Performance Modeling – Part I
Single Queues

Prof. Daniel A. Menasce
Dept. of Computer Science
George Mason University

Purpose of Models

- Provide a way to derive performance metrics from model parameters.
- Examples of performance metrics:
  - Response time
  - Throughput
  - Availability
- Types of parameters:
  - Workload intensity (e.g., arrival rates)
  - Service demands.

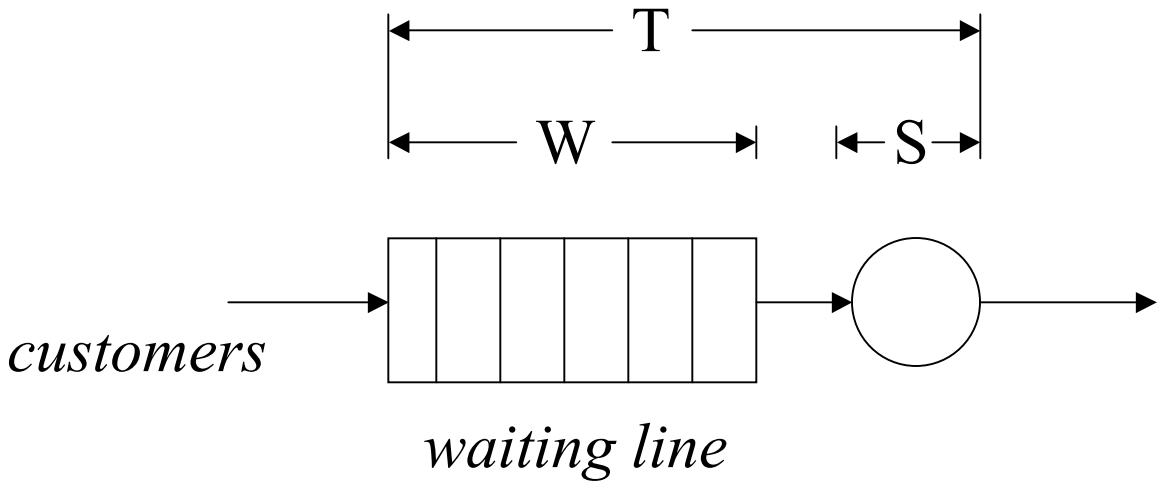
Type of Models

- Simulation: mimic flow of transactions through a system.
  - Distribution-driven
  - Trace-driven
- Analytic: set of formulas or computational algorithms.
  - Exact
  - Approximate
- Hybrid

When to Use?

- Use Exact Analytic Models Whenever Possible.
- Use Approximate Analytic Models:
  - For first-cut analysis
  - If validated by simulation
  - To reduce combinations of input parameters to simulation models.
- Use Simulation:
  - If there is no tractable analytic model.

Single Queue



$$T = W + S$$

Example of An Analytic Model M/G/1 Queue

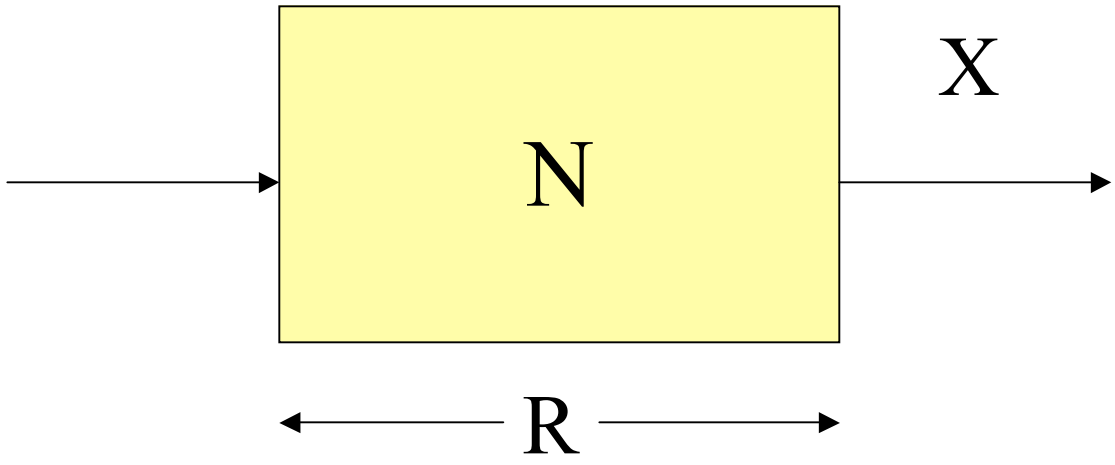
- Single server.
- Arrival process is Poisson (interarrival times are exponentially distributed).
- Service time is arbitrarily distributed.

$$T = E[S] + \frac{\lambda\, E[S^2]}{2(1-\rho)} = E[S] + \frac{\rho E[S](1+C_s^2)}{2(1-\rho)}$$

Where

$$\rho = \lambda E[S] < 1$$

Little's Law



The average number of customers in a "black box" is equal to the average time spent in the box multiplied by the throughput of the box.

$$N = R \, x \, X$$

Little's Law Example I

- An NFS server was monitored during 30 min and the number of I/O operations performed during this period was found to be 32,400. The average number of active requests (N_{req}) was 9.
- What was the average response time per NFS request at the server?

Little's Law Example I

- An NFS server was monitored during 30 min and the number of I/O operations performed during this period was found to be 32,400. The average number of active requests (N_{req}) was 9.
- What was the average response time per NFS request at the server?

"black box" = NFS server

$$X_{server} = 32{,}400 / 1{,}800 = 18 \text{ requests/sec}$$

$$R_{req} = N_{req} / X_{server} = 9 / 18 = 0.5 \text{ sec}$$

Little's Law Example II

- A large portal service offers free email service. The number of registered users is two million and 30% of them send send mail through the portal during the peak hour. Each mail takes 5.0 sec on average to be processed and delivered to the destination mailbox. During the busy period, each user sends 3.5 mail messages on average. The log file indicates that the average size of an e-mail message is 7,120 bytes.
- What should be the capacity of the spool for outgoing mails during the peak period?

Little's Law Example II

- A large portal service offers free email service. The number of registered users is two million and 30% of them send send mail through the portal during the peak hour. Each mail takes 5.0 sec on average to be processed and delivered to the destination mailbox. During the busy period, each user sends 3.5 mail messages on average. The log file indicates that the average size of an e-mail message is 7,120 bytes.
- What should be the capacity of the spool for outgoing mails during the peak period?

AvgNumberOfMails = Throughput x ResponseTime

= (2,000,000 x 0.30 x 3.5 x 5.0) / 3,600 =

2,916.7 mails

AvgSpoolFile = 2,916.7 x 7,120 bytes = 19.8 MBytes

Little's Law Example III

- A Web-based brokerage company runs a three-tiered site. The site is used by 1.1 million customers. During the peak hour, 20,000 users are logged in simultaneously. The e-commerce site processes 3.6 million business functions per hour on a peak-load hour.
- What is the average response time of an e-commerce function during the peak hour?

Little's Law Example III

- A Web-based brokerage company runs a three-tiered site. The site is used by 1.1 million customers. During the peak hour, 20,000 users are logged in simultaneously. The e-commerce site processes 3.6 million business functions per hour on a peak-load hour.
- What is the average response time of an e-commerce function during the peak hour?

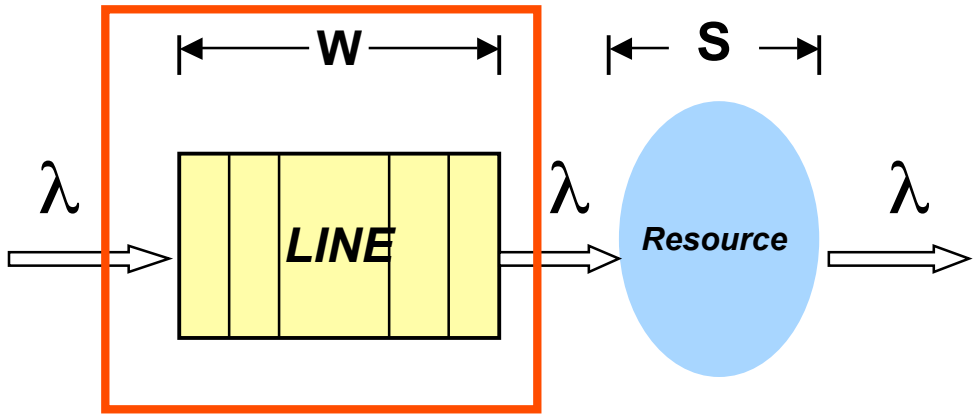
Black box = E-commerce site

AverageResponseTime = AvgNumberOfUsers / SiteThroughput

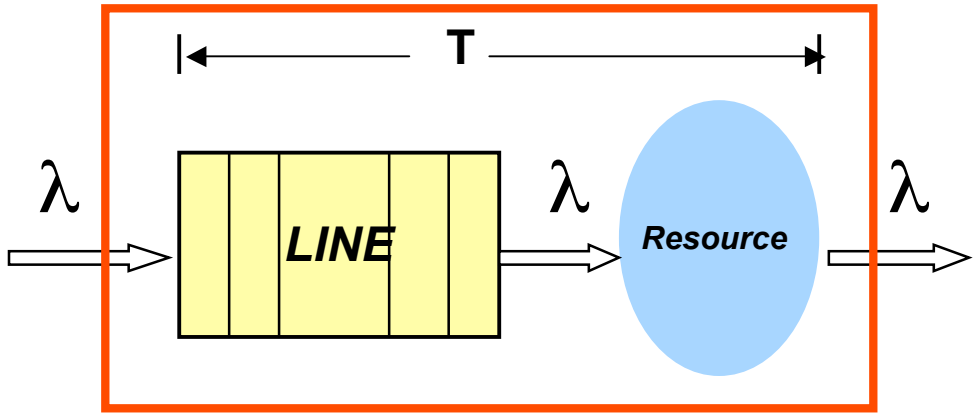
= 20,000 / (3,600,000 / 3,600) = 20 sec

Using Little's Law in the M/G/1 Queue

$$E[N_q] = \frac{\rho^2(1+C_s^2)}{2(1-\rho)}$$

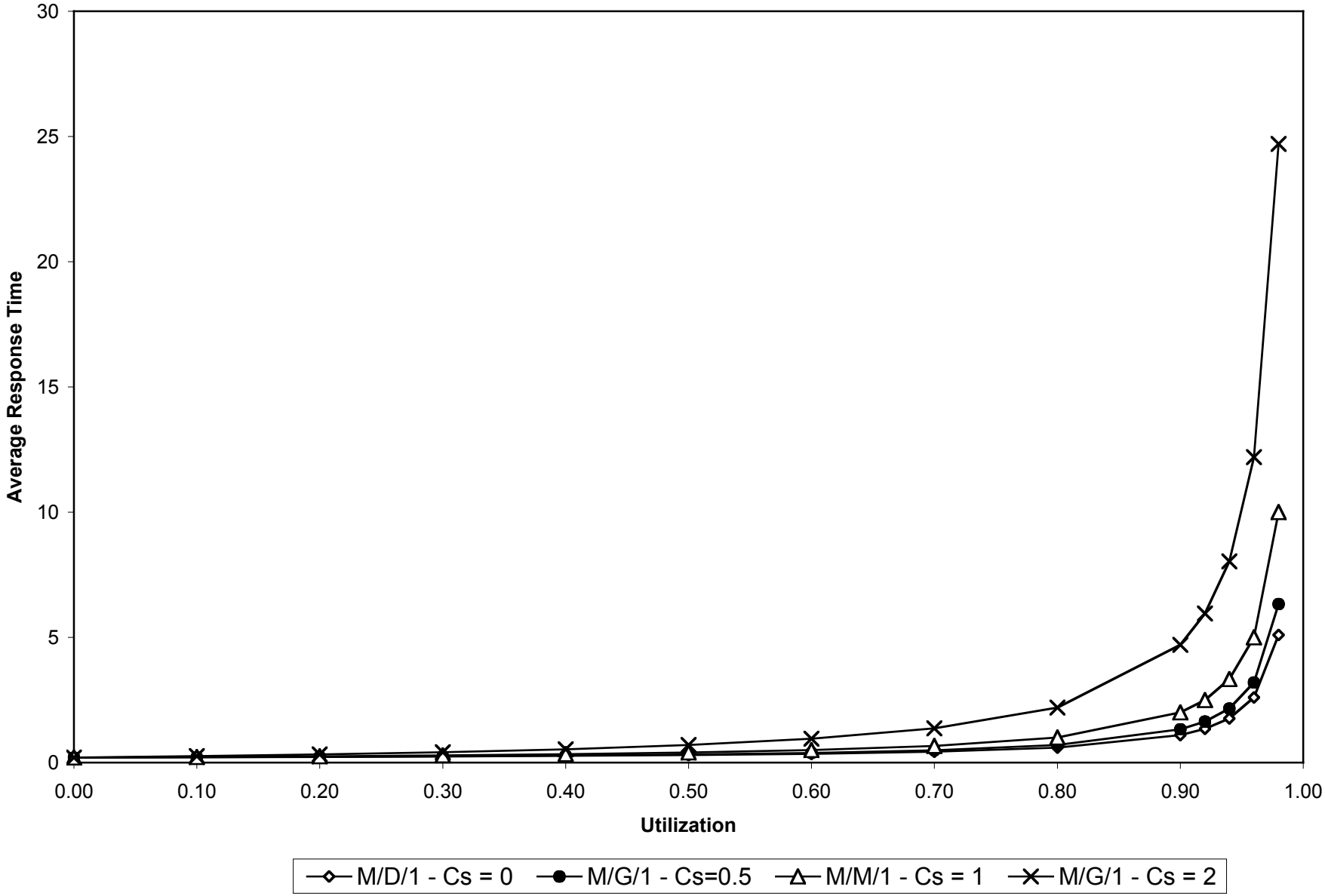


$$E[N] = \rho + \frac{\rho^2(1+C_s^2)}{2(1-\rho)}$$



Exercise

- Plot the response time for M/G/1 as a function of ρ for M/M/1, M/D/1, and distributions with coefficient of variation equal to ½ and 2. Assume that E[S] = 0.2. Vary λ accordingly.
- What conclusions do you take from looking at the graphs?



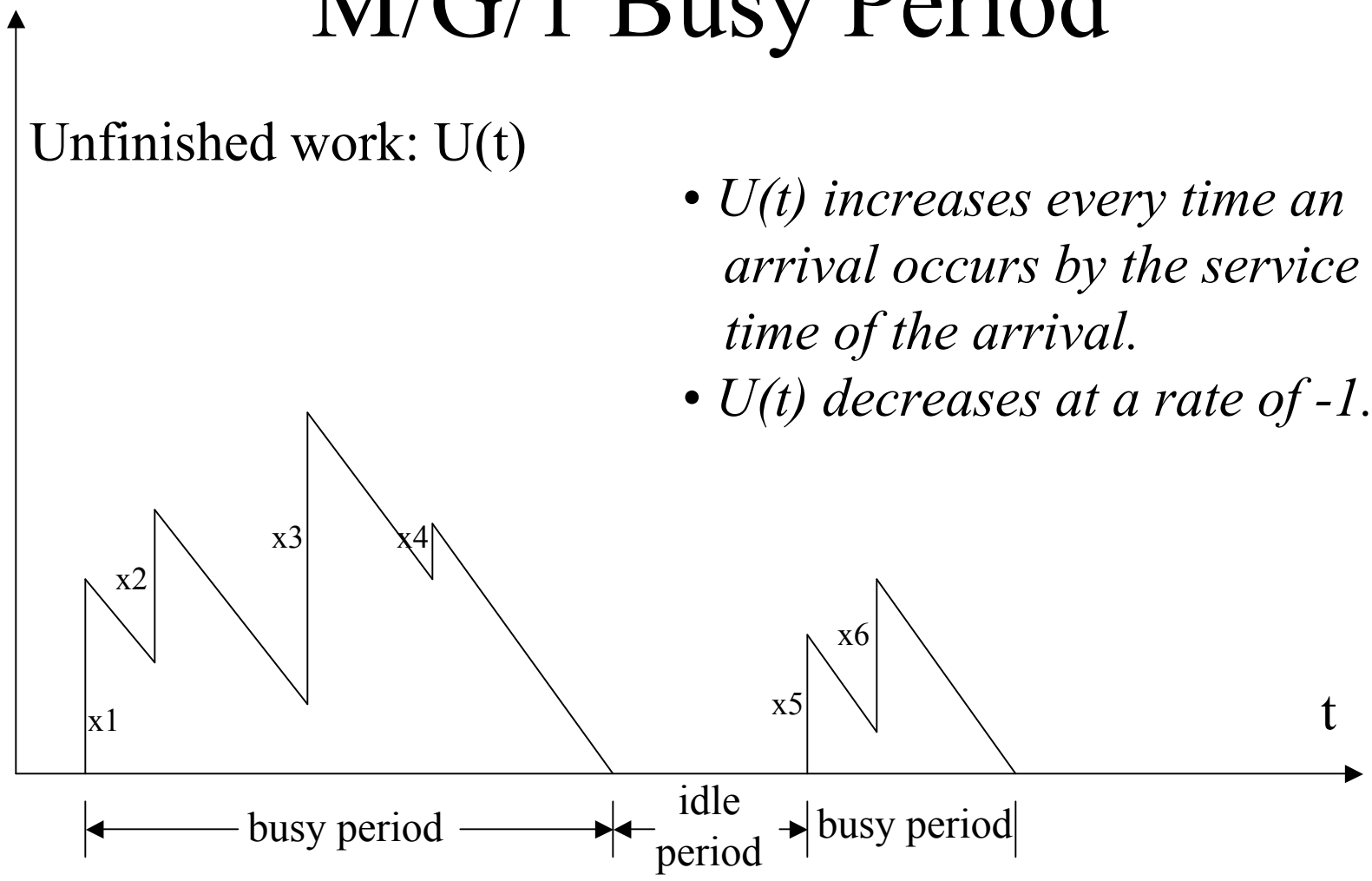
M/G/1, M/M/1, and M/D/1

M/G/1: $$W = \frac{\rho E[S](1 + C_s^2)}{2(1 - \rho)}$$

M/D/1: $$W = \frac{\rho E[S]}{2(1 - \rho)}$$

M/M/1: $$W = \frac{\rho E[S]}{(1 - \rho)}$$

M/G/1 Busy Period



- *U(t) increases every time an arrival occurs by the service time of the arrival.*
- *U(t) decreases at a rate of -1.*

M/G/1 Busy Period

- Time elapsed since the server becomes busy until it becomes idle again.

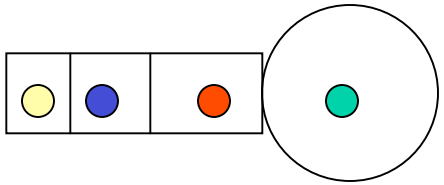
$$E[B] = \frac{E[S]}{1-\rho}$$

- Expected number of customers served in a busy period.

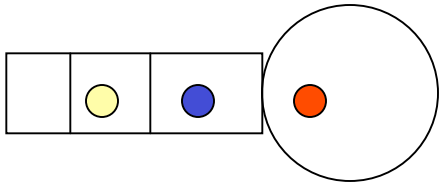
$$E[N_b] = \frac{1}{1-\rho}$$

M/G/1 With Vacation

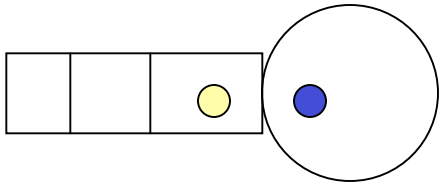
(1)



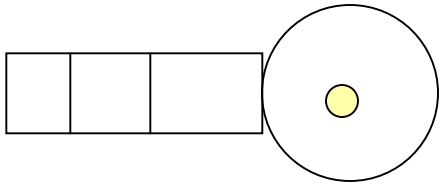
(2)



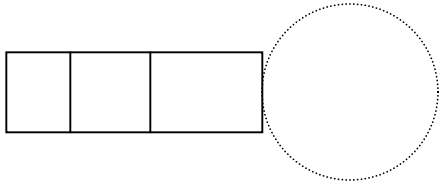
(3)



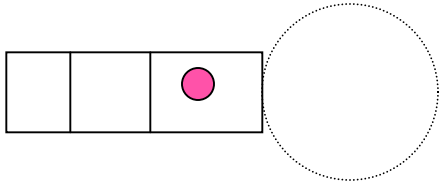
(4)



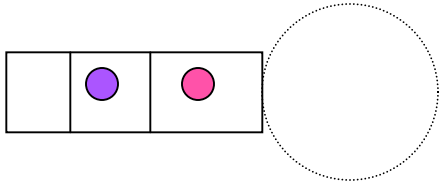
(5)



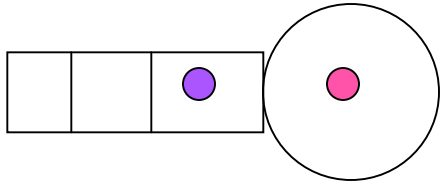
(6)



(7)



(8)



M/G/1 With Vacation

- The server goes on "vacation" for a time V (V is a generally distributed r.v.) once the server becomes idle.
- Application: polling systems.

$$W = \underbrace{\frac{\rho E[S](1+C_s^2)}{2(1-\rho)}}_{\text{regular M/G/1 waiting time}} + \frac{E[V^2]}{2E[V]}$$

Example of M/G/1 with Vacation

- A system serves requests that arrive according to a Poisson process at a rate of 0.2 requests/sec. The request processing time characteristics are: E[S] = 3.5 sec and Cs = 0.3. When there no requests to be processed, the system goes into self-diagnosis mode, which lasts for an average of 1 sec with a coefficient of variation of 2. After self-diagnosis, the system goes back to serve requests. If no requests are queued, a new self-diagnosis mode is started. What is the average waiting time of a request?

Solution to M/G/1 with Vacation Example

E[V]	W - M/G/1 no Vacation	W - M/G/1 Vacation
0.0	4.451	4.451
0.5	4.451	5.201
1.0	4.451	5.951
1.5	4.451	6.701
2.0	4.451	7.451
2.5	4.451	8.201
3.0	4.451	8.951
3.5	4.451	9.701
4.0	4.451	10.451
4.5	4.451	11.201
5.0	4.451	11.951
5.5	4.451	12.701

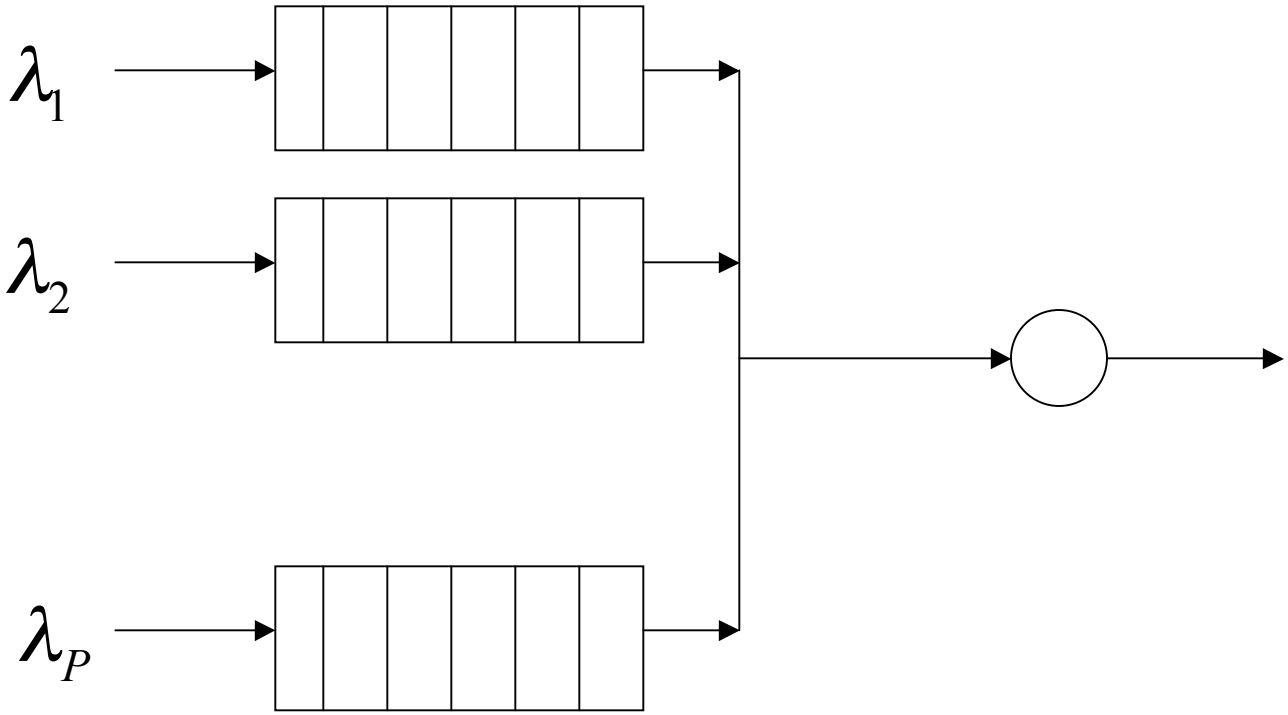
$$\sigma_V^2 = E[V^2] - (E[V])^2$$

$$\frac{\sigma_V^2}{(E[V])^2} = \frac{E[V^2]}{E[V]E[V]} - 1$$

$$\frac{(1 + C_V^2)E[V]}{2} = \frac{E[V^2]}{2E[V]}$$

M/G/1 with Priorities

- P static priorities (p= 1, …, P).
- P is the highest priority.
- FCFS within each priority queue.



M/G/1 with Non-Preemptive Priorities

$$W_p = \frac{W_0}{(1-\Pi_p)(1-\Pi_{p+1})}$$

$$W_0 = \frac{1}{2}\sum_{p=1}^{P} \lambda_p E[S_p^2]$$

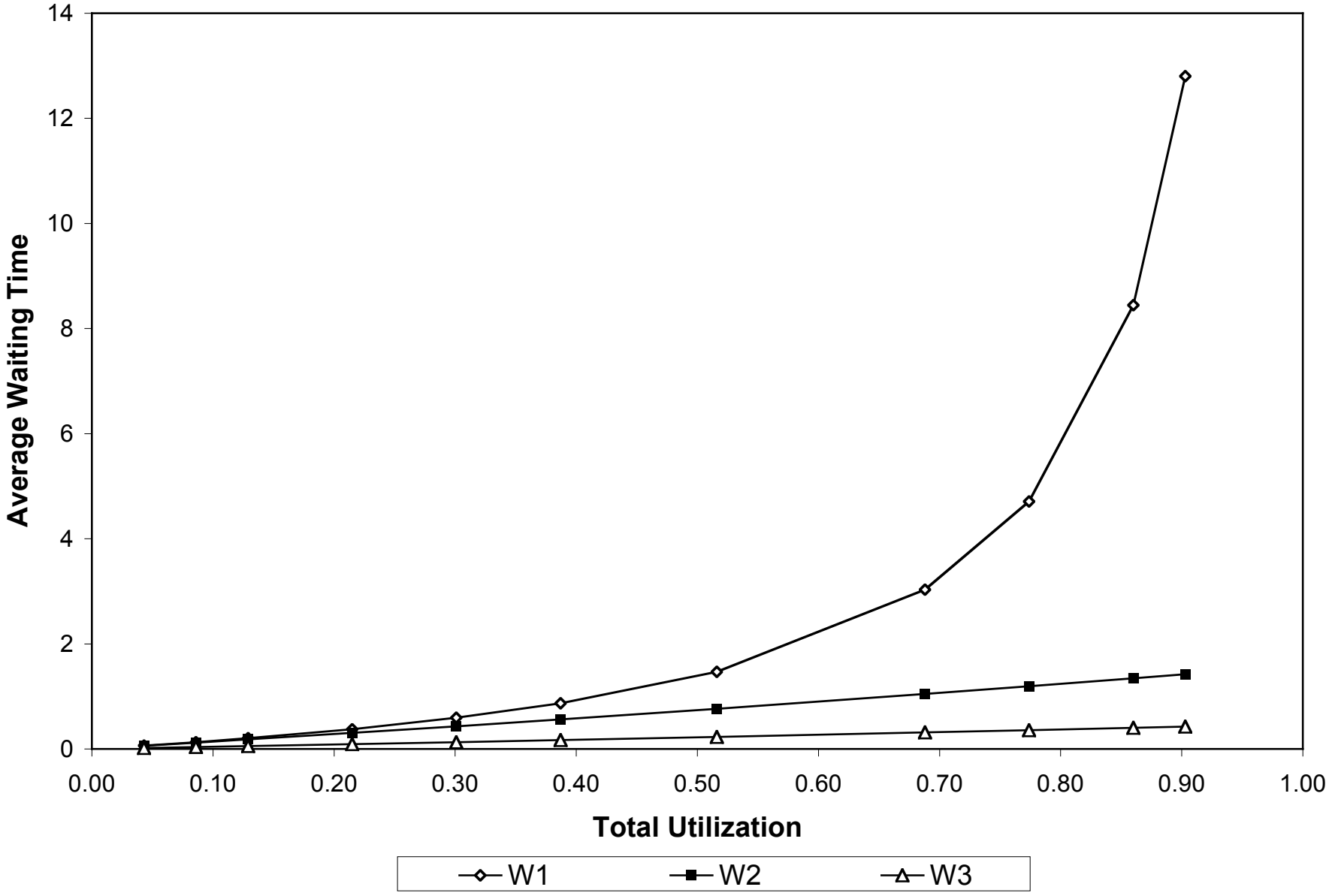
$$\rho = \sum_{p=1}^{P} \lambda_p E[S_p] = \sum_{p=1}^{P} \rho_p$$

$$\Pi_p = \sum_{i=p}^{P} \rho_i$$

Example of M/G/1 with Static Priorities

- A router receives requests at a rate of l requests/sec from a Poisson process. 50% of them are of priority 1, 30% of priority 2, and 20% of priority 3.

Priority	E[S] (sec)	E[S2] (p)
1	0.5	0.375
2	0.4	0.400
3	0.3	0.180



M/G/1 with Preemptive Resume Priorities

$$T_p = \frac{E[S_p](1-\Pi_p) + \sum_{i=p}^{P} \lambda_i E[S_i^2]/2}{(1-\Pi_p)(1-\Pi_{p+1})}$$

$$\Pi_p = \sum_{i=p}^{P} \rho_i$$

Comparing Preemptive vs. Non-Preemptive M/G/1 Queues

a	b	c	d	e	f	g=b*f	Non-preemptive	Preemptive
Priority	Lambda	E[S]	ρ	Π	E[S2]		T	T
1	1	0.50	0.500	0.775	0.375	0.3750	2.103	2.293
2	0.5	0.40	0.200	0.275	0.240	0.1200	0.790	0.543
3	0.3	0.25	0.075	0.075	0.094	0.0281	0.533	0.265

M/G/1 Distributions

Response Time: $$L_T(s) = L_S(s)\frac{s(1-\rho)}{s-\lambda+\lambda L_S(s)}$$

Laplace Transforms:

$$L_S(s) = E[e^{-s\tilde{S}}] = \int_{x=0}^{\infty} e^{-sx} f_S(x)dx$$

$$L_T(s) = E[e^{-s\tilde{T}}] = \int_{t=0}^{\infty} e^{-st} f_T(t)dt$$

M/M/1 Response Time Distribution

Laplace Transform of the response time for an exponentially distributed service time:

$$L_S(s) = \frac{1/E[S]}{s + 1/E[S]}$$

$$L_T(s) = L_S(s)\frac{s(1-\rho)}{s - \lambda + \lambda L_S(s)}$$

$$= \frac{s(1-\rho)/E[S]}{s^2 + s/E[S] - \lambda s}$$

M/M/1 Response Time Distribution

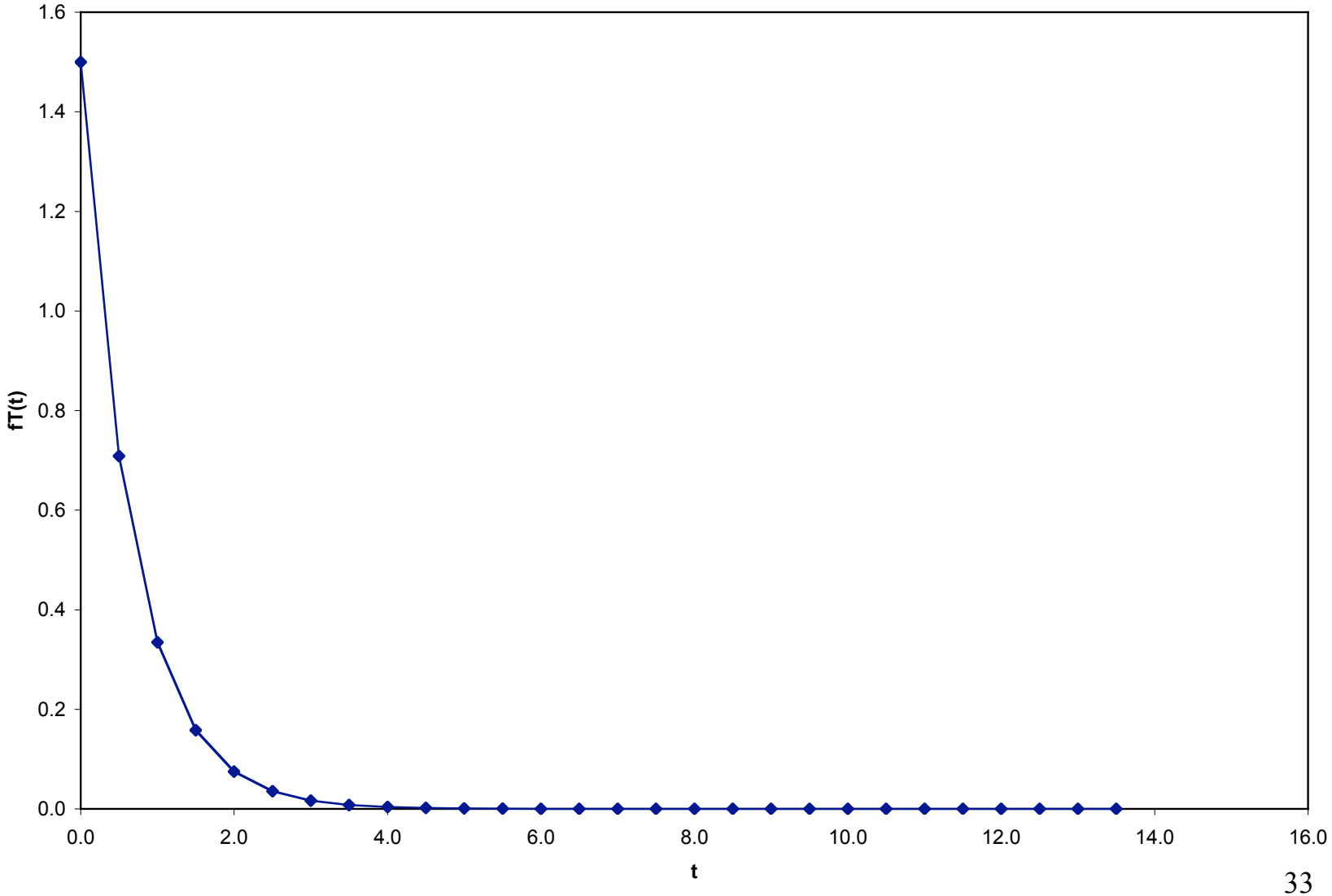
$$L_T(s) = L_S(s)\frac{s(1-\rho)}{s-\lambda+\lambda L_S(s)}$$

$$= \frac{s(1-\rho)/E[S]}{s^2 + s/E[S] - \lambda s}$$

Probability density function for the response time:

$$f_T(t) = \frac{1}{E[S]}(1-\rho)e^{-(1-\rho)t/E[S]}$$

p.d.f. for Response Time for M/M/1



M/G/1 Distributions

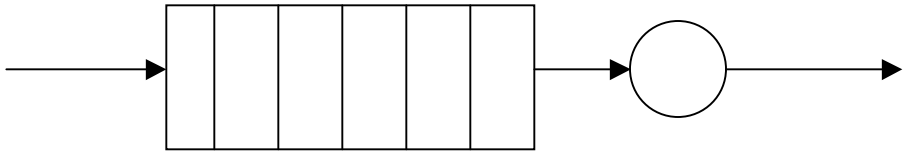
Waiting Time: $$L_W(s) = \frac{s(1-\rho)}{s - \lambda + \lambda L_S(s)}$$

Laplace Transforms:

$$L_S(s) = E[e^{-s\tilde{S}}] = \int_{x=0}^{\infty} e^{-sx} f_S(x)dx$$

$$L_T(s) = E[e^{-s\tilde{T}}] = \int_{t=0}^{\infty} e^{-st} f_T(t)dt$$

G/G/1 Queue



$$\rho = \lambda E[S] < 1$$

$$p_0 = 1 - \rho$$

An Approximation for G/G/1

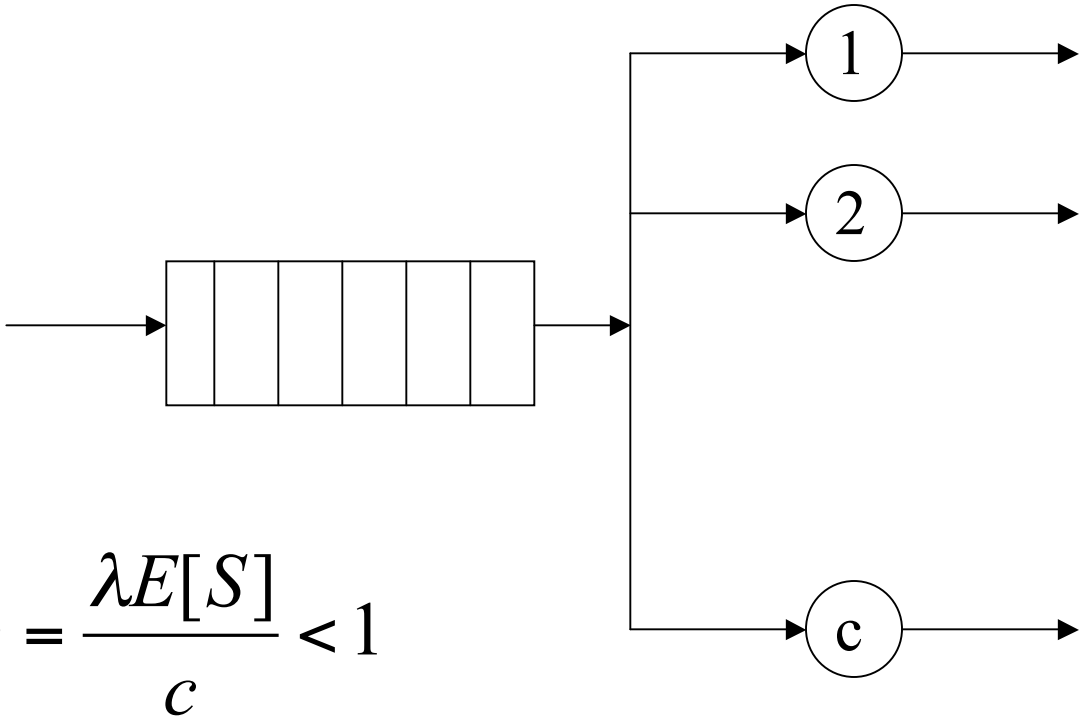
$$W \approx \frac{C_a^2 + \rho^2 C_s^2}{1 + \rho^2 C_s^2} \times \frac{\rho(1 + C_s^2)}{2(1-\rho)/E[S]}$$

C_a^2 : coefficient of variation of the interarrival time.

Approximation is exact for M/G/1, good for G/M/1, and fair for G/G/1.

The approximation improves as ρ increases.

G/G/c Queue



$$\rho = \frac{\lambda E[S]}{c} < 1$$

An Approximation for G/G/c

$$W \approx \frac{C(\rho, c)}{c(1-\rho)/E[S]} \times \frac{C_a^2 + C_s^2}{2}$$

where $C(\rho, c) = \dfrac{(c\rho)^c / c!}{(1-\rho)\sum_{n=0}^{c-1} \dfrac{(c\rho)^n}{n!} + \dfrac{(c\rho)^c}{c!}}$ is Erlang's C formula.

Approximation is exact for M/M/c.

The error increases with C_a and C_s.

The M/M/c Queue

$$W = \frac{C(\rho, c)}{c(1-\rho)/E[S]}$$

where $C(\rho, c) = \dfrac{(c\rho)^c / c!}{(1-\rho)\sum_{n=0}^{c-1} \dfrac{(c\rho)^n}{n!} + \dfrac{(c\rho)^c}{c!}}$ is Erlang's C formula.