COMP312/DATA304/DATA474
Simulation & Stochastic Models

Queuing Systems: Definitions and Notation

Alejandro C. Frery

T1 2023

VICTORIA UNIVERSITY OF WELLINGTON
TE HERENGA WAKA

School of Mathematics and Statistics
New Zealand

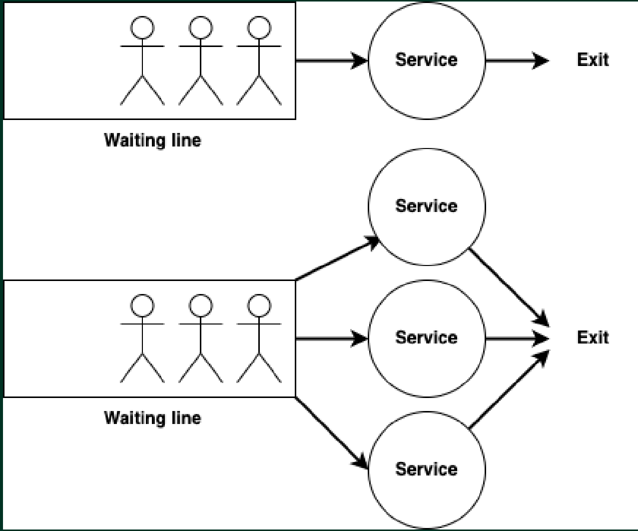
What is it about?

We will see the usual terminology with which we refer to queueing systems, and we will introduce the Poisson Process as their basic driving model.

Queuing Systems

Definition

Queueing systems consist of inputs (requests, calls, customers) that wait in line to be served by a service center.



Types of arrivals

Inputs can arrive in, typically, three regimes:

Total Control: the intervals between arrivals are perfectly known as, for instance, in the last stages of a production line.

Partial Control: the intervals between arrivals are *almost* deterministic, e.g., cargo ships that are dispatched daily at 7:00 AM.

Uncontrolled: customers in a coffee shop.

Under lack of control, the Poisson distribution is commonly used to describe the number of arrivals in a certain period of time, and the Exponential distribution is the model of choice for the time between arrivals.

Types of arrivals

Inputs can arrive in, typically, three regimes:

Total Control: the intervals between arrivals are perfectly known as, for instance, in the last stages of a production line.

Partial Control: the intervals between arrivals are *almost* deterministic, e.g., cargo ships that are dispatched daily at 7:00 AM.

Uncontrolled: customers in a coffee shop.

Under lack of control, the Poisson distribution is commonly used to describe the number of arrivals in a certain period of time, and the Exponential distribution is the model of choice for the time between arrivals.

Types of arrivals

Inputs can arrive in, typically, three regimes:

Total Control: the intervals between arrivals are perfectly known as, for instance, in the last stages of a production line.

Partial Control: the intervals between arrivals are *almost* deterministic, e.g., cargo ships that are dispatched daily at 7:00 AM.

Uncontrolled: customers in a coffee shop.

Under lack of control, the Poisson distribution is commonly used to describe the number of arrivals in a certain period of time, and the Exponential distribution is the model of choice for the time between arrivals.

Types of arrivals

Inputs can arrive in, typically, three regimes:

Total Control: the intervals between arrivals are perfectly known as, for instance, in the last stages of a production line.

Partial Control: the intervals between arrivals are *almost* deterministic, e.g., cargo ships that are dispatched daily at 7:00 AM.

Uncontrolled: customers in a coffee shop.

Under lack of control, the Poisson distribution is commonly used to describe the number of arrivals in a certain period of time, and the Exponential distribution is the model of choice for the time between arrivals.

Types of arrivals

Inputs can arrive in, typically, three regimes:

Total Control: the intervals between arrivals are perfectly known as, for instance, in the last stages of a production line.

Partial Control: the intervals between arrivals are *almost* deterministic, e.g., cargo ships that are dispatched daily at 7:00 AM.

Uncontrolled: customers in a coffee shop.

Under lack of control, the Poisson distribution is commonly used to describe the number of arrivals in a certain period of time, and the Exponential distribution is the model of choice for the time between arrivals.

Service: lines

- We may have a single server, or multiple servers.
- When there are multiple servers, there may be a single waiting line, or a waiting line for each server, or combinations of both strategies.
- The service time can be deterministic or random.

Service: lines

- We may have a single server, or multiple servers.
- When there are multiple servers, there may be a single waiting line, or a waiting line for each server, or combinations of both strategies.
- The service time can be deterministic or random.

Service: lines

- We may have a single server, or multiple servers.
- When there are multiple servers, there may be a single waiting line, or a waiting line for each server, or combinations of both strategies.
- The service time can be deterministic or random.

Service: disciplines

Requests may be acknowledged in a variety of disciplines:

- **FIFS:** First-In, First-Served[1].
- **LIFS:** Last-In, First-Served.
- **SIRO:** Served In Random Order.
- **PP:** Priority and Preemptive (as in a hospital Emergency Room).
- **PNP:** Priority and Non-Preemptive (as in a bank).

[1] What's the difference between FIFS and FIFO?

Service: disciplines

Requests may be acknowledged in a variety of disciplines:

FIFS: First-In, First-Served[1].

LIFS: Last-In, First-Served.

SIRO: Served In Random Order.

PP: Priority and Preemptive (as in a hospital Emergency Room).

PNP: Priority and Non-Preemptive (as in a bank).

[1] What's the difference between FIFS and FIFO?

Service: disciplines

Requests may be acknowledged in a variety of disciplines:

FIFS: First-In, First-Served[1].

LIFS: Last-In, First-Served.

SIRO: Served In Random Order.

PP: Priority and Preemptive (as in a hospital Emergency Room).

PNP: Priority and Non-Preemptive (as in a bank).

[1] What's the difference between FIFS and FIFO?

Service: disciplines

Requests may be acknowledged in a variety of disciplines:

FIFS: First-In, First-Served[1].

LIFS: Last-In, First-Served.

SIRO: Served In Random Order.

PP: Priority and Preemptive (as in a hospital Emergency Room).

PNP: Priority and Non-Preemptive (as in a bank).

[1] What's the difference between FIFS and FIFO?

Requests may be acknowledged in a variety of disciplines:

FIFS: First-In, First-Served[1].

LIFS: Last-In, First-Served.

SIRO: Served In Random Order.

PP: Priority and Preemptive (as in a hospital Emergency Room).

PNP: Priority and Non-Preemptive (as in a bank).

[1] What's the difference between FIFS and FIFO?

Customer's behavior in the waiting line

Customers may show a variety of behaviors, e.g.:

Balking: if the line is longer than M customers, some customers will refuse to join the system.

Rejection: the line has a maximum capacity, over which no new customers can join.

Reneging: customers in line wait until a maximum time $T_{\max}$, and leave if not served.

Jockeing: in multi-line systems, some customers may choose to move to shorter lines.

Customer's behavior in the waiting line

Customers may show a variety of behaviors, e.g.:

Balking: if the line is longer than M customers, some customers will refuse to join the system.

Rejection: the line has a maximum capacity, over which no new customers can join.

Reneging: customers in line wait until a maximum time T_{max}, and leave if not served.

Jockeing: in multi-line systems, some customers may choose to move to shorter lines.

Customer’s behavior in the waiting line

Customers may show a variety of behaviors, e.g.:

Balking: if the line is longer than M customers, some customers will refuse to join the system.

Rejection: the line has a maximum capacity, over which no new customers can join.

Reneging: customers in line wait until a maximum time T_{max}, and leave if not served.

Jockeing: in multi-line systems, some customers may choose to move to shorter lines.

Customer's behavior in the waiting line

Customers may show a variety of behaviors, e.g.:

Balking: if the line is longer than M customers, some customers will refuse to join the system.

Rejection: the line has a maximum capacity, over which no new customers can join.

Reneging: customers in line wait until a maximum time $T_{\max}$, and leave if not served.

Jockeing: in multi-line systems, some customers may choose to move to shorter lines.

Kendall Notation

Kendall (1953) proposed describing queueing models using two letters (A/S) and a number (c) denoting:

A: the model that describes the time between arrivals of customers, e.g., M – Markovian or Memoryless; D – Degenerate or Deterministic; E_k – Erlang law with shape parameter k; PH – Phase-type distribution; GI – General (to be specified) independent and identically distributed;

S: the distribution that characterizes the service times (same as above);

c: number of service channels.

Example: the M/M/1 queue.

Kendall Notation

Kendall (1953) proposed describing queueing models using two letters (A/S) and a number (c) denoting:

- **A:** the model that describes the time between arrivals of customers, e.g., M – Markovian or Memoryless; D – Degenerate or Deterministic; E_k – Erlang law with shape parameter k; PH – Phase-type distribution; GI – General (to be specified) independent and identically distributed;
- **S:** the distribution that characterizes the service times (same as above);
- **c:** number of service channels.

Example: the M/M/1 queue.

Kendall Notation

Kendall (1953) proposed describing queueing models using two letters (A/S) and a number (c) denoting:

- **A:** the model that describes the time between arrivals of customers, e.g., M – Markovian or Memoryless; D – Degenerate or Deterministic; E_k – Erlang law with shape parameter k; PH – Phase-type distribution; GI – General (to be specified) independent and identically distributed;
- **S:** the distribution that characterizes the service times (same as above);
- **c:** number of service channels.

Example: the M/M/1 queue.

Kendall (1953) proposed describing queueing models using two letters (A/S) and a number (c) denoting:

- **A:** the model that describes the time between arrivals of customers, e.g., M – Markovian or Memoryless; D – Degenerate or Deterministic; E_k – Erlang law with shape parameter k; PH – Phase-type distribution; GI – General (to be specified) independent and identically distributed;
- **S:** the distribution that characterizes the service times (same as above);
- **c:** number of service channels.

Example: the M/M/1 queue.

Kendall Notation

Kendall (1953) proposed describing queueing models using two letters (A/S) and a number (c) denoting:

A: the model that describes the time between arrivals of customers, e.g., M – Markovian or Memoryless; D – Degenerate or Deterministic; E_k – Erlang law with shape parameter k; PH – Phase-type distribution; GI – General (to be specified) independent and identically distributed;

S: the distribution that characterizes the service times (same as above);

c: number of service channels.

Example: the M/M/1 queue.

Extension by Lee

Lee (1966) extended the Kendall notation by adding a letter (D) and two numbers (K/N) denoting:

- **D:** The queue discipline (e.g., FIFS, LIFS, PNP, PP).
- **K:** capacity of queue, the maximum number of customers allowed in the queue.
- **N:** The size of the population from which the customers come (small populations significantly affect the system behaviour).

Example: $M/M/3/FIFS/10/\infty$.

Extension by Lee

Lee (1966) extended the Kendall notation by adding a letter (D) and two numbers (K/N) denoting:

D: The queue discipline (e.g., FIFS, LIFS, PNP, PP).

K: capacity of queue, the maximum number of customers allowed in the queue.

N: The size of the population from which the customers come (small populations significantly affect the system behaviour).

Example: $M/M/3/FIFS/10/\infty$.

Extension by Lee

Lee (1966) extended the Kendall notation by adding a letter (D) and two numbers (K/N) denoting:

D: The queue discipline (e.g., FIFS, LIFS, PNP, PP).

K: capacity of queue, the maximum number of customers allowed in the queue.

N: The size of the population from which the customers come (small populations significantly affect the system behaviour).

Example: $M/M/3/FIFS/10/\infty$.

Extension by Lee

Lee (1966) extended the Kendall notation by adding a letter (D) and two numbers (K/N) denoting:

- **D:** The queue discipline (e.g., FIFS, LIFS, PNP, PP).
- **K:** capacity of queue, the maximum number of customers allowed in the queue.
- **N:** The size of the population from which the customers come (small populations significantly affect the system behaviour).

Example: $M/M/3/FIFS/10/\infty$.

Extension by Lee

Lee (1966) extended the Kendall notation by adding a letter (D) and two numbers (K/N) denoting:

- **D:** The queue discipline (e.g., FIFS, LIFS, PNP, PP).
- **K:** capacity of queue, the maximum number of customers allowed in the queue.
- **N:** The size of the population from which the customers come (small populations significantly affect the system behaviour).

Example: $M/M/3/FIFS/10/\infty$.

The Kendall-Lee notation

If only A/S/c are specified, then we assume a General Independent service, infinite capacity queue, and infinite population: $A/S/c/GI/\infty/\infty$.

References

Kendall, D. G. (1953), 'Stochastic processes occurring in the theory of queues and their analysis by the method of the imbedded Markov chain', *The Annals of Mathematical Statistics* **24**(3), 338–354.

Lee, A. (1966), *Applied Queueing Theory*, Macmillan, London.