Resources with queue

07/08/08

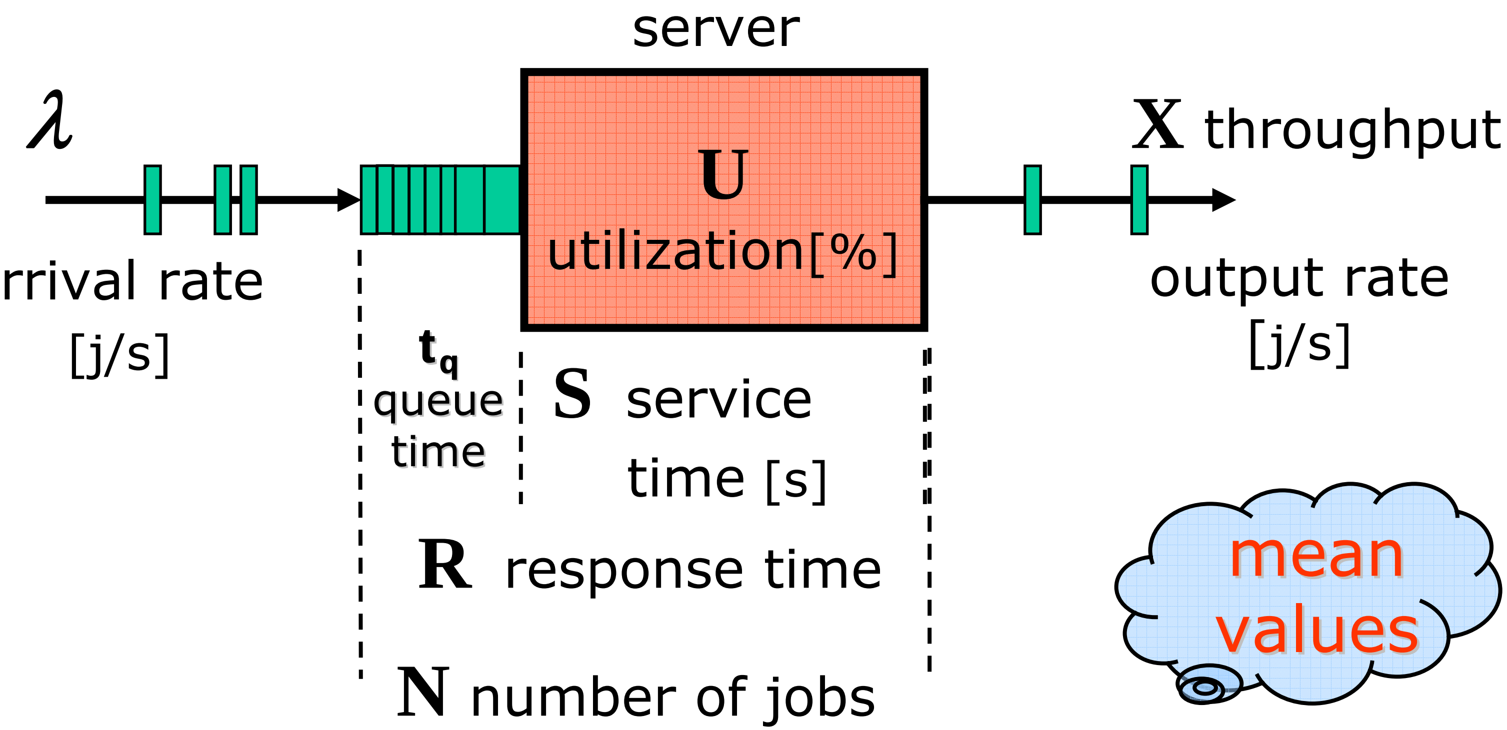
outline

- M/M/1, derivation of performance indices
- case study: router performance
- case study: packet loss
- case study: wireless channel
- M/M/2, M/M/n
- case study: load balancing schemes
- M/M/1/k
- case study: finite buffer capacity
- M/G/1
- case study: impact of variance on the performance

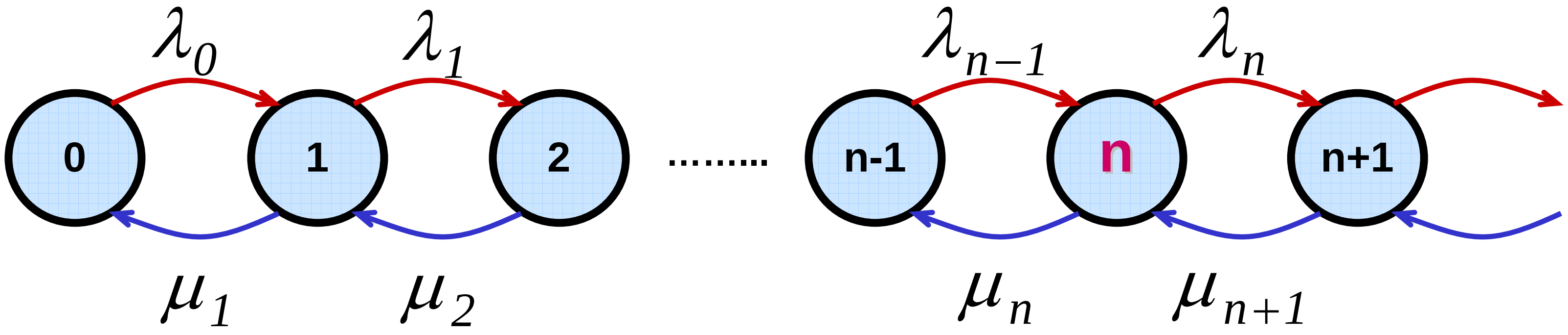
M/M/1

arr.Markov/serv.Markov/1server

resource with queue



M/M/1 birth-death model



balance equations $\quad \mu_1 p_1 = \lambda_0 p_0 \qquad n=0$

$$\underbrace{\lambda_{n-1} p_{n-1} + \mu_{n+1} p_{n+1}}_{\text{in n}} = \underbrace{\lambda_n p_n + \mu_n p_n}_{\text{out from n}} \qquad \forall n \geq 1$$

$$\lambda_0 p_0 = \mu_1 p_1 \qquad p_1 = \frac{\lambda_0}{\mu_1} p_0$$

$$\sum_{i=0}^{\infty} p_i = 1$$

$$p_2 = \frac{\lambda_1}{\mu_2} p_1 = \frac{\lambda_0 \lambda_1}{\mu_1 \mu_2} p_0$$

$$p_n = \frac{\lambda_0 \lambda_1 \ldots . \lambda_{n-1}}{\mu_1 \mu_2 \ldots . \mu_n} p_0$$

limiting probabilities

$$\sum_{i=0}^{\infty} p_i = 1 \qquad \lambda_0 = \lambda_1 = \ldots \lambda_n = \lambda \qquad \mu_0 = \mu_1 = \ldots \mu_n = \mu$$

$$p_0 = \frac{1}{1 + \sum_{k=1}^{\infty} \left(\lambda / \mu \right)^k} \qquad \sum_{k=0}^{\infty} \left(\frac{\lambda}{\mu} \right)^k \; geom.\, series \rightarrow \frac{1}{1 - \lambda / \mu} \quad if \; \left| \frac{\lambda}{\mu} \right| < 1$$

Utilization = λ / μ < 1

$$p_0 = 1 - \frac{\lambda}{\mu} = 1 - U \qquad p_n = \left(1 - \frac{\lambda}{\mu} \right) \left(\frac{\lambda}{\mu} \right)^n = (1 - U)\, U^n$$

scheduling algorithms independence

- the queue length distribution p_n is independent of the scheduling algorithm (FIFO, LIFO, PS); the balance equations are the same
- as a consequence, also the mean response time R (mean time spent in system) is independent of the scheduling algorithm (remember the Little law N=XR)
- the distribution of R depends on the scheduling algorithm
-

tail probability

probability that there are at least *n* jobs in the system (queue + service)

$$P[N \geq n] = \sum_{k=n}^{\infty} p_k = \sum_{k=n}^{\infty} (1-U)\,U^k = (1-U)\sum_{k=n}^{\infty} U^k =$$

$$= (1-U)\,U^n \frac{1}{1-U} = U^n$$

N: number of customers in the system

$$N = \sum_{n=0}^{\infty} n p_n = \sum_{n=0}^{\infty} n(1-U)U^n = (1-U) \sum_{n=0}^{\infty} nU^n$$

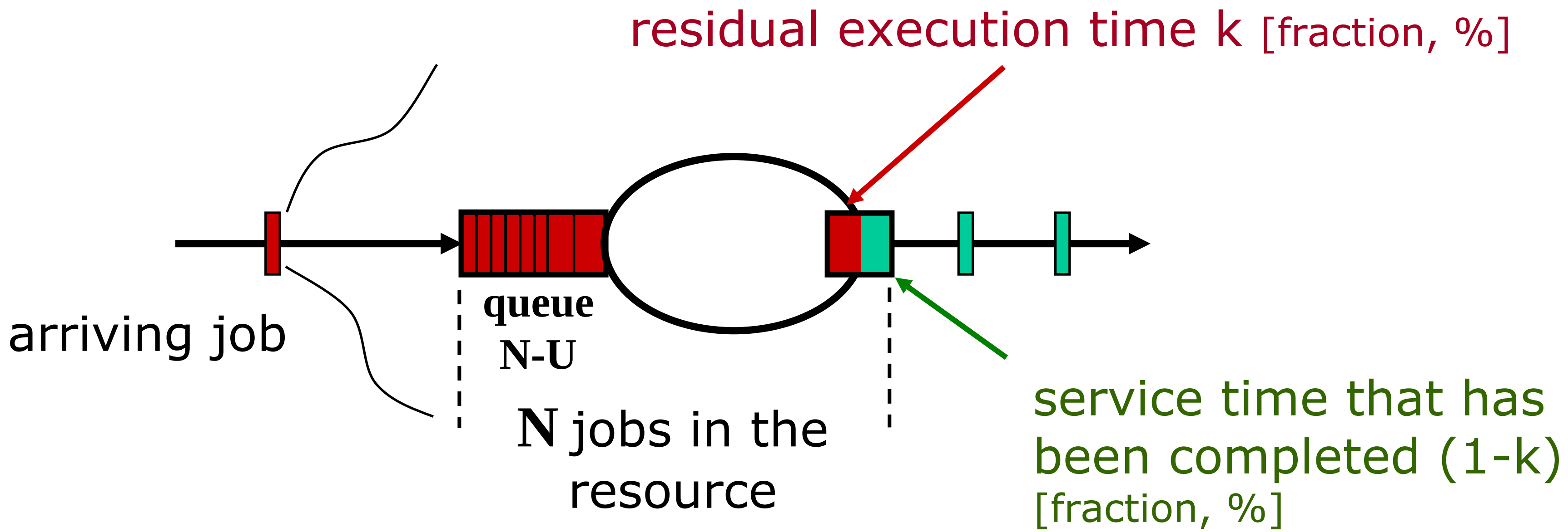
$$= (1-U)\left(\sum_{n=0}^{\infty} (n+1)U^n - \sum_{n=0}^{\infty} U^n \right)$$

$$\sum_{n=0}^{\infty} U^n = \frac{1}{1-U} \qquad \frac{d}{dU}\left(\sum_{n=0}^{\infty} U^n \right) = \frac{d}{dU}\left(\frac{1}{1-U} \right)$$

$$\sum_{n=0}^{\infty} nU^{n-1} = \frac{1}{(1-U)^2} \quad then$$

$$\mathbf{N} = (1-U)\left(\frac{1}{(1-U)^2} - \frac{1}{1-U} \right) = \frac{1}{1-U} - 1 = \frac{\mathbf{U}}{\mathbf{1-U}}$$

response time R



R= S (service time required by the job) **+**

(N-U) S (service time required by the jobs enqueued ahead) **+**

U (k S) (service time remaining)

= S + NS - U(1-k)S

response time R, mean time spent in the system

$$R_n = S + (n-1)S + S = (n+1)S \quad \textit{memoryless}$$

$$\mathbf{R} = \sum_{n=0}^{\infty}(n+1)S\, p_n = \sum_{n=0}^{\infty}(n+1)S(1-U)U^n = S(1-U)\sum_{n=0}^{\infty}(n+1)U^n$$

$$= S(1-U)\frac{1}{(1-U)^2} = \frac{\mathbf{S}}{\mathbf{1-U}}$$

R *has* exp *distribution*

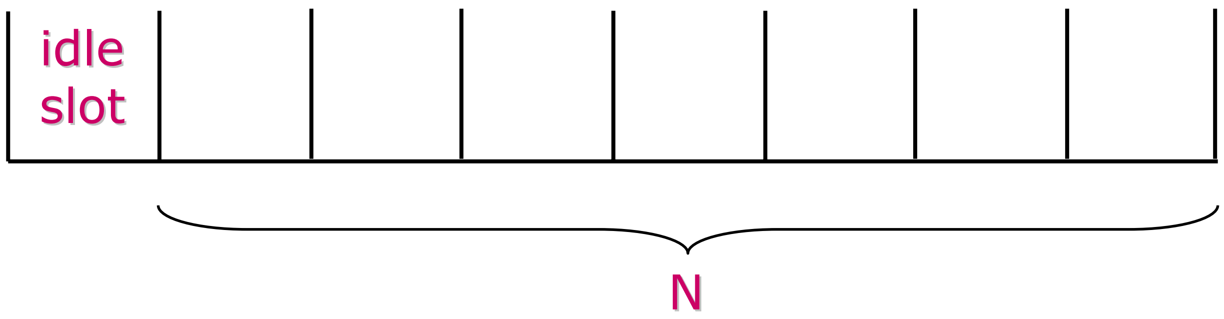
$$f_R(t) = (\mu-\lambda)e^{-(\mu-\lambda)t} = \left[\mu(1-U)\right]e^{-\left[\mu(1-U)\right]t} = R^{-1}\, e^{-R^{-1}t}$$

response time R, intuitive derivation

$$N \ \ number\ of\ \ jobs\ in\ the\ system\ (in\ queue\ and\ in\ execution)$$

$$U = 1 - idle\ fraction = 1 - \frac{1}{N+1} = 1 - \frac{S}{(N+1)S} = 1 - \frac{S}{R}$$

$$UR = R - S \qquad S = R(1-U) \quad \Rightarrow \quad R = \frac{S}{1-U}$$



performance indices

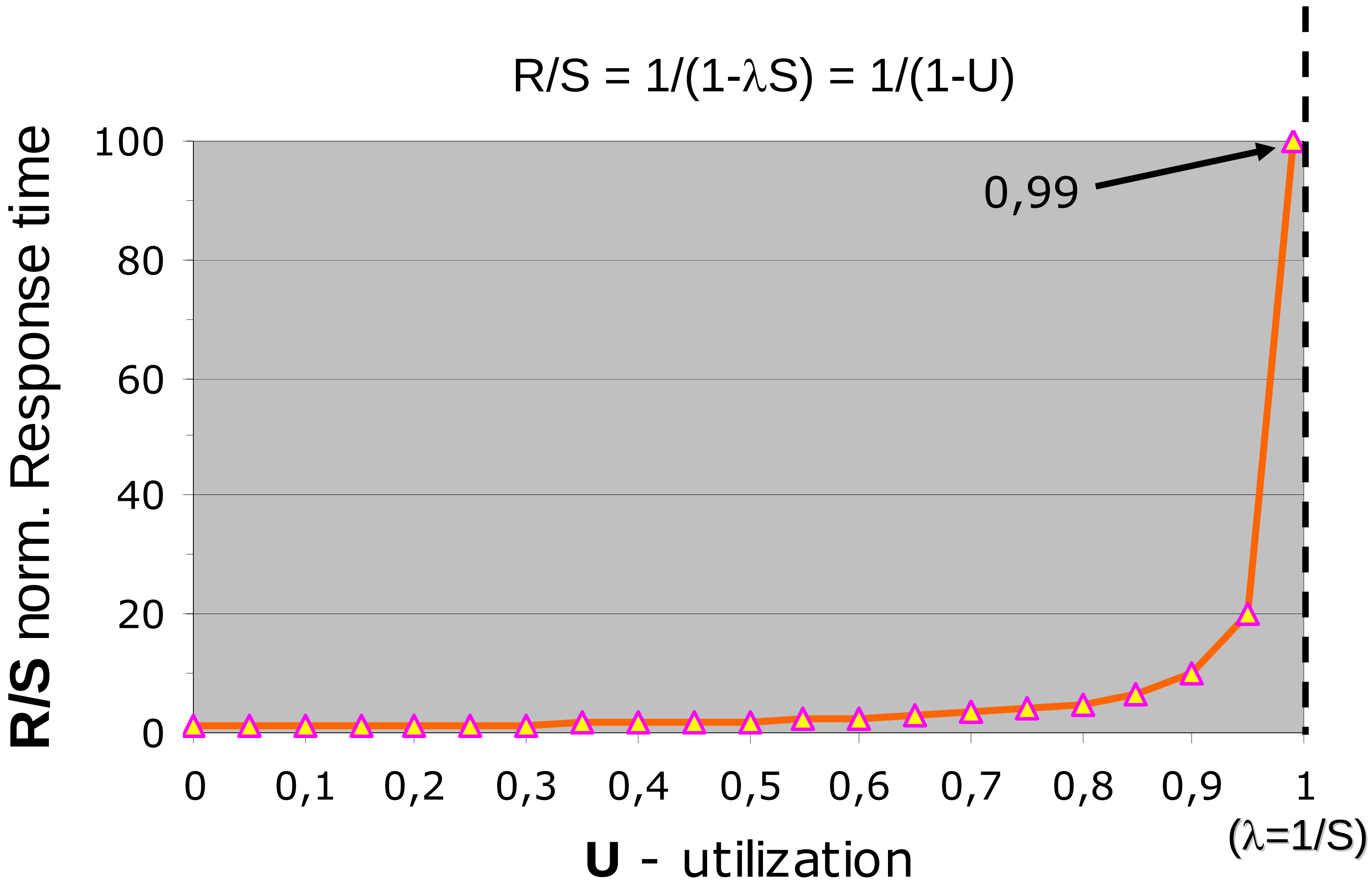
- **U** utilization [%]
- **S** average service time [s]
- λ arrival rate [j/s]
- **X** throughput (departure rate) [j/s]
- **R** response time [s]

- **N** number of jobs [j]
- **V** visits to a resource

$$U_i = \lambda_i S_i \qquad D_i = V_i S_i \qquad \boxed{R_i = \frac{S_i}{1 - U_i}} \qquad N_i = \frac{U_i}{1 - U_i}$$

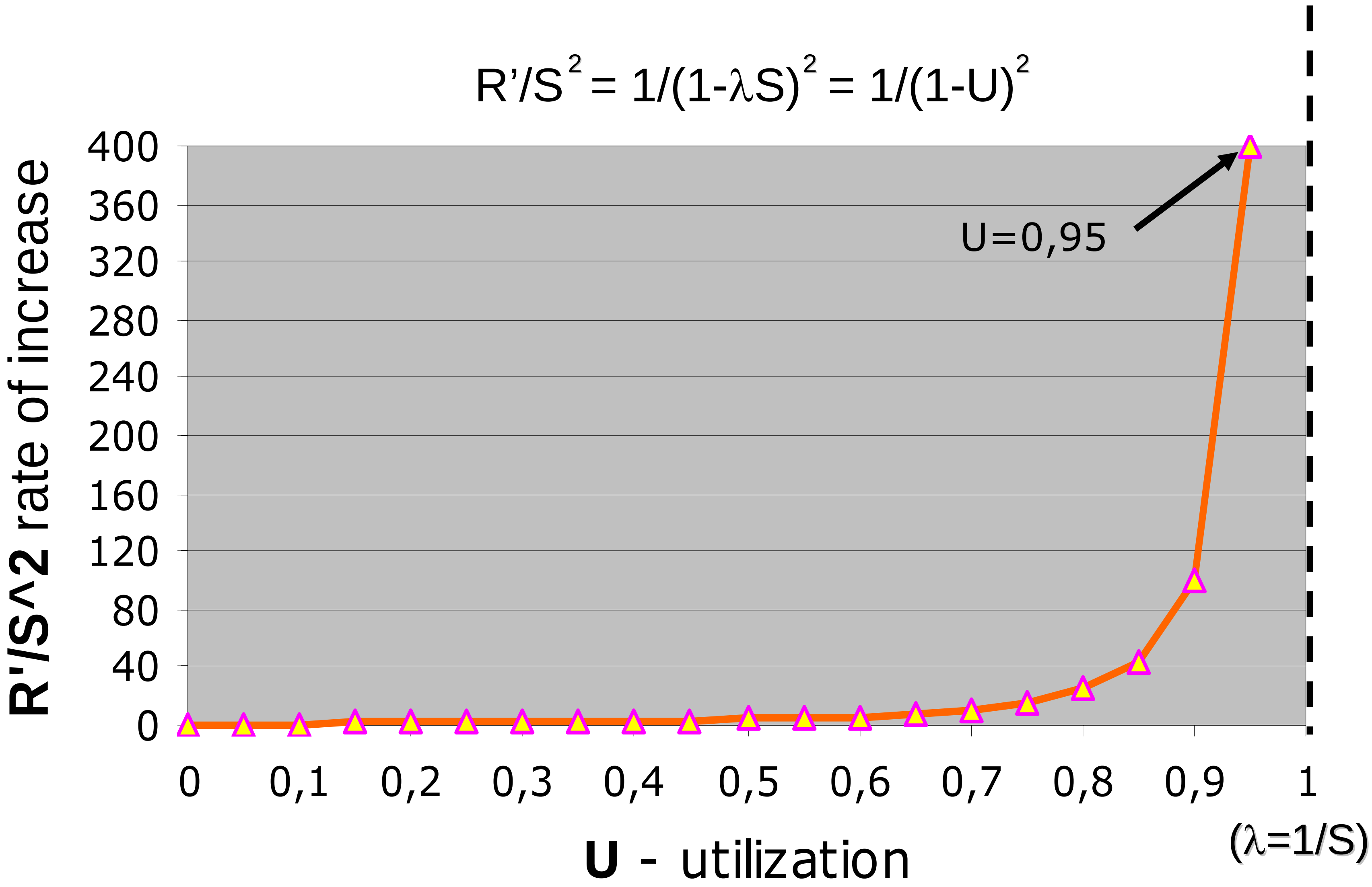
Little's result $N_i = X_i R_i \qquad X_i = X V_i$

R - normalized response time



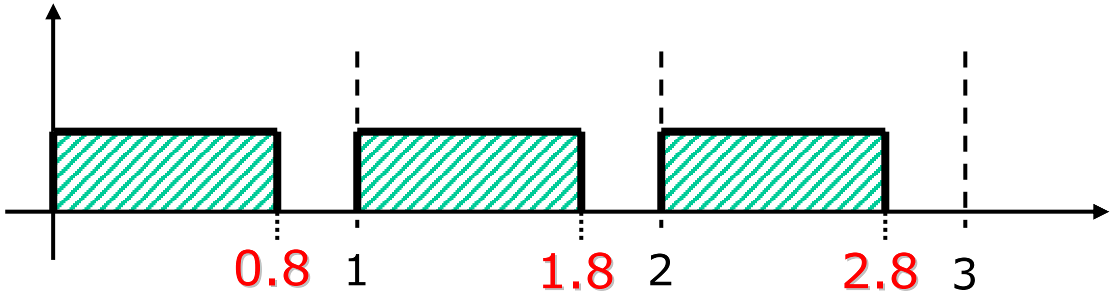
R – normalized rate of increase

$$R'/S^2 = 1/(1-\lambda S)^2 = 1/(1-U)^2$$



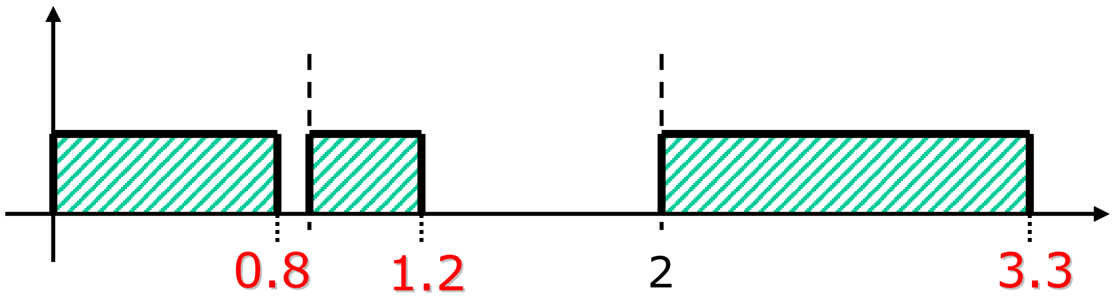
influence of distributions

deterministic arrival and service times



$$U = 0.8 \qquad N = 0.8$$

exponentially distributed arrival and service times



$$U = 0.8 \qquad N = \frac{U}{1-U} = \frac{0.8}{0.2} = 4$$

system power Φ

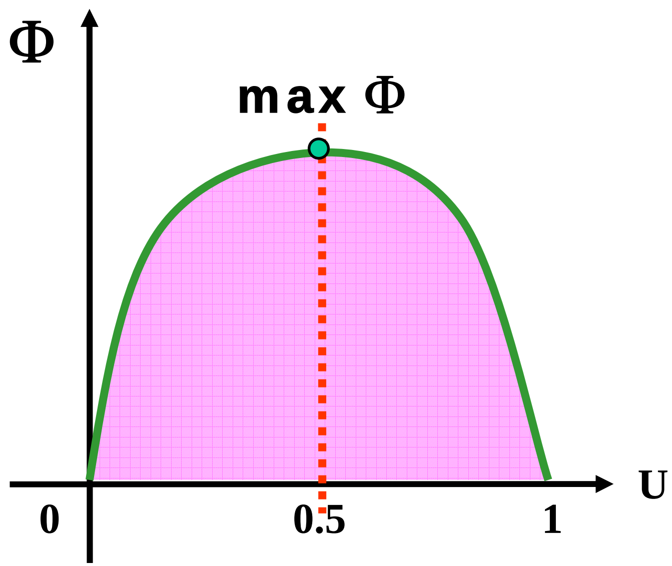
$$\Phi = \frac{\text{throughput}}{\text{response time}} = \frac{\lambda}{R} = \frac{\lambda(1-\lambda S)}{S}$$

$$\max\Phi \Rightarrow \Phi' = 0 = \frac{1}{S}[1-2\lambda S] \Rightarrow \lambda^{opt} = \frac{1}{2S}$$

$$U^{opt} = \lambda^{opt} S = 1/2$$

$$N^{opt} = \frac{U^{opt}}{1-U^{opt}} = 1$$

$1/2$ queue $\quad 1/2$ in service

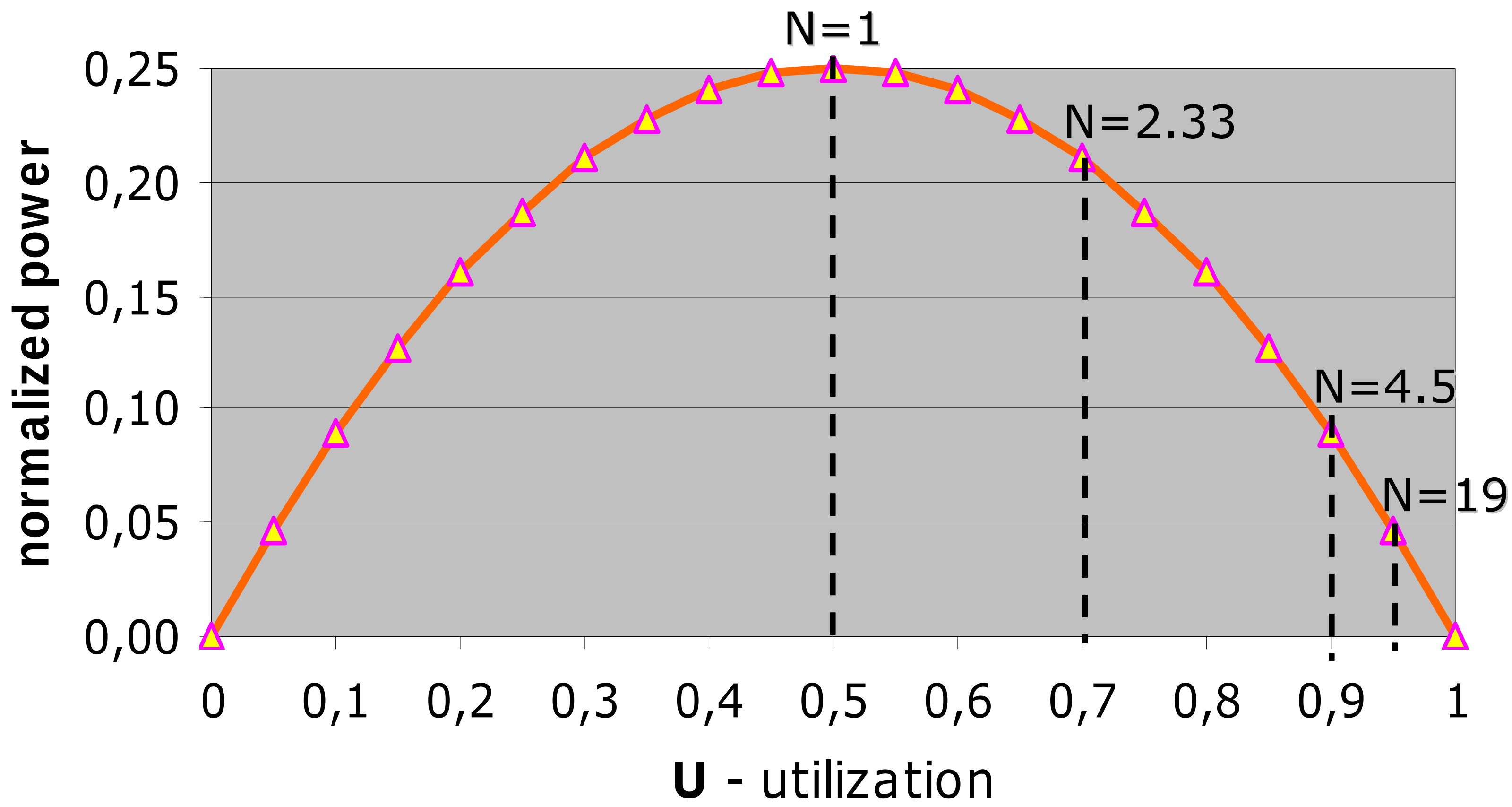


ϕ power

$$\Phi = \frac{\text{throughput}}{\text{response time}} = \frac{\lambda}{R} = \frac{\lambda(1-\lambda S)}{S}$$

norm.power = ϕ S^2 = U(1-U)

ϕ'=0 λ^{opt} = 1/ 2S U=0.5 N=1



case study: wireless communication channel (1)

traffic from several devices to be sent on a wireless communication channel

- total volume of traffic multiplexed on the channel 390,000 characters per minute (cpm)
- capacity of the wireless comm. channel: 57.6 Kbps

- average transmission time per character?

case study: wireless communication channel (2)

- channel transmit 8-bit characters, so the maximum capacity is:

  μ = transm.rate/charc.length =57600/8=7200 cps

- traffic arrival rate: λ=390,000/60= 6500 cps
- channel utilization $U = \lambda/\mu$ =6500/7200= 0.9027 (~90%)
- avg. no. N of characters waiting to be transmitted and in transmission:

  $$N = U / (1-U) = 9.277 \text{ char.}$$

- avg. no. N_q of char. in queue

  $$N_q = N-U = U^2 /(1-U) = 8.374 \text{ char.}$$

- avg. transmission time $R = N/\lambda = S/(1-U) = 1.427$ ms
- 90 percentile $\Pi(90) \cong 2.3$ R = 3.282 ms

case study: router (1)

variable length packets arriving from several links are time multiplexed over a single digital link

- arrival rate: λ=4 pkt/sec (240 pkt/min)
- link transmission rate: 6.4 Kb/sec, 0.8 KB/sec
- avg pkt length: 176 Bytes
- infinite buffer capacity

- performance indices ?
- probability that $N \geq 10$ packets are waiting for transmission?

case study: router (2)

- packet transmission time S:
  S = pkt.length/transm.rate =176/800=0.22 sec
- line utiliz. $U = \lambda S = 4 \times 0.22 = 0.88$ (88%)
- avg. no. N of packets in the router:
  $$N = U / (1-U) = 7.33 \text{ pkt}$$
- avg. no. N_q of pkt in queue
  $$N_q = U /(1-U) = 6.45 \text{ pkt}$$
- response time $R = S/(1-U) = 1.83$ sec
- 90 percentile $\Pi(90) \cong 2.3\, R = 4.209$ sec

case study: router (3)

- 10 or more packets are in queue if and only if 11 or more packets are in the router:

$$P[N \geq 11] = \sum_{k=11}^{\infty} p_k = \sum_{k=11}^{\infty} (1-U)\,U^k = U^{11} = 0.245$$

case study: packet loss in the router buffers (1)

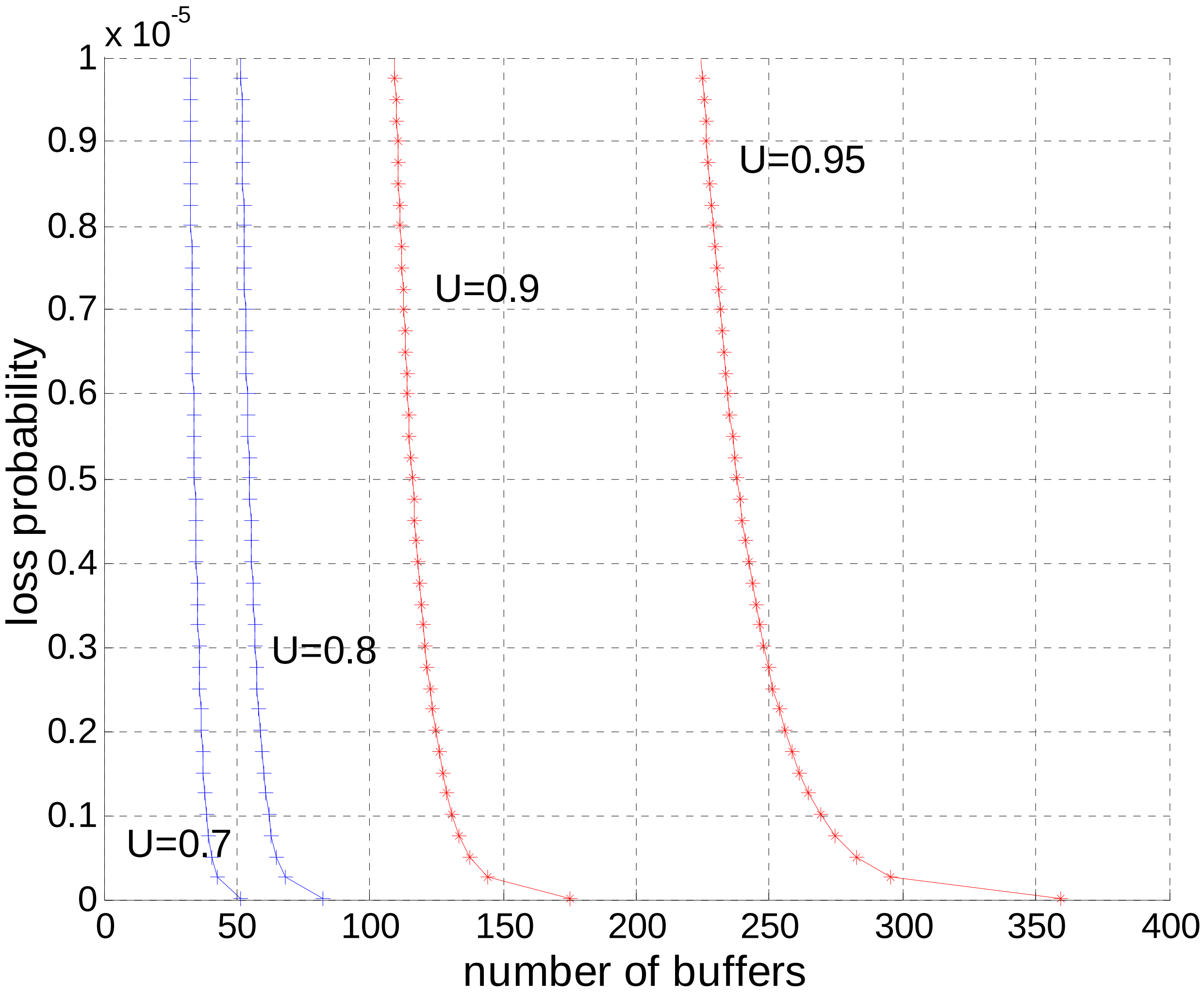
- p_n: prob. that the queue length is > n, it represents the prob. of packet loss with a buffer of n dimensions
- packet loss should be < ε

$$p_n[N \geq n] = U^n \qquad\qquad U^n \leq \varepsilon$$

$$n \log U \leq \log \varepsilon \qquad\qquad n \geq \frac{\log \varepsilon}{\log U} = \frac{-\log \varepsilon}{-\log U} = \text{cost } \log \frac{1}{\varepsilon}$$

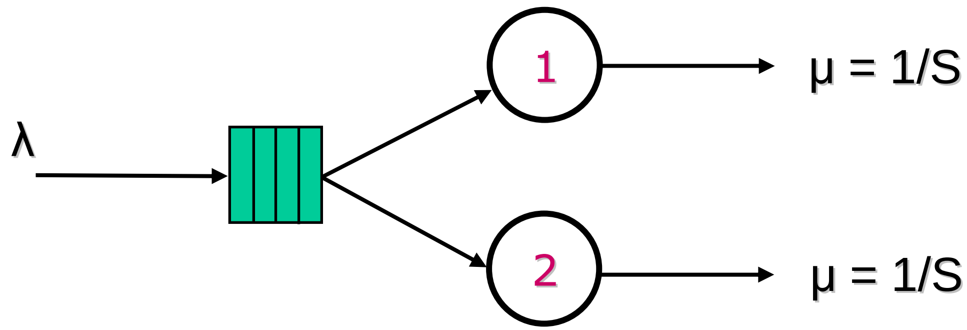
- to limit the packet loss (< ε) it is required that the buffer size n should be greater than the value given by a logarithmic function of 1/ε (usually $\varepsilon = 10^{-6}$, U=0.8)

case study: packet loss in the router buffers (2)



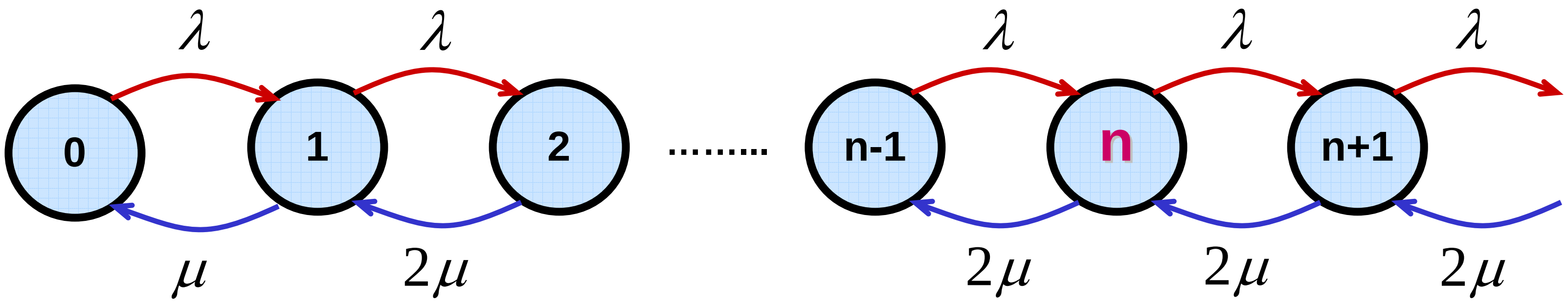
M/M/2
M/M/n

M/M/2 arr.Markov/serv.Markov/2 servers



- arrival rate λ
- 2 servers, each with a service rate μ, share a common queue
- a request is sent to the free server

M/M/2 birth-death model



$$\lambda p_0 = \mu p_1 \qquad p_1 = \frac{\lambda}{\mu} p_0$$

$$utilization\ of\ each\ server\ \ U = \frac{\lambda}{2\mu} < 1$$

$$\lambda p_1 = 2\mu p_2 \qquad p_2 = \frac{\lambda}{2\mu}\, p_1 = \frac{\lambda}{2\mu}\frac{\lambda}{\mu} p_0$$

where is the intelligence of the scheduler?

$$p_3 = \frac{\lambda}{2\mu}\, p_2 = \left(\frac{\lambda}{2\mu}\right)^2 \frac{\lambda}{\mu} p_0$$

$$\sum_{i=0}^{\infty} p_i = 1$$

…………..

$$p_n = \frac{\lambda}{2\mu}\, p_{n-1} = \left(\frac{\lambda}{2\mu}\right)^{n-1} \frac{\lambda}{\mu} p_0$$

$$p_0 + \sum_{i=1}^{\infty} \left(\frac{\lambda}{2\mu}\right)^{i-1} \frac{\lambda}{\mu} p_0 = 1$$

M/M/2

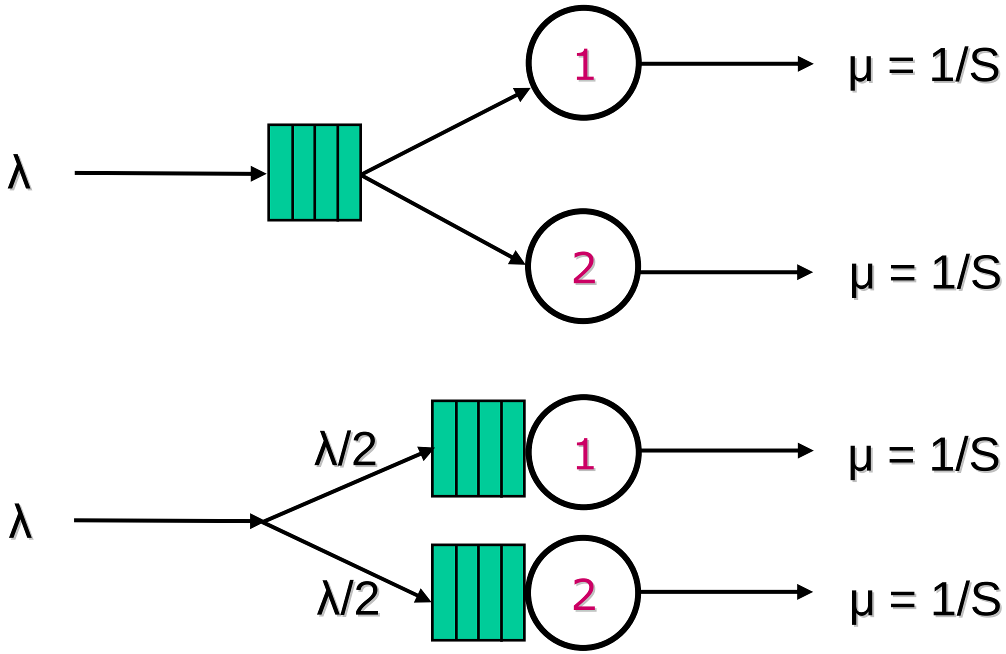
$$p_0 = \frac{1}{1+\frac{\lambda}{\mu}\sum_{i=0}^{\infty}\left(\frac{\lambda}{2\mu}\right)^i} = \frac{1}{1+\frac{\lambda}{\mu}\frac{1}{1-\frac{\lambda}{2\mu}}} = \frac{1}{1+\frac{\lambda\ 2\mu}{\mu(2\mu-\lambda)}} = \frac{2\mu-\lambda}{2\mu+\lambda} =$$

$$= \frac{1-\frac{\lambda}{2\mu}}{1+\frac{\lambda}{2\mu}} = \frac{1-U}{1+U}$$

$$N = \sum_{k\geq 0} k\ p_k = 2U + \frac{U(2U)^2}{2!}\frac{p_0}{(1-U)^2} = \frac{2U}{1-U^2}$$

$$R = \frac{N}{\lambda} = \frac{2U}{1-U^2}\frac{1}{\lambda} = \frac{S}{1-U^2}$$

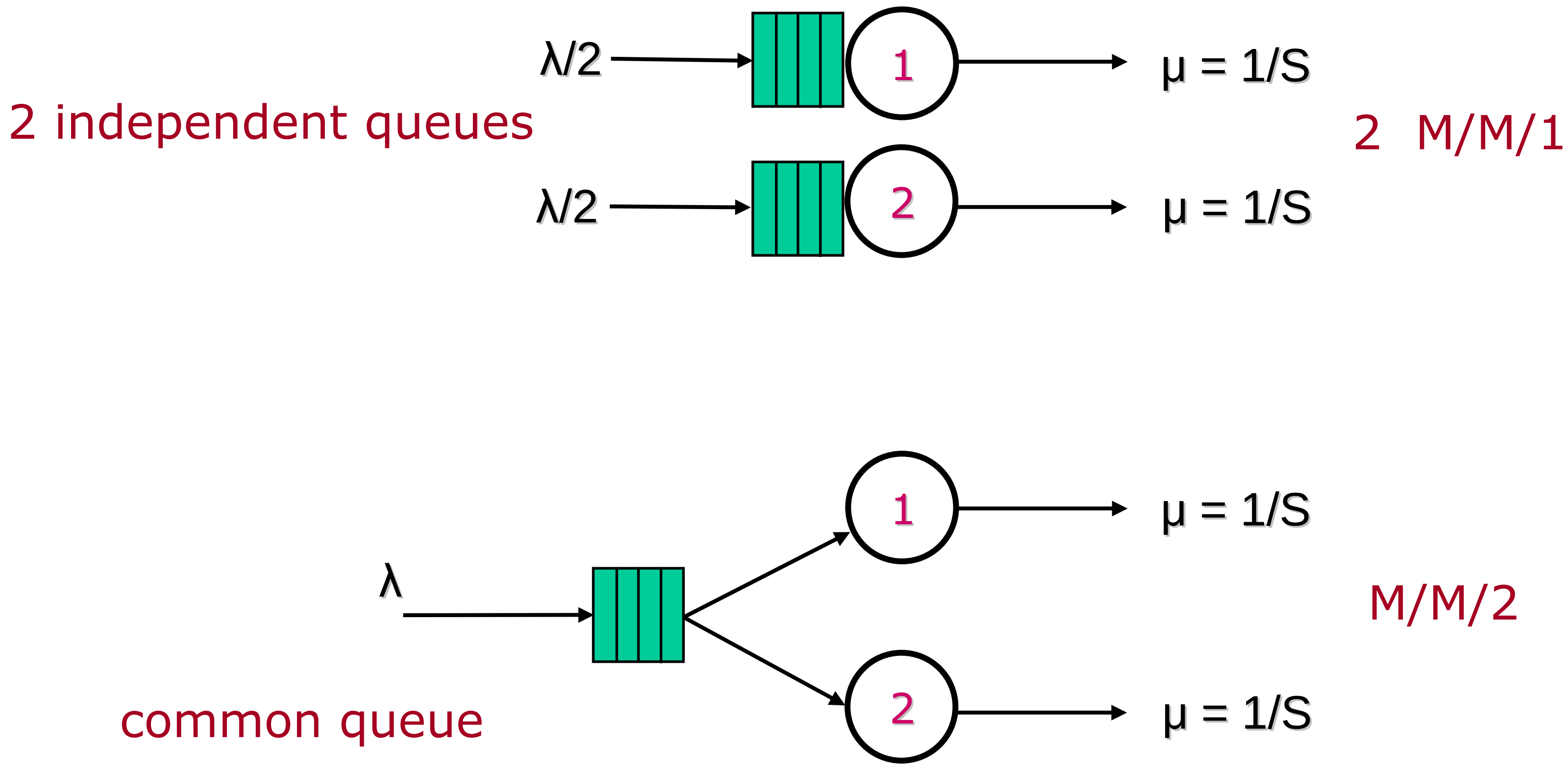
M/M/2 vs 2 M/M/1



$$R_{2\ \mathrm{M/M/1}} > R_{\mathrm{M/M/2}} \qquad \frac{S}{1-\frac{\lambda}{2}S} > \frac{S}{1-\left(\frac{\lambda}{2}S\right)^2}$$

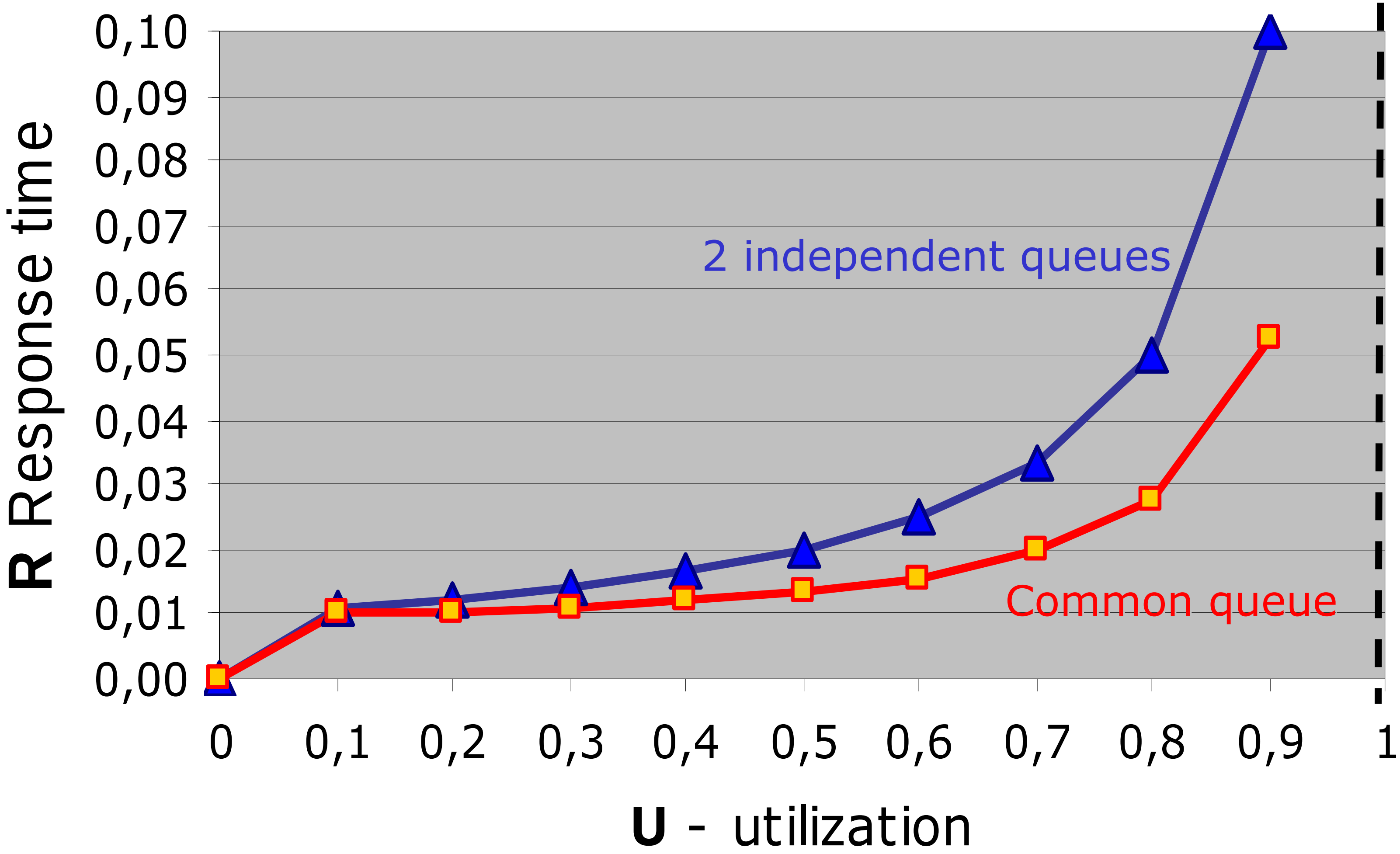
- common queue organization is better than separate queue organization

case study: load balancing schemes



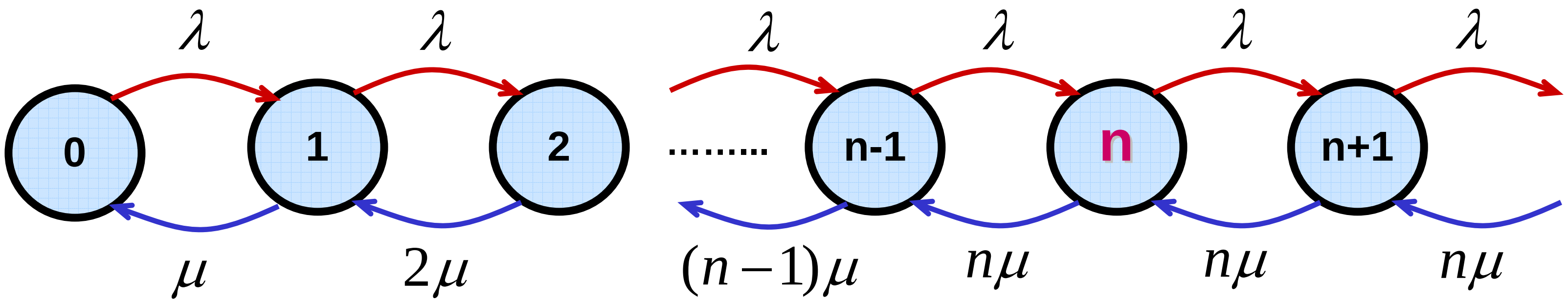
- same workload λ=0-200 req/s
- same processors capacity S=0.010 s

case study: load balancing



- U=0.99 each server (λ=99x2=198 req/s, λ/2 each server)
- $R_{M/M/2}$ = 0.5025s $R_{2\ M/M/1}$ = 1s (+100%)

M/M/n arr.Markov/serv.Markov/n servers



$$p_1 = \frac{\lambda}{\mu} p_0 \qquad U = \frac{\lambda}{n\mu} < 1$$

$$p_2 = \frac{\lambda}{2\mu} p_1 = \frac{\lambda}{2\mu}\frac{\lambda}{\mu} p_0$$

$$p_3 = \frac{\lambda}{3\mu} p_2 = \frac{\lambda}{2\mu}\frac{\lambda}{3\mu}\frac{\lambda}{\mu} p_0$$

....

$$p_{n-1} = \frac{\lambda}{n\mu} p_{n-1} = \left(\frac{\lambda}{\mu}\right)^{n-1} \frac{1}{(n-1)!} p_0$$

$$p_n = \frac{\lambda}{n\mu} p_{n-1} = \left(\frac{\lambda}{\mu}\right)^{n} \frac{1}{n!} p_0$$

$$p_{n+1} = \frac{\lambda}{n\mu} p_{n-1} = \left(\frac{\lambda}{\mu}\right)^{n+1} \frac{1}{n!n} p_0$$

....

$$p_{n+k} = \frac{\lambda}{n\mu} p_{n+k-1} = \left(\frac{\lambda}{\mu}\right)^{n+k} \frac{1}{n!n^k} p_0$$

M/M/n

$$\sum_{i=0}^{\infty} p_i = 1 \qquad \sum_{i=0}^{n} p_i + \sum_{i=n+1}^{\infty} p_i = 1$$

$$U = \frac{\lambda}{n\mu} < 1$$

$$\sum_{i=0}^{n} \left(\frac{\lambda}{\mu}\right)^i \frac{1}{i!} p_0 + \sum_{i=n+1}^{\infty} \left(\frac{\lambda}{\mu}\right)^i \frac{1}{n!n^{i-n}} p_0 = 1$$

$$p_0 = \frac{1}{\sum_{i=0}^{n} \left(\frac{\lambda}{\mu}\right)^i \frac{1}{i!} + \sum_{i=1}^{\infty} \left(\frac{\lambda}{\mu}\right)^{n+i} \frac{1}{n!n^i}}$$

$$= \frac{1}{\sum_{i=0}^{n} \left(\frac{\lambda}{\mu}\right)^i \frac{1}{i!} + \left(\frac{\lambda}{\mu}\right)^n \frac{1}{n!} \sum_{i=1}^{\infty} \left(\frac{\lambda}{n\mu}\right)^i}$$

$$= \frac{1}{\sum_{i=0}^{n} \left(\frac{\lambda}{\mu}\right)^i \frac{1}{i!} + \left(\frac{\lambda}{\mu}\right)^n \frac{1}{n!} \frac{\lambda}{n\mu} \frac{1}{1-\frac{\lambda}{n\mu}}}$$

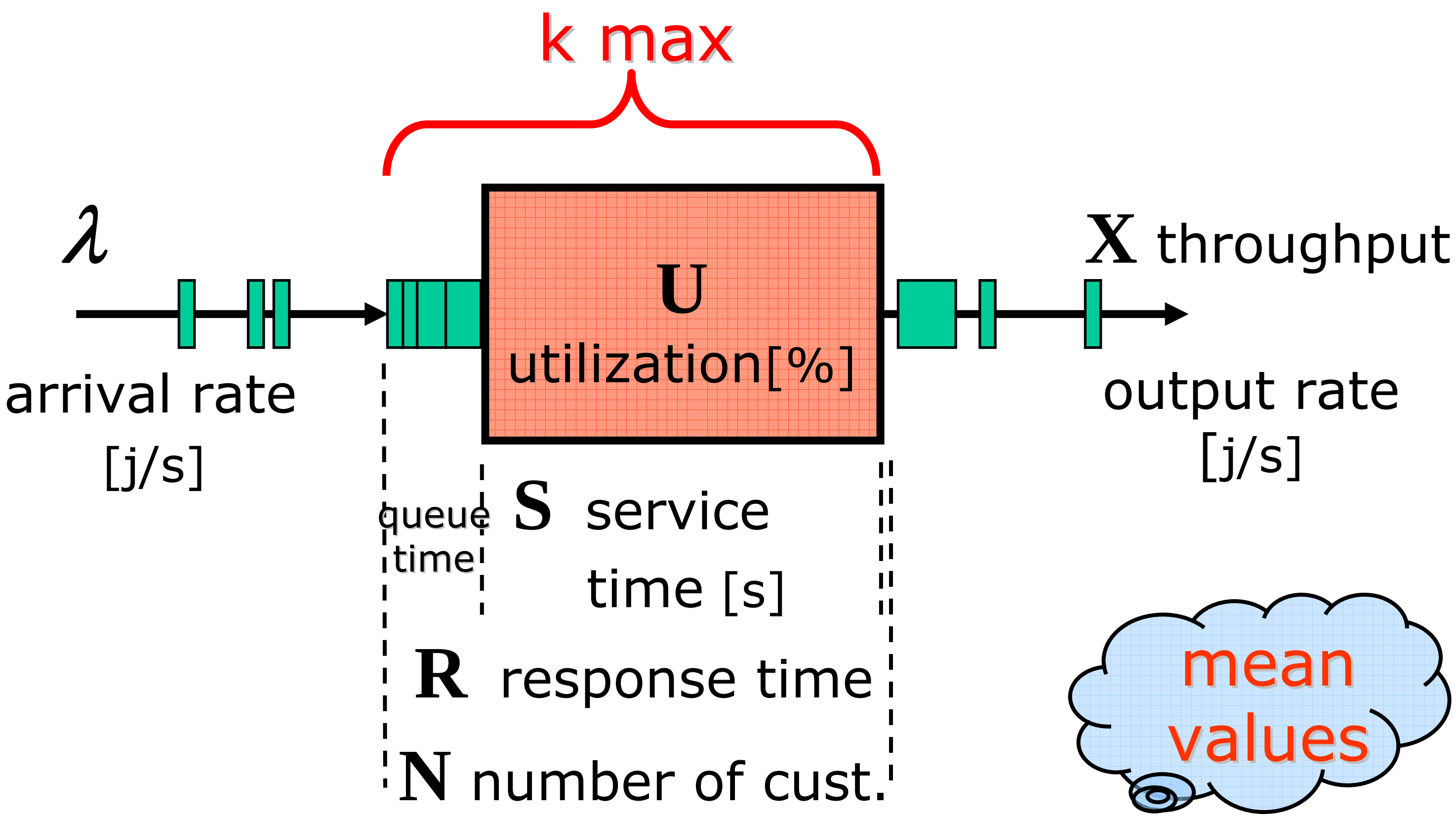
$$= \frac{1}{\sum_{i=0}^{n} \left(\frac{\lambda}{\mu}\right)^i \frac{1}{i!} + \left(\frac{\lambda}{\mu}\right)^n \frac{1}{n!} \frac{\lambda}{n\mu - \lambda}}$$

$$= \frac{1}{\sum_{i=0}^{n-1} \frac{(nU)^i}{i!} + \frac{(nU)^n}{n!} \frac{1}{1-U}}$$

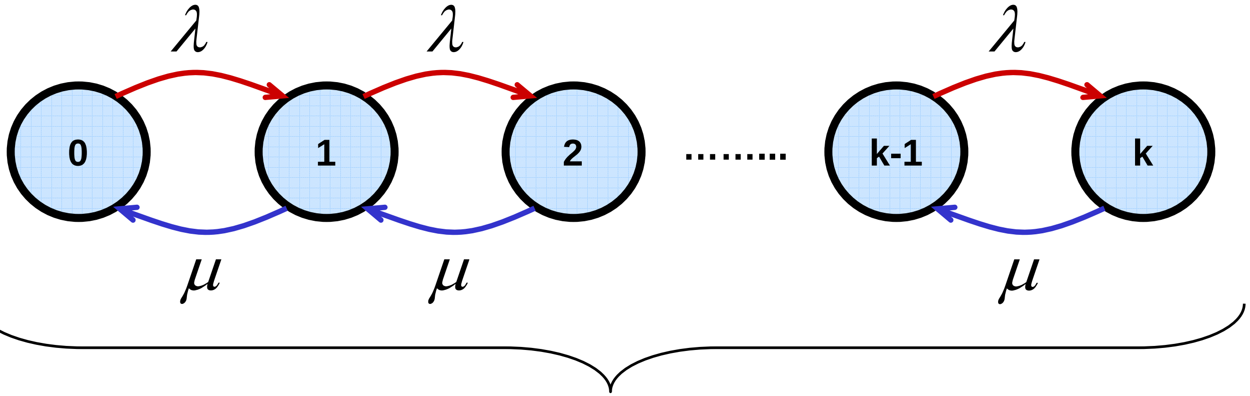
M/M/1/k

finite capacity queue

- the number of buffer positions is limited to k
- requests arriving when N=k are lost



M/M/1/k Markov/Markov/1 server/k users



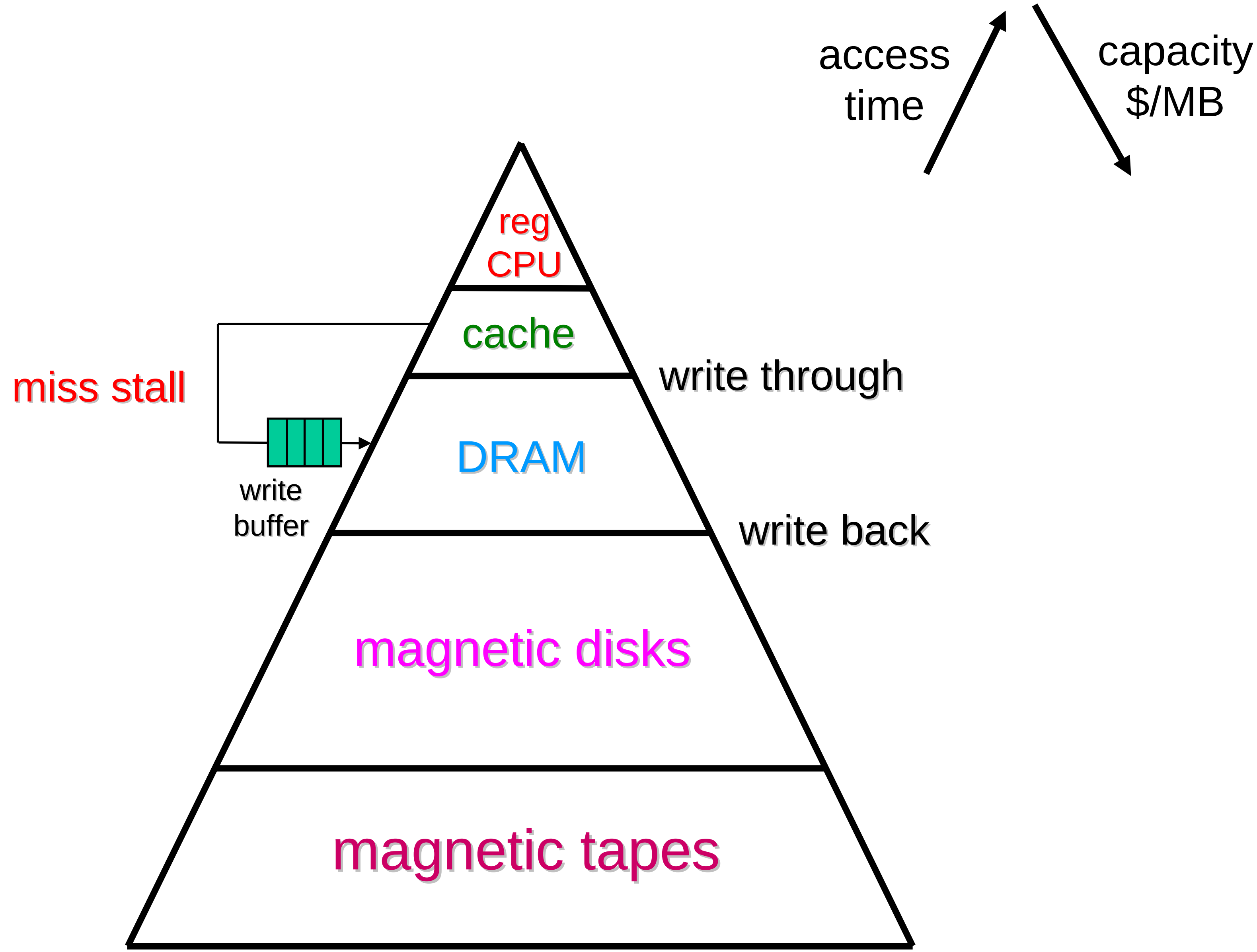
K max number of customers in the system

$$p_0 = \frac{1-U}{1-U^{k+1}} \qquad p_n = \frac{1-U}{1-U^{k+1}}\, U^n$$

$$N = \frac{U}{1-U} - \frac{(k+1)\,U^{k+1}}{1-U^{k+1}}$$

p_k loss probability

memory hierarchy



case study: finite buffer capacity (1)

consider the same system (router and workload) of case study 1 but with a limited buffer positions

- identify the number of buffer positions k such that the probability that all are full is $<$ 0.5% (0.005)
- compute the performance indices

case study: finite buffer capacity (2)

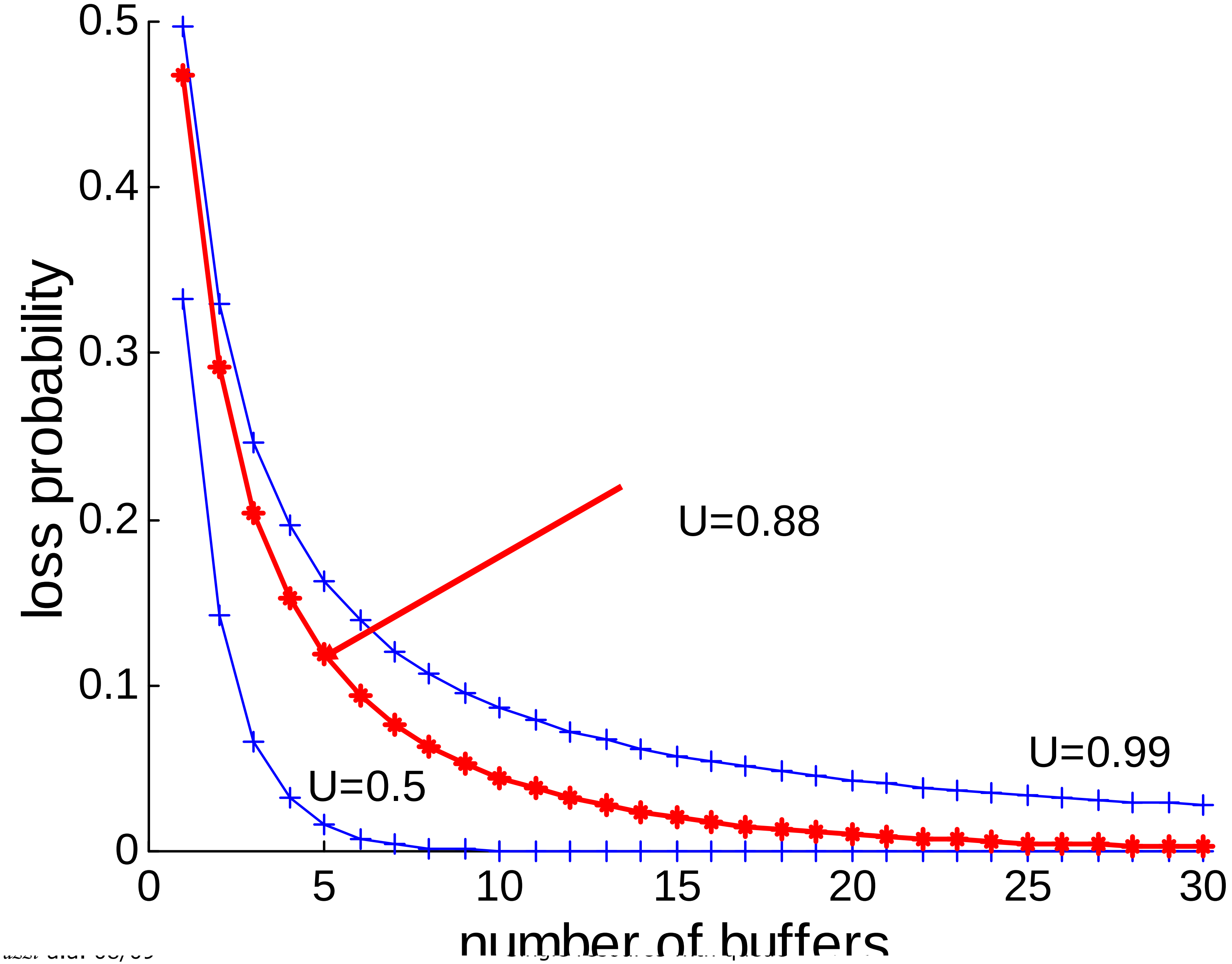
- line utiliz. $U = \lambda\, S = 4 \times 0.22 = 0.88$ (88%)
- p_k prob. that all k buffers are full

  $P[N \geq 20] = 0.009989 \approx 1\%$ (19 buffers full)

  $P[N \geq 25] = 0.005095$ (24 buffers full)

  $P[N \geq 26] = 0.004464 < 0.5\%$ (25 buffers full)
- $N = 6.449$ pkt [7.33 ∞]
- $R = N/[\lambda(1-p_{26})] = 1.62$ sec [1.83 ∞]
- $X = \lambda(1-p_{26}) = 238.88$ pkt/min [240 ∞]
- $p_{26} = 0.0044$ (0.446% of pkts is lost!)

case study: prob. of packet loss (3)



M/G/1

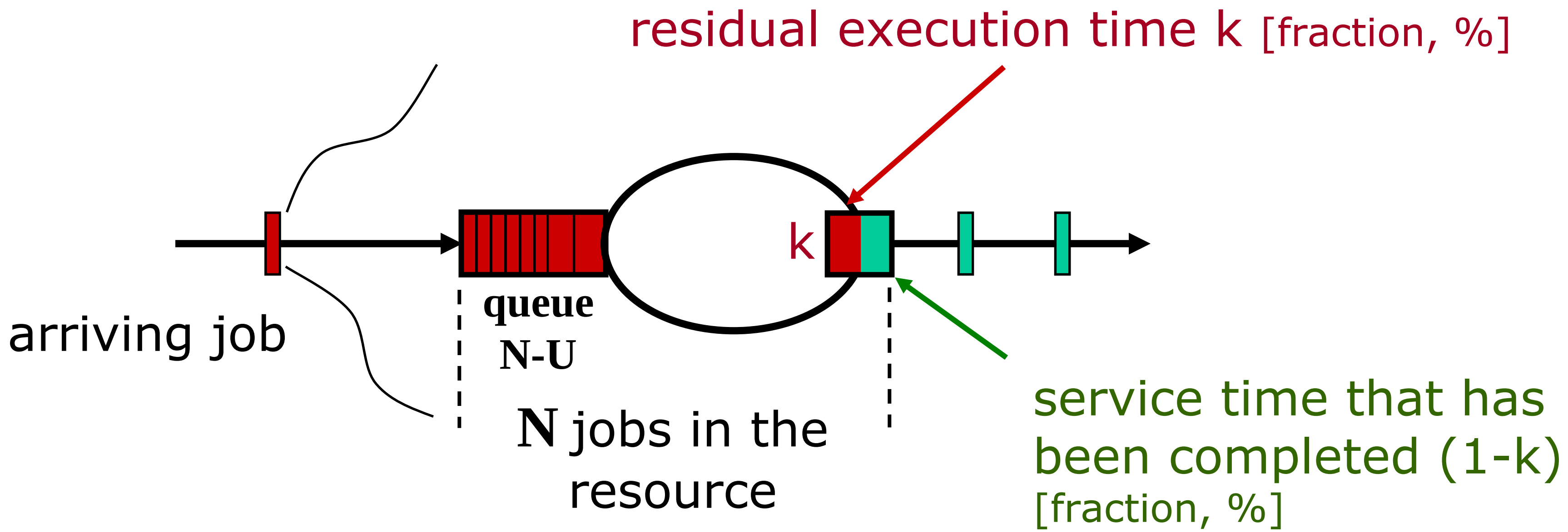
M/G/1 arr.Markov/serv.General/1server

- single server queueing system
- arrival process: Poisson, avg. arrival rate λ
- general service distribution, S avg., (the first two moments E[S] and E[S^2] exist and are finite)
- scheduling discipline: FCFS
- the average number of jobs N in the system depends only on the first two moments of the service time distribution

Pollaczek-Khinchin mean value formula

$$E[N] = N = U + \frac{U^2(1+c^2)}{2(1-U)}$$

response time R



R = S (service time required by the job) **+**

(N-U) S (service time required by the jobs enqueued ahead) **+**

U (k S) (service time remaining)

= S + NS - U(1-k)S

response time R

$$R = S + SN - S(1-k)U \quad by\ \ Little\ \ law$$

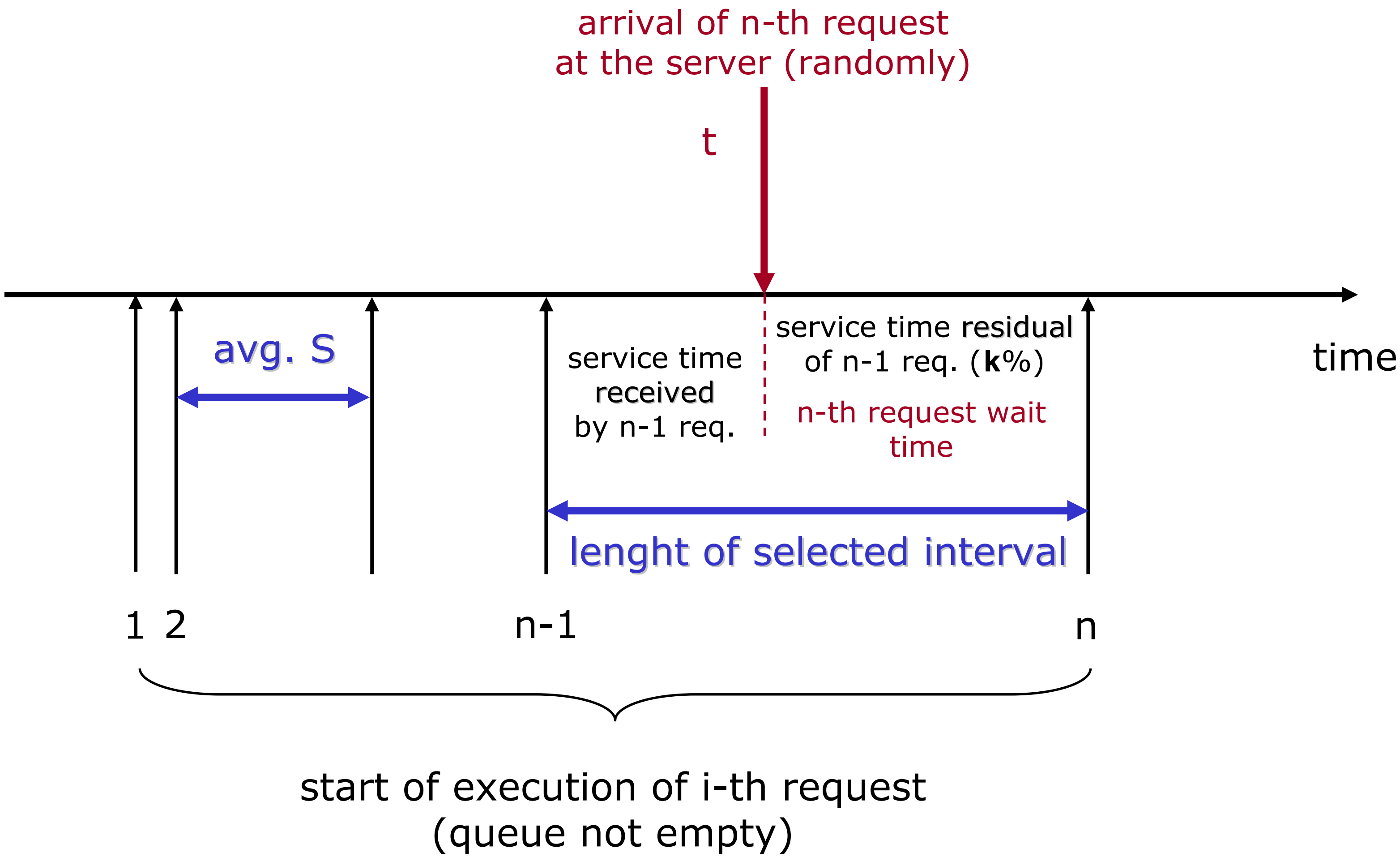
$$= S + S\,XR - S(1-k)U$$

$$= \frac{S}{1-U}\ \left[1-(1-k)U\right] \qquad k : fraction\ of\ residual\ service\ time$$

$$k = \frac{1}{2}(1+c^2)$$

$$R = S + \frac{US\,(1+c^2)}{2\,(1-U)} \qquad Pollaczek - Khinchine$$

residual service time



residual execution time

- the interval of time in which the new request of service arrive at a busy server is not the average service time
- a long interval is more likely to be selected at random than a short one
- [N(t),Rec(t)] state of the system
- N(t): number of jobs in the system, discrete, numerable
- Rec(t): amount of service time received, continuous

residual execution time

- S: avg service time (interval lenght)
- f(s): density function
- f(s)ds: probability that an interval of lenght included in [s, s+ds] is present in the observation period T
- f(s)ds T/S : avg. number of intervals with this lenght in the observation period T
- the probability that an interval of lenght s will be selected is proportional to s f(s) ds
- P[s<X≤s+ds]=k s f(s)ds $\qquad$ $f_S(s)$= k s f(s)ds

$$\int_0^\infty f_S(s)\,ds = \int_0^\infty k s f(s)\,ds \qquad 1 = k\int_0^\infty s\, f(s)\,ds \qquad k = \frac{1}{S}$$

$$\textit{density of the selected interval}\quad f_S(s) = \frac{s\, f(s)}{S}$$

residual execution time

- the average lenght of the selected interval is

$$S\,selected = \int_0^\infty s\, f_S(s)\, ds \;=\; \int_0^\infty s\, \frac{s\, f(s)}{S} ds \;= \frac{m_S^2}{m_S^1} = \frac{m_S^2}{S} = \frac{\sigma^2 + S^2}{S}$$

- the average residual time r is half of this interval lenght

$$r\, avg.residual\, time = \frac{1}{2} Selected = \frac{1}{2}\frac{m_S^2}{S} = \frac{\sigma^2 + S^2}{2S} = \frac{S}{2}(c^2 + 1)$$

K completion coefficient

$$k : completion\ \ coefficient$$

$$kS\ \ (residual\ exec.\ time) = \frac{m_2}{2S} = \frac{\sigma^2 + S^2}{2S} = \frac{S}{2}(1 + c^2)$$

$$k = \frac{1}{2}(1 + c^2) \qquad\qquad for\, c = 0\ \ no\ \ variance\ \ k = \frac{1}{2}$$

$$for\ \ c > 1\ \ k > 1\ \ hyper\exp \qquad\qquad for\ \ c < 1\ \ k < 1\ \ hypo\exp$$

for $c = 1 \quad \Rightarrow$ exp $\quad R = S\left[1 + \frac{2U}{2(1-U)}\right] = \frac{S}{1-U}$

N: number of jobs in the system

- the number of jobs in the system can be derived from R applying Little law N=λR

$$N = \lambda R = \lambda S \left[1 + \frac{U(1+c^2)}{2(1-U)} \right]$$

$$= U \left[1 + \frac{U(1+c^2)}{2(1-U)} \right]$$

$$for\ c = 1 \qquad N = \frac{U}{1-U}$$

case study: influence of variability

- consider a web site with average service time S=2.4 sec. utilized at 80%
- compute the average response time R for different service time distributions:
  a) Pareto
  b) Two-stage hyperexponential
  c) Exponential
  d) Three-stage Erlang (hypoexponential)
  e) Constant

case study: S=2.4s

	Service time distribution	Coeff. Variation	Response time (s)	Queueing time (s)
a	Pareto α=2.1	2.18	**30.05**	27.65
b	Two-stage hyperexponential	1.526	18.40	16.00
c	Exponential	1	12.00	9.60
d	Three-stage Erlang	0.577	8.80	6.40
e	Constant	0	**7.20**	4.80