5 TRANSPORTATION QUEUEING

Randolph W. Hall

5.1 Introduction

Since the time that humans first gathered into societies, there have been queues. They have existed whenever people have demanded more of a service than that service could provide. Though queueing is by no means new, the study of queues is relatively recent, dating only to the beginning of the twentieth century and the work of A.K. Erlang (Brockmeyer *et al*, 1948). Erlang's investigations centered on determining capacity requirements for telephone systems, a then very new technology. Even to this day, much of the research in queueing has been directed at applications in communication. The first textbook on the subject, *Queues, Inventories and Maintenance,* was written in 1958 by Morse. The first textbook focusing on queueing applications in transportation (*Applications of Queueing Theory*) was written by Newell in 1971.

Research on queueing in transportation has evolved in its own distinct direction, in part due to the influence of Newell's work, and in part due to the unique aspects of transportation systems. Unlike applications of queueing in communication or production, queues in transportation tend to be much more predictable and, as a consequence, much of the research on queues in transportation has been directed at non-stationary (time varying) systems. Non-stationarities arise in transportation because:

- People prefer to travel at set times of the day and week, largely corresponding to their work schedules. These demand surges create much of the queueing in transportation, and
- In many transportation systems (e.g., mass transit, trucking, railroads and intersections), customers are served in bulk.

From the standpoint of capacity provision, transportation often relies on major investments in infrastructure, such as roadways, runways or railroad lines. Infrastructure intensive systems have only limited latitude for adjusting capacity to responsd to fluctuating demand. Thus, queues recur at known times when customers arrive at a faster rate than the infrastructure can accommodate.

Another unique aspect of transportation is that the customer service mechanism is often defined by the spacing between vehicles along a guideway, and not by how quickly a person or piece of equipment can process customers. Thus, the time to serve a customer is determined by the customer's behavior. A queueing system also behaves as a continuum of serial servers, with extremely short service times, interacting with each other. Therefore, the system model depends not only on the number of customers that queue at a particular location, but the physical length of that queue, and whether that queue spills back into other servers. These phenomena are the core subject matter of *traffic flow theory,* covered in Chapter 6.

Finally, transportation is different from most other queueing applications because the service mechanism is frequently government owned. As a consequence, pricing normally is not used to level out demand patterns, and there tends to be much less flexibility in varying capacity to match fluctuating demand.

Most textbooks in queueing theory emphasize modeling stochastic characteristics of queues that occur in steady-state (i.e., the probability distribution for the state of the system is not time dependent). Unfortunately, for the reasons mentioned above, this theory is not always relevant to transportation. Instead, queueing models in transportation are more likely to concentrate on the non-stationary characteristics of queueing, as well as on the optimization of system design and system control. Examples include:

- Determining the best cycle length and phase lengths for traffic signals.
- Evaluating the consequences of adding lanes or changing the geometric configuration of a highway on "recurrent" (peak period) and "non-recurrent" (incident produced) delay.
- Optimizing the frequency at which buses or trucks should be dispatched along a route, taking cost of operation and service quality into consideration.
- Determining how many service vehicles are needed to respond to randomly occuring demand that is spread over a service region.

5.2 Elements of a Queueing System

Queueing systems are defined by three elements: customers, servers and queues. Customers are the persons or things that await service. They can be travelers, or the vehicles that they travel in. Customers can also be the good, piece of freight or container that is being shipped. The server is the resource that provides the service to the customer. It could be a piece of roadway, a bus, or gate in an airport, to name a few examples. The queue is the group of customers waiting to be served, along with the place they are waiting. Queues can occur as orderly lines, but they also can be groups of customers spread out in a terminal waiting area or perhaps a warehouse. All queueing systems have customers and servers, though occasionally they don't have queues. This occurs when the system refuses to accept customers when they cannot be served immediately.

The performance of the queueing system is defined by the arrival process, service process and queue discipline. The arrival process represents the time pattern by which customers enter the queueing system. Arrival processes in transportation are usually non-stationary, meaning the average arrival rate varies in some predictable way. Arrival processes also exhibit some level of stochastic variation, which is usually represented by the probability distribution for the inter-arrival time (the time separation between two successive arrivals). The service process represents the time and resources required to serve a customer. Service process, like arrival processes, exhibit stochastic variations and often non-stationary patterns (when capacity varies by time). The service time can also depend on the type of customer. The queue discipline is the rule for sequencing customers. Typically, this is a first-come-first-serve pattern. However, other disciplines are used to account for priorities, or to group customers for efficiency (such as a traffic signal, which groups by turn pattern).

Queueing systems are important in transportation because of their effects on customers, and because of the cost of providing the service. The dominant effect is delay, which might be measured in such ways as "time in system", "average speed," or "waiting time." Fundamentally, queueing analysis is used to determine the difference between how long it takes to complete a trip, and how long it would have taken if there were no queueing or congestion. The following are examples of the performance measures that can be predicted with queueing models or measured in the field:

Throughput: Rate at which customers are processed by the system

Crowding or Congestion: Separation between customers, or density of customers (e.g., vehicles per lane-mile of roadway).

Lost Customers: Number of customers that do not travel because of queueing.

Queue Percentage: Percentage of customers that encounter a queue prior to receiving service (instead of being served immediately).

Service Cost: The annual or per customer expense of providing the service.

Productivity: In some cases, the productivity of the server depends on the amount of queueing and whether the system is saturated.

In some instances, queues are stochastic, reflecting a momentary surge in demand or drop in capacity. In others, queues are predictable, following a regular daily pattern. And in some cases queues are perpetual, being present whenever a facility is open for business. One of the objectives in designing a queueing system is remove perpetual and predictable queues, and then to minimize the occurrence of stochastic queues.

5.3 History of Research on Transportation Queueing

Nearly all of the published research on queueing in transportation is motivated by a modal application, such as vehicles on roadways or mass transit. Nevertheless, there is considerable cross-over in concepts and methods between modes. This section provides a few examples.

Traffic: Vehicular Flow on Highways

Controlled access highways were first constructed in the 1930s and 1940s, and only became widely available in the United States in the 1950s and 1960s. Even in the 1990s, they are uncommon in many parts of the world. Research on queueing on highways paralleled this pattern, with the 1950s and 60s seeing a surge of activity, with more or less steady activity ever since. To this day, problems in highway traffic flow have influenced our understanding of queueing phenomena more than any other mode of transportation (for instance, see Hankin and Wright, 1958; Lovas, 1994; and Older, 1968; as examples of how vehicular traffic research has influenced modeling of pedestrian traffic). Its three greatest contributions have been: (1) modeling speed and capacity as functions of vehicle concentrations, (2) modeling the formation and size of queues with shock waves, and (3) application of cumulative diagrams to represent non-stationary phenomena. Secondarily, it has stimulated thinking on congestion pricing, though this research has yet to be applied in a significant way.

Queues on highways are typically manifest in slowed, rather than completely stopped, traffic, making queues difficult both to count and model. It was observed as early as 1935 (Greenshields), that traffic has a natural tendency to slow as the vehicle concentration (vehicles per unit length of roadway) increases, because vehicles naturally reduce speed to provide safe spacing. Extremely large concentrations only occur under jammed conditions, when both vehicle speeds and vehicle flows (product

of concentration and speed) are small. Vehicle flows are maximized at moderate concentrations, when vehicle speeds are only slightly impeded by congestion. The maximum flow value is referred to as the highway capacity.

Lighthill and Whitham (1955) and Richards (1956) used speed/concentration curves in their kinematic wave theory to model the formation of queues behind roadway bottlenecks – that is, places where capacity is lower than upstream or downstream sections. The end of a queue is modeled as a shock-wave, representing an abrupt change in traffic density and speed. So long as traffic arrives at the bottleneck at a faster rate than its capacity, the shock-wave grows upstream.

The 1950s is notable for introducing concepts from physics into the study of traffic queues, as in the kinematic wave theory of Lighthill and Whitman, and and also the thermodynamic theories of Newell (1955). It also was a period that established traffic science, and more generally transportation science, as a field of research that blends empirical and theoretical investigation. This is especially evident in the work of Wardrop (1952), Edie (1956), Edie and Foote (1958) and Edie (1961), and Herman *et al* (1959). Edie and Foote's investigations are especially famous, and are based on extensive data collection on traffic flows and speeds in the Holland and Lincoln tunnels in New York.

Non-stationary phenomena are critical to analysis of queueing on highways, due to peaking of traffic during commute periods. This type of queueing is sometimes called "recurrent congestion", as it occurs on a daily basis. Recurrent congestion is distinguished from "non-recurrent congestion", representing delay caused by random occurrences, such as accidents. Considerable research has been devoted to analyzing the effects of random incidents on highway, often by the same basic methods as non-stationary phenomena. However, research on vehicular queueing usually does not account for random variations in inter-arrival or service times, as is common in mainstream queueing literature. Queueing caused by this type of randomness is viewed as secondary relative to queues caused by accidents or queues caused by non-stationary traffic patterns.

Cumulative diagrams have been a part of the traffic science literature for some time as a representation of non-stationary phenomena. They are used to show the cumulative count of vehicles passing a point along a roadway, but they are applied more generally in queueing to represent cumulative counts of arriving and departing customers. They can be used to measure vehicle concentrations, queue sizes, travel times and delays. They are used to model empirically observed processed (i.e., based on actual counts) and also to deterministically model average system performance. Finally, they are used to represent non-recurrent incidents by randomizing event times, durations and magnitudes. The methodology is documented in the texts by Newell (1971, 1982) and Hall (1991), and later in this chapter.

According to Newell (1993), empirically based cumulative curves first appeared in published literature in 1960 (Edie and Foote), and were first used as a predictive tool in 1965 (Gazis and Potts), though they had already been used for some time within state transportation departments. May and Keller (1967) represented traffic as a continuously flowing fluid within a cumulative diagram to model the formation and dissipation of a queue caused by peaking in traffic flows. More recently, in 1993, Newell merged the concepts of cumulative diagrams with wave theory, relying on a three-dimensional version of the cumulative diagram (traffic is a function of both time and space; Makigami *et al*, 1971).

Roads in most countries have been financed through the imposition of taxes, most commonly paid when purchasing fuel. As a consequence, road users do not ordinarily pay additional charges on costly roads. And it is very rare for roadway charges to be related to how heavily the roadway is utilized or the amount of congestion on the roadway. As a consequence, economists have argued that roadways are overutilized during peak periods. (This is because drivers impose more delay on other vehicles during congested periods than they personally incur.)

Vickrey (1963, 1969) proposed that queues can be eliminated through the application of a continuously variable toll, and that all road users would benefit (despite that added toll). Beckmann *et al* (1956), Beckmann (1965) and Dafermos and Sparrow (1971) proposed route based tolls to influence traveler routes, and to optimize use of roadway capacity on primary and parallel routes. Numerous papers have been written since, but the basic approach has remained constant. Prices are set such that travelers optimally equilibrate across travel times and travel routes, greatly reducing or eliminating queueing. The equilibration is based on a combination of direct cost and indirect cost (representing the inconvenience of traveling on a secondary time or at a non-preferred time). In general, however, the models are highly speculative, as realistic data are not available to verify their underlying behavioral assumptions, and because pricing policies are dictated by politics, technology and practicality more than idealized toll structures.

Traffic: Signalized Intersections

Signalized intersections operate as bulk service systems, in which the server alternates between different customer types. A customer type represents a vehicular path through the intersection, defined by a "from direction", a "to direction" and possibly by a lane. Unlike bulk service systems in production, intersections allow different customer types to be served simultaneously, provided that their trajectories do not intersect, allowing for many ways to combine trajectories into flow patterns.

The queueing delay for any trajectory through an intersection depends on the signal's cycle length, green phase length (portion of cycle that signal is green for the trajectory), and the synchronization of the green phase with the pattern of vehicle arrivals. It also depends on intersection parameters, such as vehicle service rates

during the green phase and the average arrival rate. The usual pattern is that queues accumulate during a red phase, dissipate at a rate matching signal capacity at the start of the green phase and, after the queue vanishes and until the signal turns red again, vehicles are served as they arrive.

Research on intersections has centered on optimizing cycle length, phase lengths, phase patterns and signal offsets (representing time lags between adjacent intersections). Cycle lengths are typically extended when it is necessary to increase an intersection's capacity. This is because capacity losses occur at each phase change; hence, enlarging the cycle length reduces the capacity lost per unit time. If arrival rates are small, cycle lengths are set shorter, so as to minimize cycle delays. [If rates are very small, traffic may be better served by a stop sign or uncontrolled intersection, further reducing cycle delays at the expense of lower capacity (Tanner, 1962; Cheng and Allam, 1992).] Phase lengths are apportioned according to arrival rates and service rates.

As general practice, phase lengths must be at least large enough to serve all vehicles that arrive in a cycle, and should sometimes be even longer if the arrival rate is much larger for a traffic stream than others. Offsets are set to provide synchronization between intersections. Ideally, a signal should enter its green phase as the vehicles begin arriving from an upstream signal. These vehicles arrive in "platoons" (i.e., clusters of vehicles), which have a tendency to disperse as they travel away from an intersection (Pacey, 1956; Grace and Potts, 1964). When intersections are spaced far apart, platoon dispersion (as well as turning traffic) makes it impossible and perhaps unnecessary to synchronize traffic signals. Closely spaced intersections, on the other hand, can be synchronized to minimize cyclic delays and provide for a smoother progression of traffic (e.g., Allsop, 1970; Robertson, 1969; Little *et al*, 1981).

Synchronization is easily accommodated on isolated one-way streets. However, perfect synchronization is usually impossible on two-way streets (in which case opposing directions may arrive at different times) or in signal grids (in which case crossing streets may require different synchronizations). In any case, synchronization demands identical, or integer-multiple, cycle lengths, to ensure that settings do not drift apart. This in itself forces a compromise, as traffic levels at some intersections may demand longer cycle lengths than others.

Grids of signals can also experience blocking effects. This can occur when signals are closely spaced and poorly synchronized, and is exacerbated by poor driver behavior. When a signal operates close to saturation, vehicles may queue back to the preceding intersection. When the preceding intersection turns green, they are blocked from passing through the intersection because the downstream segment is already occupied. The situation worsens when the signals are out of phase with each other, and can be especially problematic in a tight grid of intersections. Intersection

blocking in Manhattan is the source of the term "gridlock", which has lately become synonymous with any form of queueing.

Essential trade-offs between cycles length, phase length and queue time were captured as early as 1941 in the work of Clayton, but has since been enhanced through consideration of stochastic effects and different signal configurations and control policies. Most of the literature treats arriving and departing vehicles as fluids, flowing at constant rates within time intervals. In some cases, these rates are stochastic, and in others arrival rates may vary within a cycle (accounting for effects of upstream signals). In Webster's classic work (1958), arrival patterns were simulated, and empirical relationships were statistically estimated for waiting time as a function of signal parameters. Newell (1965) examines signal through analytical expressions in which queue parameters are random variables, but once these parameters are determined the intersection behaves as a deterministic/fluid system. He, along with Miller (1963), examined the effects of spillover from one traffic cycle to the next, which can significantly exacerbate queueing when operating close to capacity.

Transit and Trucking

Mass transit and truck systems have similar characteristics in that they serve "customers" (people in the case of transit, and shipments in the case of trucking) in groups (called bulk service). Bulk service also occurs in production systems, such as batch chemical processes, printing, and metal stamping, and therefore research on queueing systems is somewhat intertwined among these applications. In all cases, the basic issues are to determine when bulk services should occur, how many customers should be served in each bulk service, and which customers should be served. The decisions are optimized against cost objectives (e.g., cost of providing the service), customer service objectives (e.g., average time waiting or average time in inventory), and throughput objectives (e.g., ensuring that customers can be served as fast as they arrive). Unlike traffic signals, bulk service in trucking and transit occurs virtually instantaneously, as the vehicle departs. Furthermore, bulk service models for transit and trucking usually do not consider what happens to the resource (vehicle) when it completes its service.

Perhaps the most famous and widely used model is the Wilson economic-order-quantity model, which was developed in the early 20th century. The basic premise is that a total cost function (sum of inventory and set-up cost) is minimized by optimizing the number of customers served in each bulk service. The resulting equation provides a square-root relationship between order quantity and the arrival rate of customers. Similar ideas can be found in the transportation literature, most notably in the work of Newell (1971), Blumenfeld *et al* (1985), Burns *et al* (1985) and Hall (1996). Newell demonstrated how to optimize the interval between dispatches for non-stationary/deterministic systems. The other three papers determined how the Wilson model can be applied in transportation contexts,

accounting for inventory at both the source and destination of a trip, synchronization with arrival and departure processes at the trip ends, and multiple-stop vehicle routes. (These topics are covered in depth in Section 5.5)

One of the most interesting application papers in queueing is Oliver and Samuel's (1967) study of mail processing. This is one of a few papers that examines sortation in terminals and transportation to and from the terminal as a linked process. But the paper is most significant for determining how capacity should be determined within a serial queueing system under non-stationary demand. Their fundamental conclusion was that staffing should be allocated in a way that evens out capacities, thus providing minimal queueing once the customer has passed through the initial server.

A second area of interest is real-time control of routes, governing the release of vehicles from stops in response to random arrival rates. Again, the earliest work in this area falls outside of the transportation literature (Bailey, 1954; Neuts, 1967). More recent work includes Powell (1985), Powell and Humblet (1984), and Powell (1986), who investigated a variety of policies for dispatching or canceling services based on the elapsed time from the previous service and the number of customers waiting. Similar policies have been investigated for transfer terminals by Hall *et al* (2001) and cyclic truck routes (Hall, 2002). These contributions fall in the tradition of dispatching policies form the production literature.

A final area concerns schedule control of vehicles traveling on routes with multiple stops. Here, the application is almost exclusively transit. In this context, it is usually impossible to hold vehicles at stops if there are insufficient customers. First, it would be unwise to base a dispatch policy on just one stop when the bus will later serve many downstream locations. Second, most transit systems advertise a schedule that is relied on by customers. Finally, the majority of the service cost is incurred whether the vehicle is in motion or stopped, so there is little cost advantage in holding a vehicle or canceling a trip.

In routes providing frequent service (headways of 10 minutes or less), the objective in schedule control is largely to ensure consistency in headways (time separation between vehicle arrivals or departures). Customers on short-headway lines typically do not consult schedules before arriving at their stops, and therefore arrival patterns are reasonably stationary relative to the schedule. Second, as demonstrated in Osuna and Newell (1971), average waiting time increases with the square of the coefficient of variation in the headway (ratio of standard deviation to the mean). Completely random Poisson vehicle arrivals generate twice the average wait of deterministic arrivals. In fact, waiting time can be worse then the Poisson case, as vehicles on frequent lines have a tendency to bunch. Headways on very frequent lines are inherently unstable: when a bus falls slightly behind schedule, it tends to pick up more passengers, causing it to slow further, until it eventually bunches with the trailing bus (Newell, 1975, Barnett, 1974). This can be controlled,

to some degree, by slowing down a trailing bus when it is catching up with the preceding bus. However, the added delay for passengers already on the trailing bus limits the applicability of this (and other) control strategies, except at the very start of lines.

The behavior of infrequent lines differs substantially from frequent lines. Customers generally do consult schedules, making arrival patterns non-stationary. Therefore, waiting time is not defined by the headway, but instead by the random deviations in the bus arrivals at the stop, along with the customer's selected arrival time relative to the schedule. Finally, because late bus generally do not pick up additional passengers, schedules tend to be much more stable.

Aircraft and Airports

As in road transportation, a fundamental issue in air transportation is accommodating peak traffic loads. And though techniques such as fluid models have been applied in air transportation (e.g., Newell, 1979), a separate branch of research has evolved in which stochastic phenomena are explicitly modeled. Unlike highway traffic, the number of customers (represented by aircraft) that may reside in a queue is relatively small, making it relatively easy to measure the system state as a discrete entity, and also making round-off errors introduced in fluid models somewhat more significant. Consequently, this line of research is linked more directly to mainstream queueing research.

Air transport research is dominated by the phenomena of runway queues. Runways are traditionally a weak link in the air transport system, likely due to the high cost and environmental constraints in their construction, and safety requirements in operation. A complication in modeling runway queues is that the service time for an aircraft depends on the type of preceding aircraft, which is defined by speed and size (creating wake effects that can impose safety risks to trailing aircraft), and whether it is taking off or landing. Therefore, as in many production systems, it can be advantageous to sequence customers in a way that optimizes throughput (Newell, 1979).

Stochastic modeling of runway queues is represented in the work of Gallagher and Wheeler (1958), Koopman (1972), Peterson *et al* (1995a,b) and Odoni and Roth (1983). Odoni and Roth, for instance, developed an approximation for the time constant within an exponential decay function, representing the difference between the expected state of the system at a time t and the limiting state as t goes toward ∞. Newell (1982) is also notable for development of relaxation times, representing the approximate time for a system to reach steady derive. Newell demonstrated that the relaxation time goes toward infinity as the arrival rate approaches capacity, and that steady-state equations are inherently inaccurate for systems that operate close to capacity, even if arrival rates fluctuate only slightly. These were derived from diffusion models, and were not intended for a specific modal application.

Runway queues have also been evaluated within the context of "ground holding", which is a form of network flow control in which the release of aircraft from a departure airport is based on congestion and weather at the destination airport. Ground holding is advantageous because queues are shifted from the airspace to the airports, saving operating costs and enhancing safety. Stochastic programming methods have been used to study the problem (Odoni, 1987; Andreatta and Romanin-Jacur, 1987; Richetta and Odoni, 1993).

A second queueing application is the baggage claim process (Horonjeff, 1967; Ghobrial *et al*, 1982; Robuste and Daganzo, 1992). Here, a service is not completed until two events occur: the arrival of the passenger (or passenger group) and the arrival of the bag (or bag group). Hence, the service is defined by the maximum of a set of random variables. Horonjeff's analysis is based of actual bag and passenger arrival patterns, which are expressed relative to the time that an aircraft begins disembarking passengers. Ghobrial *et al* offer an extension, in which the time required to retrieve a bag is a function of the passenger density surrounding the baggage carousel, and Robuste and Daganzo examine baggage sortation and containerization strategies (similar issues arise in rail and ocean terminals).

Railways

Railways are somewhat unique as transportation modes in two ways: shipments are grouped into long serial units during transportation, and vehicles have no steering capability. Each situation has led to research on queueing.

Train transportation exhibits strong scale economies, meaning that the cost per unit declines substantially when trains operate in longer lengths. However, it is unusual for a single origin/destination pair to generate sufficient traffic to create a long train. Therefore, different origins and destinations must somehow be grouped together. This is accomplished in classification terminals. Each arriving train brings cars from a common set of origins. The train is then broken apart and sorted according to groups of destinations. The sorted cars are finally formed into outbound trains. The process is sometimes repeated multiple times, and sometimes pre-sorting at one terminal to reduce work at a downstream terminal.

The classic work on train sortation can be found in the book by Beckmann *et al* (1956), who modeled the expected number of train breaks (and associated service time) as a function of the number of sortation categories and their probabilities. More detailed models were not developed until the 1970s, and is represented in the work of Petersen (1977a,b), Turnquist and Daskin (1982), and Daganzo *et al* (1983). These authors developed models representing the time required to process a train based on how cars are grouped into sortation blocks on outbound trains.

Because trains cannot be steered, and because the guideway is restricted to narrow track, vehicles can only pass each other at prescribed locations (sidings), and

with the assistance of switching. This contrasts with most other forms of transportation, where vehicles can pass by steering into another lane or otherwise outside the trajectory of the other vehicle. Most railroads are designed to have either one track (shared by opposing directions) or two tracks (one for each direction). In the first case, trains must be switched into sidings to allow faster trains to overtake slower trains (e.g., a passenger train passing a slower freight train), or whenever trains meet from opposing directions, no matter how fast they are traveling. With two tracks, sidings are only needed to allow faster trains to pass slower trains.

Queueing research has centered on design, including: (1) provision of one or two tracks, (2) separation between sidings, (3) operating policies, with respect to speed, passing priority and train scheduling. Railroads must consider whether the benefits of operational flexibility and reduced delay justify the added expense of constructing additional track or sidings. This investment is typically only justified when traffic levels are sufficient. Research on the subject is represented by Frank (1966), Petersen (1974) and Welch and Gussow (1986). A common technique is to utilize time-space diagrams (the vehicle trajectory, showing position as a function of time) to identify train "interference" (i.e., the intersection of vehicle trajectories). Petersen determines the interference frequency as a function of the train separations and speeds, and associates these with interference delays siding locations. This research is closely related to the traffic flow literature, both in its use of time-space diagrams and in its modeling of interference.

Spatial Queueing

A final application area spans transportation and location science, and concerns queueing for spatially separated resources, such as police or fire service. The general question is to allocate resources in a way that minimizes a measure of response time, while staying within an available budget. In some cases, the resources reside at fixed bases (e.g., fire), and in other cases the resources are mobile (e.g., police). Versions also exist where the customer travels to the server, rather than the server traveling to the customer.

The response time typically includes a combination of travel time (from where the resource is located to where it is needed), call processing time, and queueing time. One of the interesting phenomena is that when the system gets busy, it is the travel time that suffers rather than queueing time. This is because when nearby resources are busy, a more distant resource is dispatched instead – creating a longer response time. Simultaneously, the throughput degrades, as it takes longer to serve calls when travel distance increases.

Much of the work on the topic can be attributed to a series of projects conducted by the RAND Corporation in New York City in the early 1970s. Examples of research in the area include Chaiken and Larson (1972), Green and Kolesar (1989), Ignall *et al* (1978), Kolesar (1975), Kolesar and Blum (1973), Kolesar *et al* (1975),

Larson (1972) Rider, (1976). The work is most notable for how it has blended empiricism, theory, and application. This includes modeling response distance as a square-root function of the average territory served by each resource, explicitly representing resource allocation and call rates as non-stationary functions, precisely modeling service time distributions and verifying results against actual performance.

5.4 Representation of Queueing Processes

Cumulative diagrams and fluid models are the most important contributions of transportation to the queueing literature, and this is our emphasis here. They have been applied to all modes of transportation, and are useful in displaying and modeling queueing phenomena, and in system optimization.

Basic Concepts

A cumulative diagram indicates how many customers (often vehicles) have passed a point in the transportation system as a function of time (measured from an initialization time). A cumulative arrival diagram indicates how many customers have entered the system, and a cumulative departure diagram indicates how many customers have left the system. Figure 5.1 provides an example empirical cumulative diagram. In an empirical diagram, individual customers are represented by steps in the curve, corresponding to the time instants when events occurred (either an arrival or a departure). Additional curves can be created, is desired, for intermediate points, as when customers pass through serial servers.

Cumulative diagrams are important because they provide many performance measures in one simple picture. Let:

$A(t)$ = cumulative arrivals from time 0 to time t
$D_s(t)$ = cumulative departures from the system from time 0 to time t

The number of customers in the system at any time t is simply:

$L_s(t)$ = number of customers in the system at time t
$$A(t) - D_s(t) \tag{5.1}$$

And the total time spent by customers in the system up to time t is:

$W(t)$ = total time spent by customers up to time t
$$= \int_0^t L_s(\tau)d\tau + \int_0^t [A_s(\tau)-D_s(\tau)]d\tau \tag{5.2}$$

Two critical performance measures are the average number of customers in the system and the average time in system per customer. The average number of customers is easily derived from W(t):

$\mathbf{L}(t)$ = average customers in system, time 0 to time t
$$\mathbf{L}(t) = W(t)/t \tag{5.3}$$

In cases where the system begins and ends in an empty state (i.e., $L_s(0) = L_s(t)=0$), the average waiting time is also easily defined:

$\mathbf{W}(t)$ = average time in system from time 0 to time t
$$\mathbf{W}(t) = W(t)/A(t) \tag{5.4}$$

Combining these expressions, it can be seen that

$$t\mathbf{L}(t) = A(t)\mathbf{W}(t) \tag{5.5a}$$
$$\mathbf{L}(t) = [A(t)/t]\mathbf{W}(t) \tag{5.5b}$$

Equation 5.5b is a special case of Little's formula (1961), which states that the average number of customers in the system asymptotically approaches the average time in system multiplied by the customer arrival rate for a wide class of systems.

All of these results are clearly seen in a cumulative diagram, as Figure 5.1 illustrates. The number of customers in system (queue size) is the vertical separation between the cumulative curves, and the total waiting time is the area between the curves. The average time in system is the average horizontal separation and the average customers in system is the average vertical separation. If customers are processed in a FCFS order, the diagram also shows the time in system for individual customers, also measured by the horizontal separation. If the sequence is not FCFS, then another graphical device, such as a GANTT chart, is needed to show the time in system for individual customers.

Fluid Models

In a fluid model, individual customers are represented as a continuously flowing fluid rather than discrete entities. This has the effect of smoothing out the steps in the arrival and departure curves. Fluid models are often used to predict the future performance of queueing systems, or just to simplify the representation of observed phenomena.

In a fluid model, arrival rate, $\lambda(t)$, and departure rate, $\mu(t)$, are defined by the derivatives of their corresponding cumulative curves:

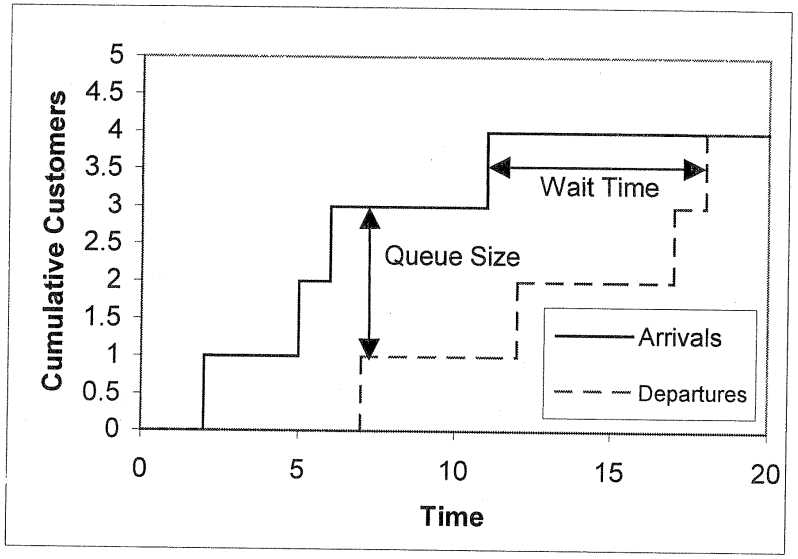

Figure 5.1 Cumulative Diagram

$$\lambda(t) = dA(t)/dt \quad (5.6a)$$
$$\mu(t) = dD(t)/dt \quad (5.6b)$$

In bulk service systems, the service rate can be undefined; otherwise it reflects three factors: (1) the speed at which customers can be processed by the server, (2) the size of the queue, and (3) the rate at which customers arrive. Servers ordinarily operate at their fastest rate when queues are present, and operate at the same rate at which customers arrive when queues are not present. Exceptions exist, as service capacity can be variable, depending on demand, and service times can sometimes change as queue lengths change.

To illustrate fluid models, we consider a simple system in which the service rate is limited to a capacity c, but service times are very short. This might represent queueing at a highway toll plaza, for instance. The arrival process and departure process are both non-stationary, but are assumed to be deterministic, for purposes of illustration. Under these conditions, Figure 5.2 illustrates how the queues would evolve over a period of peak arrivals. The system is shown to evolve through a series of four phases:

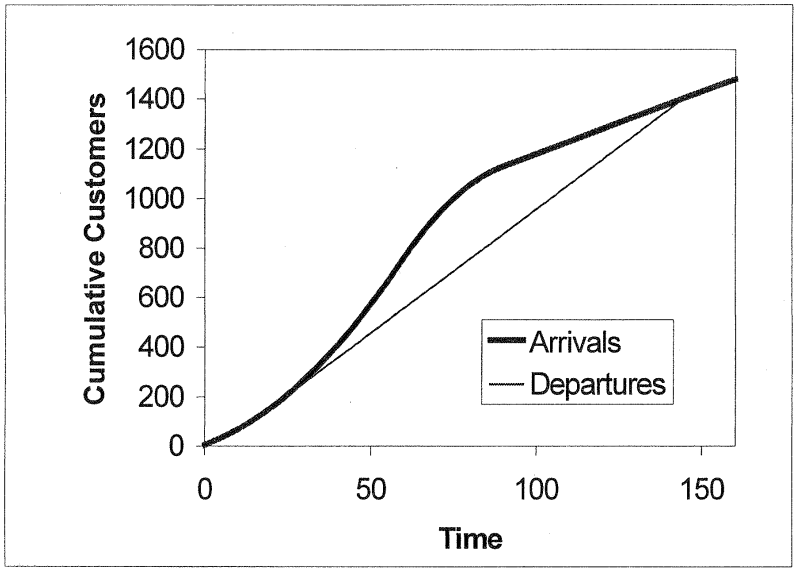

Figure 5.2 Cumulative Fluid Model

Phase 1: Stagnant

$$A(t) = D_s(t) \qquad \lambda(t) \le c \qquad \mu(t) = \lambda(t) \qquad dL(t)/dt = 0 \quad (5.7)$$

Phase 1 represents the initial period when customers can be processed as fast as they arrive (time 0 to time 20 in the figure).

Phase 2: Growth

$$A(t) > D_s(t) \qquad \lambda(t) > c \qquad \mu(t) = c \qquad dL(t)/dt = \lambda(t) - c > 0 \quad (5.8)$$

Phase 2 represents the period in which the queue grows because customers cannot be served as fast as they arrive (time 20 to time 80 in the figure).

Phase 3: Decline

$$A(t) > D_s(t) \qquad \lambda(t) \le c \qquad \mu(t) = c \qquad dL(t)/dt = \lambda(t) - c \le 0 \quad (5.9)$$

Phase 3 begins when the queue reaches its maximum length, which occurs when the arrival rate drops down to capacity. It ends when the queue vanishes. (Time 80 to time 140 in the figure.)

Phase 4: Stagnant

$$A(t) = D_s(t) \qquad \lambda(t) < c \qquad \mu(t) = \lambda(t) \qquad dL(t)/dt = 0 \tag{5.10}$$

Phase 4 is when the queue is again stagnant at 0, with customers arriving slower than they can be served. An interesting phenomenon is that $\mu(t)$ exhibits a discontinuity at the time the queue vanishes (5.3), dropping suddenly from c to the current arrival rate. Thus, the departure rate pattern is highly asymmetrical in queueing systems.

Analysis Through Cumulative Diagrams

Through perturbation analysis, it is possible to optimize the design of the queueing system. It is relatively straight forward, for instance, to model the effects of changing system capacity. Increasing capacity has a non-linear effect on time in system, as it causes both the duration of the queue (length of Phase 2 and 3), and the magnitude of the queue to decline. And when capacity exceeds the maximum arrival rate, the queue vanishes. Comparison of departure curves can be used to select a capacity from a set of discrete options.

Cost trade-offs can be evaluated through use of marginal analysis, in which capacity is continuously varied. We define a total cost function as:

$$\begin{aligned} C &= \text{capacity cost} + \text{waiting cost} \\ C &= \alpha c + \beta w(c) \end{aligned} \tag{5.11}$$

where:

α = capacity cost per unit capacity
β = waiting cost per unit customer time
$W(c)$ = total waiting time when capacity equals c

A necessary condition for optimality is that cost must not decrease if the capcity is changed by a small amount Δc. The change in cost, ΔC, if c is increased by Δc can be written as the sum of the change in capacity cost and the change in waiting cost:

$$\Delta C = a\Delta c + \beta[W(c+\Delta c) - W(c)] \tag{5.12}$$

The change in waiting time (the term within the brackets) can be calculated from the cumulative diagrams. Assume, as in Figure 5.3, that one predictable queue occurs per time period. Then, for small values of Δc, the change in waiting time can be approximated from the area of the triangle shown in the figure. That is:

$$W(c+\Delta c) - W(c) \approx \tfrac{1}{2}\, T(c)\, [T(c)\Delta c] \tag{5.13}$$

ΔC represents the marginal change in cost, which must equal zero at the optimum (provided that $W(c)$ is continuously differentiable). Substitution of Eq. 5.13 in Eq. 5.12 provides the following optimality criterion:

$$T^*(c) = \sqrt{2\alpha/\beta} \tag{5.14}$$

Eq. 5.14 states that the optimal capacity is represented by the duration of the queueing period – time from when the queue first forms until it vanishes – and not by the arrival rates during the queueing period. The optimal duration increases with the square root of the capacity cost (when capacity is expensive, longer duration queues can be tolerated) and decreases with the square-root of the waiting cost (when waiting is expensive, queues should be shorter in duration).

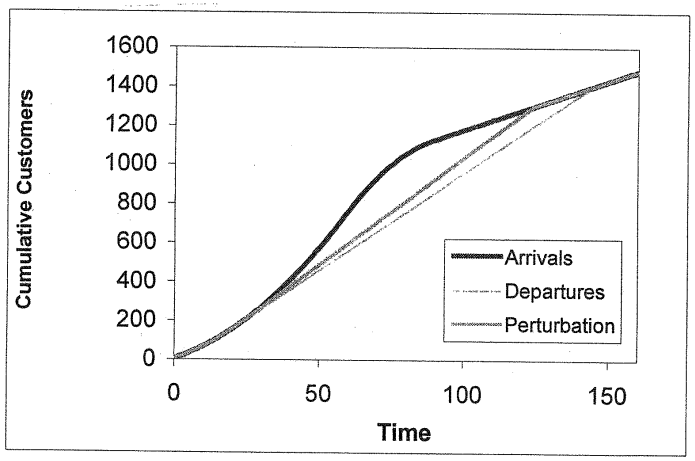

Figure 5.3 Marginal Analysis

Extensions

The cumulative modeling approach has been extended in a variety of ways.

- Investigation of the combined effects of stochastic variability and non-stationarity, principally through application of diffusion models.
- Optimization of other system attributes, such as staffing plans and time-off scheduling.
- Measuring the effects of incidents that cause capacity to decline over short intervals
- Estimating effects of behavioral responses, causing arrival rates to be a function of queue lengths, waiting times or tolls.
- Evaluating queueing in bulk-service systems, such as signalized intersections, transit and trucking.

Bulk service will be examined in some depth in the following section. But first, we note that incidents often have a pronounced effect on system performance. This is especially true when incidents occur around the time that a queue begins to form, as it affects everyone who arrives over the queue's entire duration. The effect is not nearly so great when an incident occurs later, as it only affects those customers that arrive later. As a consequence, queue management demands special care during Phase 2, both to prevent harmful incidents, and to persuade customers, if possible, to arrive at other times.

5.5 Bulk Service Models

Economic Order Quantity (EOQ) and Economic Production Quantity (EPQ) models have been used for many years in transportation and manufacturing to optimize cycle lengths, load sizes and batch quantities for bulk service. While research in this area today is focused on complex scheduling systems, many of the underlying assumptions of the EOQ/EPQ models have been retained, especially in transportation applications.

This section describes how the EOQ/EPQ methodology is applied, taking both input processes and output process into account. To this end, a set of "characteristic cumulative diagrams" is developed to represent a range of scenarios. The principal assumptions are: (1) input and output processes occur at constant and deterministic rates (in some scenarios, rates are allowed to alternate between "on" and "off" phases through batch processing). (2) Set-up and order costs are independent of batch size. (3) Batches can be initiated instantaneously when the queue size drops to zero. (4) Queueing costs are linear functions of the average queue size.

The systems considered will have three components: an input process, a bulk transportation system, and customers. The models explicitly represent bulk transportation of goods, but they are easily adapted to represent other transportation systems, such as traffic signals and buses. Hence, the input represents a production process. The section is organized to demonstrate the effects of: (1) Synchronization of input batch sizes with output batch sizes; and (2) Coordination of input and output when there are multiple customer or product types.

Basic Methodology

The general approach is to represent total cost per unit time as the sum of a queue cost and a "set-up" cost. The queue cost equals the average queue level multiplied by a queue cost parameter. The set-up cost equals the number of set-ups or orders per unit time (the demand rate divided by the batch size) multiplied by the cost per set-up.

The following parameters are used to represent the system. In some cases, these parameters are subscripted to denote an individual customer or product.

d = output rate (items/time)
p = input rate (items/time)
S = input set-up cost (money/set-up)
A = transportation "set-up" cost (money/order)
h = queue cost (money/customer per unit time).

To simplify expressions, transportation lead time (i.e., the transportation time from origin to destination) is assumed to be zero. With respect to optimizing batch sizes, this assumption results in no loss in generality, provided that lead times are independent of the other parameters. While it is not difficult to incorporate a "pipeline" cost to represent lead-time, the cumulative diagrams lose clarity. Queueing cost is also assumed to be identical at source and destination, again with the intention of highlighting principles. For similar reasons, the time to perform the set-up is assumed to be negligible relative to the run time. Finally, batch sizes are assumed to be unconstrained.

The decision variables are the order and production batch sizes, which in turn define the order and production cycles:

Q_p = production batch size
Q_t = transportation order quantity
T_p = production cycle time = Q_p/d
T_t = transportation cycle time = Q_t/d.

While in most cases the production and order quantities are held constant, it will, in some instances, be less costly to allow for varying quantities.

Queue holding costs are defined by the cumulative production at the source and cumulative demand at the customer (or customers).

P(t) = cumulative production from time 0 to time t
D(t) = cumulative demand from time 0 to time t.
I(t) = customers in the system at time t = P(t)-D(t) .

The order and set-up costs depend on P(t) and D(t), as achieving a small queue requires more frequent set-ups and orders.

Dispatching Rule We now define a general characteristic of batch transportation systems under optimal control. The characteristic is a necessary condition for optimality when the following four conditions apply, but as a matter of practice applies more broadly:

(1) transportation set-up cost is fixed with respect to shipment size,
(2) queue cost is a linear function of the total queue in the system (i.e., P(t)-D(t)),
(3) vehicle size is unlimited,
(4) P(t) and D(t) are non-decreasing and represent a single product.

Let:

T(t) = cumulative items dispatched from the manufacturer, from time 0 to time t.

Then at the time of any dispatch:

T(t) = D(t) immediately before dispatch
T(t) = P(t) immediately after dispatch.

In words, the dispatching rule states that a shipment should be sent as soon as the queue is exhausted at the customer, and that the order quantity (i.e., shipment size) should be identical to the queue on-hand at the manufacturer: P(t)-D(t). Visually, this rule is manifest in the cumulative graphs presented later through the staircase pattern for T(t), which alternately "bounces" between D(t) and P(t).

The optimality of the dispatching rule can be proved by contradiction. From any solution that violates the rule, it is possible to construct a solution which obeys the rule, with equal or lower cost. Specifically, if T(t) does not equal D(t) immediately before dispatch, then the shipment can be delayed until T(t) = D(t), with no increase in queue cost, and a possible decrease in transportation cost (if two shipments can be consolidated). If T(t) does not equal P(t) immediately after dispatch, then the shipment size could be increased, with no change in queue cost,

and a possible decrease in transportation cost (if a subsequent shipment can be eliminated).

Queue Models

This section creates a set of seven characteristic cumulative diagrams, each representing a different cyclic queueing pattern. In a subsequent section, these curves are used as building blocks for developing EOQ/EPQ models. While the diagrams represent production/distribution, they are easily adapted to represent other situations in transportation.

The average queue level equals the average separation between the cumulative production and cumulative demand curves, which is determined by calculating the area of separation and dividing by the elapsed time. The separation depends on the batch sizing policies, both in production and transportation. In its simplest form, production and demand are characterized by Figure 5.4 or 5.5. Figure 5.4 is the textbook version of the EOQ model, as it assumes instantaneous production and transportation. Figure 5.5 is the textbook version of the EPQ model, as it assumes production occurs at some set rate, and transportation occurs continuously, and not in batch.

In a more general sense, average queue level may be defined by any of the following types of cumulative production and demand diagrams, which will be called the "characteristic curves." (Recall that, in all cases, constant demand is assumed.) The set of cases is not completely exhaustive, but does encompass most reasonable patterns that apply to direct transportation routes in production/distribution.

1. Instantaneous Production/Batch Distribution (Synchronized) This is the textbook EOQ model (Figure 5.4).

$$\text{Average Queue Level} = Q_t/2$$

2. Instantaneous (or Constant) Distribution/Batch Production As in Figure 5.5, production is immediately available for consumption, eliminating batch size inventories in distribution. Figure 5.5 is equivalent to the textbook EPQ model.

$$\text{Average Queue Level} = (Q_p/2)(1-d/p)$$

3. Constant Production/Batch Distribution As in Figure 5.6, production and demand occur at a constant rate. Inventories exist at both point of production and

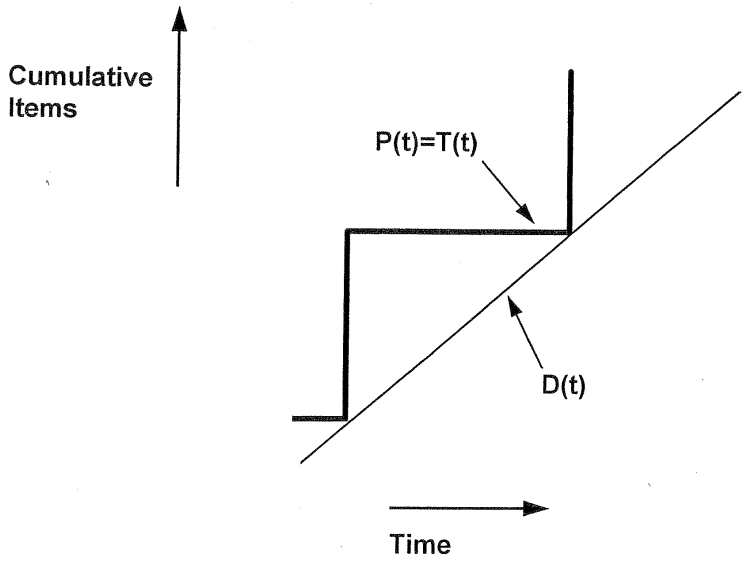

Figure 5.4 Instantaneous Production/Batch Distribution

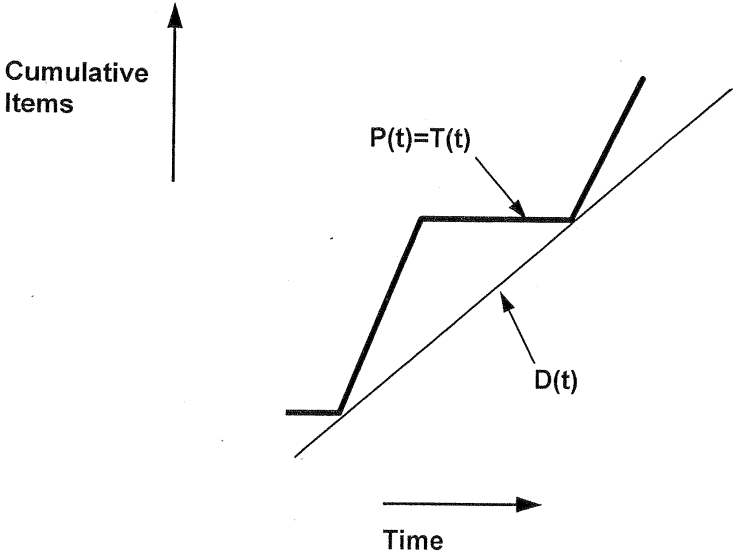

Figure 5.5 Instantaneous Distribution/Batch Production

point of demand as a result of distribution batch sizes with constant sizes and constant separation.

$$\text{Average Queue Level} = Q_t$$

4. Batch Production/Batch Distribution In all of these cases, the product is both manufactured in batches and transported in batches.

4a. Synchronized/lot-for-lot As in Figure 5.7, transportation is synchronized with production, so that a dispatch occurs as soon as a batch is manufactured. The average queue level at the customer is the transportation batch size divided by two. The average queue level at the manufacturer is one-half the production batch size, multiplied by the proportion of time that the machine is running (d/p).

$$\text{Average Queue Level} = (Q_p/2)(d/p) + Q_t/2 = (Q/2)(1 + d/p),$$

where $Q = Q_t = Q_p$ (due to lot-for-lot production). If d=p, the machine runs continuously and the average queue level is the same as for case 3. If p>>d, production is effectively instantaneous, and the average queue level is the same as case 1. Finally the ratio of average queue level relative to case 2 (instantaneous distribution) is (p+d)/(p-d). Hence, if d<<p, average queue levels are approximately the same. As d approaches p, the ratio approaches infinity, indicating that the EPQ model greatly underestimates queue level in batch distribution when production and demand rates are similar.

4b. Synchronized/multiple transportation lots Due to this case's complexity, the queue model will be presented later within the context of a specific system scenario (Scenario F).

4c. Non-synchronized/lot-for-lot As in Figure 5.8, the transportation and production cycle lengths are identical. However, transportation is not scheduled to coincide with the end of a production run. (This may occur if multiple products, each with a different start/end time, are transported in each cycle.)

$$\text{Average Queue Level} = (Q_p/2)(d/p) + Q_t/2 + \tau d = (Q/2)(1 + d/p) + \tau d$$

where τ is the time lag between the end of the production run and the time of dispatch ($Q=Q_t=Q_p$).

4d. Non-synchronized As in Figure 5.9, production and transportation both occur in batches, but are not synchronized. As a result, the average queue is the sum of case 2 and case 3 (Blumenfeld *et al*, 1985).

$$\text{Average Queue Level: } (Q_p/2)(1-d/p) + Q_t$$

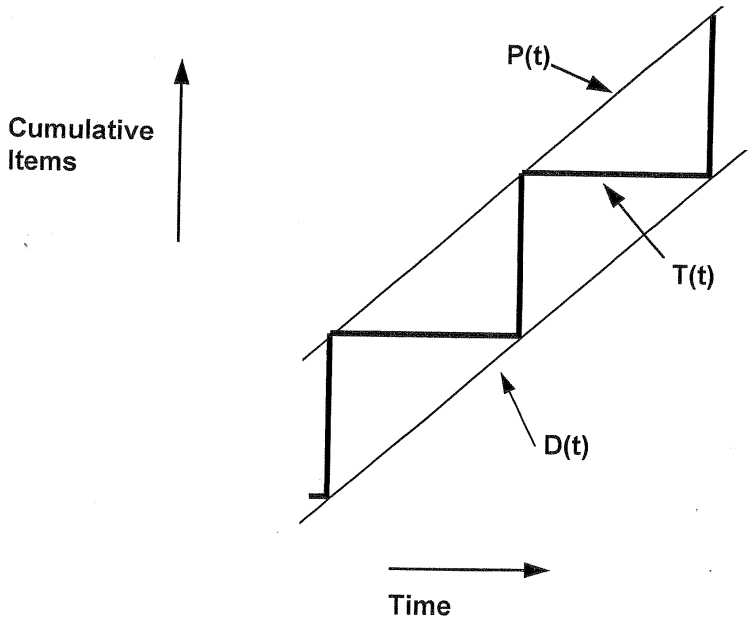

Figure 5.6 Constant Production/Batch Distribution

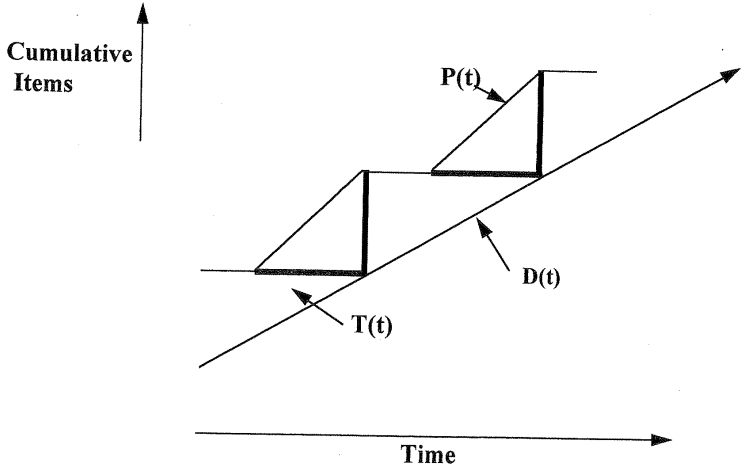

Figure 5.7 Synchronized Lot-for-Lot

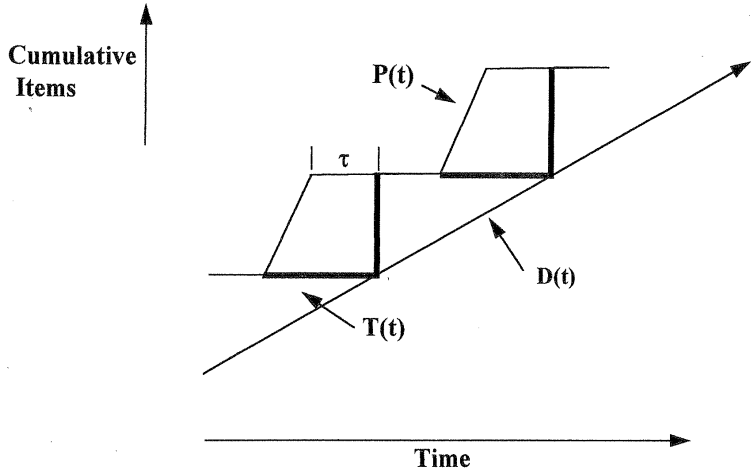

Figure 5.8 Non-Synchronized Lot-for-Lot

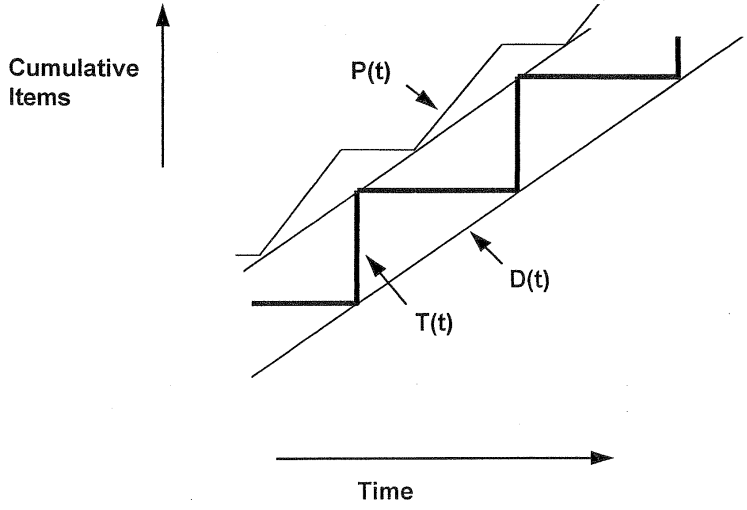

Figure 5.9 Non-Synchronized

Set-Up Cost Models

The set-up cost per unit time is the cost per set-up (or order), multiplied by the number of set-ups (or orders) per unit time. For consistency with the queue models, it is necessary to derive the number of set-ups per unit time as a function of the batch size (or sizes). In the classic EOQ and EPQ models, this function is simply the following:

f_p = production set-up frequency

$$f_p = d/Q_p = 1/T_p \qquad (5.15a)$$

f_t = transportation "set-up" frequency

$$f_t = d/Q_t = 1/T_t \ . \qquad (5.15b)$$

These models are adequate when a single product is manufactured/distributed, but more precision is needed for multiple products. In the transportation process, in particular, it is customary to serve multiple products within the same batch. Hence, for any origin/destination pair, there is a single transportation cycle length, which is identical for all products:

$$f_t = d_1/Q_{1t} = d_2/Q_{2t} = \ ... = d_i/Q_{it} = ... \qquad (5.16)$$

where the first subscript on d and Q denotes product number, and where Q is interpreted as the shipment size per dispatch. Equivalently, a "composite product" can be defined, where demand is the sum across all products, expressed in a common unit (such as weight or dollar value). Then Eq. 5.1 would apply, provided that Q and d are interpreted in this common unit. In Case 4b, where batch sizes vary within a production cycle, a further modification is needed. The transportation set-up frequency will be the number of orders per production cycle (n) multiplied by the production set-up frequency (d/Q_p).

Problem Dimensions

The number of potential variations to the EOQ and EPQ model is quite enormous. Our purpose is to present a range of scenarios, and later discuss the implications of the more significant variations on cost. This will be accomplished by identifying the "characteristic cumulative diagram" that applies to the scenario, computing total cost, and optimizing the production batch size and transportation order quantity.

The scenarios are defined at two levels. At the top level, the defining attributes are the number of customers and the number of plants. At the lower level, scenarios are defined by the number of machines within each plant and the number of products:

Top-level Attributes

1) Single Customer/Single Plant
2) Multiple Customer/Single Plant
3) Single Customer/Multiple Plants
4) Multiple Customer/Multiple Plants

Lower-level Attributes

a) Single Machine/Single Product
b) Multiple Machines/One Product per Machine
c) Single Machine/Multiples Products per Machine
d) Multiple Machines/Multiple Products per Machine .

Attributes (a) and (b) do not require production changeovers; hence, production is continuous, and the transportation order quantity is the only decision variable. Attributes (c) and (d) demand changeovers between products; hence, both production batch size and order quantity must be optimized. Table 1 summarizes the scenarios covered in the section, which are constructed by combining attributes. The first three are fairly straight-forward, and do not entail schedule interactions among products. The second three are more complex.

Cost Analysis: Simple Scenarios

This section develops cost models for three simple scenarios, which illustrate the effects of accounting for: (1) queue costs at both the manufacturer and customer; (2) consolidation of multiple products from multiple machines; and (3) costs for unsynchronized systems. These are classified as simple cases because all treat one product at a time.

A. Single Customer/ Single Plant (Queue at Manufacturer and Customer) In this scenario, one machine operates at a constant rate (equaling the demand rate), producing a single product, without interruption, for a single customer. Set-ups do not occur because product change-overs are not needed. Hence, the only decision variable is the transportation order quantity.

The cumulative diagram in Figure 5.6 (constant production/batch distribution) characterizes the situation. The objective function, and its optimal solution, are then:

$$C(Q_t) = A(d/Q_t) + Q_t h \qquad (5.17a)$$

$$Q_t^* = \sqrt{Ad/h} \qquad (5.17b)$$

$$C^* = 2\sqrt{Ahd}. \qquad (5.17c)$$

B. Single Plant/Multiple Machines/Single Customer (Consolidation Effect) In this scenario, each machine produces a single product at a constant rate, for which demand also occurs at a constant rate. The products are manufactured at a single

plant, and distributed to a single customer. Unlike the prior scenario, different products are consolidated in the transportation process. This situation illustrates a major difference between EPQ and EOQ models. Whereas batch production does not allow different products to be processed simultaneously (rather, alternating phases are needed), batch transportation virtually mandates simultaneous service. That is, from the standpoint of cost minimization, it is cheaper to consolidate products in the same vehicle than to transport each product separately.

Blumenfeld *et al* (1985) examined this situation, and introduced the concept of a composite product to represent the portfolio of product characteristics contained in the load. Hence, Figure 5.6 is interpreted as the demand among all products sent between the manufacturer and customer. The cost model, and optimized results, are shown below. Cycle length is used as the decision variable, rather than batch size, because batch size varies among products:

$$C(T_t) = A/T_t + T_tHD \quad (5.18a)$$
$$T_t^* = \sqrt{A/HD} \quad (5.18b)$$
$$C^* = 2\sqrt{AHD}, \quad (5.18c)$$

where:

$$HD = \Sigma\, h_j d_j . \quad (5.19)$$

C. Multiple Plants and Customers (Unsynchronized) In a system with multiple plants and customers, it may be impossible to synchronize transportation and production cycles due to scheduling conflicts. As a result, larger queues must be held at the manufacturer to buffer against cyclic fluctuations. In this scenario, batch production and batch distribution are assumed. The system is decomposed to individual plant/customer/product combinations, assuming the absence of synchronization, as in Figure 5.9. The cost model, and optimized results, are shown below:

$$C(Q_p,Q_t) = S(d/Q_p) + A(d/Q_t) + [(Q_p/2)(1-d/p) + Q_t]h \quad (5.20a)$$
$$Q_t^* = \sqrt{Ad/h} \quad (5.20b)$$
$$Q_p^* = \sqrt{2Sd/h(1-d/p)} \quad (5.20c)$$
$$C^* = 2\sqrt{Ahd} + \sqrt{2Shd(1-d/p)} . \quad (5.20d)$$

In this case, the production and distribution results are decoupled. Further, the production batch size is identical to the textbook EPQ model. The transportation order quantity, on the other hand, is identical to Eq. 5.17b. Hence, the base comparisons for the transportation order quantity are the same as those presented in the single customer/single plant scenario.

More Complicated Scenarios

Within this section, cost analysis is shown for three more complicated scenarios, to illustrate issues involving multiple products and multiple customers. In the first example, a single machine produces a single product to serve multiple customers. In the second, a single machine produces multiple products for a single customer. In the last, a single machine produces multiple products for multiple customers, with one product per customer.

Within the framework of EOQ/EPQ modeling, it is impossible to fully account for complex scheduling systems. In the examples, schedule conflicts are avoided by assuming either (or both) of the following: (1) products are manufactured sequentially in a common rotation cycle, or (2) production rate greatly exceeds demand. Within a rotation cycle, the production rate for a machine is assumed to be the same as the total demand for the machine. The large production rate case will only be used for multiple customer scenarios.

D. Single Machine/Multiple Customers In this scenario, a single machine produces a single product at a constant rate for multiple customers, without interruption. Though set-ups do not occur, production must still be divided into time segments, corresponding to customers. Consequently, the average queue depends both on the time to produce and the time to consume a quantity. These values are different because the production rate, by the necessity to serve multiple customers, must exceed the demand rate of any one customer. This effectively results in production batches without the need for production set-ups. Hence, the characteristic queue curve for any one customer is a batch production/batch distribution case (Figure 5.7), but the production set-up has a cost of zero.

For an individual customer, the cost can be expressed as:

$$C(Q_t) = A(d/Q_t) + h(Q_t/2)(1 + d/p) . \quad (5.21)$$

Assume that customers are served in a common rotation cycle (length T_t), and that the production rate matches the sum of the demand rates. Because there is a common product, further assume that the queue holding cost is the same for all customers. Using T_t as the decision variable, the total cost for the rotation can be expressed as:

$$C(T_t) = nA'/T_t + \Sigma\,[h(T_td_i/2)(1 + d_i/\Sigma\, d_j)] \quad (5.22)$$
$$C(T_t) = nA'/T_t + h(T_t/2)(nd' + E(d_i^2)/d')$$
$$C(T_t) = nA'/T_t + h(T_t/2)d'(n + 1 + C^2),$$

where:

d' = average demand rate
A' = average value of A_i, among customers i=1,...,n
C = coefficient of variation of the demand rate
$E(d_i^2)$ = average of enclosed quantity.
The optimized values of T_t and $C(T_t)$ are then:

$$T_t^* = \sqrt{2nA'/hd'(n+1+C^2)} \quad (5.23a)$$
$$C(T_t^*) = \sqrt{2nA'hd'(n+1+C^2)}\,. \quad (5.23b)$$

Note that if n=1, C must equal zero, and the model reduces to the same form as Eq. 5.17, or the simple single plant/single customer case. As n approaches infinity, the model converges toward something like the classic EOQ model, with $T^* = \sqrt{2A'/hd'}$, and $C(T^*)/n = \sqrt{2A'hd'}$. However, they are based on averages among all customers, not individual customer values. The scenario demonstrates that when the demand for an individual customer falls well below the production capacity, the queue model is much like the classic EOQ

If the production capacity greatly exceeds the demand rate, it might be reasonable to optimize order quantities on an individual customer basis. Eq. 5.21 could then serve as the objective function, resulting in the following solution:

$$Q^* = \sqrt{2Ad/h(1+d/p)} \quad (5.24a)$$
$$C(Q^*) = \sqrt{2Ahd(1+d/p)}\,. \quad (5.24b)$$

If p>>d, these results reduce to the exact same form as the classic EOQ.

E. Single Machine/Multiple Products: Single Customer In this scenario, demand occurs at a constant rate for each product, but production is cycled among products on a single machine, with set-ups and changeovers. First, products are assumed to be produced at the same rate, with the same queue holding cost. Later, this assumption is relaxed. As stated at the beginning of the chapter, set-up times are assumed to be negligible.

Figure 5.9 is the characteristic cumulative diagram for individual products. Given that each product must be produced at a different time (recall, a single machine is used), it is impossible to synchronize all products with distribution. The aggregate queue diagram for the rotation cycle (Figure 5.10) is more revealing. The similarity to Figure 5.6 is a striking feature of Figure 5.10, for it suggests that a rotation cycle can bear the same queue cost as simple single product cycles. That is, queues are built up at a rate p-d during a production phase, and depleted to zero at a

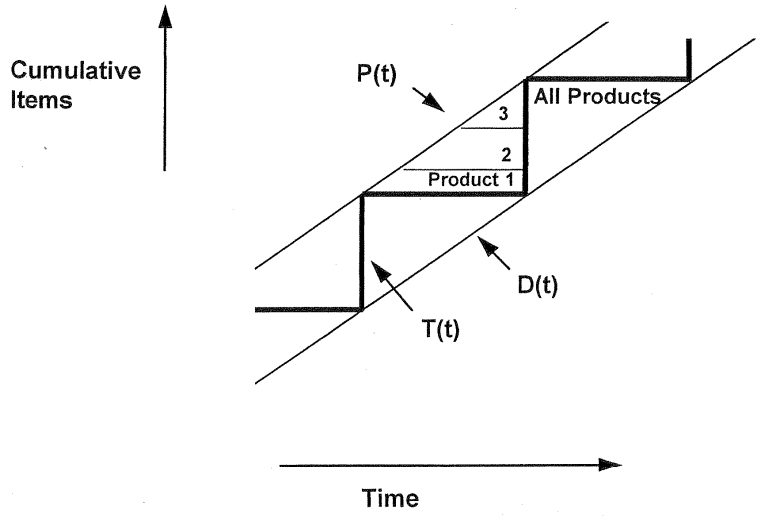

Figure 5.10 Single Machine/Multiple Products/Single Customer

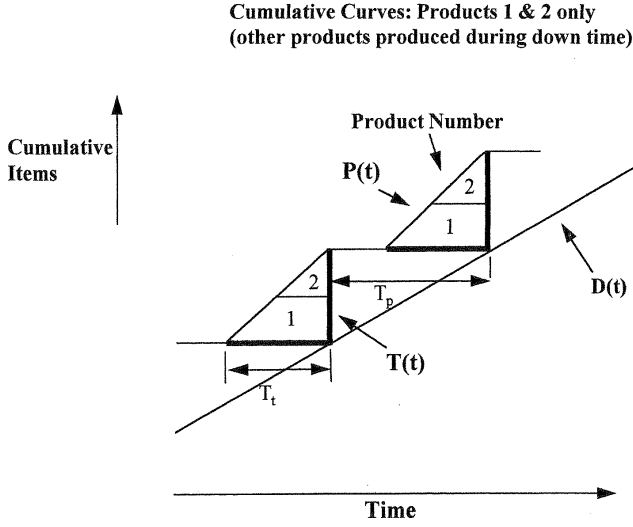

Figure 5.11 Single Machine/Multiple Products/Single Customer: Decoupled Production and Distribution Cycles

rate d when production is cycled off. The batch transfer process acts to consolidate products into the same load independently of their position within the rotation. Hence, the first product in the rotation, which must wait nearly a full cycle before dispatch, is transported at the same time as the last product in the rotation.

The cost formulation can now be represented as follows:

$$C(T) = (A+nS')/T + Tnhd', \quad (5.25a)$$

Where

$$S' = \text{average of } S_j, \text{ among customers } i=1,...,n \quad (5.25b)$$

The optimized result is then:

$$T^* = \sqrt{(A+nS')/nhd'} \quad (5.26a)$$
$$C(T^*) = 2\sqrt{(A+nS')nhd'}. \quad (5.26b)$$

These results are the same as Scenario A (single plant/single customer), with the exceptions that the "set-up cost" includes both the order cost and the combined set-up cost across all products, and that the demand is the total demand across all products.

In some instances, it is preferable to decouple production and transportation cycles, with the latter occurring more frequently than the former. These decisions can be totally decoupled when one ignores the round-off errors that result when a dispatch occurs in the middle of a product's production run. The average queue at the manufacturer is then one-half the distribution batch size. The average aggregate queue at the customer is one-half the production batch size (Figure 5.11). Again assuming a rotation cycle, the total cost is the following:

$$C(T_p,T_t) = A/T_t + nS'/T_p + (T_t/2)nhd' + (T_p/2)nhd'. \quad (5.27)$$

The optimized results are then:

$$T_t^* = \sqrt{2A/nhd'} \quad (5.28a)$$
$$T_p^* = \sqrt{2S'/hd'} \quad (5.28b)$$
$$C(T_t^*,T_p^*) = \sqrt{2Anhd'} + n\sqrt{2S'hd'}. \quad (5.28c)$$

To be implemented, the cycle lengths must be adjusted so that the manufacturing cycle is an integer multiple of the transportation cycle.

F. Single Machine/Multiple Products: One Product per Customer In this final scenario, each production batch serves a single customer, and is fully synchronized with distribution. As soon as a production run is completed, all queue for the given

product is dispatched to the customer. Production can either occur on a lot-for-lot basis, or with multiple distribution lots per production cycle.

Simple Rotation Cycle In a simple rotation cycle, production and transportation are synchronized with the same cycle length for all products/customers. The queue pattern for this scenario is batch production/batch distribution, synchronized lot-for-lot (Figure 5.7). The total cost for a cycle is then:

$$C(T) = n(A' + S')/T + \Sigma (T/2)h_i d_i(1 + d_i/p_i) \quad (5.29)$$

The optimized cycle length and cost are:

$$T^* = \sqrt{2(A'+S')/[HD+E(h_i d_i^2/p_i)]} \quad (5.30a)$$
$$C(T^*) = n\sqrt{2(A'+S')[HD+ E(h_i d_i^2/p_i)]} \quad (5.30b)$$

As a point of contrast, Scenario D (single machine/single product/multiple customers) did not include set-up costs, and the production rate was simply the sum of the demand rates. This leads to a relatively higher set-up cost in Eq. 5.29, and a slightly modified queue holding cost. Hence, the optimal cycle length is longer for the multiple product scenario (F) than the single product scenario (D).

As a second point of contrast, Scenario E (single machine/multiple products/single customer) uses only one transportation set-up per cycle, and queue cost is larger. Hence, the optimal cycle length is longer for the multiple customer scenario (F) case than the single customer scenarios (E).

Large Production Capacity If the production capacity greatly exceeds the demand rate, it might be reasonable to optimize order and production quantities on an individual customer basis. The following could then serve as the objective function for an individual product:

$$C(T) = (A + S)/T + (T/2)hd(1 + d/p). \quad (5.31)$$

The optimized results are then:

$$T^* = \sqrt{2(A+S)/hd(1+d/p)} \quad (5.32a)$$
$$C(T^*) = \sqrt{2(A+S)hd(1+d/p)} \quad (5.32b)$$

Allowing for Multiple Dispatches If queue holding costs are sufficiently high, it might be reasonable to provide multiple dispatches per production cycle. Let:

I_0 = queue in the system at the start of the production run
Q_i = size of transportation batch i (i = 1,2,...), within a production run

The initial queue, I_0, is exhausted at the moment that batch 1 is transported. Hence, Q_1 equals the production during the time required to consume I_0 units of queue:

$$Q_1 = p(I_0/d) . \tag{5.33}$$

Similarly, all subsequent batch sizes are dictated by the prior batch sizes, in the following fashion:

$$Q_i = p(Q_{i-1}/d) = I_0(p/d)^i . \tag{5.34}$$

I_0 can now be derived, by recognizing that the sum of the transportation batch sizes within a cycle must equal the production batch size:

$$Q_p = \Sigma Q_i = I_0 \Sigma (p/d)^i , \tag{5.35a}$$

or

$$I_0 = Q_p/[\Sigma (p/d)^i] . \tag{5.35b}$$

Referring to Figure 5.12, the average queue level can now be characterized as the sum of a base level, I_0, and an EPQ type quantity:

$$\text{Average Queue Size} = I_0 + Q_p[(p-d)/p]/2 . \tag{5.36}$$

With I_0 as given in Eq. 5-35b, the average queue size becomes:

$$C(T_p) = (mA+S)/T_p + hT_p[1/\Sigma_{i=1}^{m} (p/d)^i + (1-d/p)/2] , \tag{5.37}$$

where m is the number of transportation cycles per production cycle. Through a combination of search techniques and calculus, it is not difficult to optimize m and T_p within the above expression.

Summary of More Complicated Scenarios Introduction of scheduling considerations complicates EOQ and EPQ calculations in several ways. First, to avoid schedule conflicts, either a rotation cycle must be optimized, or simplifying assumptions must be made with respect to production capacity. Second, the combination of batch production and batch distribution results in somewhat non-standard forms for the queue equations. Third, both production and transportation set-up costs must be considered when optimizing cycle length.

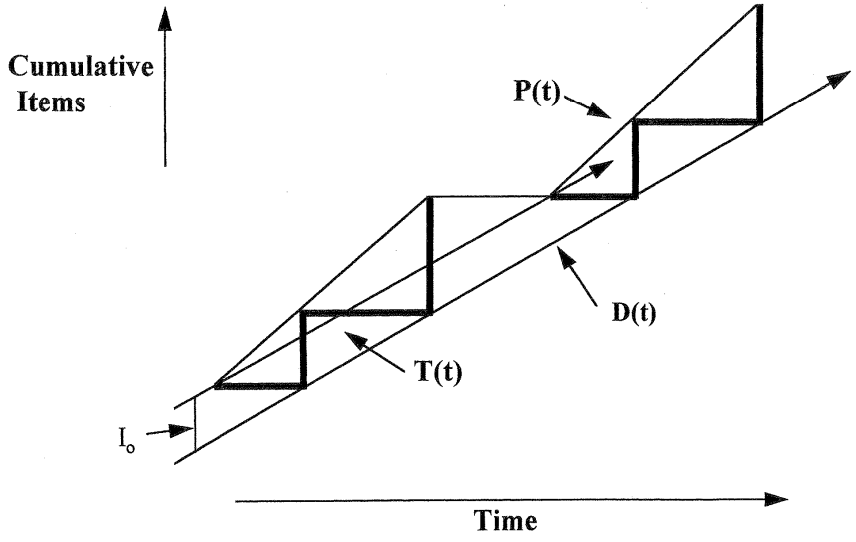

Figure 5.12 Synchronized/Multiple Transportation Lots

Extensions

The scenarios presented in this section served to illustrate a methodology, and to illustrate the complexity of accurately accounting for queue sizes when input and output processes are discontinuous. Many extensions have been covered in the literature, including the following:

Random Cycle Length Batch processes seldom occur precisely on schedule. Consequently, the headways between batches vary randomly, causing average and maximum queues sizes to increase. This occurs because customers are more likely to arrive during longer headways, and because the average wait for a long headway is greater than the average wait for a short headway. In the special case where customers arrive at random relative to batch times, the average wait is given by:

$$E(W) = [E(h)/2][1 + C^2(H)] \tag{5.38}$$

Where E(h) is the mean headway and C(H) is the headway coefficient of variation. If batches occur with the randomness of a Poisson process, E(W) = E(h), which reflects the memoryless property of the exponential distribution (the headway distribution for a Poisson process).

Non-Stationary Demand Headways between batch services should vary in relationship to the demand rate. Larger demand invites shorter headways, according to an inverse square-root relationship. In some systems, however, the total waiting time per dispatch should stay constant for all demand rates. For example, if demand increases by a factor or 2, then headway should decrease by a factor of $\sqrt{2}$, batch size should increase by a factor of $\sqrt{2}$, and the product stays constant. Another common characteristic of optimal batching is that the arrival time at the time of service equals the ratio of the number of customers served to the time until the subsequent dispatch.

Multiple Stop Transportation Routes Scenarios can be further delineated by transportation characteristics, principally, whether or not transportation equipment is shared among customers and plants. Sharing, in the form of multiple-stop pick-up and delivery routes, can provide substantial savings in transportation and queue cost in low demand systems (Burns *et al*, 1985; Daganzo, 1985; Hall, 1985). This naturally adds complexity, as it may be desirable to serve some customers less frequently than others, yet also ensure that their service intervals are synchronized so that all shipments within a territory occur on a common schedule.

Capacity Considerations Capacity is important in two ways. First, batch sizes may be limited by the size of available vehicles and, second, the batch service system may be limited in the total rate at which customers are processed. Either factor leads to solutions that violate the "Dispatching Rule" presented in this section. In the former case, the optimal *feasible* batch size is generally the minimum of two values: the vehicle size or the cost minimizing batch size, as determined in this section. In the latter case, the batch size may need to be enlarged, to reduce the batch frequency, and reduce loss times when initiating batches. This is especially relevant to traffic signals, where cycle lengths are typically defined by capacity considerations rather than set-up costs.

Real-time Control Random variations in demand can make it desirable to alter headways and batch sizes in real-time. When the number of customers is insufficient, a headway can be extended or a batch can be cancelled. Dispatch times might also be altered to provide greater consistency in headways, thus minimizing its coefficient of variation and reducing waiting time.

5.6 Future Directions

As it has in the past, future research on queueing in transportation is likely to respond to innovations in the methods of transportation. Technologies for automating and controlling vehicle movements on highways has already stimulated queueing research, addressing delays and capacities associated with lane-following strategies, lane-assignment and entrance/exit processes. Changes in aircraft routing and control, possibly allowing aircraft to travel in free-space rather than on prescribed paths, is also likely to stimulate original research.

Future research will also be directed at gaps in the literature. A notable example is the paucity of research on queueing within terminals, and on the interactions between sorting processes and transportation processes. Relatively little is known on how terminal queues interact with vehicular queues. Yet the problem grows in importance, as more shipments are transported through parcel transportation companies, in which sortation is a critical cost driver.

Finally, despite the considerable accomplishments in understanding the behavior of queues on roadways, researchers have been largely unsuccessful in actually eliminating vehicular queues. It appears inevitable, as observed long ago, that in the absence of road pricing queues will exist. Developing and testing pricing methods for roadways, and then creating a mechanism by which they can be implemented, is perhaps the most important challenge to the field. But success in this area demands far more than an understanding of the mathematics of queues; it demands accurate representations of human behavior, along with knowledge of the institutional and technical aspects of toll collection.

One clear aspect of research on queueing in transportation is that the most significant papers have offered a blend of empiricism and theory, and have been innovative in exploring new applications. It is simply insufficient to develop the mathematical theorems. The papers that best explain important "real-world" phenomena, or provide generalizable methods for system design and operation, have been the most significant, and will likely continue to be in the future.

5.7 Acknowldgement

Portions of the chapter appeared earlier in Transportation Research (Hall, 1996). These excerpts are reprinted with the permission of the publisher, Pergammon Press. Research was supported in part by National Science Foundation grant DMI-9732878.

5.8 References

Allsop, R.E. (1970). Optimisation techniques for reducing delay to traffic in signalised road networks. Ph.D. Thesis, University of London.

Andreatta, G. and Romanin-Jacur, G. (1987). Aircraft flow management under congestion. *Transportation Science*, **21**, 249-253.

Bailey, N.T.J. (1954). On queueing processes with bulk service, Journal of the Royal Stastical Society B, **16**, 80-87.

Barnett, A. (1974). On controlling randomness in transit operations. *Transportation Science*, **8**, 102-116.

Beckmann, M.J., McGuire, C.B. and Winsten, B. (1956). *Studies in the Economics of Transportation*,. Yale University Press, New Haven, Connecticut.

Beckmann, M.J. (1965). On optimal tolls for highways,tunnels and bridges, In: *Vehicular Traffic Science* (Edie, Herman and Rothery, eds.), 331-341. Elsevier, New York.

Blumenfeld, D.E., Burns, L.D., Diltz, J.D. and Daganzo, C.F. (1985). Analyzing trade-offs between transportation, inventory and production costs on freight networks. *Transportation Research*, **19B**, 361-380.

Blumenfeld, D.E., Burns, L.D. and Daganzo, C.F. (1991). Synchronizing production and transportation schedules. *Transportation Research*, **25B**: 23-27.

Burns, L.D., Hall, R.W., Blumenfeld, D.E. and Daganzo, C.F. (1985). Distribution strategies that minimize transportation and inventory cost. *Operations Research*, **33**, 469-490.

Chaiken, J. and Larson, R. (1972). Methods for allocating urban emergency units: a survey. *Management Science*, **19**, 110-130.

Cheng, T.E.C. and Allam, S. (1992). A review of stochastic modelling of delay and capacity at unsignalized interjections. *European Journal of Operations Research*, **60**, 247-259.

Clayton, A.J.H. (1941). Road traffic calculations, *Journal of Institute of Civil Engineers*, **16**, 247-284, 558-594.

Dafermos, S. C. and Sparrow, F.T. (1971). Optimal resource allocation and toll patterns in a user-optimized transportation network. *Journal of Transportation Economic Policy*, **5**, 198-200.

Daganzo, C.F. (1982). Supplying a single location from heterogeneous sources. *Transportation Research*, **19B**: 409-420.

Daganzo (1989). On the coordination of inbound and outbound schedules at transportation terminals, Institute of Transportation Studies Research Report, Berkeley, California.

Daganzo, C.F., Dowling, R.G. and Hall, R.W. (1983). Railroad classification yard throughput: the case of multistage triangular sorting. *Transportation Research*, **17A**, 95-106.

Edie, L.C. (1956). Traffic delays at toll booths. *Journal of Operations Research*, **4**, 107-138.

Edie, L.C. (1961). Car-following and steady-state theory for non-congested traffic. *Operations Research*, **9**, 66-76.

Edie, L.C., and Foote, R.S. (1958). Traffic flow in tunnels, *Proceedings of the Highway Research Board*, **37**, 334-44.

Edie, L.C., and Foote, R.S. (1960). Effect of shock waves on tunnel traffic flow. *Proceedings of the Highway Research Board*, **39**, 492-505.

Frank, O. (1966). Two-way traffic in a single line of railway. *Operations Research*. **14**, 801-811.

Gallagher, H.P. and Wheeler, R.C. (1958). Nonstationary queueing probabilities for landing congested aircraft. *Operations Research*, **6**, 264-275.

Ghobrial, A., Daganzo, C.F. and Kazimi, T. (1982). Baggage claim area congestion at airports: an empirical model of mechanized claim device performance. *Transportation Science*, **16**, 246-260.