Queueing Systems: Lecture 1

Amedeo R. Odoni
October 6, 2004

Topics in Queueing Theory

- Introduction to Queues
- Little's Law
- Markovian Birth-and-Death Queues
- The M/M/1 and Other Related Queues
- The M/G/1 Queue and Extensions
- Priority Queues
- Some Useful Bounds
- Congestion Pricing
- Queueing Networks; State Representations
- Dynamic Behavior of Queues

Lecture Outline

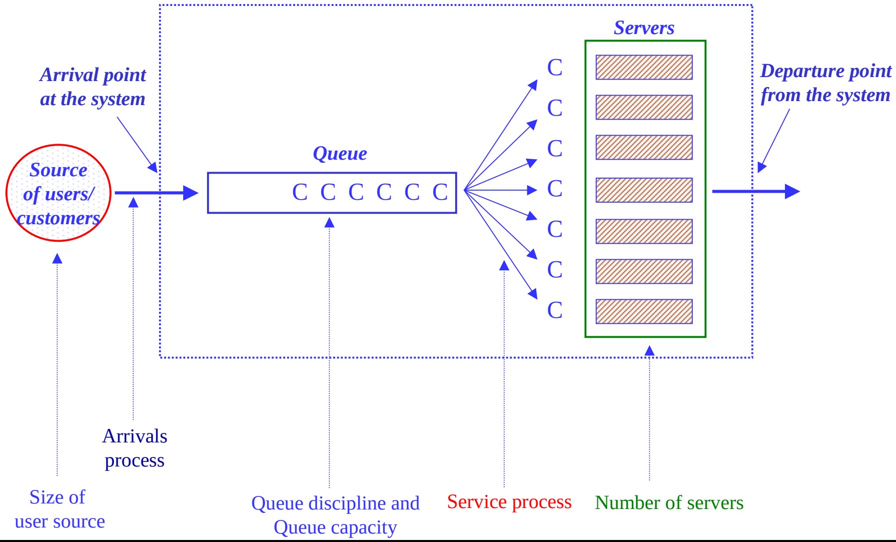
- Introduction to queueing systems
- Conceptual representation of queueing systems
- Codes for queueing models
- Terminology and notation
- Little's Law and basic relationships

Reference: Chapter 4, pp. 182-193

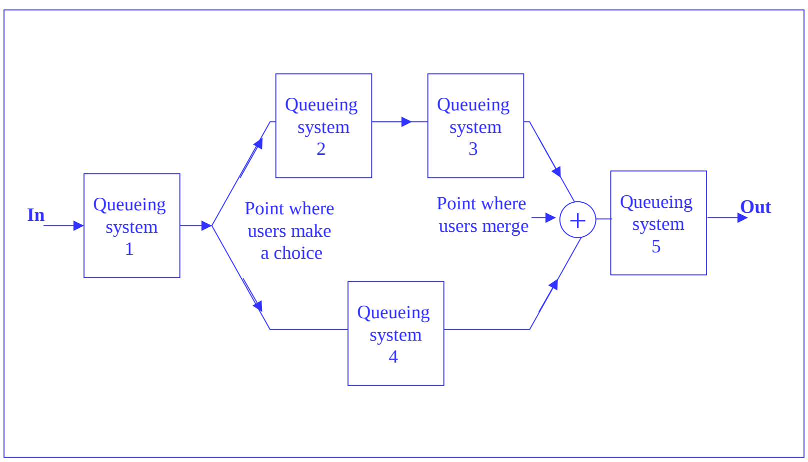
Queues

- Queueing theory is the branch of operations research concerned with waiting lines (delays/congestion)
- A queueing system consists of a user source, a queue and a service facility with one or more identical parallel servers
- A queueing network is a set of interconnected queueing systems
- Fundamental parameters of a queueing system:
  - Demand rate - Capacity (service rate)
  - Demand inter-arrival times - Service times
  - Queue capacity and discipline (finite vs. infinite; FIFO/FCFS, SIRO, LIFO, priorities)
  - Myriad details (feedback effects, "balking", "jockeying", etc.)

A Generic Queueing System



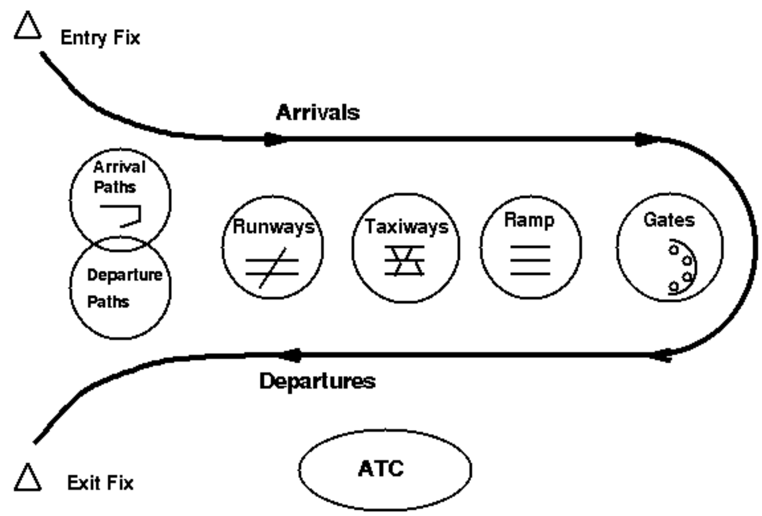
Queueing network consisting of five queueing systems



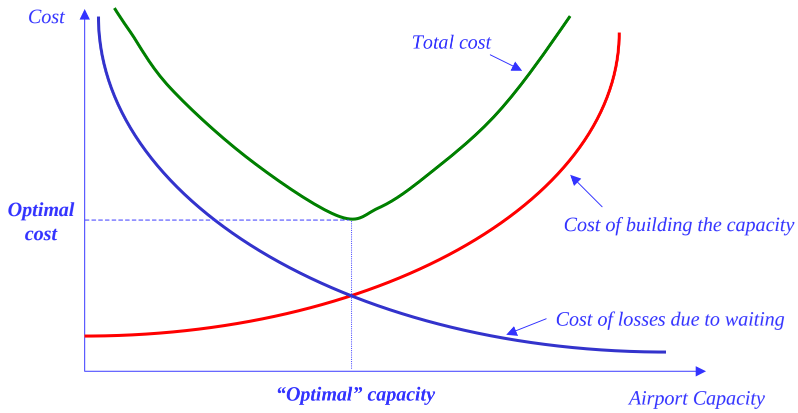
Applications of Queueing Theory

- **Some familiar queues:**
  - _ Airport check-in; aircraft in a holding pattern
  - _ Automated Teller Machines (ATMs)
  - _ Fast food restaurants
  - _ Phone center's lines
  - _ Urban intersection
  - _ Toll booths
  - _ Police or other spatially distributed urban services
- **Level-of-service (LOS) standards**
- **Economic analyses involving trade-offs among operating costs, capital investments and LOS**
- **Congestion pricing**

The Airside as a Queueing Network



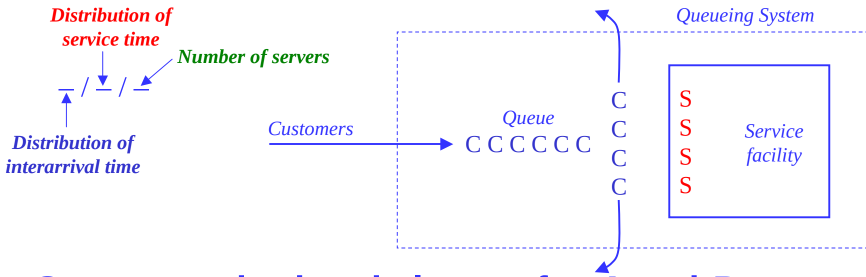
Queueing Models Can Be Essential in Analysis of Capital Investments



Strengths and Weaknesses of Queueing Theory

- Queueing models necessarily involve approximations and simplification of reality
- Results give a sense of order of magnitude, changes relative to a baseline, promising directions in which to move
- Closed-form results essentially limited to "steady state" conditions and derived primarily (but not solely) for birth-and-death systems and "phase" systems
- Some useful bounds for more general systems at steady state
- Numerical solutions increasingly viable for dynamic systems
- Huge number of important applications

A Code for Queueing Models: *A*/*B*/*m*



- **Some standard code letters for *A* and *B*:**
  - _ *M: Negative exponential* (*M stands for memoryless*)
  - _ *D: Deterministic*
  - _ *E_k:kth-order Erlang distribution*
  - _ *G: General distribution*

Terminology and Notation

- ***Number in system*: number of customers in queueing system**
- ***Number in queue* or "*Queue length*": number of customers waiting for service**
- ***Total time in system* and *waiting time***
- **$N(t)$ = number of customers in queueing system at time t**
- **$P_n(t)$ = probability that $N(t)$ is equal to n at time t**
- **λ_n: mean arrival rate of new customers when $N(t) = n$**
- **μ_n: mean (total) service rate when $N(t) = n$**

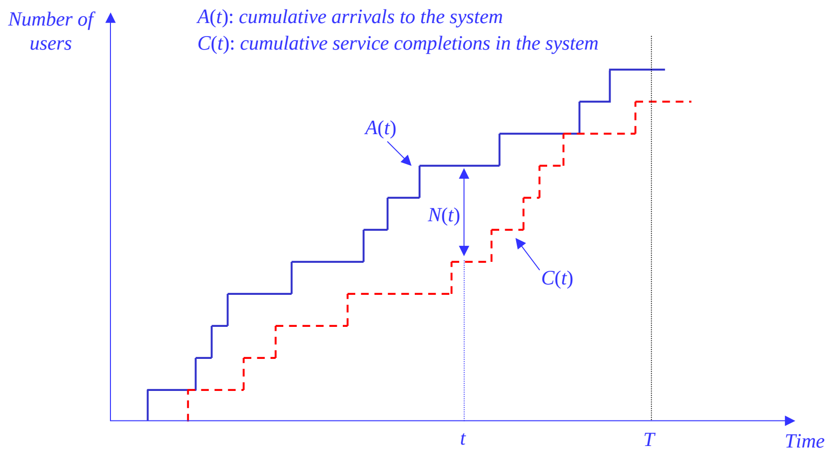
Terminology and Notation (2)

- ***Transient state***: state of system at t is influenced by the state of the system at t = 0
- ***Steady state***: state of the system is independent of initial state of the system
- m: number of servers (parallel service channels)
- If λ_n and the service rate per busy server are constants, λ and μ, respectively, then $\lambda_n = \lambda$, $\mu_n = \min(n\mu, m\mu)$; in this case:
  - Expected inter-arrival time = $1/\lambda$
  - Expected service time = $1/\mu$

Some Expected Values of Interest at Steady State

- ***Given:***
  - λ = arrival rate
  - μ = service rate per service channel
- ***Unknowns:***
  - L = expected number of users in queueing system
  - L_q = expected number of users in queue
  - W = expected time in queueing system per user ($W = E(w)$)
  - W_q = expected waiting time in queue per user ($W_q = E(w_q)$)
- 4 unknowns $\Rightarrow$ We need 4 equations

Little's Law



$$L_T = \frac{\int_0^T N(t)dt}{T} = \frac{A(T)}{T} \cdot \frac{\int_0^T N(t)dt}{A(T)} = \lambda_T \cdot W_T$$

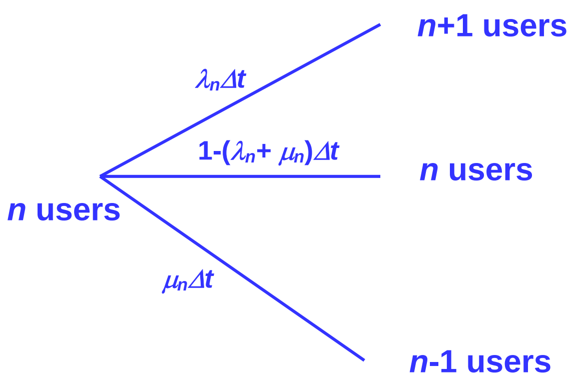
Relationships among L, L_q, W, W_q

- **Four unknowns: L, W, L_q, W_q**
- **Need 4 equations. We have the following 3 equations:**
  - $L = \lambda W$ (Little's law)
  - $L_q = \lambda W_q$
  - $W = W_q + \frac{1}{\mu}$
- **If we can find any one of the four expected values, we can determine the three others**
- **The determination of L (or other) may be hard or easy depending on the type of queueing system at hand**
- $L = \sum_{n=0}^{\infty} nP_n$ (P_n : probability that n customers are in the system)

Birth-and-Death Queueing Systems

1. m parallel, identical servers.
2. Infinite queue capacity (for now).
3. Whenever *n* users are in system (in queue plus in service) arrivals are Poisson at rate of λ_n per unit of time.
4. Whenever *n* users are in system, service completions are Poisson at rate of μ_n per unit of time.
5. FCFS discipline (for now).

The Fundamental Relationship



$P_n(t)$ = Prob [*n* users in system at time *t*]

$$P_n(t+\Delta t) = P_{n+1}(t)\cdot\mu_{n+1}\cdot\Delta t + P_{n-1}(t)\cdot\lambda_{n-1}\cdot\Delta t + P_n(t)\cdot[1-(\mu_n+\lambda_n)\cdot\Delta t]$$

The differential equations that determine the state probabilities

$$P_n(t+\Delta t) = P_{n+1}(t)\cdot\mu_{n+1}\cdot\Delta t + P_{n-1}(t)\cdot\lambda_{n-1}\cdot\Delta t + P_n(t)\cdot[1-(\mu_n+\lambda_n)\cdot\Delta t]$$

After a simple manipulation:

$$\frac{dP_n(t)}{dt} = -(\lambda_n+\mu_n)\cdot P_n(t) + \lambda_{n-1}\cdot P_{n-1}(t) + \mu_{n+1}\cdot P_{n+1}(t) \qquad (1)$$

(1) applies when *n* = 1, 2, 3,….; when *n* = 0, we have:

$$\frac{dP_0(t)}{dt} = -\lambda_0\cdot P_0(t) + \mu_1\cdot P_1(t) \qquad (2)$$

• The system of equations (1) and (2) is known as the Chapman-Kolmogorov equations for a birth-and-death system

The "state balance" equations

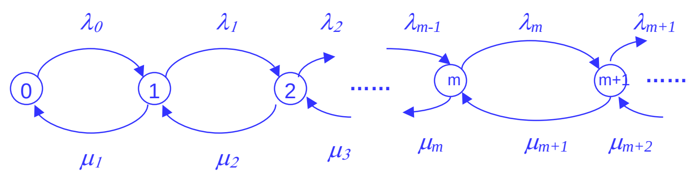
- **We now consider the situation in which the queueing system has reached "steady state", i.e., *t* is large enough to have $P_n(t) = P_n$, independent of t, or** $\frac{dP_n(t)}{dt} = 0$
- **Then, (1) and (2) provide the state balance equations:**

$$\lambda_0\cdot P_0 = \mu_1\cdot P_1 \qquad n=0 \qquad (3)$$

$$(\lambda_n+\mu_n)\cdot P_n = \lambda_{n-1}\cdot P_{n-1} + \mu_{n+1}\cdot P_{n+1} \qquad n=1,2,3,.. \qquad (4)$$

- **The state balance equations can also be written directly from the state transition diagram**

Birth-and-Death System: State Transition Diagram



- We are interested in the characteristics of the system under equilibrium conditions ("steady state"), i.e., when the state probabilities $P_n(t)$ are independent of t for large values of t
- Can write system balance equations and obtain closed form expressions for P_n, L, W, L_q, W_q

Solving.....

Solving (3) and (4), we have:

$$P_1 = \frac{\lambda_0}{\mu_1} \cdot P_0; \quad P_2 = \frac{\lambda_1}{\mu_2} \cdot P_1 = \frac{\lambda_1 \cdot \lambda_0}{\mu_2 \cdot \mu_1} \cdot P_0 \qquad etc.$$

and, in general,

$$P_n = \frac{\lambda_{n-1} \cdot \lambda_{n-2} \cdot \ldots\ldots \cdot \lambda_1 \cdot \lambda_0}{\mu_n \cdot \mu_{n-1} \cdot \ldots\ldots \cdot \mu_2 \cdot \mu_1} \cdot P_0 = K_n \cdot P_0$$

But, we also have: $1 = \sum_{n=0}^{\infty} P_n = P_0 \cdot (1 + \sum_{n=1}^{\infty} K_n)$

Giving, $$P_0 = \frac{1}{1 + \sum_{n=1}^{\infty} K_n}$$

Condition for steady state:

$$\sum_{n=1}^{\infty} K_n < \infty$$