53$$

$$n = 3: \quad p_b = \frac{\frac{27}{6}}{1 + 3 + 4.5 + 4.5} = 0.35$$

$$n = 4: \quad p_b = \frac{\frac{81}{24}}{1 + 3 + 4.5 + 4.5 + 3{,}375} = 0.20$$

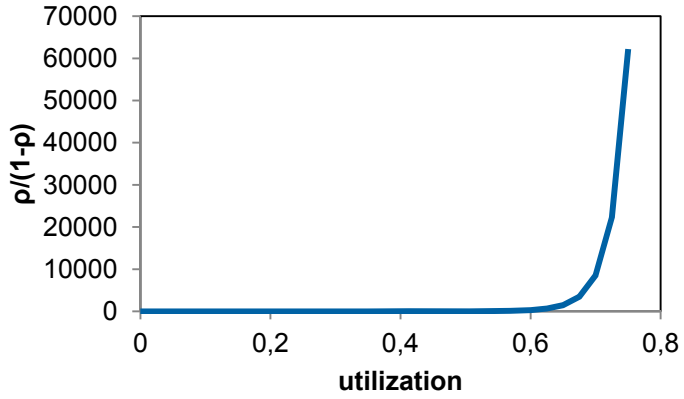
$$n = 5: \quad p_b = \frac{\frac{243}{120}}{1 + 3 + 4.5 + 4.5 + 3{,}375 + 2.025} = 0.11$$

$$n = 6: \quad p_b = \frac{1{,}0125}{18.4 + 1.0125} = 0.052$$

$$n = 7: \quad p_b = \frac{0.44}{18.4 + 1.0125 + 0.44} = 0.022$$

Waiting Time vs. Utilization

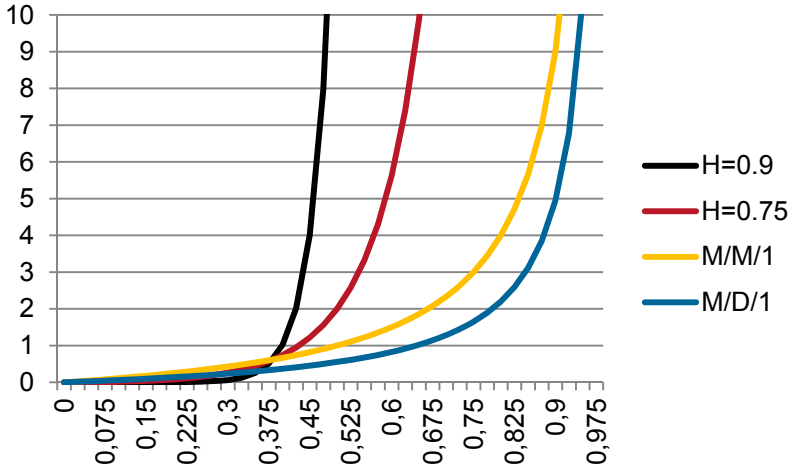
- Waiting queue length in an M/M/n system
  - $\Omega = p_w \frac{\rho}{1-\rho}$



- This system behavior is fundamental. Problems with system stability occur below 100%. Only totally deterministic systems can handle 100% load without stability problems.

Internet Traffic

- The traffic in the Internet behaves extremely variant (high variation, higher than “Markov-traffic”!) and is self-similar (Hurst-parameter H).
- Waiting queue length
  - $q = \frac{\rho\text{^}1/2(1-H)}{(1-q)^{H/(1-H)}}$



- Poisson behaviors are very common in session arrivals or dial-in processes.