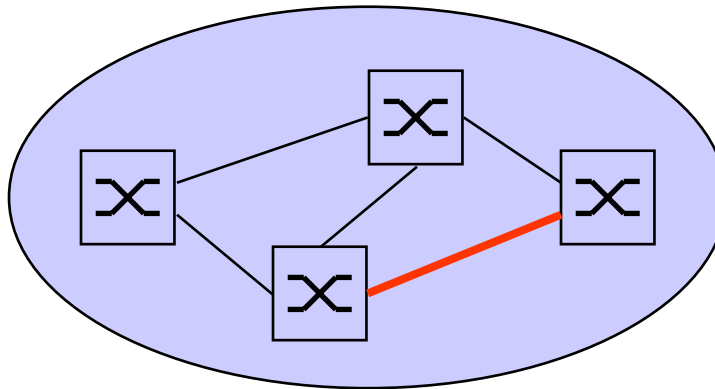# Examples - Traffic Models

Contents

- Model for telephone traffic
- Packet level model for data traffic
- Flow level model for elastic data traffic
- Flow level model for streaming data traffic

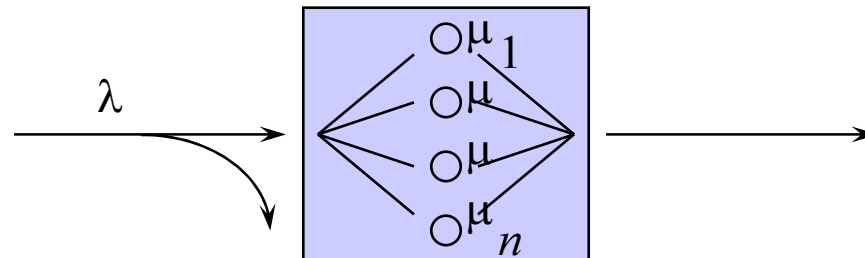# Classical model for telephone traffic (1)

- **Loss models** have traditionally been used to describe (circuit-switched) telephone networks
  - Pioneering work made by Danish mathematician A.K. Erlang (1878-1929)
- Consider a link between two telephone exchanges
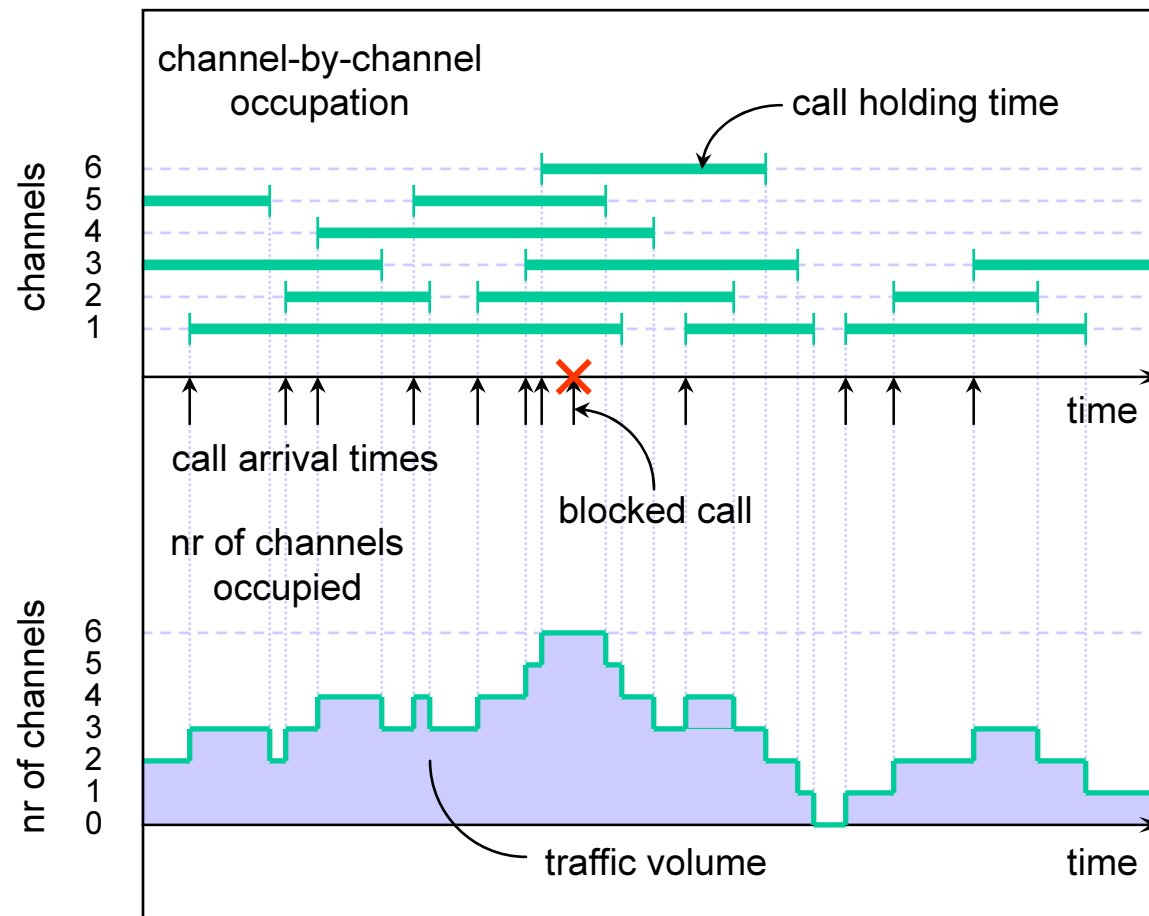 - traffic consists of the ongoing telephone calls on the link

Classical model for telephone traffic (2)

- Erlang modelled this as a **pure loss system** ($m = 0$)
  - customer = call
    - λ = call arrival rate (calls per time unit)
  - service time = (call) holding time
    - $h = 1/\mu$ = average holding time (time units)
  - server = channel on the link
    - n = nr of channels on the link

Traffic process

channel-by-channel occupation

call holding time

channels

6
5
4
3
2
1

call arrival times

blocked call

nr of channels occupied

6
5
4
3
2
1
0

nr of channels

traffic volume

time

5

Traffic intensity

- The strength of the offered traffic is described by the traffic intensity a

- By definition, the **traffic intensity** a is the product of the arrival rate λ and the mean holding time h:

$$a = \lambda h$$

- The traffic intensity is a **dimensionless** quantity. Anyway, the unit of the traffic intensity a is called **erlang** (**erl**)
- By Little's formula: traffic of one erlang means that one channel is occupied on average

- Example:
  - On average, there are 1800 new calls in an hour, and the average holding time is 3 minutes. Then the traffic intensity is

$$a = 1800 * 3 / 60 = 90 \text{ erlang}$$

Blocking

- In a loss system some calls are lost
  - a call is lost if all n channels are occupied when the call arrives
  - the term **blocking** refers to this event
- There are two different types of blocking quantities:
  - **Call blocking** B_c = probability that an arriving call finds all n channels occupied = the fraction of calls that are lost
  - **Time blocking** B_t = probability that all n channels are occupied at an arbitrary time = the fraction of time that all n channels are occupied
- The two blocking quantities are not necessarily equal
  - Example: your own mobile
  - But if calls arrive according to a Poisson process, then $B_c = B_t$
- Call blocking is a better measure for the quality of service experienced by the subscribers but, typically, time blocking is easier to calculate

Call rates

- In a loss system each call is either **lost** or **carried.** Thus, there are three types of call rates:
  - λ_{offered} = arrival rate of all call attempts
  - λ_{carried} = arrival rate of carried calls
  - λ_{lost} = arrival rate of lost calls

$$\lambda_{\text{offered}} \qquad \lambda_{\text{carried}}$$
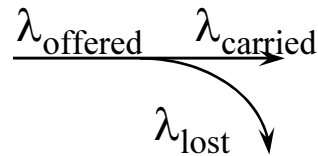
$$\lambda_{\text{lost}}$$

$$\lambda_{\text{offered}} = \lambda_{\text{carried}} + \lambda_{\text{lost}} = \lambda$$

$$\lambda_{\text{carried}} = \lambda(1 - B_{\text{c}})$$

$$\lambda_{\text{lost}} = \lambda B_{\text{c}}$$

8

> # Traffic streams

- The three call rates lead to the following three traffic concepts:
  - **Traffic offered** $a_{\text{offered}} = \lambda_{\text{offered}} h$
  - **Traffic carried** $a_{\text{carried}} = \lambda_{\text{carried}} h$
  - **Traffic lost** $a_{\text{lost}} = \lambda_{\text{lost}} h$
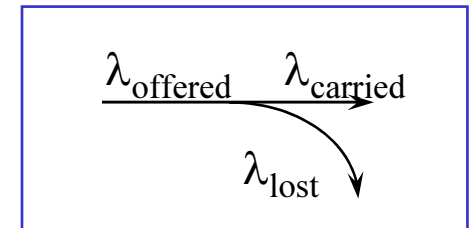
$$a_{\text{offered}} = a_{\text{carried}} + a_{\text{lost}} = a$$

$$a_{\text{carried}} = a(1 - B_{\text{c}})$$

$$a_{\text{lost}} = a B_{\text{c}}$$

- Traffic offered and traffic lost are hypothetical quantities, but traffic carried is **measurable**, since (by Little's formula) it corresponds to the average number of occupied channels on the link

Teletraffic analysis (1)

- System capacity

  - n = number of channels on the link

- Traffic load

  - a = (offered) traffic intensity

- Quality of service (from the subscribers' point of view)

  - B_c = call blocking = probability that an arriving call finds all n channels occupied

- Assume an **M/G/n/n loss system**:

  - calls arrive according to a **Poisson process** (with rate λ)
  - call holding times are independently and identically distributed according to **any distribution** with mean h

Teletraffic analysis (2)

- Then the quantitive relation between the three factors (system, traffic, and quality of service) is given by **Erlang's formula**:

$$B_{\mathrm{c}} = \mathrm{Erl}(n, a) := \frac{\dfrac{a^n}{n!}}{\displaystyle\sum_{i=0}^{n} \dfrac{a^i}{i!}}$$

$$n! = n \cdot (n-1) \cdot \ldots \cdot 2 \cdot 1, \quad 0! = 1$$

- Also called:
  - Erlang's B-formula
  - Erlang's blocking formula
  - Erlang's loss formula
  - Erlang's first formula

Example

- Assume that there are $n = 4$ channels on a link and the offered traffic is $a = 2.0$ erlang. Then the call blocking probability B_c is

$$B_c = \text{Erl}(4,2) = \frac{\frac{2^4}{4!}}{1 + 2 + \frac{2^2}{2!} + \frac{2^3}{3!} + \frac{2^4}{4!}} = \frac{\frac{16}{24}}{1 + 2 + \frac{4}{2} + \frac{8}{6} + \frac{16}{24}} = \frac{2}{21} \approx 9.5\%$$

- If the link capacity is raised to $n = 6$ channels, then B_c reduces to
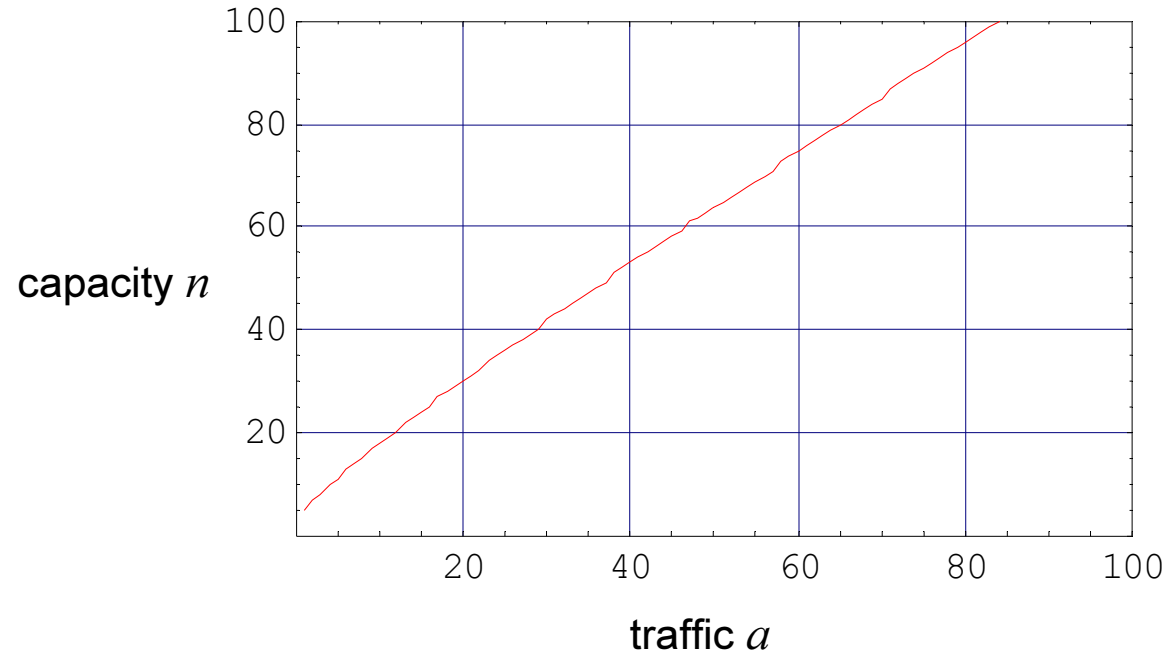
$$B_c = \text{Erl}(6,2) = \frac{\frac{2^6}{6!}}{1 + 2 + \frac{2^2}{2!} + \frac{2^3}{3!} + \frac{2^4}{4!} + \frac{2^5}{5!} + \frac{2^6}{6!}} \approx 1.2\%$$

Capacity vs. traffic

- Given the quality of service requirement that $B_c < 1\%$, the required capacity n depends on the traffic intensity a as follows:

$$n(a) = \min\{i = 1,2,\ldots \mid \mathrm{Erl}(i,a) < 0.01\}$$

capacity n

traffic a

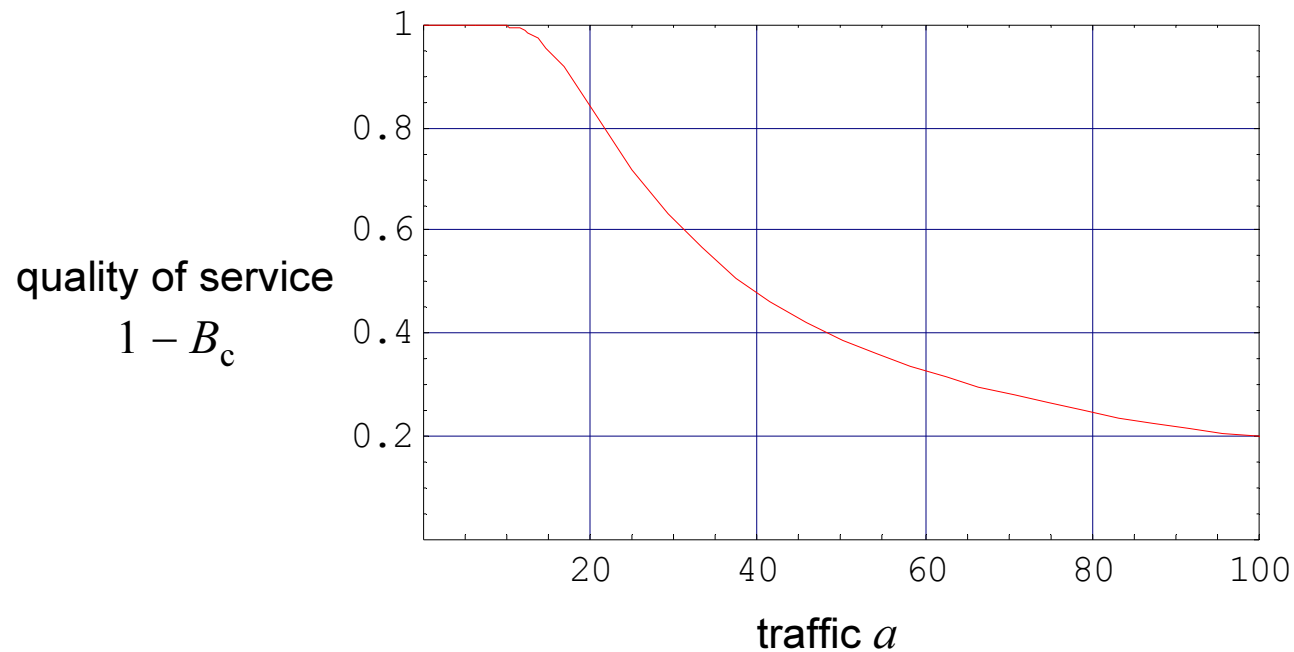## Quality of service vs. traffic

- Given the capacity $n = 20$ channels, the required quality of service $1 - B_c$ depends on the traffic intensity a as follows:

$$1 - B_c(a) = 1 - \mathrm{Erl}(20, a)$$

quality of service $1 - B_c$

traffic a

14

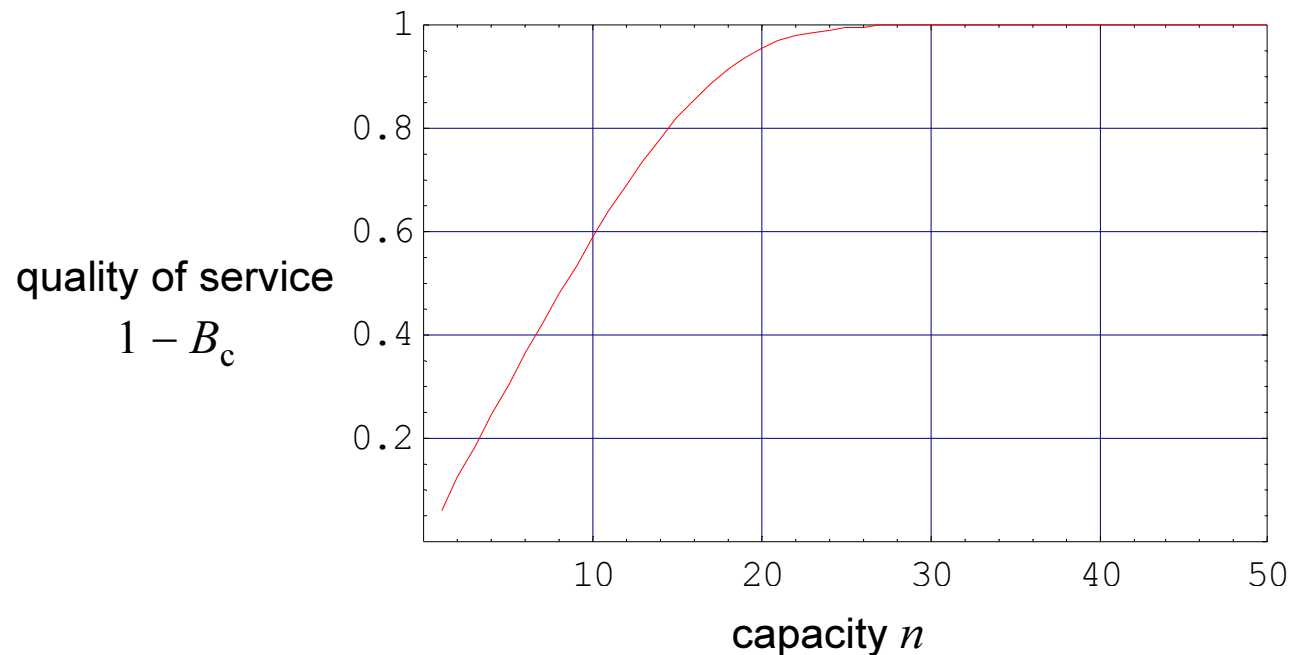## Quality of service vs. capacity

- Given the traffic intensity $a = 15.0$ erlang, the required quality of service $1 - B_c$ depends on the capacity n as follows:

$$1 - B_c(n) = 1 - \text{Erl}(n, 15.0)$$

quality of service
$1 - B_c$

capacity n

Contents

- Model for telephone traffic
- Packet level model for data traffic
- Flow level model for elastic data traffic
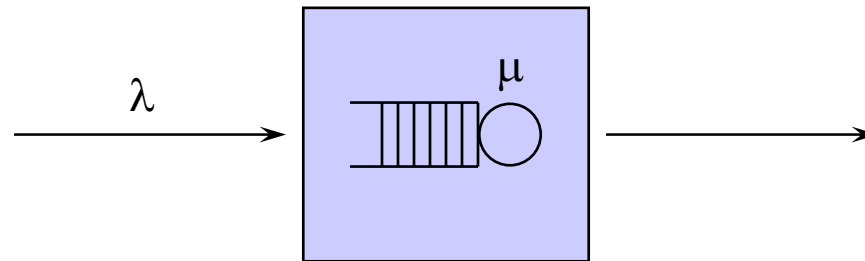- Flow level model for streaming data traffic

Packet level model for data traffic (1)

- **Queueing models** are suitable for describing (packet-switched) data traffic at packet level
  - Pioneering work made by many people in 60's and 70's related to ARPANET, in particular *L. Kleinrock* (`http://www.lk.cs.ucla.edu/`)
- Consider a link between two packet routers
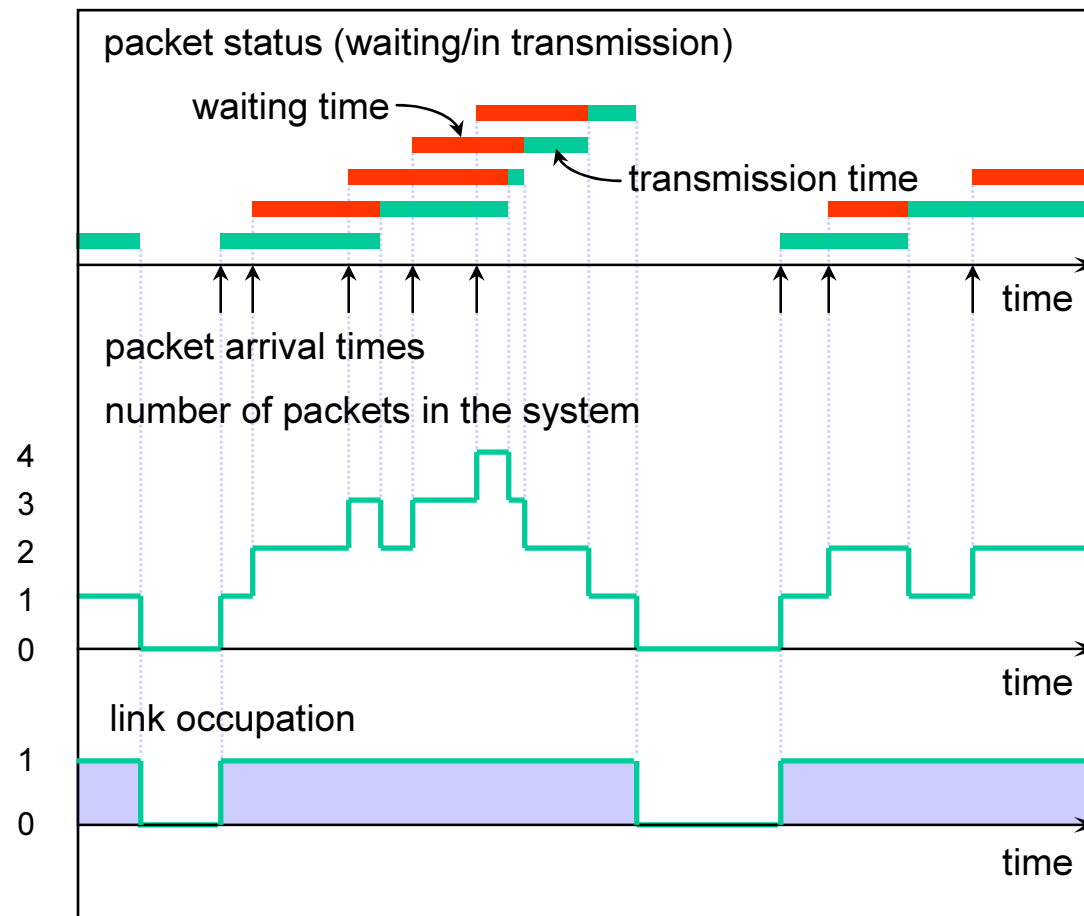 - traffic consists of data packets transmitted along the link
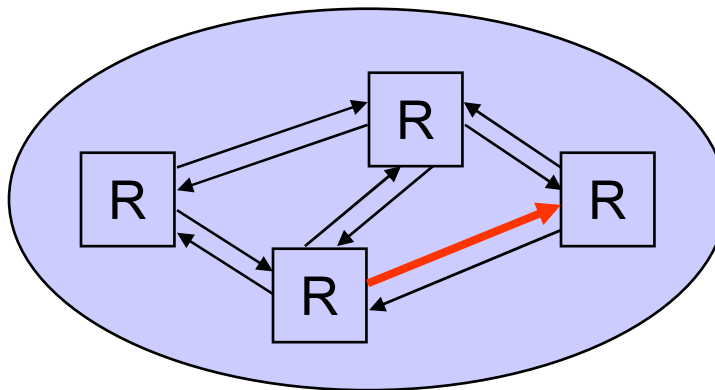
Packet level model for data traffic (2)

- This can be modelled as a **pure queueing system** with a single server ($n = 1$) and an infinite buffer ($m = \infty$)
  - customer = packet
    - λ = packet arrival rate (packets per time unit)
    - L = average packet length (data units)
  - server = link, waiting places = buffer
    - C = link speed (data units per time unit)
  - service time = packet transmission time
    - $1/\mu = L/C$ = average packet transmission time (time units)

18

Traffic process

packet status (waiting/in transmission)

waiting time

transmission time

packet arrival times

number of packets in the system

4
3
2
1
0

time

link occupation

1
0

time

Traffic load

- The strength of the offered traffic is described by the traffic load ρ
- By definition, the **traffic load** ρ is the ratio between the arrival rate λ and the service rate $\mu = C/L$:

$$\rho = \frac{\lambda}{\mu} = \frac{\lambda L}{C}$$

  – The traffic load is a **dimensionless** quantity
  – By Little's formula, it tells the **utilization factor** of the server, which is the probability that the server is busy

Example

- Consider a link between two packet routers. Assume that,
  - on average, 50,000 new packets arrive in a second,
  - the mean packet length is 1500 bytes, and
  - the link speed is 1 Gbps.
- Then the traffic load (as well as, the utilization) is

$$\rho = 50{,}000 * 1500 * 8 / 1{,}000{,}000{,}000 = 0.60 = 60\%$$

Delay

- In a queueing system, some packets have to wait before getting served
    - An arriving packet is buffered, if the link is busy upon the arrival
- **Delay** of a packet consists of
    - the **waiting time**, which depends on the state of the system upon the arrival, and
    - the **transmission time**, which depends on the length of the packet and the capacity of the link
- Example:
    - packet length = 1500 bytes
    - link speed = 1 Gbps
    - transmission time = 1500*8/1,000,000,000 = 0.000012 s = 12 μs

Teletraffic analysis (1)

- System capacity

  - C = link speed in kbps

- Traffic load

  - λ = packet arrival rate in pps (considered here as a variable)

  - L = average packet length in kbits (assumed here to be constant 1 kbit)

- Quality of service (from the users' point of view)

  - P_z = probability that a packet has to wait "too long", i.e. longer than a given reference value z (assumed here to be constant z = 0.00001 s = 10 μs)

- Assume an **M/M/1 queueing system**:

  - packets arrive according to a **Poisson process** (with rate λ)

  - packet lengths are independent and identically distributed according to the **exponential distribution** with mean L

Teletraffic analysis (2)

- Then the quantitive relation between the three factors (system, traffic, and quality of service) is given by the following formula:

$$P_z = \text{Wait}(C, \lambda; L, z) :=$$

$$\begin{cases} \frac{\lambda L}{C} \exp(-(\frac{C}{L} - \lambda)z) = \rho \exp(-\mu(1-\rho)z), & \text{if } \lambda L < C \ (\rho < 1) \\ 1, & \text{if } \lambda L \geq C \ (\rho \geq 1) \end{cases}$$

- Note:
  - The system is **stable** only in the former case ($\rho < 1$). Otherwise the number of packets in the buffer grows without limits.

Example

- Assume that packets arrive at rate $\lambda = 600{,}000$ pps $= 0.6$ packets/µs and the link speed is $C = 1.0$ Gbps $= 1.0$ kbit/µs.
- The system is stable since

$$\rho = \frac{\lambda L}{C} = 0.6 < 1$$

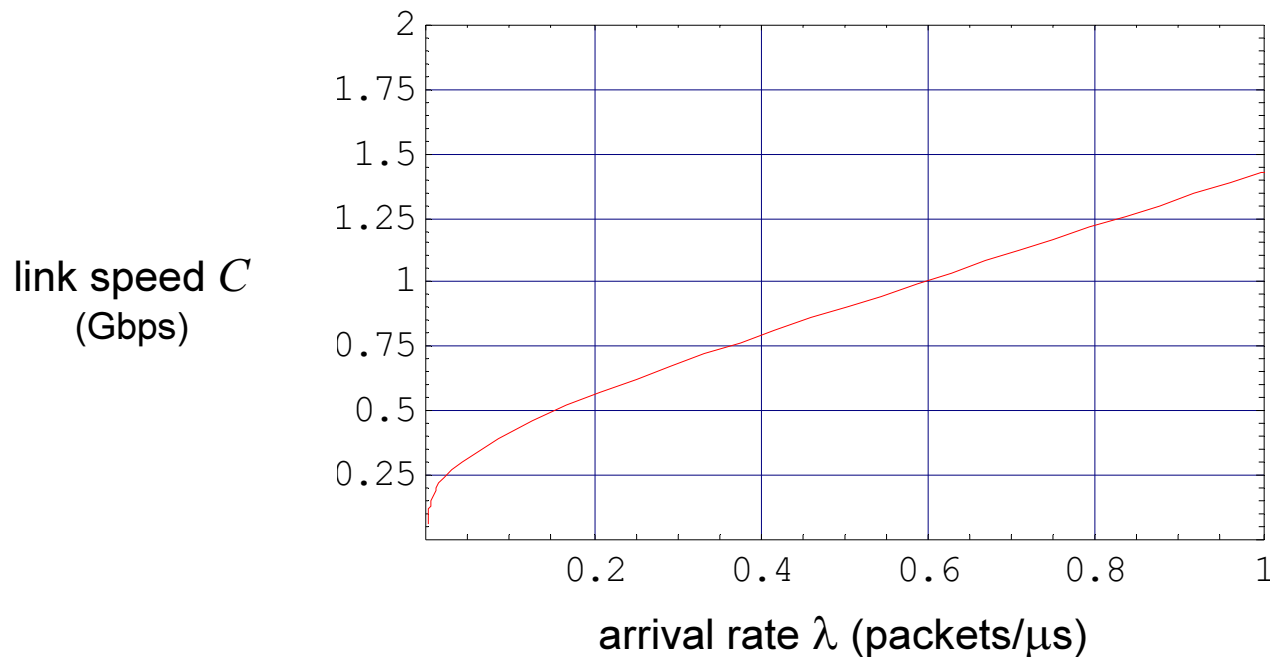- The probability P_z that an arriving packet has to wait too long (i.e. longer than $z = 10$ µs) is

$$P_z = \mathrm{Wait}(1.0, 0.6; 1, 10) = 0.6 \exp(-4.0) \approx 1\%$$

Capacity vs. arrival rate

- Given the quality of service requirement that $P_z < 1\%$, the required link speed C depends on the arrival rate λ as follows:

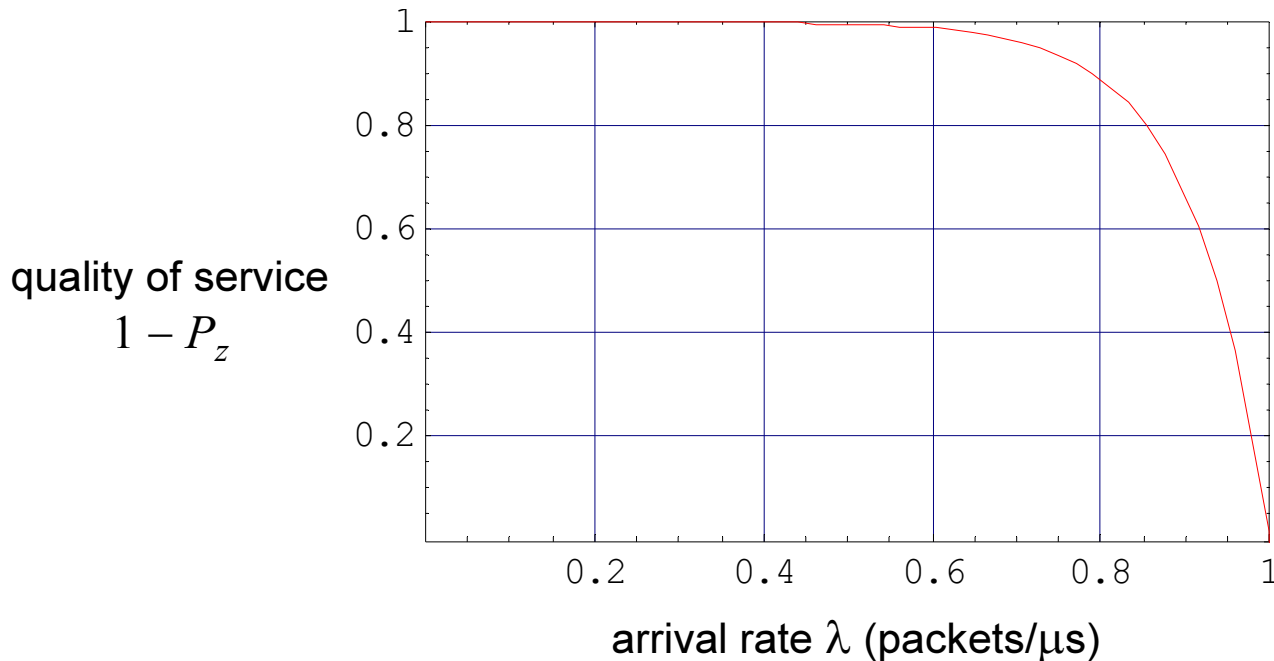$$C(\lambda) = \min\{c > \lambda L \mid \text{Wait}(c, \lambda; 1, 10) < 0.01\}$$

link speed C
(Gbps)

arrival rate λ (packets/μs)

26

Quality of service vs. arrival rate

- Given the link speed $C = 1.0$ Gbps $= 1.0$ kbit/μs, the quality of service $1 - P_z$ depends on the arrival rate λ as follows:

$$1 - P_z(\lambda) = 1 - \text{Wait}(1.0, \lambda; 1, 10)$$

quality of service
$1 - P_z$

arrival rate λ (packets/μs)

27

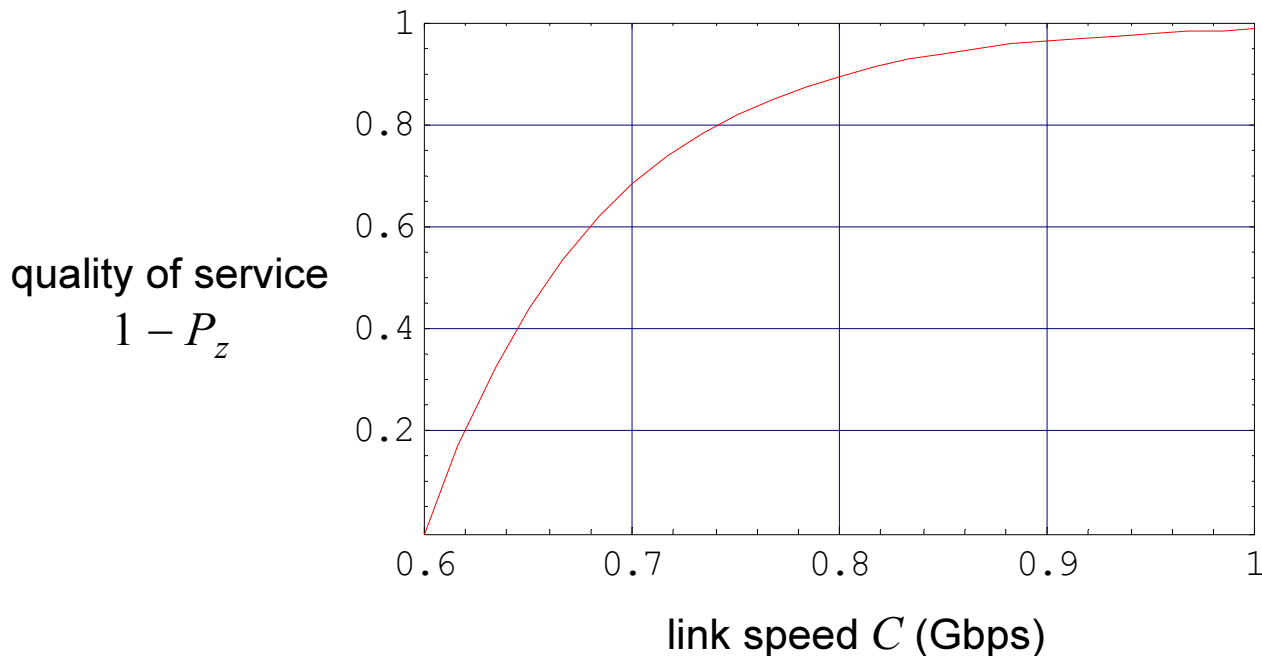## Quality of service vs. capacity

- Given the arrival rate $\lambda = 600{,}000$ pps $= 0.6$ packets/µs, the quality of service $1 - P_z$ depends on the link speed C as follows:

$$1 - P_z(R) = 1 - \text{Wait}(C, 0.6; 1, 10)$$

quality of service
$1 - P_z$

link speed C (Gbps)

28

Contents

- Model for telephone traffic
- Packet level model for data traffic
- Flow level model for elastic data traffic
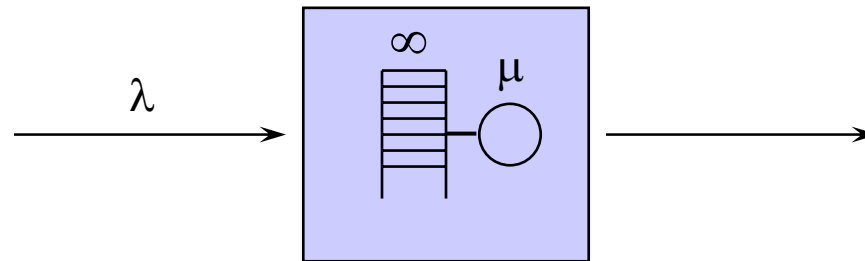- Flow level model for streaming data traffic

Flow level model for elastic data traffic (1)

- **Sharing models** are suitable for describing elastic data traffic at flow level
  - Elasticity refers to the adaptive sending rate of TCP flows
  - This kind of models have been proposed, e.g., by *J. Roberts* and his researchers (`http://perso.rd.francetelecom.fr/roberts/`)
- Consider a link between two packet routers
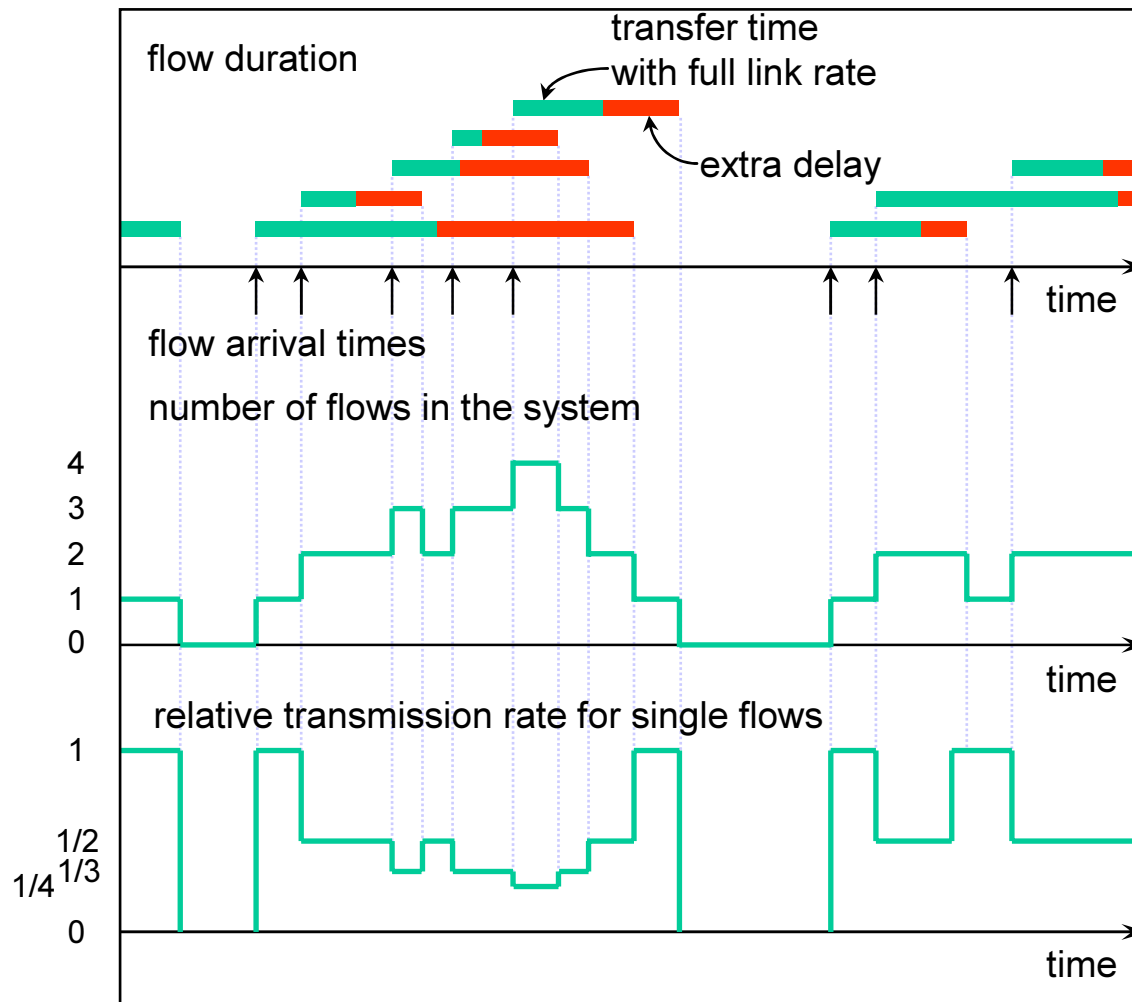 - traffic consists of TCP flows loading the link

30
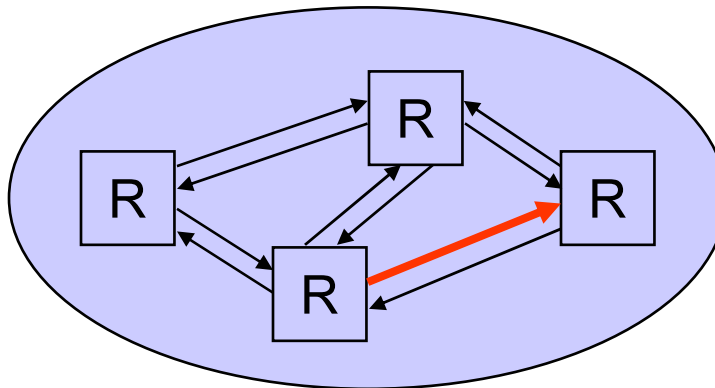
Flow level model for elastic data traffic (2)

- The simplest model is a single server ($n = 1$) **pure sharing system** with a fixed total service rate of μ
  - customer = TCP flow = file to be transferred
    - λ = flow arrival rate (flows per time unit)
    - S = average flow size = average file size (data units)
  - server = link
    - C = link speed (data units per time unit)
  - service time = file transfer time with full link speed
    - $1/\mu = S/C$ = average file transfer time with full link speed (time units)

31

Traffic process

flow duration

transfer time with full link rate

extra delay

flow arrival times

number of flows in the system

4
3
2
1
0

time

relative transmission rate for single flows

1

1/2
1/3
1/4

0

time

32

Traffic load

- The strength of the offered traffic is described by the traffic load ρ

- By definition, the **traffic load** ρ is the ratio between the arrival rate λ and the service rate $\mu = C/S$:

$$\rho = \frac{\lambda}{\mu} = \frac{\lambda S}{C}$$

  - The traffic load is (again) a dimensionless quantity
  - It tells the utilization factor of the server

3. Examples

Example

- Consider a link between two packet routers. Assume that,
  - on average, 50 new flows arrive in a second,
  - average flow size is 1,500,000 bytes, and
  - link speed is 1 Gbps.
- Then the traffic load (as well as, the utilization) is

$$\rho = 50 * 1,500,000 * 8 / 1,000,000,000 = 0.60 = 60\%$$

34

Throughput

- In a sharing system the service capacity is shared among all active flows. It follows that **all** flows get delayed (unless there is only a single active flow)

- By definition, the ratio between the average flow size S and the average total delay D of a flow is called **throughput** θ,

$$\theta = S / D$$

- Example:
    - $S = 1$ Mbit
    - $D = 5$ s
    - $\theta = S/D = 0.2$ Mbps

Teletraffic analysis (1)

- System capacity

  - C = link speed in Mbps

- Traffic load

  - λ = flow arrival rate in flows per second (considered here as a variable)

  - S = average flow size in kbits (assumed here to be constant 1 Mbit)

- Quality of service (from the users' point of view)

  - θ = throughput

- Assume an **M/G/1-PS sharing system**:

  - flows arrive according to a **Poisson process** (with rate λ)

  - flow sizes are independent and identically distributed according to **any distribution** with mean S

Teletraffic analysis (2)

- Then the quantitive relation between the three factors (system, traffic, and quality of service) is given by the following formula:

$$\theta = \mathrm{Xput}(C, \lambda; S) := \begin{cases} C - \lambda S = C(1 - \rho), & \text{if } \lambda S < C \ (\rho < 1) \\ 0, & \text{if } \lambda S \geq C \ (\rho \geq 1) \end{cases}$$

- Note:
    - The system is **stable** only in the former case ($\rho < 1$). Otherwise the number of flows as well as the average delay grows without limits. In other words, the throughput of a flow goes to zero.

Example

- Assume that flows arrive at rate $\lambda = 600$ flows per second and the link speed is $C = 1000$ Mbps $= 1.0$ Gbps.
- The system is stable since

$$\rho = \frac{\lambda S}{C} = \frac{600}{1000} = 0.6 < 1$$
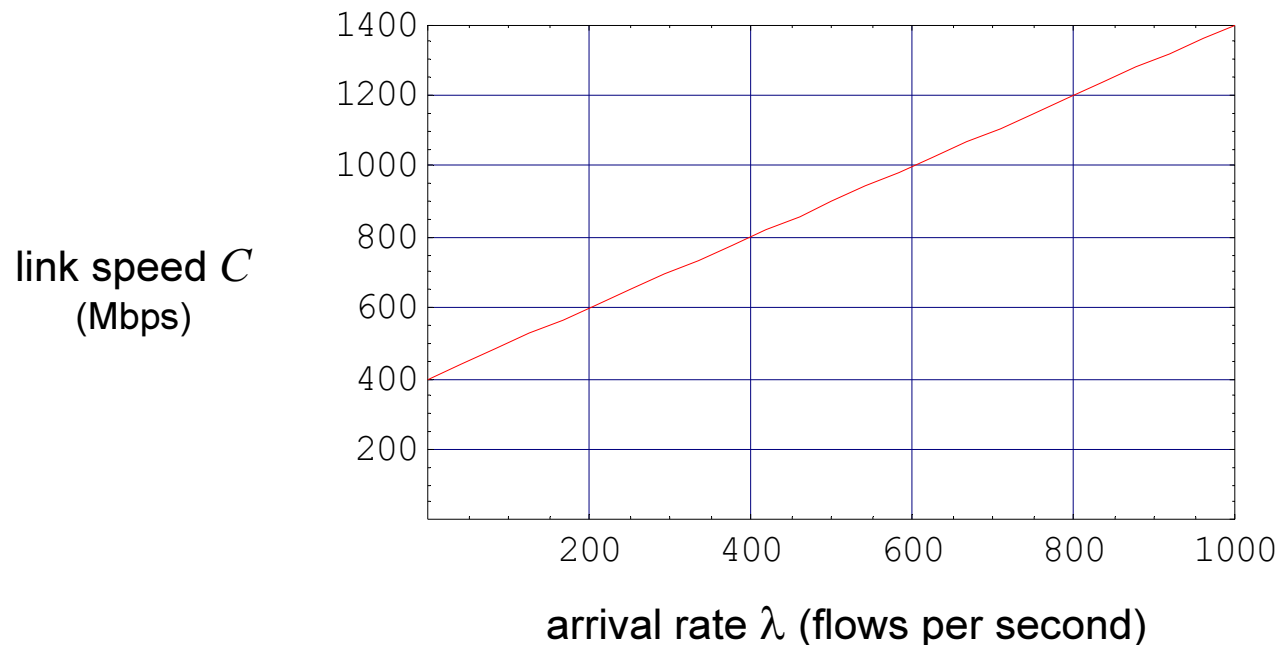
- Throughput is

$$\theta = \text{Xput}(1000, 600; 1) = 1000 - 600 = 400 \text{ Mbps} = 0.4 \text{ Gbps}$$

Capacity vs. arrival rate

- Given the quality of service requirement that $\theta \geq 400$ Mbps, the required link speed C depends on the arrival rate λ as follows:

$$C(\lambda) = \min\{c > \lambda S \mid \mathrm{Xput}(c, \lambda; 1) \geq 400\} = \lambda S + 400$$
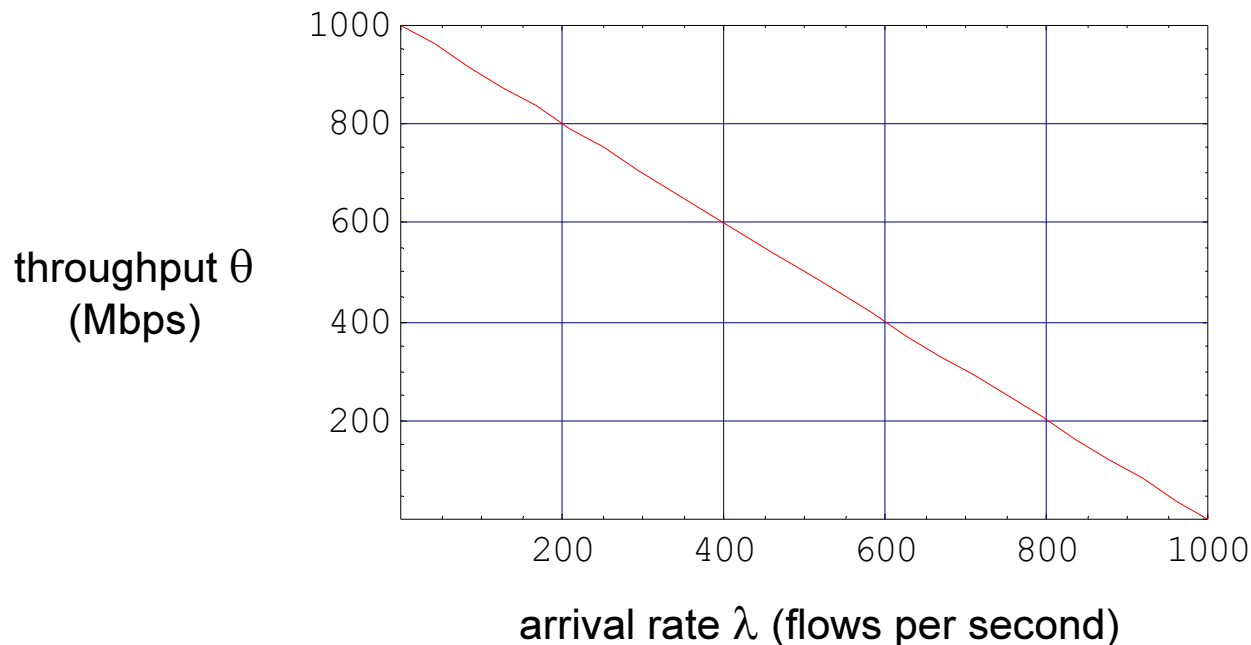
link speed C (Mbps)

arrival rate λ (flows per second)

39

Quality of service vs. arrival rate

- Given the link speed $C = 1000$ Mbps, the quality of service θ depends on the arrival rate λ as follows:

$$\theta(\lambda) = \mathrm{Xput}(1000, \lambda; 1) = 1000 - \lambda S, \quad \lambda < 1000/S$$

throughput θ
(Mbps)

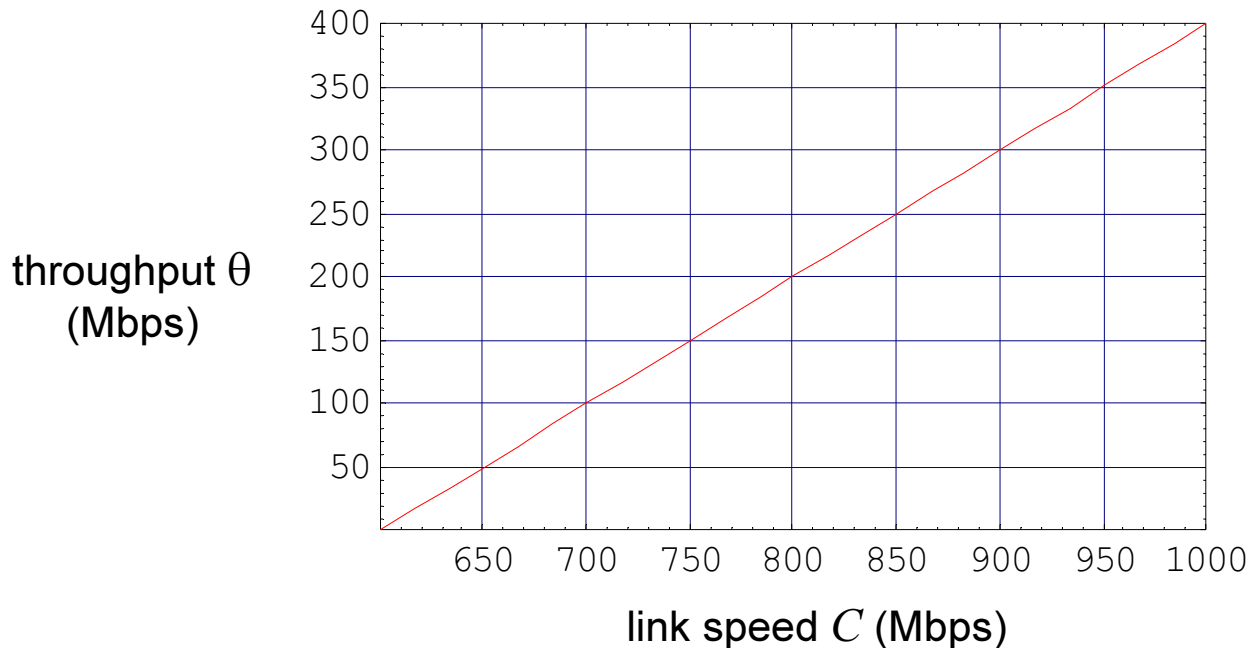arrival rate λ (flows per second)

40

3. Examples

Quality of service vs. capacity

- Given the arrival rate $\lambda = 600$ flows per second, the quality of service θ depends on the link speed C as follows:

$$\theta(C) = \mathrm{Xput}(C, 600; 1) = C - 600S, \quad C > 600S$$

throughput θ
(Mbps)

link speed C (Mbps)

41

Contents

- Model for telephone traffic
- Packet level model for data traffic
- Flow level model for elastic data traffic
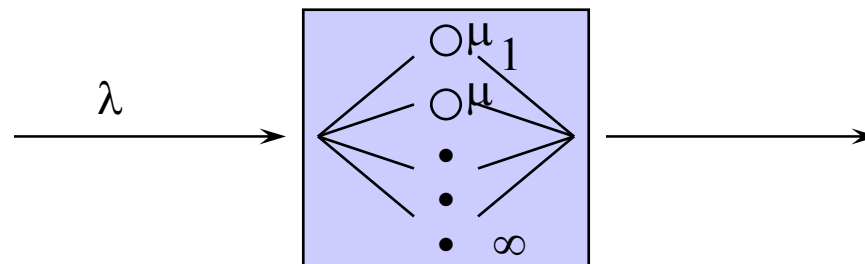- Flow level model for streaming data traffic

Flow level model for streaming CBR traffic (1)

- **Infinite system** is suitable for describing streaming CBR traffic at flow level
  - The transmission rate and flow duration of a streaming flow are insensitive to the network state
  - This kind of models applied in 90's to the teletraffic analysis of CBR traffic in ATM networks
- Consider a link between two packet routers
  - traffic consists of UDP flows carrying CBR traffic (like VoIP) and loading the link
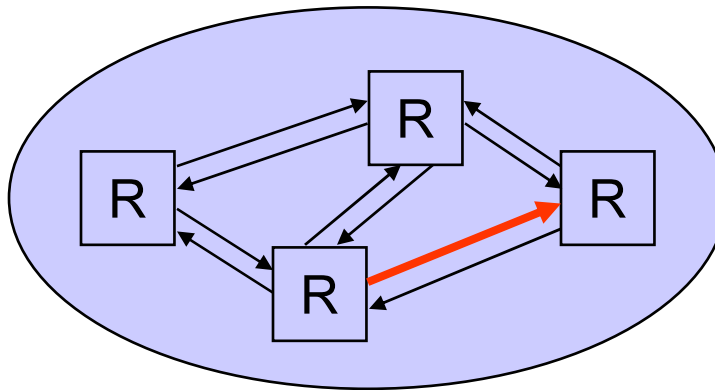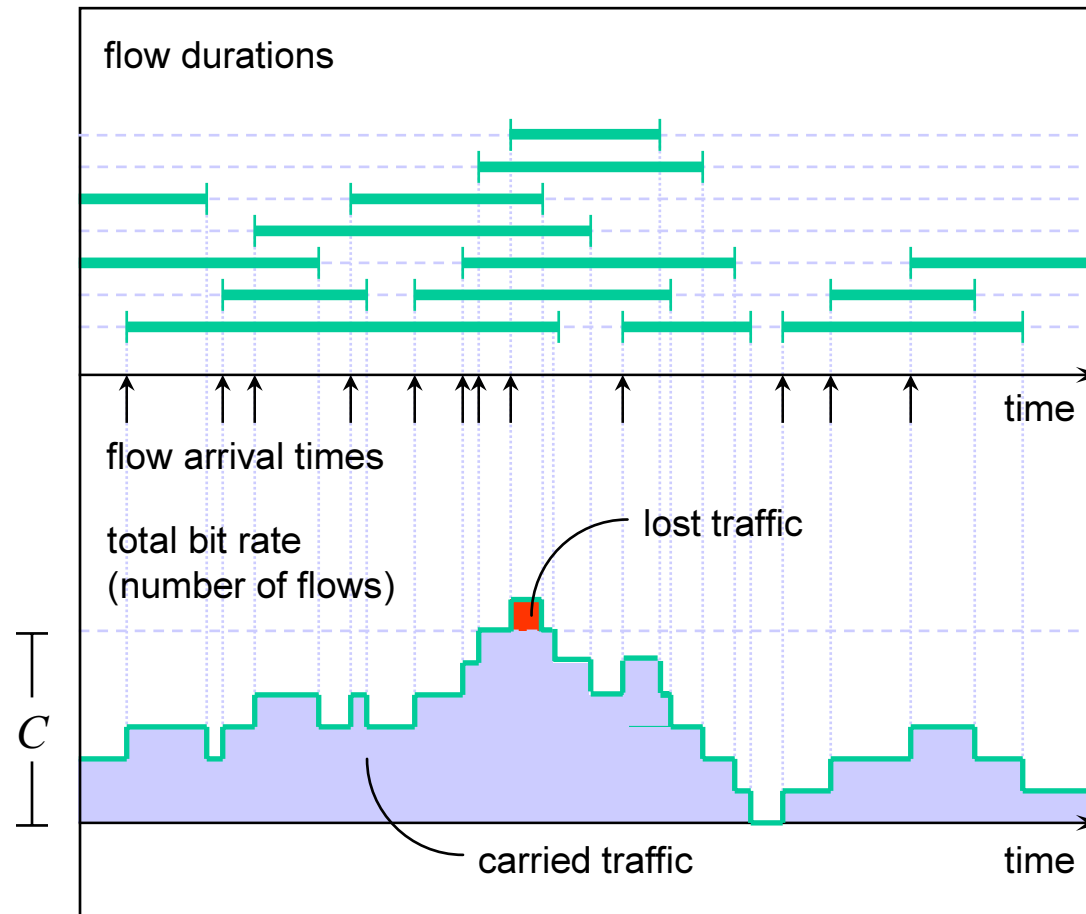
Flow level model for streaming CBR traffic (2)

- Model: an **infinite system** ($n = \infty$)
  - customer = UDP flow = CBR bit stream
    - λ = flow arrival rate (flows per time unit)
  - service time = flow duration
    - $h = 1/\mu$ = average flow duration (time units)
- **Bufferless** flow level model:
  - when the total transmission rate of the flows exceeds the link capacity, bits are lost (uniformly from all flows)

44

Traffic process

flow durations

flow arrival times

total bit rate
(number of flows)

C

lost traffic

carried traffic

time

time

Offered traffic

- Let r denote the bit rate of any flow
- The strength of offered traffic is described by average total bit rate R
  - By Little's formula, the average number of flows is

$$a = \lambda h$$

  - This may be called **traffic intensity** (cf. telephone traffic)
  - It follows that

$$R = ar = \lambda hr$$

Loss ratio

- Let N denote the number of flows in the system

- When the total transmission rate Nr exceeds the link capacity C, bits are lost with rate

$$Nr - C$$

- The average loss rate is thus

$$E[(Nr - C)^+] = E[\max\{Nr - C, 0\}]$$

- By definition, the **loss ratio** p_{loss} gives the ratio between the traffic lost and the traffic offered:

$$p_{\text{loss}} = \frac{E[(Nr - C)^+]}{E[Nr]} = \frac{1}{ar} E[(Nr - C)^+]$$

47

Teletraffic analysis (1)

- System capacity
  - $C = nr$ = link speed in kbps
- Traffic load
  - $R = ar$ = offered traffic in kbps
  - r = bit rate of a flow in kbps.
- Quality of service (from the users' point of view)
  - p_{loss} = loss ratio
- Assume an **M/G/∞ infinite system**:
  - flows arrive according to a **Poisson process** (with rate λ)
  - flow durations are independent and identically distributed according to **any distribution** with mean h

Teletraffic analysis (2)

- Then the quantitive relation between the three factors (system, traffic, and the quality of service) is given by the following formula

$$p_{\text{loss}} = \text{LR}(n,a) := \frac{1}{a} \sum_{i=n+1}^{\infty} (i-n)\frac{a^i}{i!} e^{-a}$$
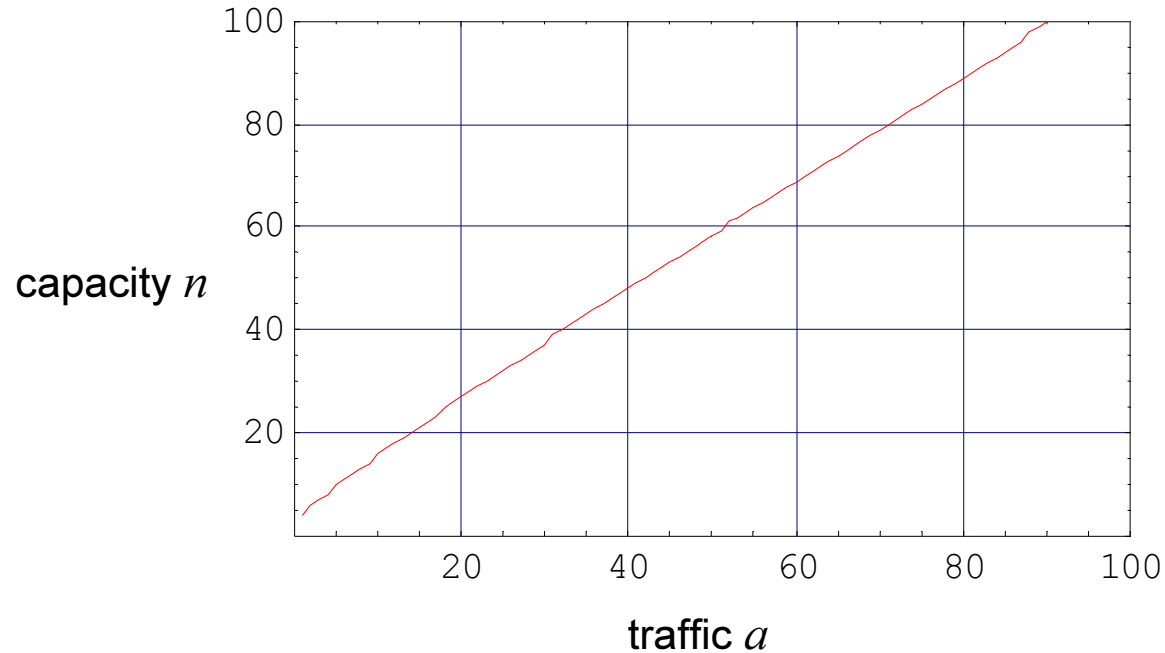
- Example:
  - $n = 20$
  - $a = 14.36$
  - $p_{\text{loss}} = 0.01$

Capacity vs. traffic

- Given the quality of service requirement that $p_{\text{loss}} < 1\%$, the required capacity n depends on the traffic intensity a as follows:

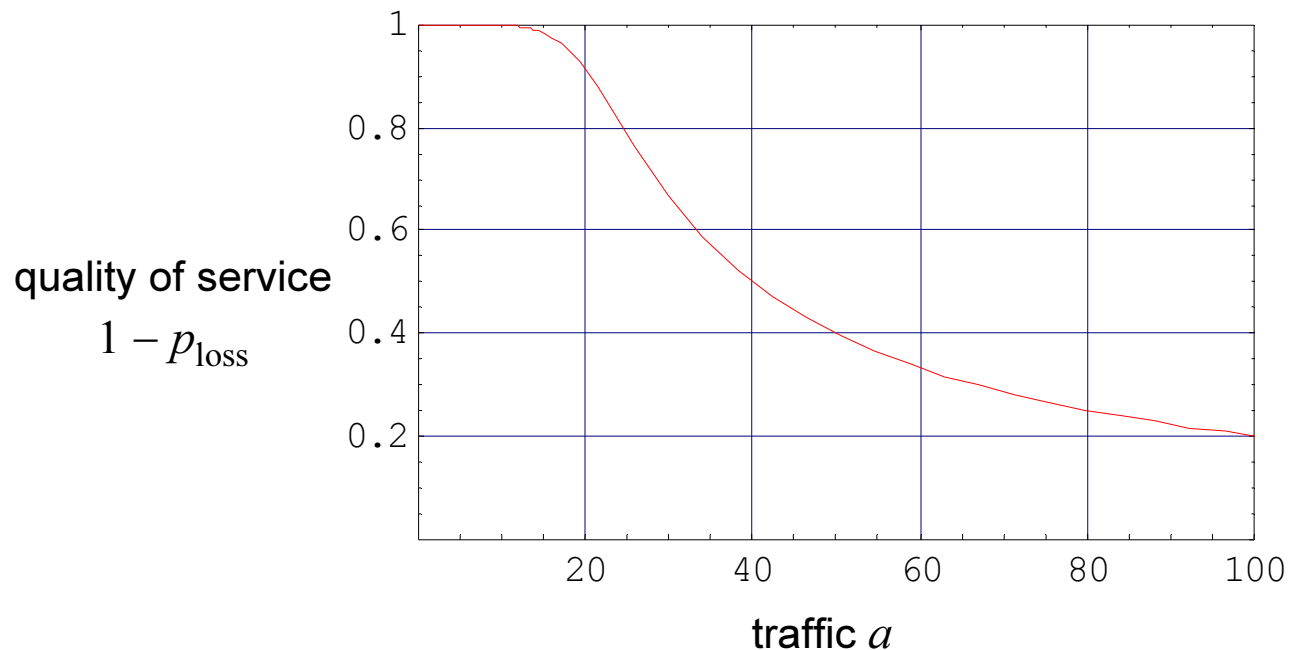$$n(a) = \min\{i = 1,2,\ldots \mid \text{LR}(i,a) < 0.01\}$$

capacity n

traffic a

50

Quality of service vs. traffic

- Given the capacity $n = 20$, the required quality of service $1 - p_{\text{loss}}$ depends on the traffic intensity a as follows:

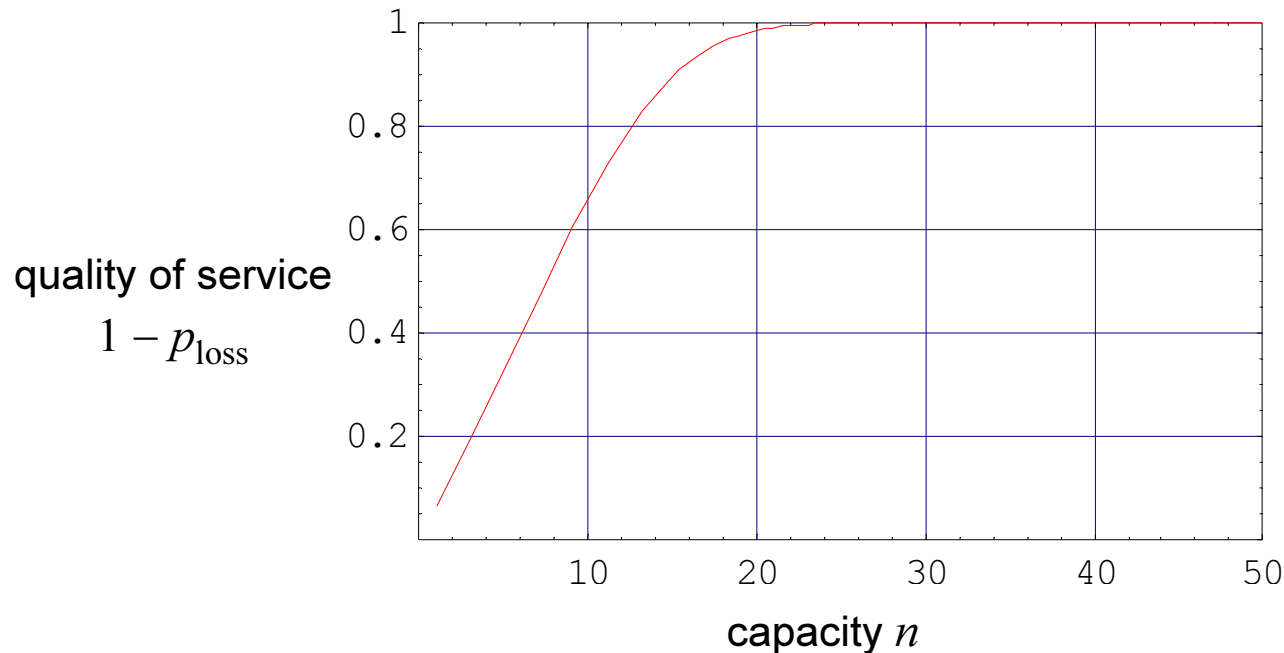$$1 - p_{\text{loss}}(a) = 1 - \text{LR}(20, a)$$

quality of service
$1 - p_{\text{loss}}$

traffic a

51

Quality of service vs. capacity

- Given the traffic intensity $a = 15.0$ erlang, the required quality of service $1 - p_{\text{loss}}$ depends on the capacity n as follows:

$$1 - p_{\text{loss}}(n) = 1 - \text{LR}(n, 15.0)$$

quality of service
$1 - p_{\text{loss}}$

capacity n

52

THE END