QUEUE

Authored by
mohammad looti

October 24, 2025

QUEUE

Primary Disciplinary Field(s): Social Psychology, Operations Research, Management Science, Industrial Engineering

1. Core Definition and Social Function

The term **queue** refers, fundamentally, to an orderly formation or sequence of individuals or entities--be they people, tasks, or data packets--awaiting a shared resource, service, or opportunity. This definition extends beyond the purely physical arrangement of a waiting line to encompass the structured temporary storage mechanism inherent in resource allocation. In a social context, a queue represents a spontaneous or mandated social contract designed to ensure fairness and efficiency in access. It is characterized by its adherence to established, often unwritten, rules of conduct, emphasizing concepts such as **first-come, first-served (FCFS)**, patience, and mutual respect among members. The psychological comfort derived from participating in an orderly queue stems from the perceived equity of the waiting process, mitigating potential conflict over scarce resources.

The necessity of the queue arises directly from a mismatch between demand and capacity: when the rate of arrival (demand) exceeds the rate of service (capacity) for a finite period, a waiting line forms. This phenomenon is ubiquitous in modern industrialized societies, occurring in settings ranging from retail checkout lines and traffic flows to computational task processing and telecommunications routing. Sociologically, the queue is a micro-institution that dictates behavior; individuals entering a queue tacitly agree to abide by common courtesy and recognized, if not explicitly verbalized, rules of conduct. Deviation from these rules--such as queue jumping or cutting in line--is generally met with social sanction, underscoring the powerful moral and ethical dimension embedded within the waiting process.

While the most intuitive understanding of a queue involves a physical line of people, the academic and technical study of queuing systems, particularly within operations research, generalizes this concept significantly. In this generalized view, a queue is not just the line itself but the entire system comprising the source population, the arrival process, the service facilities, and the waiting line mechanism. This formalized approach allows for the mathematical modeling and prediction of system performance, seeking to optimize the balance between the cost of providing sufficient service capacity and the cost (both economic and psychological) associated with customer waiting time.

2. Etymology and Historical Context

The word **queue** entered the English language from the French *queue*, meaning "tail" or "pigtail."

Its initial usage referred to a physical braid of hair or a train of animals or vehicles. The specific meaning of a "line of people waiting their turn" is primarily a development of the nineteenth century, becoming particularly prevalent in British English during the early 20th century, coinciding with the rise of mass urbanization, standardized public services, and centralized distribution systems. While waiting has always been a necessity, the formalized, orderly line became an essential feature of resource management during periods of high demand, such as wartime rationing or the distribution of essential services.

Historically, the concept of regulated waiting gained significant prominence during World War II, when effective resource management became critical. The need to allocate goods fairly and efficiently led to the establishment of visible, regulated waiting lines, making the queue a symbol of organized civic life. The formal mathematical study of queuing, however, predates this period, beginning with the work of Danish mathematician Agner Krarup Erlang. Erlang, working for the Copenhagen Telephone Company in the early 1900s, developed mathematical models to determine how many telephone operators were needed to handle incoming calls efficiently without excessive waiting times. His pioneering work laid the foundation for the entire field of Queuing Theory (also known as waiting-line theory), establishing the nomenclature and fundamental formulas still used today.

3. Queuing Theory (Operations Research)

Queuing theory is a branch of Operations Research dedicated to the mathematical study of waiting lines, which forms the technical backbone for understanding queues. The central aim of queuing theory is to calculate key performance metrics--such as average waiting time, average queue length, and server utilization--under various assumptions regarding arrival patterns and service times. These calculations are crucial for optimizing service delivery systems across industries, ensuring cost-effectiveness while maintaining customer satisfaction. The theory relies heavily on statistical models, particularly the Poisson distribution for arrival rates and the exponential distribution for service times, to simulate real-world variability.

A fundamental tool in this field is Kendall's Notation (A/B/C), which provides a standardized shorthand for classifying different queuing systems. 'A' describes the arrival distribution (e.g., M for Markovian/Poisson), 'B' describes the service time distribution (e.g., M for Markovian/Exponential), and 'C' describes the number of servers. The simplest and most foundational model is the M/M/1 system, representing Poisson arrivals, exponentially distributed service times, and a single server. Understanding these parameters allows managers and engineers to predict how changes in staffing, service speed, or customer flow will affect the overall system performance and customer experience.

Key theoretical concepts derived from queuing theory include **Little's Law**, which states that the average number of customers in a stable system (L) is equal to the average arrival rate (λ) multiplied by the average time a customer spends in the system (W), or $L = \lambda W$. This law is critical because it links waiting time directly to queue length, allowing analysts to infer one metric from the other. Furthermore, the concept of server utilization (ρ), which measures the proportion of time a server is busy, is vital; as utilization approaches 100%, queue lengths and waiting times increase exponentially, often leading to system collapse or failure to meet service level agreements.

4. Psychological Dimensions of Waiting

While Operations Research focuses on minimizing **actual waiting time**, social psychology and service management are profoundly concerned with minimizing **perceived waiting time**. Research has shown that how long people feel they have waited often differs significantly from the actual clock time, leading to the formulation of several psychological principles governing the waiting experience. The late service expert David Maister popularized several of these concepts, emphasizing that the human perception of time is subjective and highly influenced by environmental and situational factors.

A core psychological factor is the concept of **occupancy**: occupied time feels shorter than unoccupied time. Services that provide distractions, such as informative video screens, music, or engaging activities, effectively "fill" the wait time, reducing frustration. Furthermore, anxiety exacerbates perceived waiting time; uncertainty regarding the wait duration or the service outcome makes the wait feel interminable. Thus, providing reliable information, such as approximate wait times or progress updates, is a critical management strategy. This effect is especially pronounced when the service is highly valuable or emotionally charged (e.g., waiting for emergency medical care).

Other crucial psychological principles include the preference for **fairness** and **transparency**. Waiting systems perceived as unfair--where someone cuts the line or priority seems arbitrary--lead to severe customer dissatisfaction, irrespective of the absolute waiting duration. Additionally, it is generally accepted that solo waiting feels longer than group waiting, and that unexplained waits are perceived as longer than explained waits. The final moments of the waiting experience are also disproportionately influential; people tend to remember the end of an experience most vividly, meaning that a positive service initiation or a quick end to the queue can mitigate the negative perception of a long initial wait.

5. Key Characteristics of Queuing Systems

A complete queuing system is defined by five interconnected characteristics that dictate its operational behavior and performance. Understanding these characteristics is essential for modeling and improving service efficiency.

Source Population (or Calling Population): This refers to the potential pool of customers who may require service. The population can be considered either **infinite** (very large, where the departure of one customer does not significantly affect the arrival rate of others, e.g., a supermarket) or **finite** (small, where the number of potential arrivals is severely limited and affected by who is already in the system, e.g., five machines that might break down).

Arrival Process: This describes the pattern by which customers arrive at the system. The arrival rate is typically modeled using a probability distribution, most commonly the Poisson distribution, which implies random and independent arrivals. The degree of patience among customers (whether they will wait indefinitely, abandon the queue after a certain time, or balk and refuse to join) also defines the arrival process dynamics.

Queue Configuration: This refers to the physical or virtual layout of the waiting area. Common configurations include the single-line, single-server system (e.g., a typical bank teller), the single-line, multiple-server system (e.g., airport security screening), and the multiple-line, multiple-server system (e.g., standard supermarket checkouts). The configuration significantly impacts perceived fairness and actual efficiency.

Queuing Discipline: This is the rule used to select the next customer from the line for service. The discipline ensures order and predictability. The most common is **First-In, First-Out (FIFO)**, but others include Last-In, First-Out (LIFO), Priority (based on need or importance), and Shortest Processing Time (SPT).

Service Process: This defines the rate and duration of the service provided. Service time is often modeled using the exponential distribution, assuming service times are random but centered around an average. The service process also includes the number of servers, the physical layout of the service facility, and the quality of the service itself.

6. Queuing Disciplines and Efficiency

The choice of queuing discipline is a crucial management decision that balances perceived fairness against objective efficiency. While FIFO is the standard and socially accepted discipline, other methods are often utilized to optimize resource usage or address urgent needs.

The **FIFO/FCFS** discipline ensures that customers are served in the exact order of their arrival. Its primary benefit is that it is inherently perceived as the fairest system, reducing customer complaints related to inequity. However, FIFO is not always the most efficient in terms of overall system throughput, especially if there are highly variable service times. For example, a customer with a very quick request might be stuck behind a customer with a very long, complex request, delaying service for many people unnecessarily.

Conversely, systems using **Priority** discipline select customers based on pre-defined criteria, regardless of when they arrived. This is common in healthcare (triage systems) or in IT support (urgent vs. low-priority tickets). While highly effective at managing critical needs and optimizing resource allocation to high-value tasks, priority systems must be transparently managed to avoid perceptions of unfairness among low-priority customers. Similarly, the **Shortest Processing Time (SPT)** rule prioritizes tasks that can be completed quickly, which often maximizes the total number of items processed, thereby reducing the average waiting time across the entire system.

In the realm of computing and data processing, non-physical queues often use **LIFO (Last-In, First-Out)**, particularly in managing memory stacks, where the last item added is the first item to be retrieved. This contrasts starkly with social queues but proves highly efficient for recursive functions or localized data management. The decision of which discipline to implement fundamentally relies on the organization's primary objective: maximizing throughput, minimizing average wait time, or maximizing perceived customer equity.

7. Management Strategies for Customer Experience

Effective queue management moves beyond simply installing more servers; it requires a strategic focus on the customer experience and the application of psychological principles to mitigate the negative effects of waiting. Modern strategies typically focus on three areas: operational efficiency (reducing actual wait time), experience management (reducing perceived wait time), and system transparency.

One highly successful operational strategy is the adoption of the **single-line, multiple-server configuration** (often called the snake line or serpentine queue). While the total waiting time might be mathematically similar to multiple independent lines, this configuration offers several key advantages. Psychologically, it eliminates the "checkout lottery" problem, where customers must guess which line will move fastest, thereby reducing frustration. Operationally, it ensures 100% server utilization, as the person at the head of the queue automatically moves to the next available server, optimizing flow and guaranteeing FCFS fairness.

Experience management often involves diverting customer attention or providing productive activities. Retail environments frequently utilize strategic merchandising within the queue area to encourage impulse buying, turning waiting time into revenue-generating time. Service providers, such as hospitals or government offices, invest in clean, comfortable, and well-lit waiting areas, recognizing that physical comfort and environmental quality directly influence the perception of the wait. Furthermore, implementing technologies like mobile check-in and virtual queue management systems (where customers wait remotely or are notified via phone) eliminates the physical need to stand in line, shifting the burden of waiting away from the customer's active time.

Ultimately, the goal is to manage expectations. By providing accurate, real-time wait information (transparency) and ensuring that the service provided is worth the wait, organizations can significantly enhance customer satisfaction. The principle holds true that waits of unknown duration are perceived as longer than known waits, and waits that are perceived as fair are tolerated better than waits perceived as unjust. Effective queue management, therefore, is a holistic endeavor combining mathematical optimization with behavioral science.

Further Reading

Queueing Theory (Wikipedia)
Operations Research (Wikipedia)
Agner Krarup Erlang Biography
FIFO (First-In, First-Out) Principle