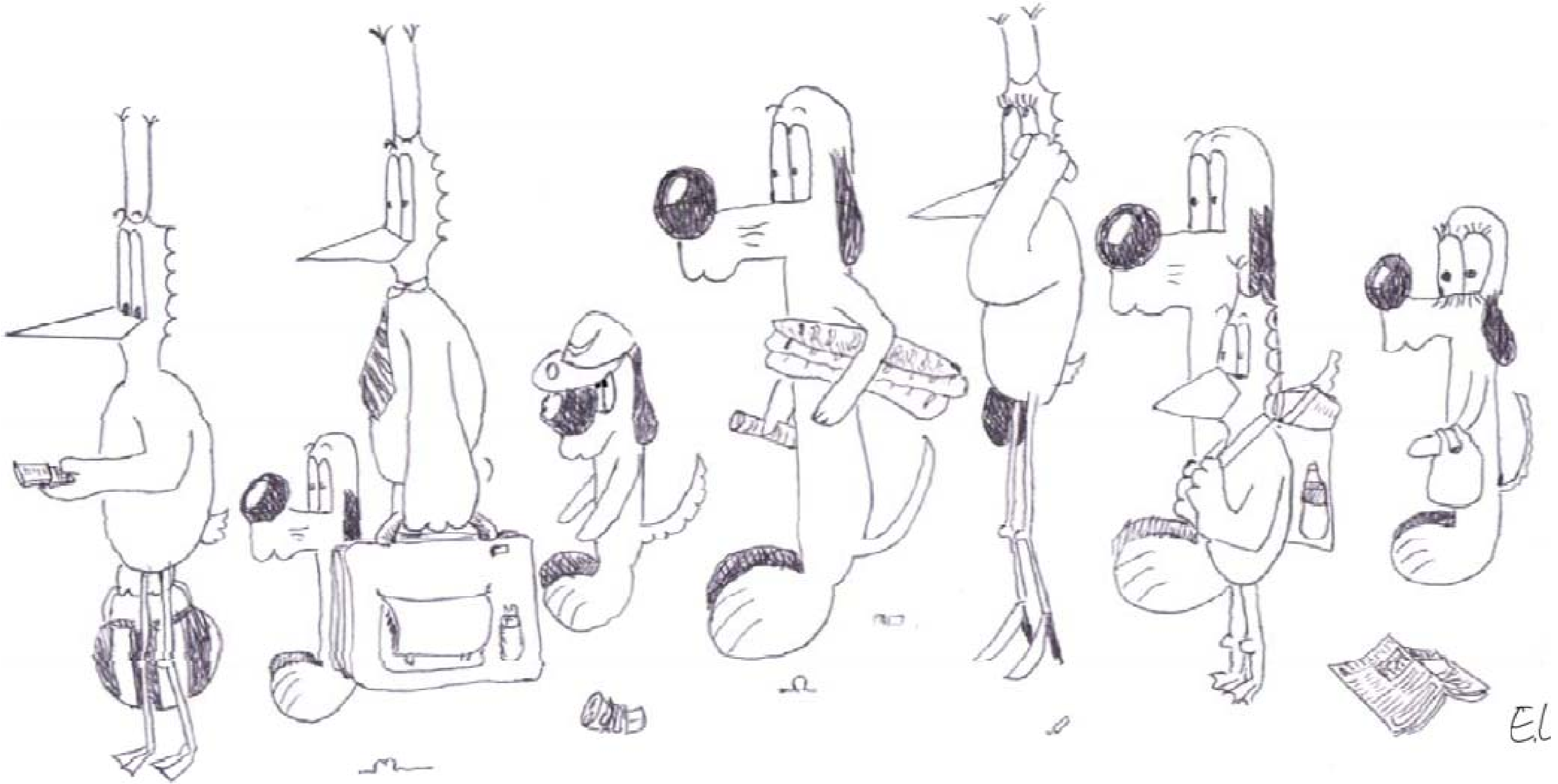
Queuing Theory
For Dummies



Jean-Yves Le Boudec

March 2019

All You Need to Know About Queuing Theory

Queuing is essential to understand the behaviour of complex computer and communication systems

In depth analysis of queuing systems is hard

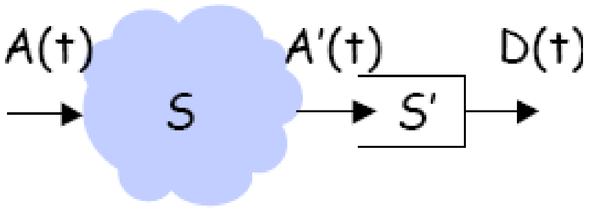
Fortunately, the most important results are easy

We will first study simple concepts

1. Deterministic Queuing

Easy but powerful

Applies to worst case and transient analysis



Example: playback buffer sizing

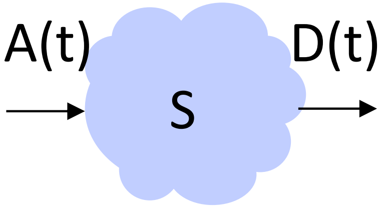
Source sends data at constant bit rate r

Network imposes a variable delay, received bit rate no longer constant

At destination, received data is stored in playback buffer, read at a constant rate r

Q: does it work ? How should the playback buffer be engineered?

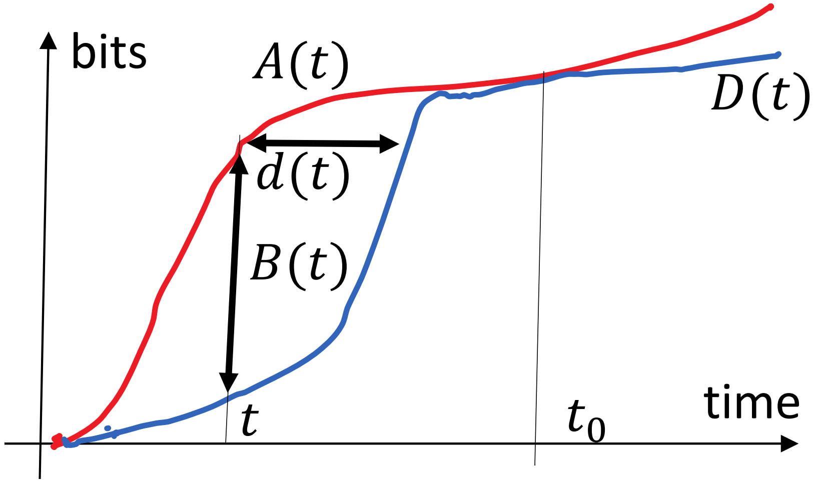
Cumulative Functions



It is convenient to use cumulative (= integral) functions:
$A(t)$ = amount of bits input to system S in $[0, t]$
$D(t)$ = amount of bits output from system S in $[0, t]$

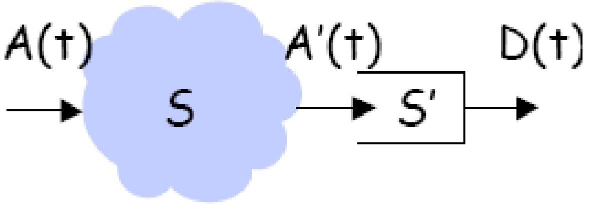
Assume no loss
$B(t)$ = backlog (buffer content) at time t

$d(t)$ = virtual delay
= delay if system is FIFO



At time t_0 the system S is empty

The playback buffer problem

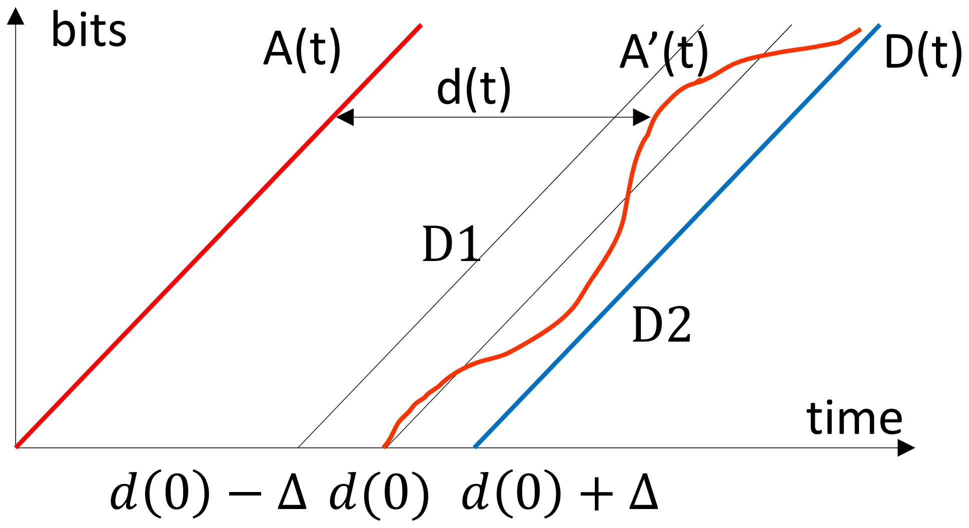


Sources sends at constant rate:

$$A(t) = rt$$

Destination wants to obtain

$$D(t) = r(t - \text{delay offset})$$



Assume FIFO network with delay jitter bounded by Δ

$$d(0) - \Delta \le d(t) \le d(0) + \Delta$$

i.e. $A'(t)$ is between the two parallel lines $D1$ and $D2$

Take $D(t)$ given by the line $D2$, i.e. (Playback Policy):

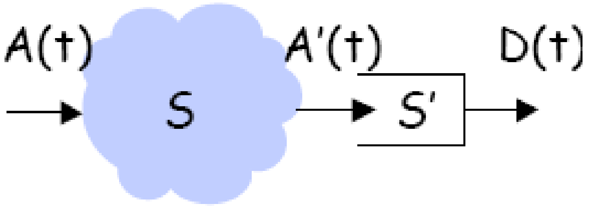
Wait for a time Δ before reading the first bit

Then read at constant rate r,

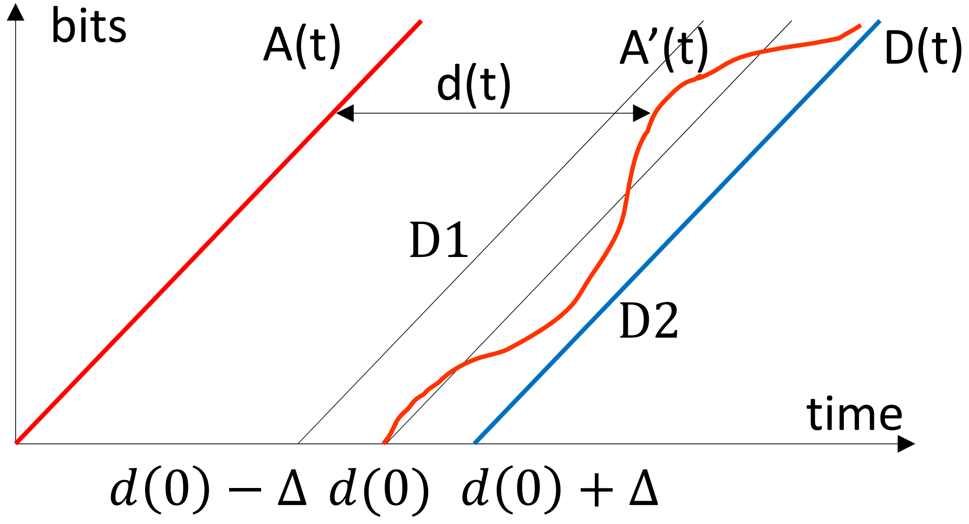
so that $D(t) = r(t - d(0) - \Delta)$

We have $D(t) \le A'(t)$ therefore: all bits arrive before time to read them (no starvation)

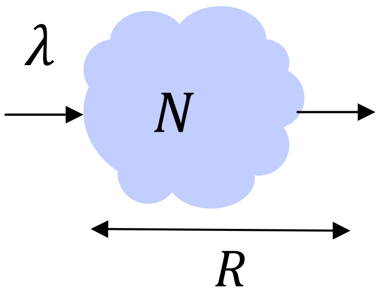
How large should the playback buffer be ?



- *A.* $r\,\Delta$
- *B.* $2\,r\,\Delta$
- C. $3\,r\,\Delta$
- *D.* $r(d(0) + \Delta)$
- E. None of the above
- F. I don’t know



2. Operational Laws
Little's law



Consider an arbitrary discrete system and call

$\lambda =$ customer arrival rate (in *customer/s*)

$R =$ average response time (in s)

(average time spent in system by an arbitrary customer)

$N =$ average number of customers in system (*customers*)

(as seen by an observer who samples the system at an arbitrary point in time)

Then [Little] $\lambda R = N$

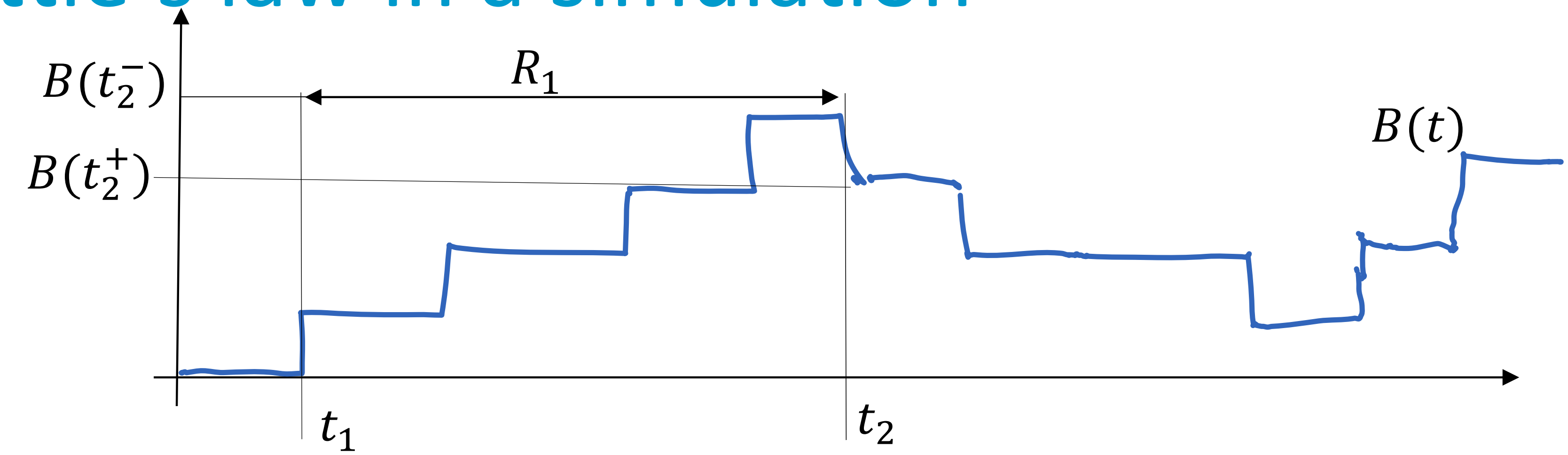
Interpretation: Say every customer pays 1 Fr per minute spent in system

A customer, in average, pays R Fr

The system, in average, receives N Fr per minute

The system is visited by λ customers per minute; the system, in average, receives λR Fr per minute $\Rightarrow \lambda R = N$

Little's law in a simulation



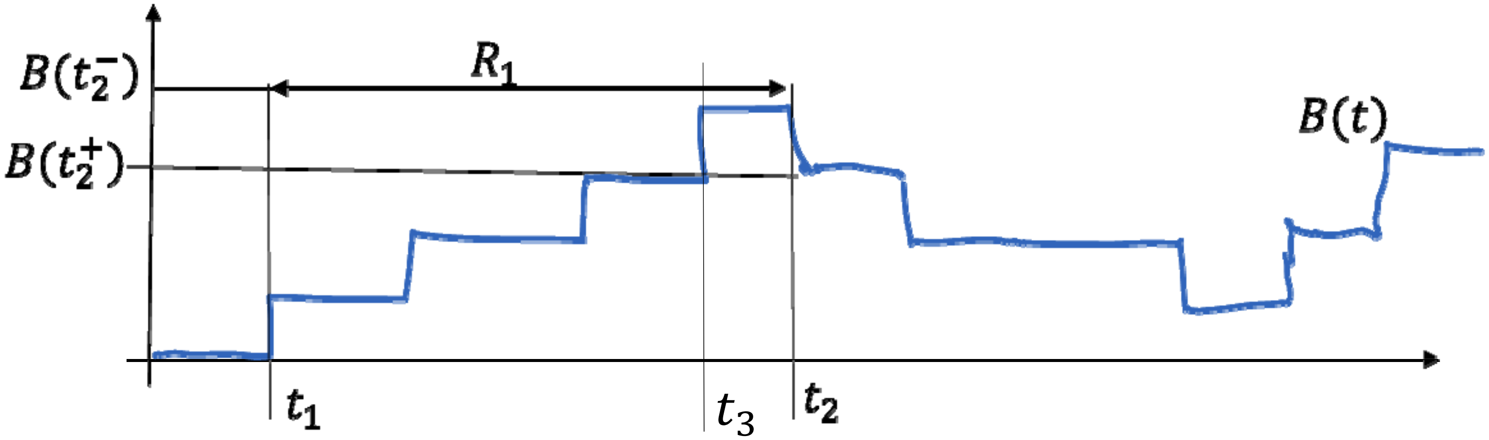
Consider a simulation where we measure R and N. We use two counters responseTimeCtr and backlogCtr, initially 0 and updated at every event. At end of simulation, we have

$$\text{responseTimeCtr} = \sum_{n=1}^{\mathbf{nbCust}} R_n \ \textbf{and} \ \text{backlogCtr} = \int_0^T B(t)dt$$

where $B(t) =$ nb customers in system at time t, $R_n =$ response time of n^{th}customer and $\mathbf{nbCust} =$ nb customers served

At end of simulation we do: $R =$ responseTimeCtr$/\mathbf{nbCust}$,
$\lambda = \mathbf{nbCust}/T$ and $N =$ backlogCtr$/T$

How is each counter updated at time t_2?



responseTimeCtr $= \sum_{n=1}^{\text{nbCust}} R_n$ **and backlogCtr** $= \int_0^T B(t)dt$

A. responseTimeCtr += $(t_2 - t_3)B(t_2^-)$
backlogCtr += $(t_2 - t_3)B(t_2^-)$

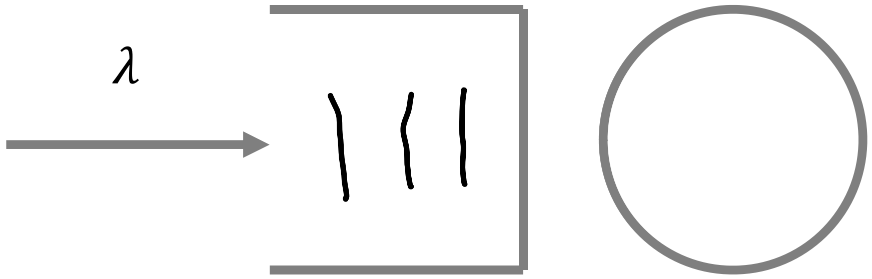
B. responseTimeCtr += $(t_2 - t_1)B(t_2^-)$
backlogCtr += $(t_2 - t_3)B(t_2^-)$

C. responseTimeCtr += $(t_2 - t_3)B(t_2^-)$
backlogCtr += $(t_2 - t_1)B(t_2^-)$

D. I don’t know

Utilization Law

Consider a single server queue and apply Little's formula to the server



$$R = \text{average service time} = \mathrm{S}$$
$$N = 0 \times P(\text{queue is empty}) + 1 \times P(\text{queue is nonempty})$$
$$= P(\text{queue is nonempty})$$

thus $\lambda S = P(\text{queue is nonempty})$

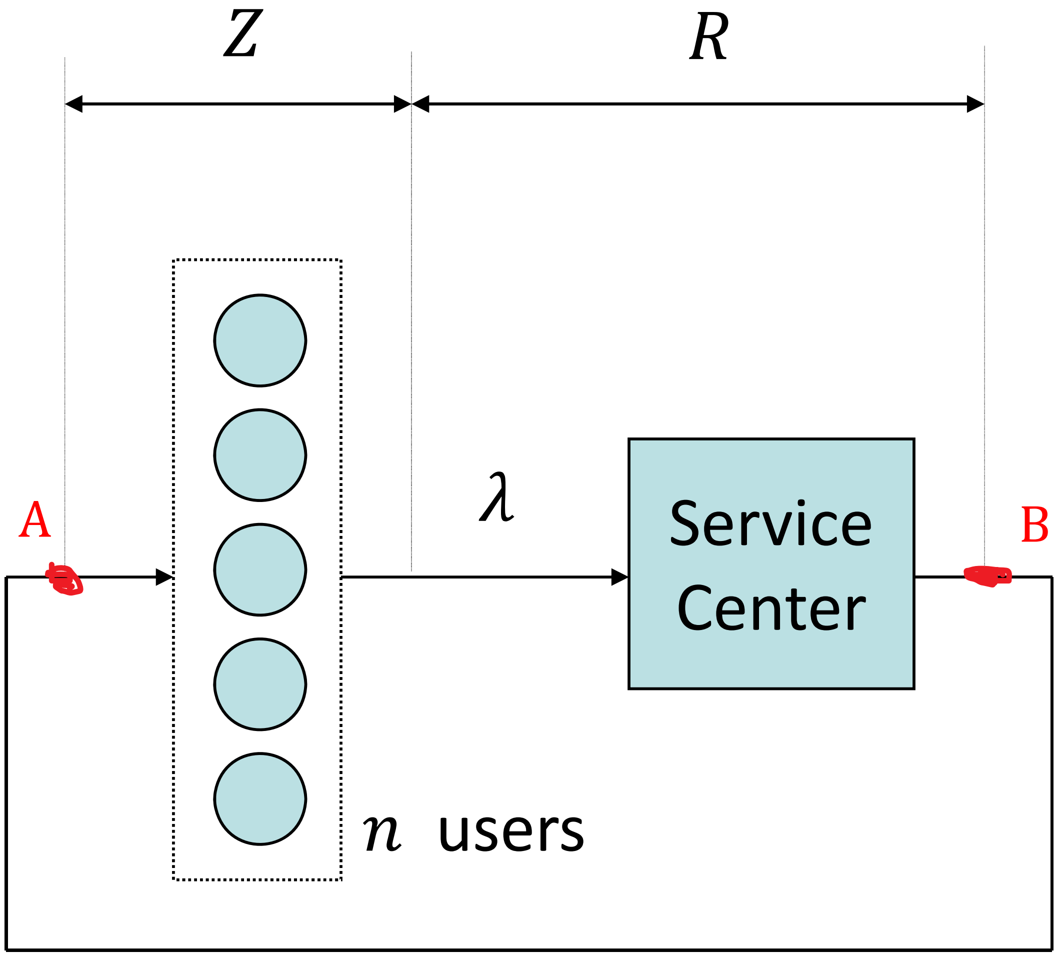
Note that λS is the utilization factor of the server

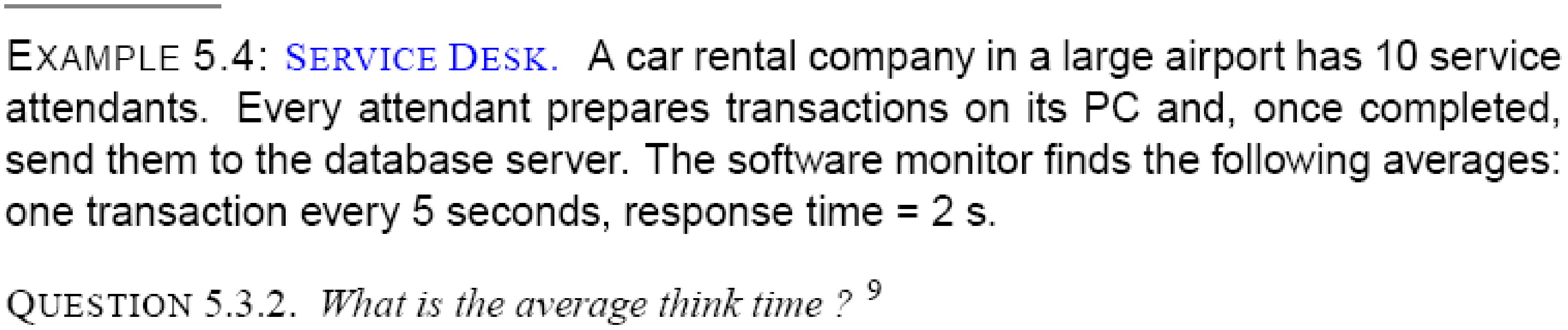
The Interactive User Model

n users alternate between think time and visit to the service center

Apply Little to the system $A \to B$

$$\lambda(\bar{Z} + \bar{R}) = n$$



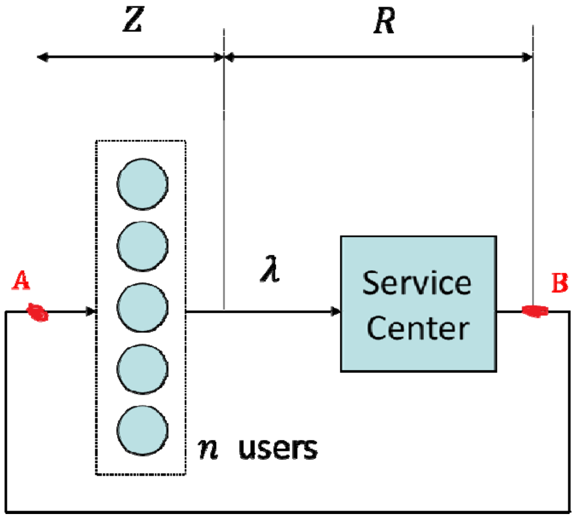


EXAMPLE 5.4: SERVICE DESK. A car rental company in a large airport has 10 service attendants. Every attendant prepares transactions on its PC and, once completed, send them to the database server. The software monitor finds the following averages: one transaction every 5 seconds, response time = 2 s.

QUESTION 5.3.2. *What is the average think time ?* [9]

EXAMPLE 5.4: SERVICE DESK. A car rental company in a large airport has 10 service attendants. Every attendant prepares transactions on its PC and, once completed, send them to the database server. The software monitor finds the following averages: one transaction every 5 seconds, response time = 2 s.

What is the average think time ?



A. 3 s

B. 6 s

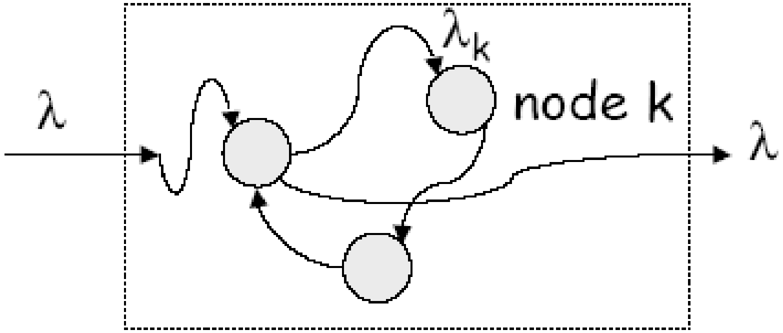
C. 12 s

D. 24 s

E. 48 s

F. I don’t know

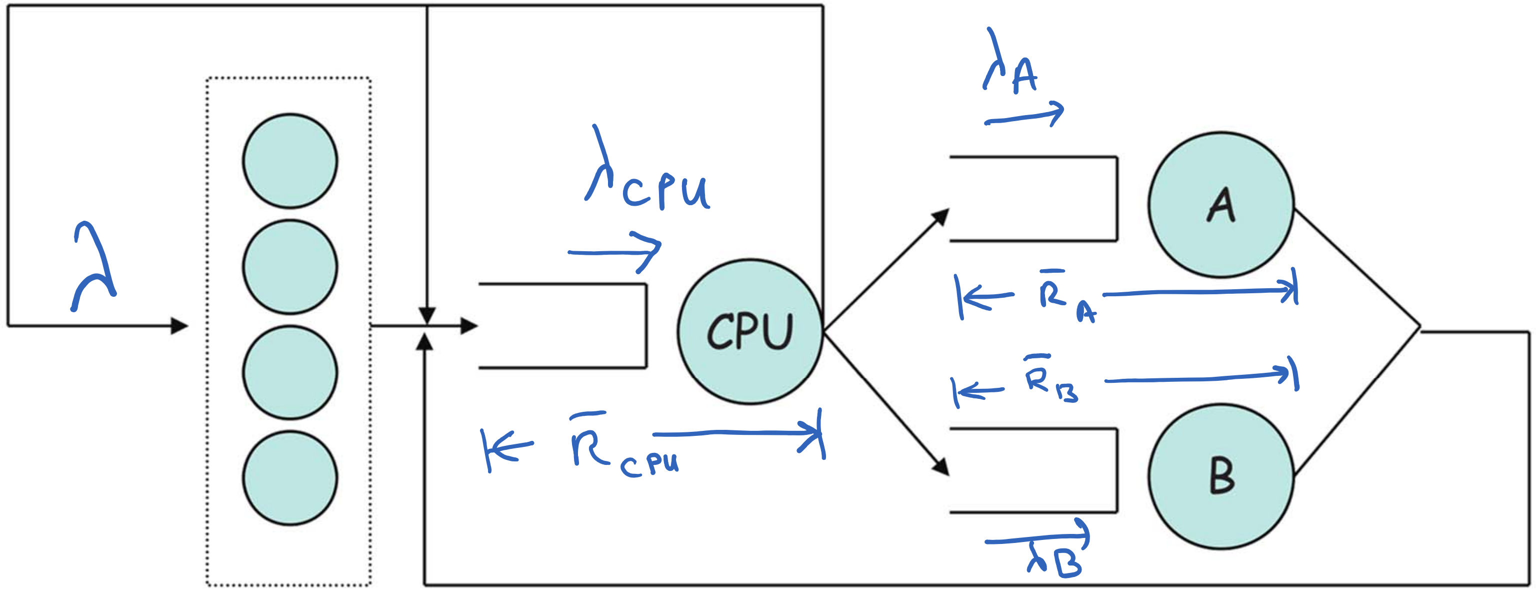
Network Laws

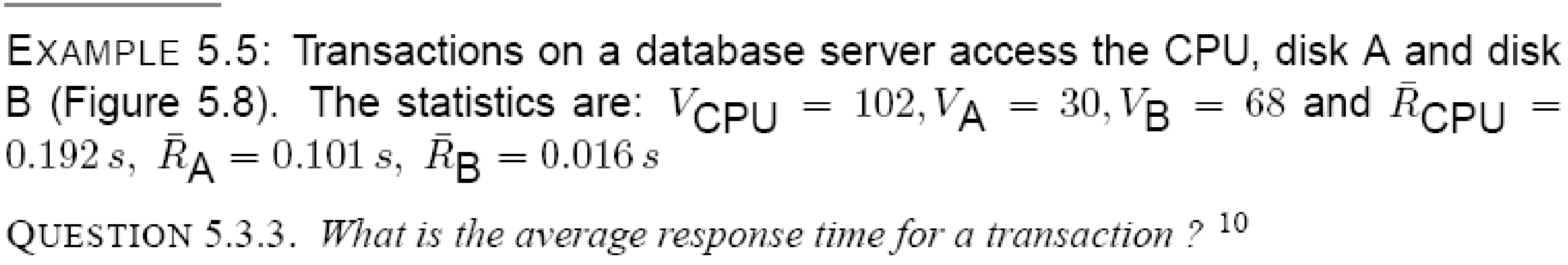


- *[Forced Flows]* $\lambda_k = \lambda V_k$, *where* λ_k *is the expected number of customers arriving per second at node* k *and* V_k *is the expected number of visits to node* k *by an arbitrary customer during its stay in the network.*
- *[Total Response Time] Let* $\bar{R}$ *[resp.* $\bar{R}_k$*] be the expected total response time* $\bar{R}$ *seen by an arbitrary customer [resp. by an arbitrary visit to node* k*].*

$$\bar{R} = \sum_k \bar{R}_k V_k$$

Example





EXAMPLE 5.5: Transactions on a database server access the CPU, disk A and disk B (Figure 5.8). The statistics are: $V_{\text{CPU}} = 102, V_{\text{A}} = 30, V_{\text{B}} = 68$ and $\bar{R}_{\text{CPU}} = 0.192\,s,\ \bar{R}_{\text{A}} = 0.101\,s,\ \bar{R}_{\text{B}} = 0.016\,s$

QUESTION 5.3.3. *What is the average response time for a transaction ?* [10]

One request visits in average 102 times CPU,20 times A and 17 times B

Forced flows: $\lambda_{\text{CPU}} = 102\,\lambda,\ \ \lambda_A = 20\,\lambda,\ \lambda_B = 17\lambda$

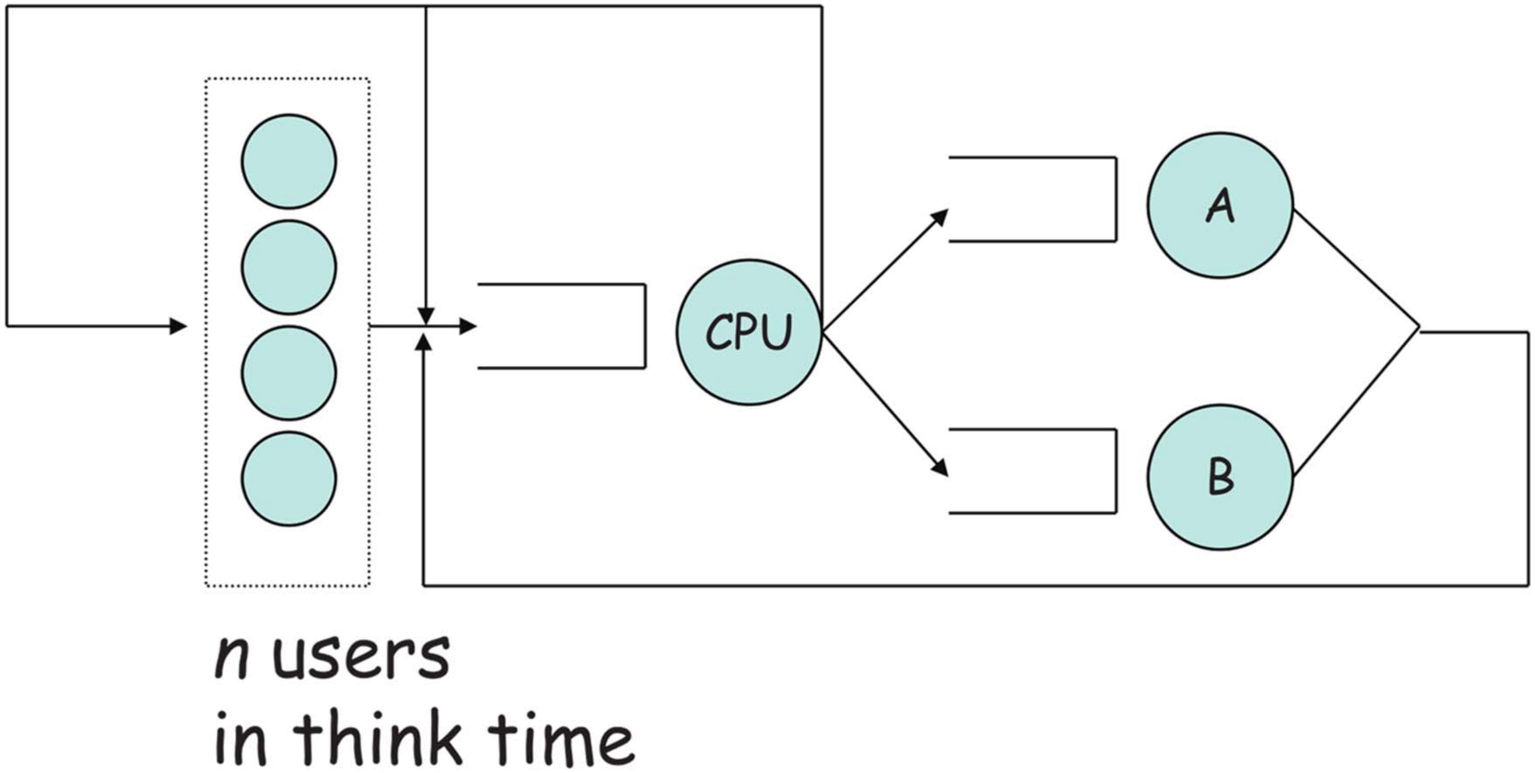
Total Response: $\bar{R} = 102\,\bar{R}_{\text{CPU}} + 20\,\bar{R}_{\text{A}} + 17\,\bar{R}_{\text{B}}$

[10] $23.7\,s$

Bottleneck Analysis

A crude, but powerful approach – often sufficient to analyze complex queuing systems

Example: what is the throughput (transactions per second) versus n?

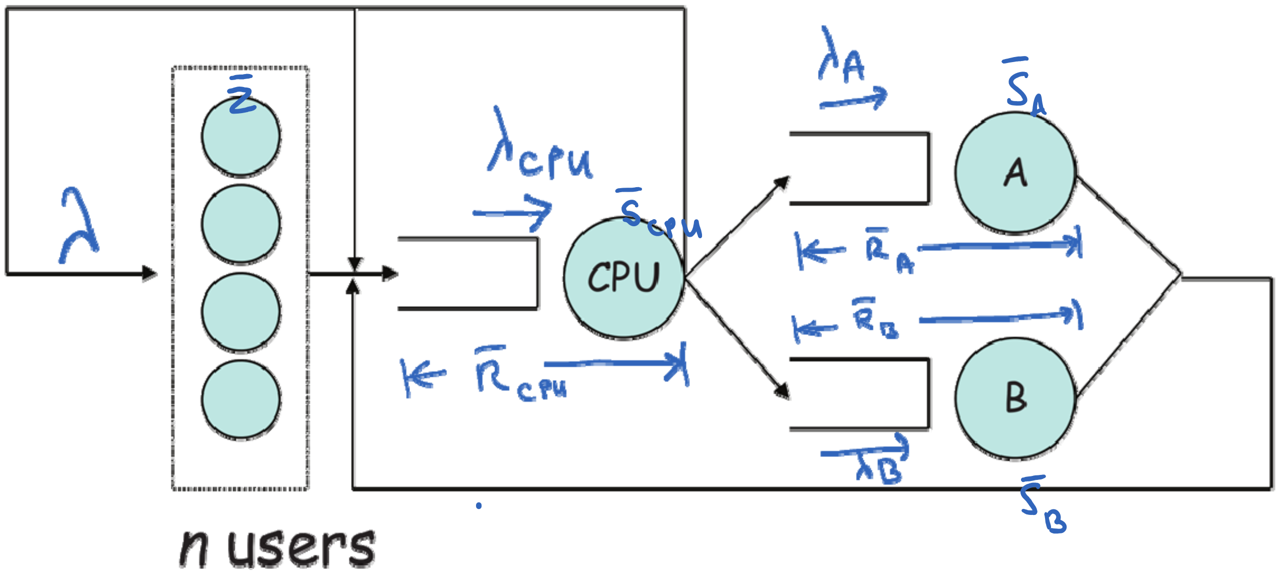


Principles of Bottleneck Analysis

Little’s Formula

Waiting time ≥ 0 -- tight at low utilization

Utilization ≤ 1 -- tight at high utilization



Little's formula:
$$\lambda(\bar{Z}+\bar{R})=n$$

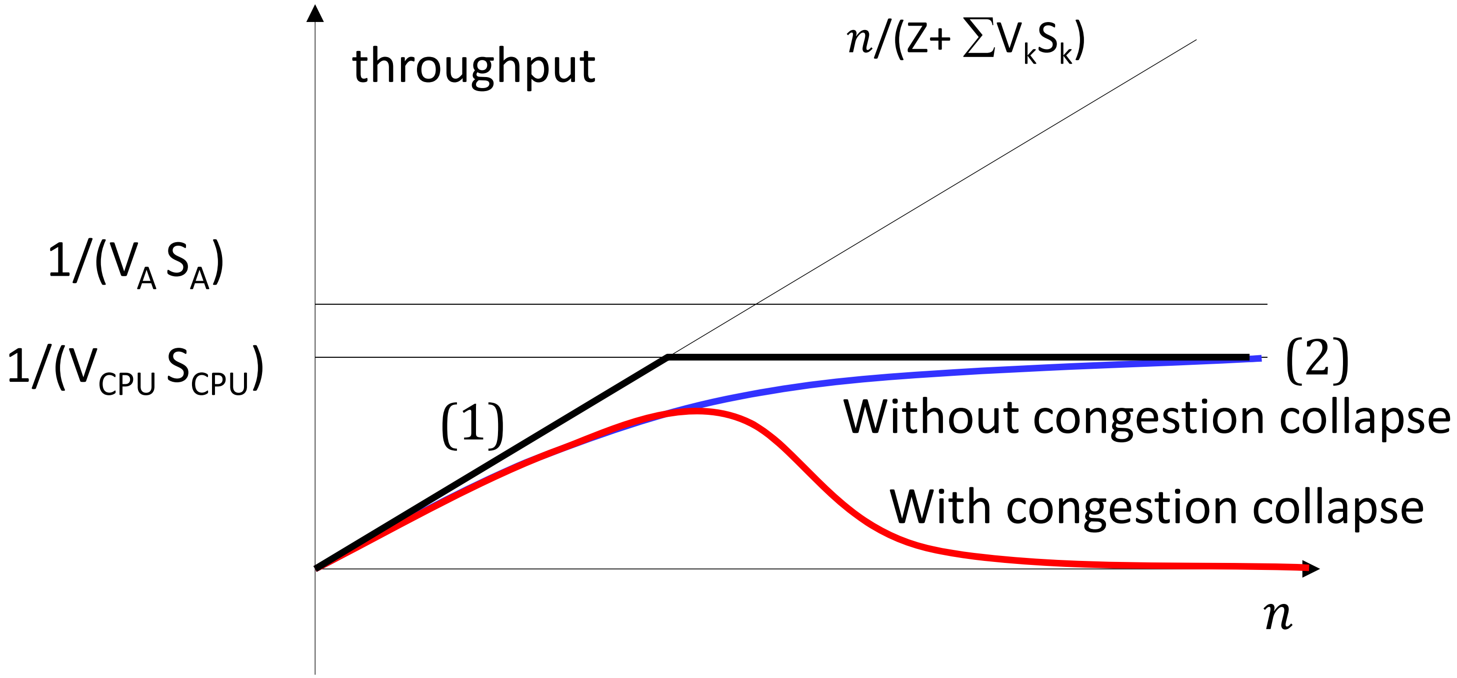
$$\lambda=\frac{n}{\bar{Z}+102\,\bar{R}_{\mathrm{CPU}}+20\,\bar{R}_{\mathrm{A}}+17\,\bar{R}_{\mathrm{B}}}$$

Waiting Time:

$$\lambda\le\frac{n}{\bar{Z}+102\,\bar{S}_{\mathrm{CPU}}+20\,\bar{S}_{\mathrm{A}}+17\,\bar{S}_{\mathrm{B}}}\quad(1)$$

Utilization: $102\,\lambda\,\bar{S}_{\mathrm{CPU}}\le 1, 20\,\lambda\,\bar{S}_{\mathrm{A}}\le 1,\ 17\,\lambda\,\bar{S}_{\mathrm{B}}\le 1$

$$\lambda\le\min\left(\frac{1}{102\bar{S}_{\mathrm{CPU}}},\frac{1}{20\,\bar{S}_{\mathrm{A}}},\frac{1}{17\,\bar{S}_{\mathrm{B}}}\right)\quad(2)$$



$$\lambda \leq \frac{n}{\bar{Z} + 102\ \bar{S}_{\mathrm{CPU}} + 20\ \bar{S}_{\mathrm{A}} + 17\ \bar{S}_{\mathrm{B}}} \quad (1)$$

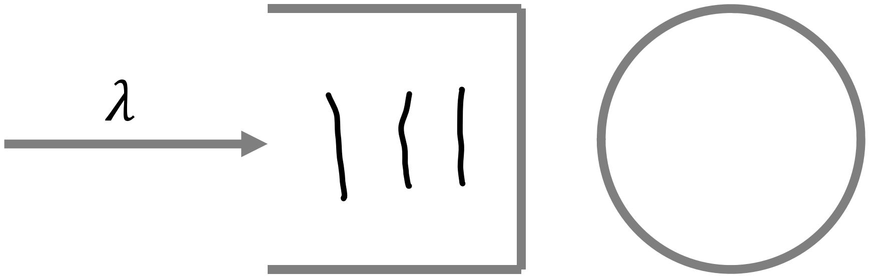
$$\lambda \leq \min(102\ \bar{S}_{\mathrm{CPU}}, 20\ \bar{S}_{\mathrm{A}}, 17\ \bar{S}_{\mathrm{B}}) \quad (2)$$

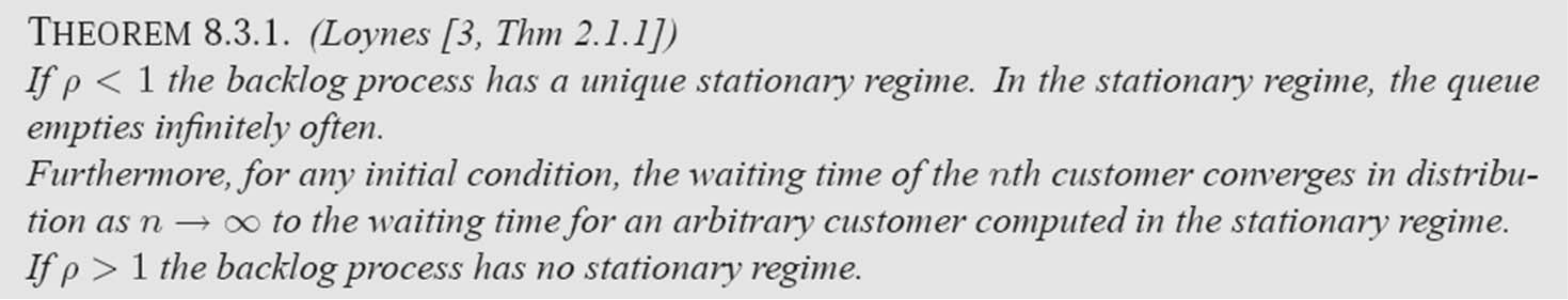
Bottleneck analysis gives the black bound

The true curve is either blue or red, depending on the presence of congestion collapse or not

A resource that achieves minimum in (2) is a *bottleneck*

3. Single Server Queue Stability





THEOREM 8.3.1. *(Loynes [3, Thm 2.1.1])*
If $\rho < 1$ the backlog process has a unique stationary regime. In the stationary regime, the queue empties infinitely often.
Furthermore, for any initial condition, the waiting time of the nth customer converges in distribution as $n \to \infty$ to the waiting time for an arbitrary customer computed in the stationary regime.
If $\rho > 1$ the backlog process has no stationary regime.

Recall that $\rho = \lambda \times \text{average service time}$

In other words

$\rho < 1 \Rightarrow$ system has a stationary regime (and converges to it)

$\rho > 1 \Rightarrow$ system has no stationary regime

If $\rho = 1$ the theorem says nothing – both cases are possible

Loynes' theorem does not assume any independence anywhere

Example: M/GI/1 queue

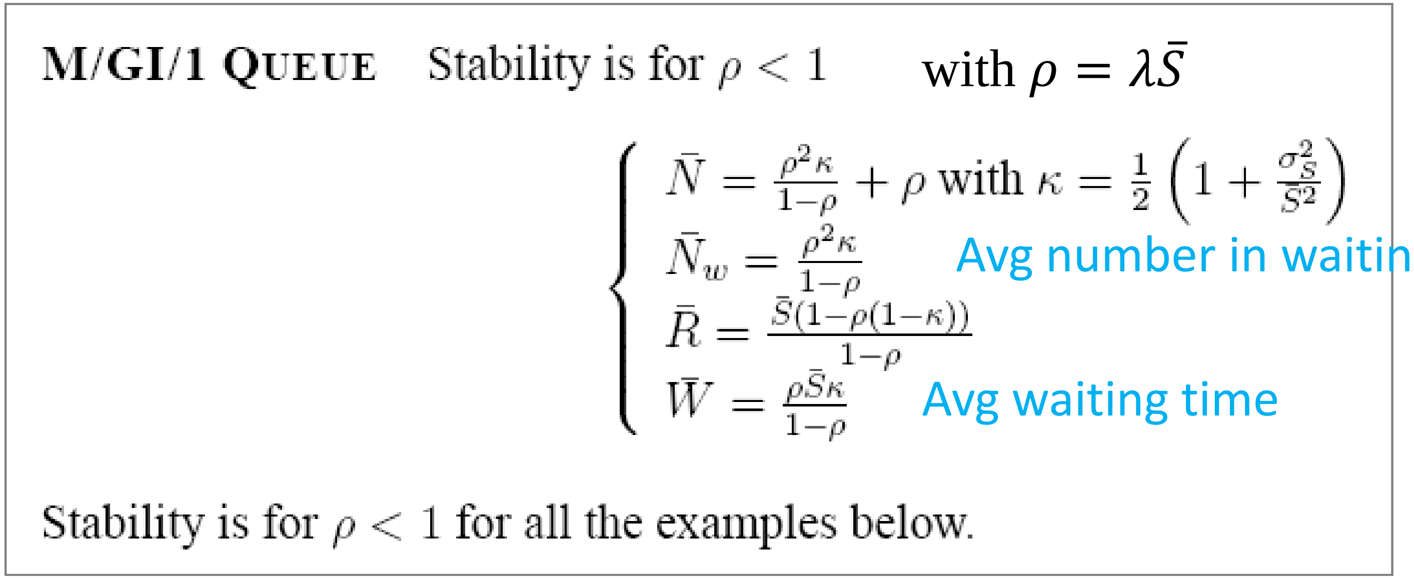
M = markov
= Poisson arrival process

1 server

Times between arrivals are exp(λ) where λ is the arrival rate

memoryless: an arrival in $[t, t + dt]$ is independent of the past and has probability $\lambda\, dt$

GI = General Independent
i.e. service times are iid with a finite variance and independent of the arrival process



M/GI/1 QUEUE Stability is for $\rho < 1$ with $\rho = \lambda \bar{S}$

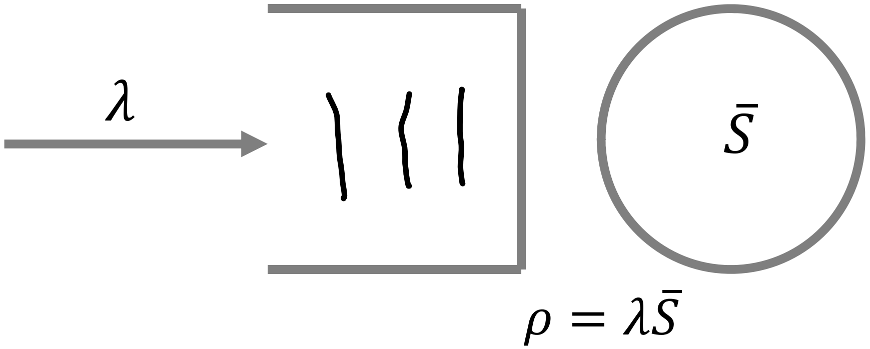
$$\begin{cases} \bar{N} = \frac{\rho^2 \kappa}{1-\rho} + \rho \text{ with } \kappa = \frac{1}{2}\left(1 + \frac{\sigma_S^2}{\bar{S}^2}\right) \\ \bar{N}_w = \frac{\rho^2 \kappa}{1-\rho} \\ \bar{R} = \frac{\bar{S}(1-\rho(1-\kappa))}{1-\rho} \\ \bar{W} = \frac{\rho \bar{S} \kappa}{1-\rho} \end{cases}$$

Avg number in waiting room

Avg waiting time

Stability is for $\rho < 1$ for all the examples below.

The formula $\bar{N} = \bar{N}_w + \rho$ is true …



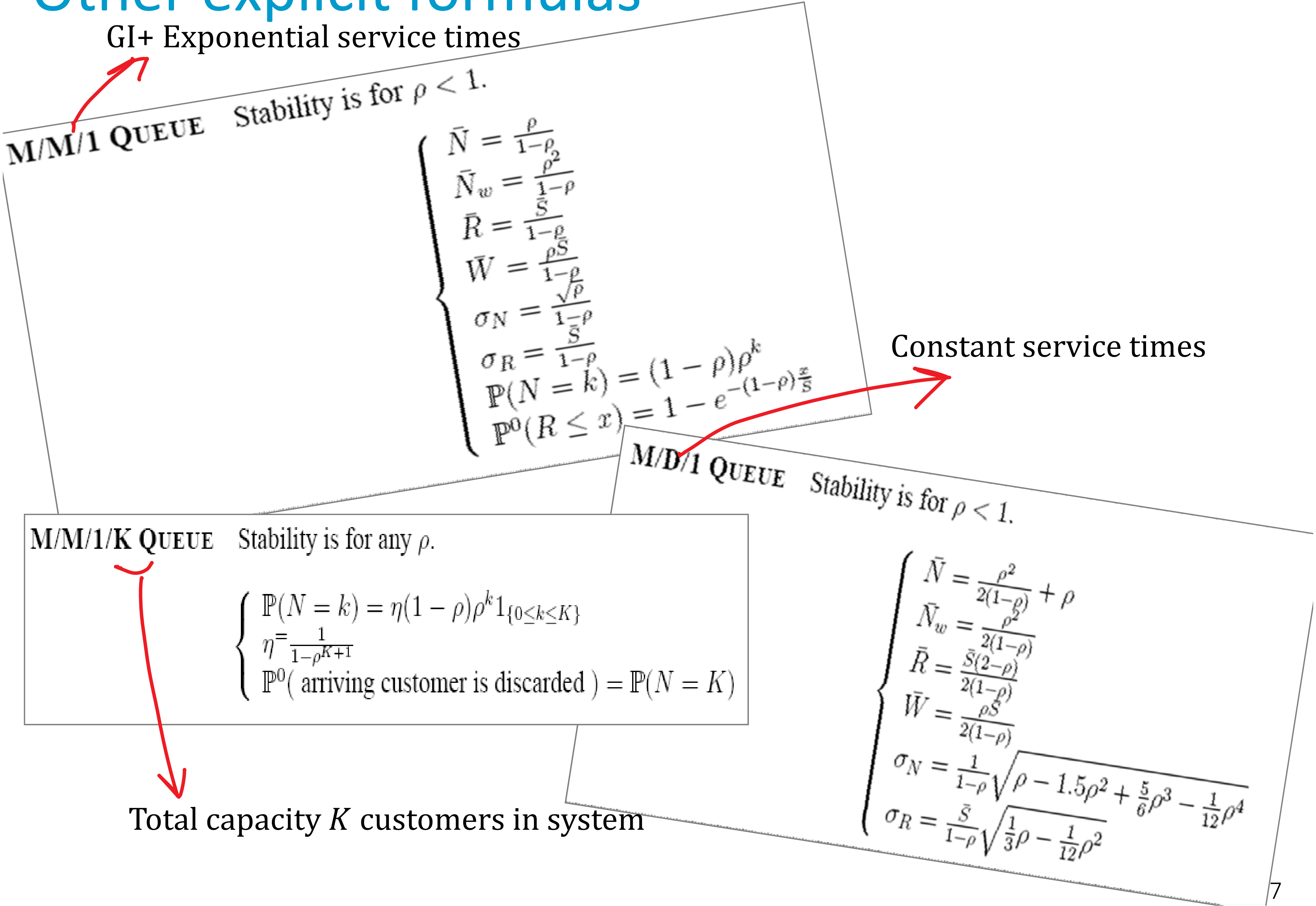
A. For the M/GI/1 queue but not for all stable single server queues

B. For all single server queues with Poisson arrivals but not for all stable single server queues

C. For all single stable server queues

D. I don’t know

Other explicit formulas



GI+ Exponential service times

M/M/1 QUEUE Stability is for $\rho < 1$.

$$\begin{cases} \bar{N} = \frac{\rho}{1-\rho} \\ \bar{N}_w = \frac{\rho^2}{1-\rho} \\ \bar{R} = \frac{\bar{S}}{1-\rho} \\ \bar{W} = \frac{\rho\bar{S}}{1-\rho} \\ \sigma_N = \frac{\sqrt{\rho}}{1-\rho} \\ \sigma_R = \frac{\bar{S}}{1-\rho} \\ \mathbb{P}(N = k) = (1-\rho)\rho^k \\ \mathbb{P}^0(R \le x) = 1 - e^{-(1-\rho)\frac{x}{\bar{S}}} \end{cases}$$

Constant service times

M/D/1 QUEUE Stability is for $\rho < 1$.

$$\begin{cases} \bar{N} = \frac{\rho^2}{2(1-\rho)} + \rho \\ \bar{N}_w = \frac{\rho^2}{2(1-\rho)} \\ \bar{R} = \frac{\bar{S}(2-\rho)}{2(1-\rho)} \\ \bar{W} = \frac{\rho\bar{S}}{2(1-\rho)} \\ \sigma_N = \frac{1}{1-\rho}\sqrt{\rho - 1.5\rho^2 + \frac{5}{6}\rho^3 - \frac{1}{12}\rho^4} \\ \sigma_R = \frac{\bar{S}}{1-\rho}\sqrt{\frac{1}{3}\rho - \frac{1}{12}\rho^2} \end{cases}$$

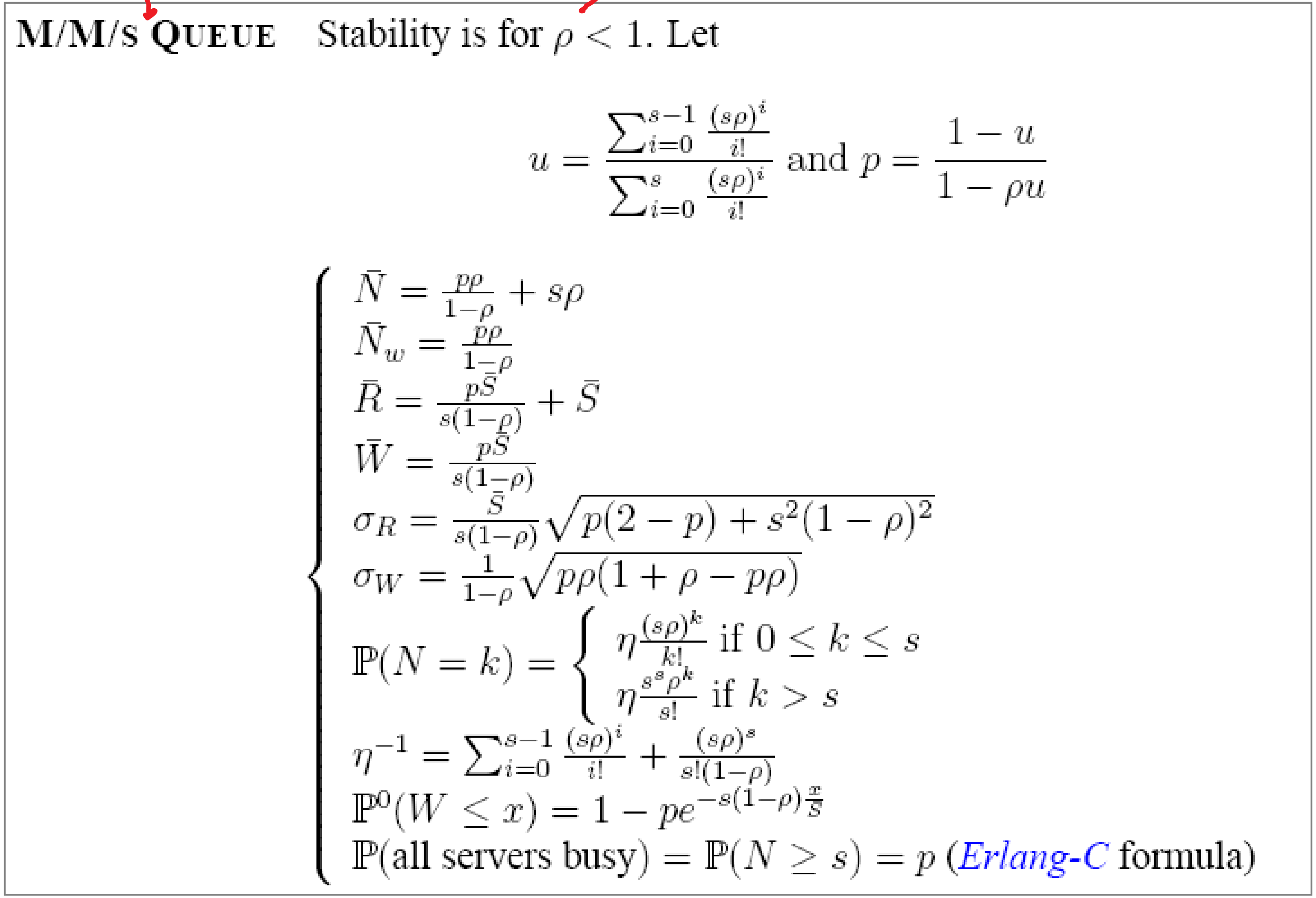
M/M/1/K QUEUE Stability is for any ρ.

$$\begin{cases} \mathbb{P}(N = k) = \eta(1-\rho)\rho^k 1_{\{0 \le k \le K\}} \\ \eta = \frac{1}{1-\rho^{K+1}} \\ \mathbb{P}^0(\text{ arriving customer is discarded }) = \mathbb{P}(N = K) \end{cases}$$

Total capacity K customers in system

s servers

$$\rho = \frac{\lambda \bar{S}}{s}$$

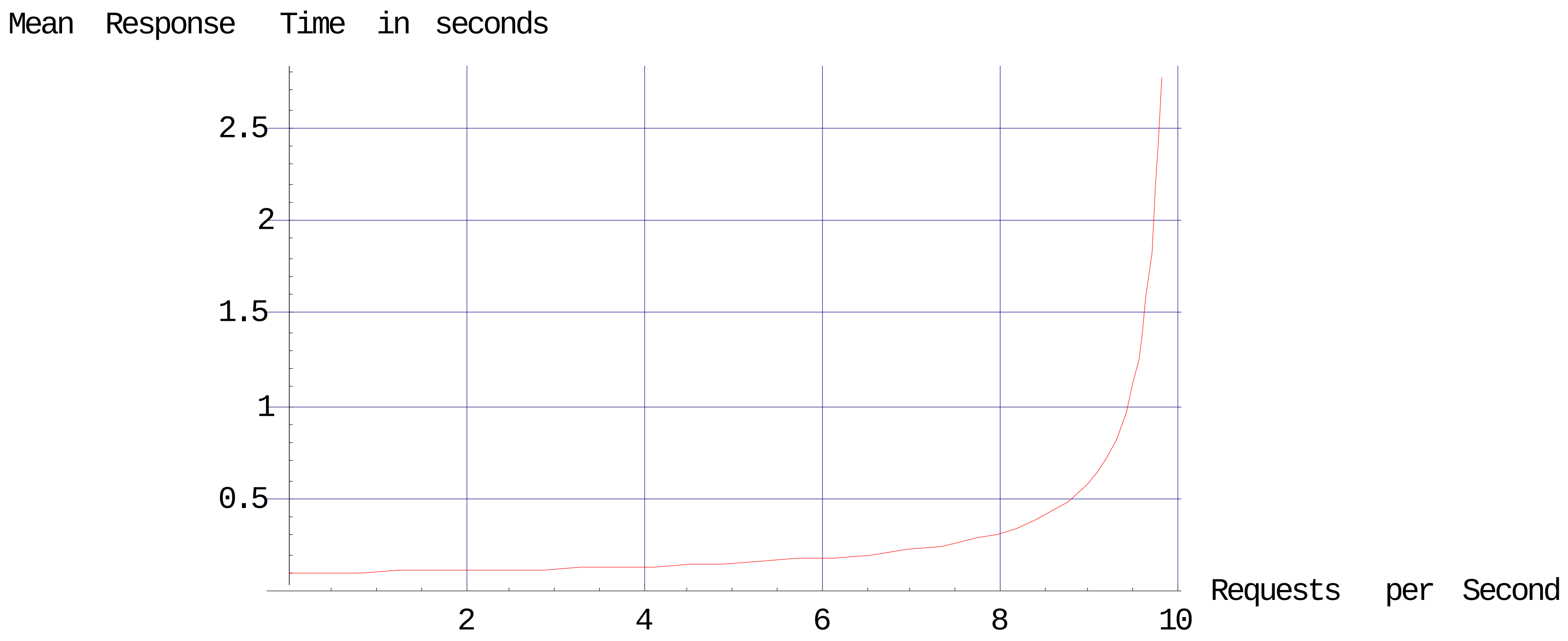


M/M/s QUEUE Stability is for $\rho < 1$. Let

$$u = \frac{\sum_{i=0}^{s-1} \frac{(s\rho)^i}{i!}}{\sum_{i=0}^{s} \frac{(s\rho)^i}{i!}} \text{ and } p = \frac{1-u}{1-\rho u}$$

$$\begin{cases}
\bar{N} = \frac{p\rho}{1-\rho} + s\rho \\
\bar{N}_w = \frac{p\rho}{1-\rho} \\
\bar{R} = \frac{p\bar{S}}{s(1-\rho)} + \bar{S} \\
\bar{W} = \frac{p\bar{S}}{s(1-\rho)} \\
\sigma_R = \frac{\bar{S}}{s(1-\rho)}\sqrt{p(2-p) + s^2(1-\rho)^2} \\
\sigma_W = \frac{1}{1-\rho}\sqrt{p\rho(1+\rho-p\rho)} \\
\mathbb{P}(N=k) = \begin{cases} \eta\frac{(s\rho)^k}{k!} \text{ if } 0 \le k \le s \\ \eta\frac{s^s\rho^k}{s!} \text{ if } k > s \end{cases} \\
\eta^{-1} = \sum_{i=0}^{s-1} \frac{(s\rho)^i}{i!} + \frac{(s\rho)^s}{s!(1-\rho)} \\
\mathbb{P}^0(W \le x) = 1 - pe^{-s(1-\rho)\frac{x}{\bar{S}}} \\
\mathbb{P}(\text{all servers busy}) = \mathbb{P}(N \ge s) = p \text{ (Erlang-C formula)}
\end{cases}$$

Non Linearity of Response Time



EXAMPLE 5.2: A database system services requests that can be modeled as a Poisson process. The time needed to process a request is 0.1 second and its standard deviation is estimated to 0.03. How does the average response time depend on the number of requests per second that can be served ? The solution is found by the M/GI/1 queue model and is plotted in Figure 5.3.

QUESTION 5.2.4. *What is the maximum load that can be served if an average response time of* 0.5 *second is considered acceptable ?* [5]

[5] 8.8 requests per second.

Which curve is for which distribution of service time?

A. Top =1
   Middle =2
   Bottom =3

B. 1,3,2

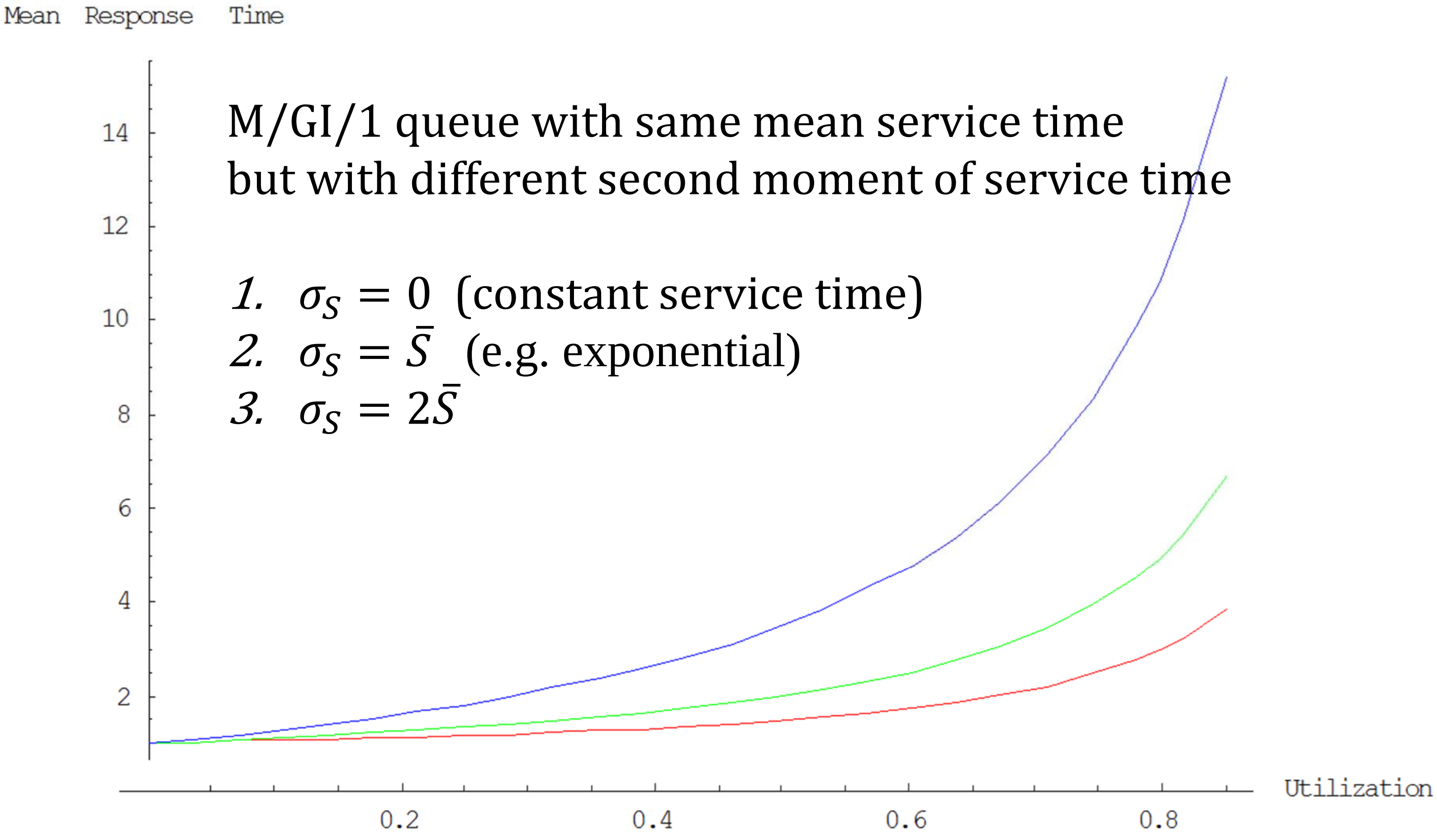
C. 2,1,3

D. 2,3,1

E. 3,1,2

F. 3,2,1

G. I don’t know



M/GI/1 queue with same mean service time
but with different second moment of service time

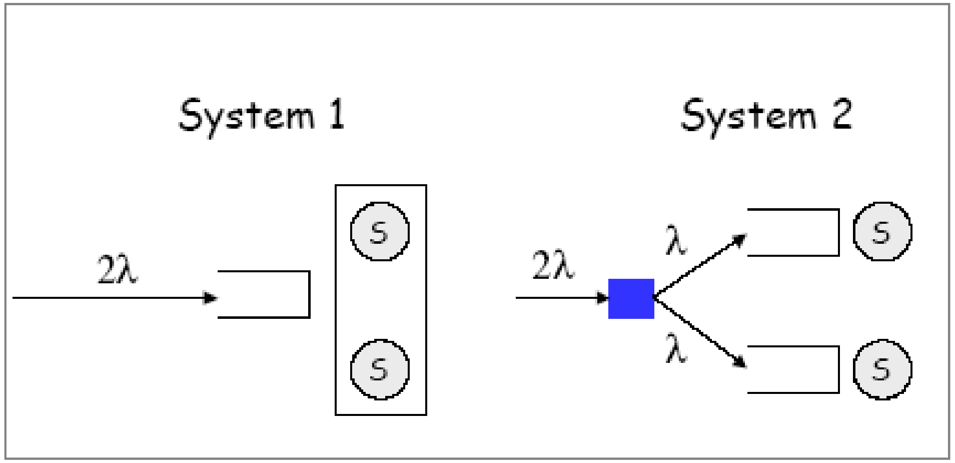
1. $\sigma_S = 0$ (constant service time)
2. $\sigma_S = \bar{S}$ (e.g. exponential)
3. $\sigma_S = 2\bar{S}$

Optimal Sharing

Assume Poisson arrivals and exponential service times.

Compare the two in terms of

- Response time
- Capacity

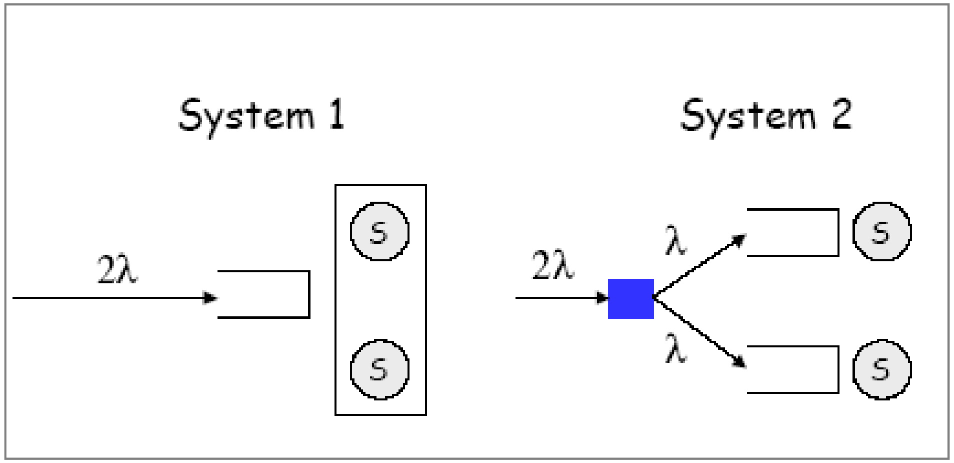


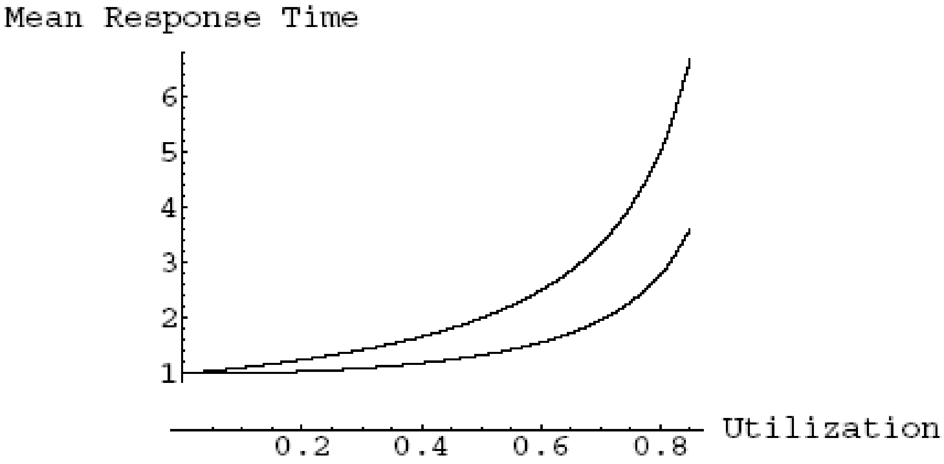
Which curve is for system 1 ?

A. The top curve

B. The bottom curve

C. I don't know





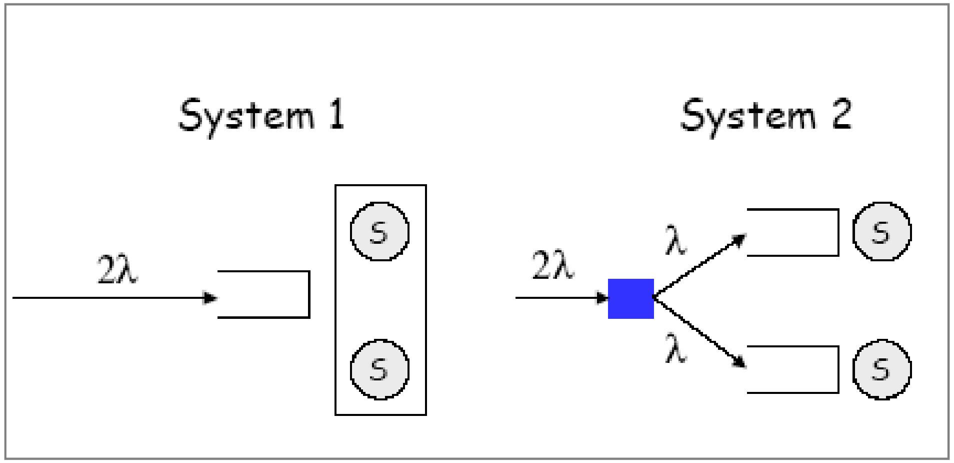
Which system has the larger capacity ? (i.e. the max λ for which system is stable ?)

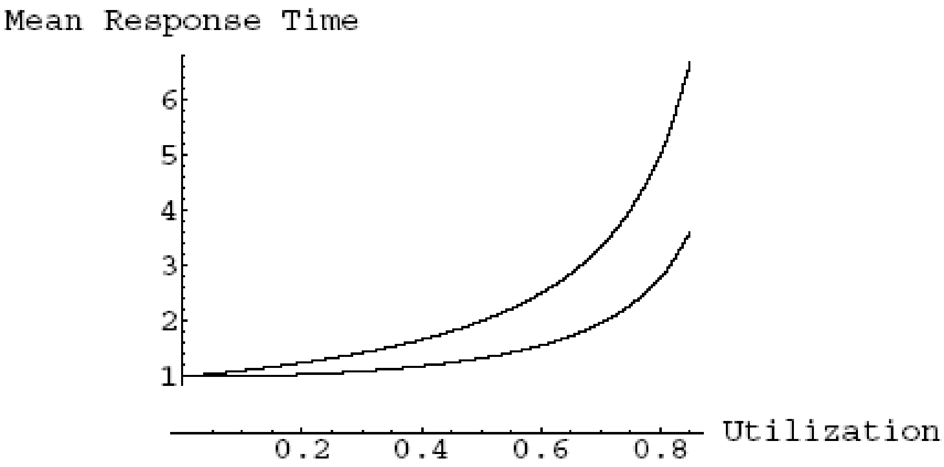
A. System 1

B. System 2

C. Both have the same

D. I don't know





For an M/M/1 queue, what is the expected response time for a customer with service time $\boldsymbol{x}$?

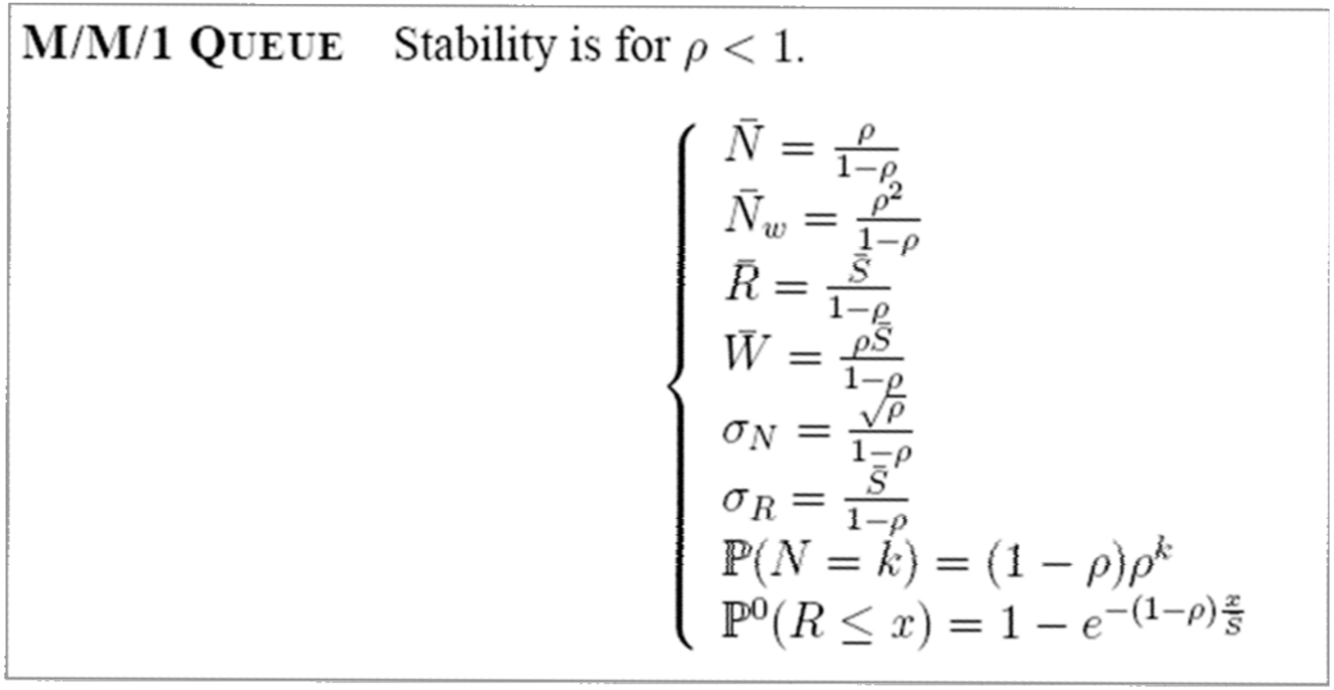
A. $\bar{S}(1-\rho\ -x\ )$

B. $x+\frac{\rho\,\bar{S}}{1-\rho}$

C. $\frac{x\rho}{1-\rho}$

D. $\frac{x\rho^2}{1-\rho}$

E. I don't know



M/M/1 QUEUE Stability is for $\rho < 1$.

$$\begin{cases} \bar{N} = \frac{\rho}{1-\rho} \\ \bar{N}_w = \frac{\rho^2}{1-\rho} \\ \bar{R} = \frac{\bar{S}}{1-\rho} \\ \bar{W} = \frac{\rho\bar{S}}{1-\rho} \\ \sigma_N = \frac{\sqrt{\rho}}{1-\rho} \\ \sigma_R = \frac{\bar{S}}{1-\rho} \\ \mathbb{P}(N=k) = (1-\rho)\rho^k \\ \mathbb{P}^0(R \le x) = 1 - e^{-(1-\rho)\frac{x}{\bar{S}}} \end{cases}$$

$x\rho$

The Processor Sharing Queue M/GI/1/PS

All queues seen so far are FIFO (a notation such as M/M/1 assumes FIFO by default)

The *processor sharing* queue M/GI/1/PS is a single server non FIFO queue where the server is equally shared between all customers present.

Models: processors, network links.

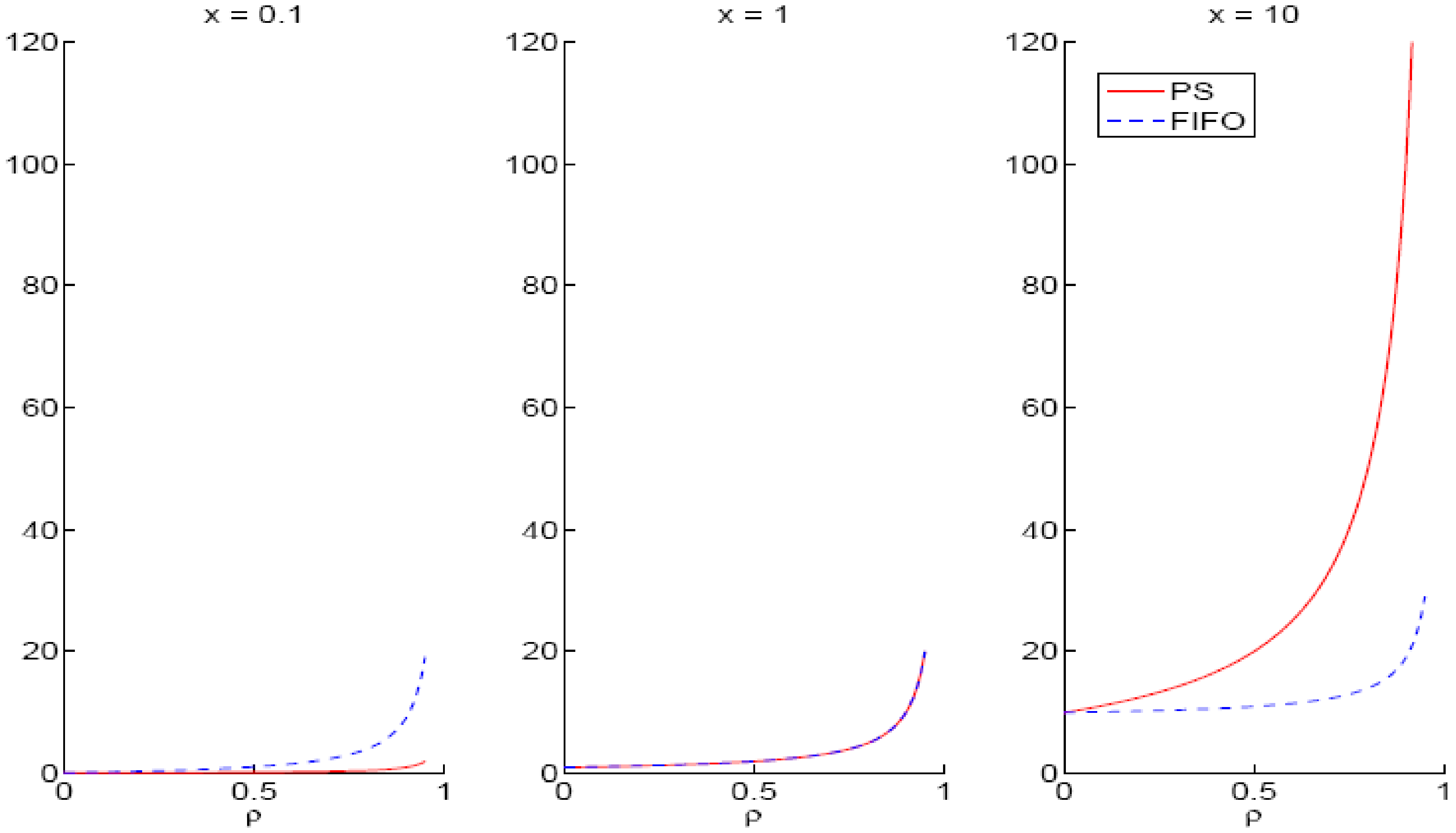
By Loynes' theorem, stability is for $\rho < 1$

The number of customers in system has a geometric distribution
$P(N(t) = k) = (1 - \rho)\rho^k$
independent of distribution of service time

Egalitarian property:

$$\mathrm{E}(\text{ response time } | \text{service time} = x) = \frac{x}{1 - \rho}$$

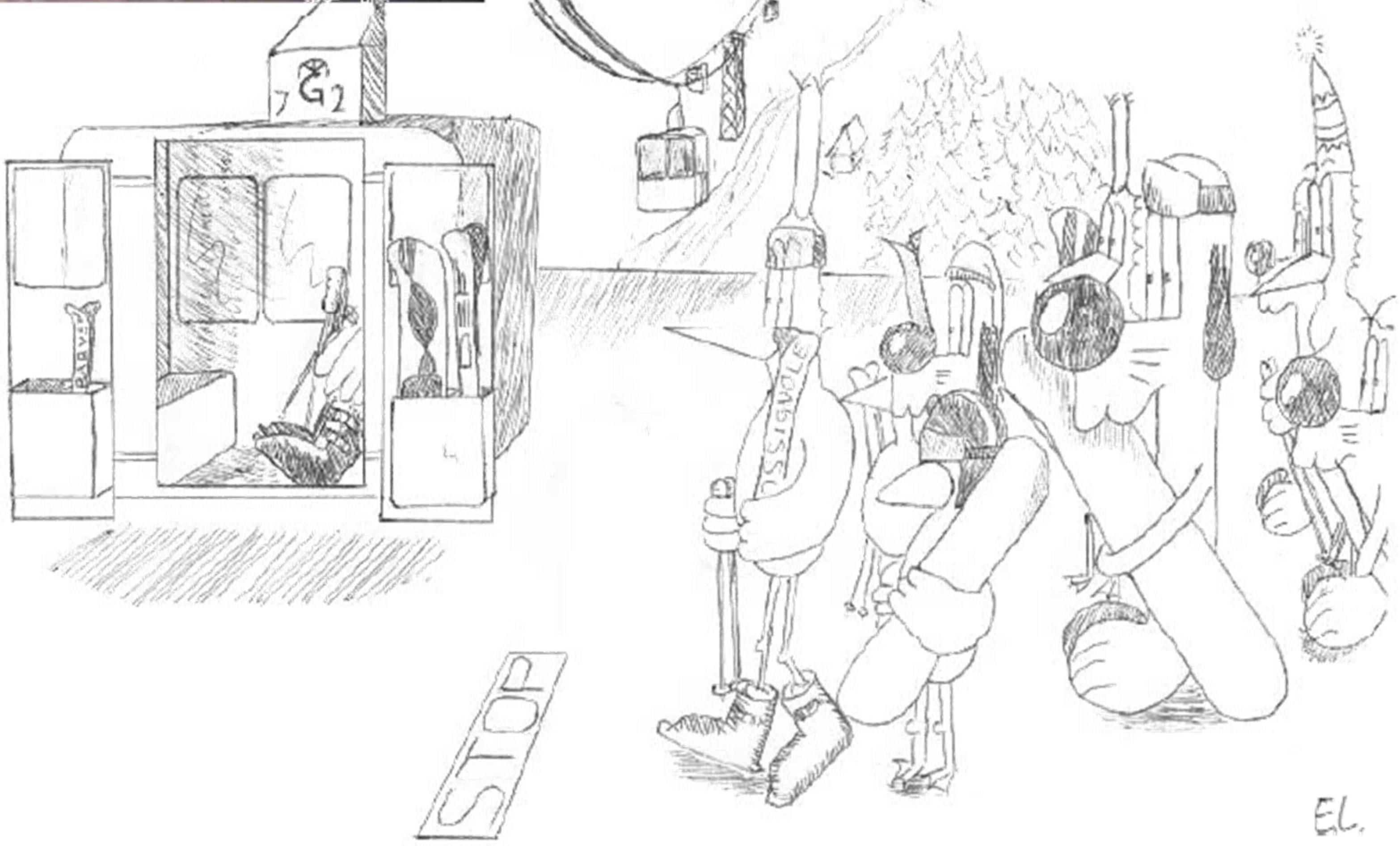
Fairness of PS: Expected response time given service time is x M/M/1/PS versus M/M/1/FIFO



4. A Case Study

Impact of capacity increase ?

Optimal Capacity ?



Methodology

Goal: evaluate impact of doubling capacity of skilift

Factors: $c =$ capacity of ski-lift in people / sec

Metric : waiting time at the lift

Load

Model 1: arrival in burst

Model 2: peak hour stationary regime

4.1. Deterministic Analysis

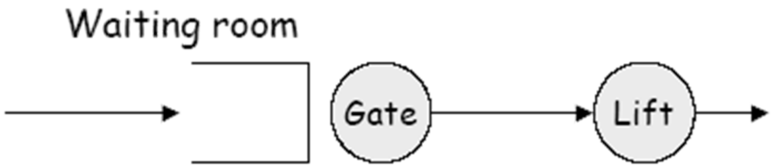
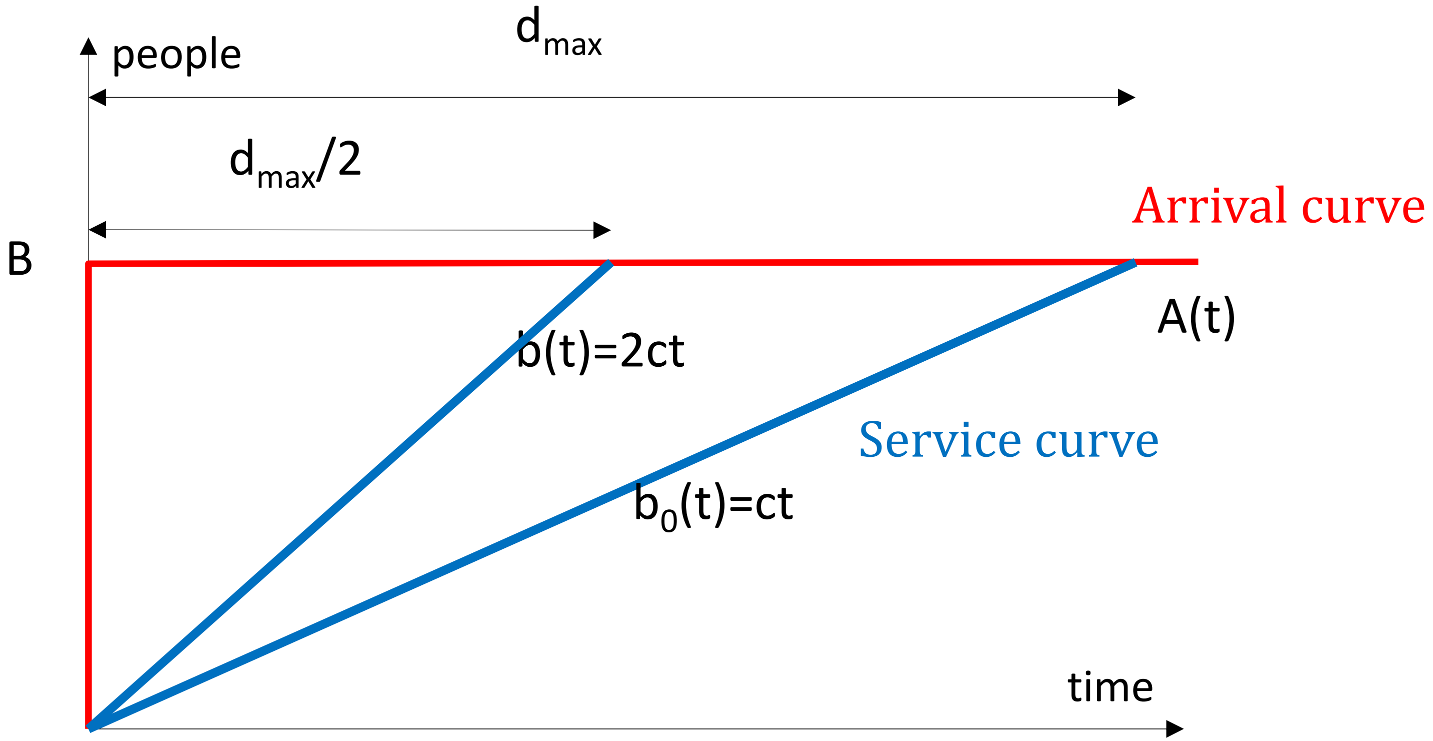


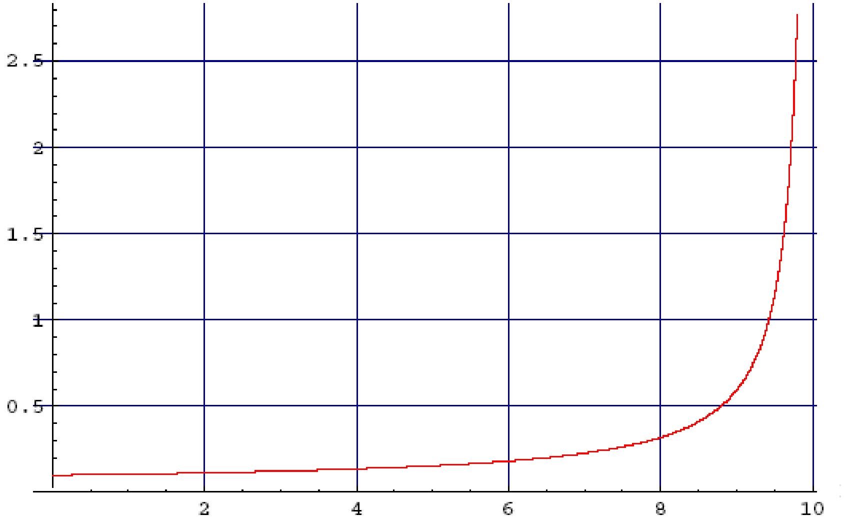
Figure 5.10: Queuing Model of Skilift



Doubling capacity ⇒ max delay is divided by 2 And same for average delay

4.2 Single Queue Analysis

Assume no feedback in the system;
we have an M/GI/1 queue

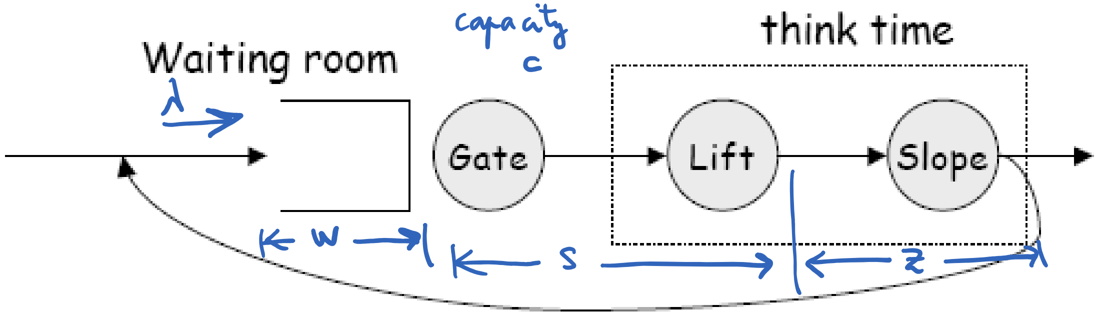


Non linearity of response time
⇒ we need to know where we were on the curve before doubling capacity

We were probably close to asymptote; doubling the capacity
⇒
very large reduction of waiting time, more than by a factor of 2

4.3 Operational Analysis

A refined model, with circulating users;



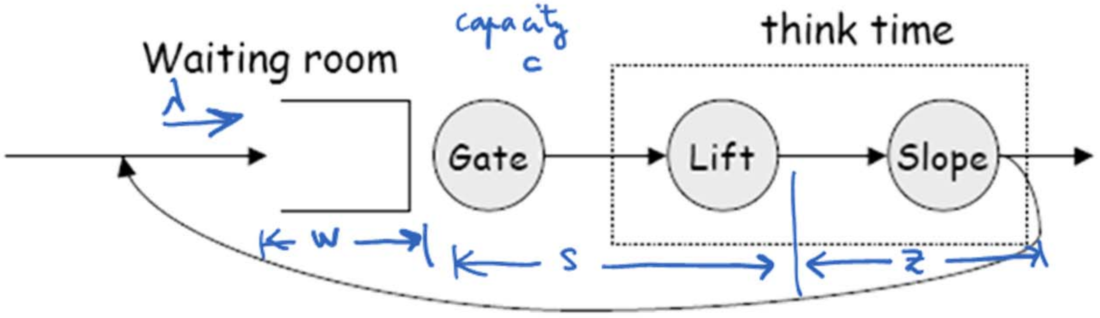
Gate has a capacity c customers/sec;

e.g. Gate has K slots, time to go through gate $\bar{G}$, $c = \frac{K}{\bar{G}}$

$\lambda, \overline{W}$ to be determined

$c, \bar{S}, \bar{Z}, \overline{N}$ known

Bottleneck Analysis



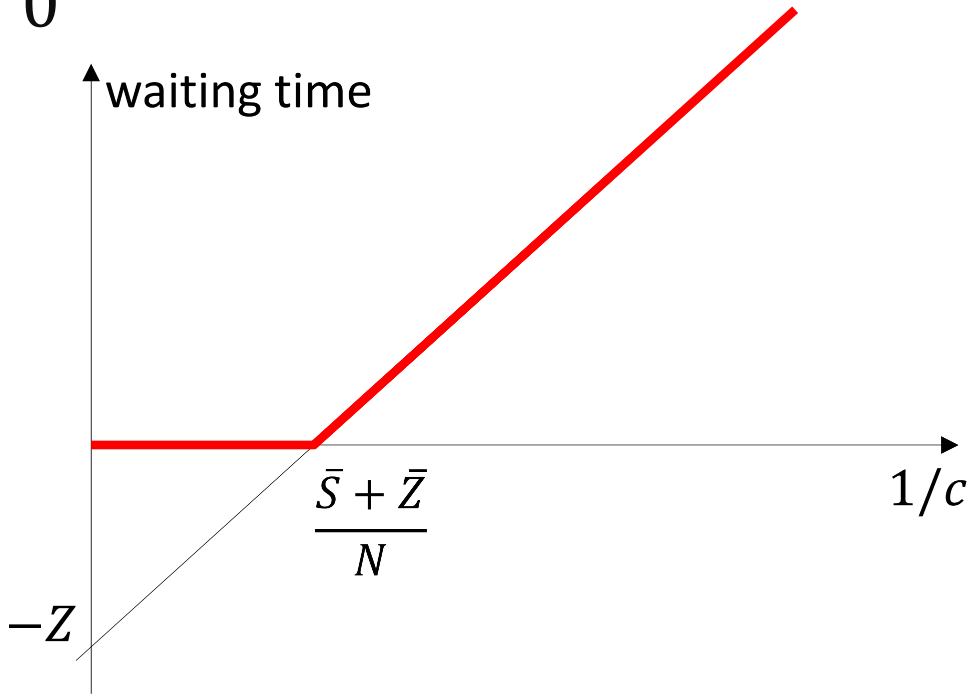
Little: $\lambda(\bar{W} + \bar{S} + \bar{Z}) = \bar{N}$

$$\Rightarrow \; \bar{W} = \frac{\bar{N}}{\lambda} - \bar{S} - \bar{Z}$$

Waiting Time: $\bar{W} \geq 0$

Utilization: $\lambda\bar{G} \leq K$ i.e. $\lambda \leq c$

$$\Rightarrow \; \bar{W} \geq \frac{\bar{N}}{c} - \bar{S} - \bar{Z} \text{ and } \bar{W} \geq 0$$



Exact solution using queuing network theory

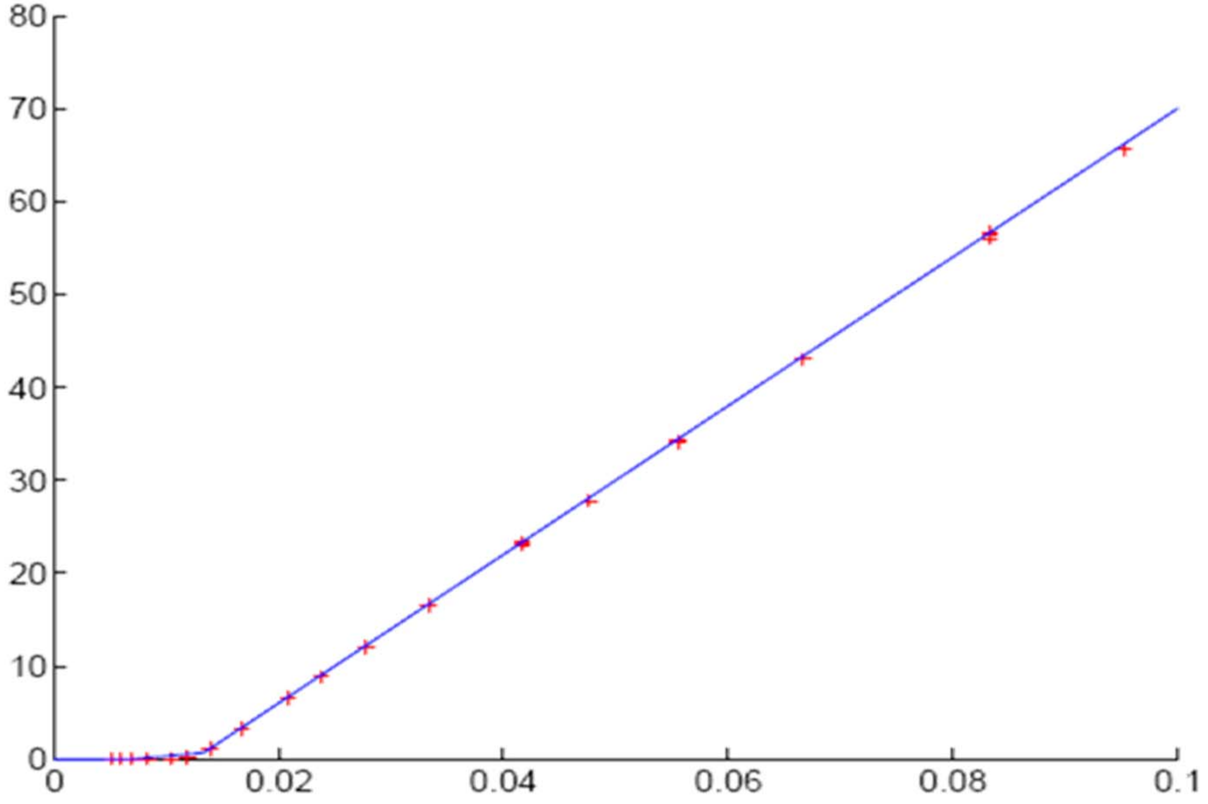


Figure 8.23: First Panel: A Model that accounts for dependency of arrival rate and waiting time. Second panel: Waiting time in minutes for this model versus $\frac{1}{c}$, where c is skilift capacity (in people per minute). The solid line is the approximation by bottleneck analysis. The crosses are obtained by analytical solution of the queuing network model in Figure 8.24, with the following parameters: population size $K = 800$ skiers; number of servers at gate $B \in \{1, 2, ...7, 8\}$; service time at gate $\bar{S} \in \{2.5, 5, 10, 20\}$ seconds; time between visits to the gate $\bar{Z} = 10$ minutes.

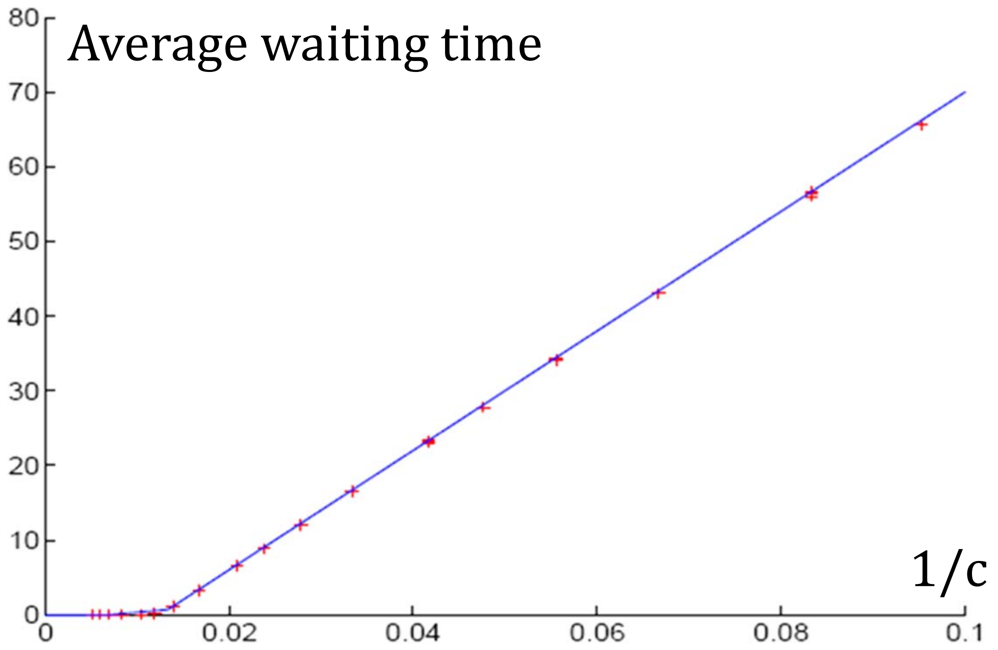
What is the impact of doubling the capacity on the average waiting time ?

A. Reduction by more than 2

B. Reduction by less than 2

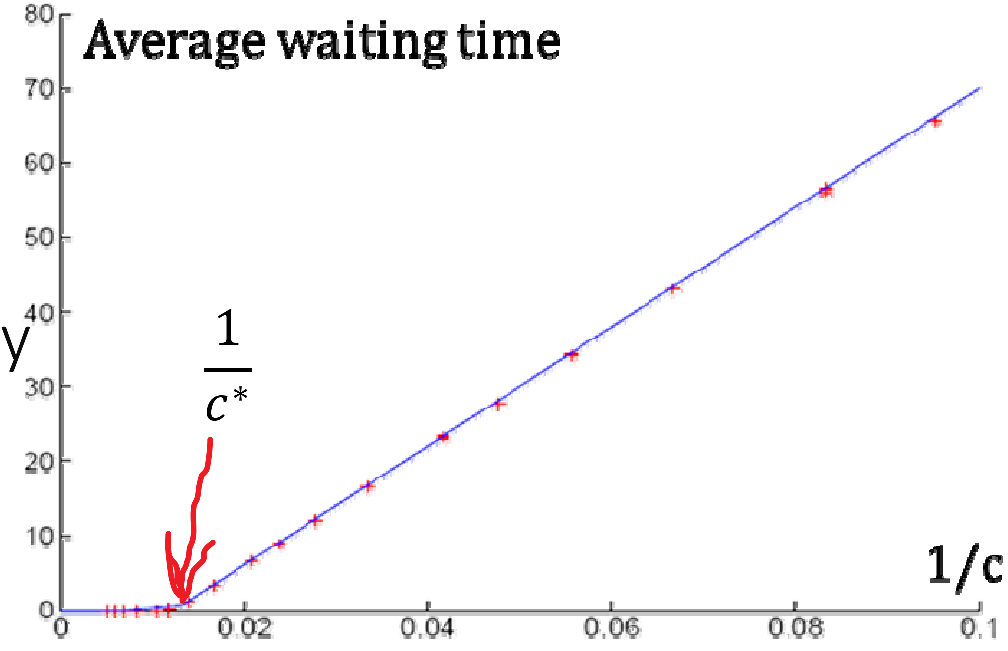
C. No reduction

D. I don’t know



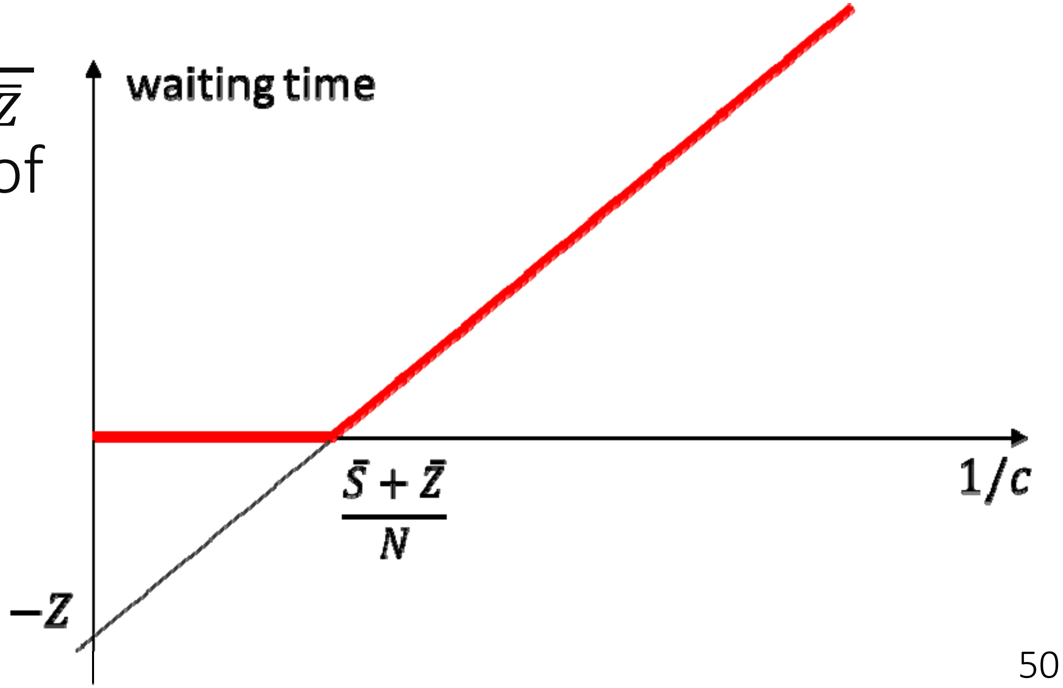
Critical Capacity

The curve shows that there is a critical capacity c^*, such that above c^* the waiting time is very small, below c^* it increases linearly



The system should be designed to operate with $c \approx c^*$

Bottleneck analysis gives $c^* \approx \frac{N}{\bar{S}+\bar{Z}}$ where N is the average number of skiers in domain and Z is the average time on slope



Conclusions

Queuing is essential in communication and information systems

M/M/1, M/GI/1, M/GI/1/PS and variants have closed forms

Little’s formula and other operational laws are powerful tools, not just for queuing systems

Bottleneck analysis and worst case analysis are usually very simple and often give good insights