CHAPTER

3

Route and Network Analysis

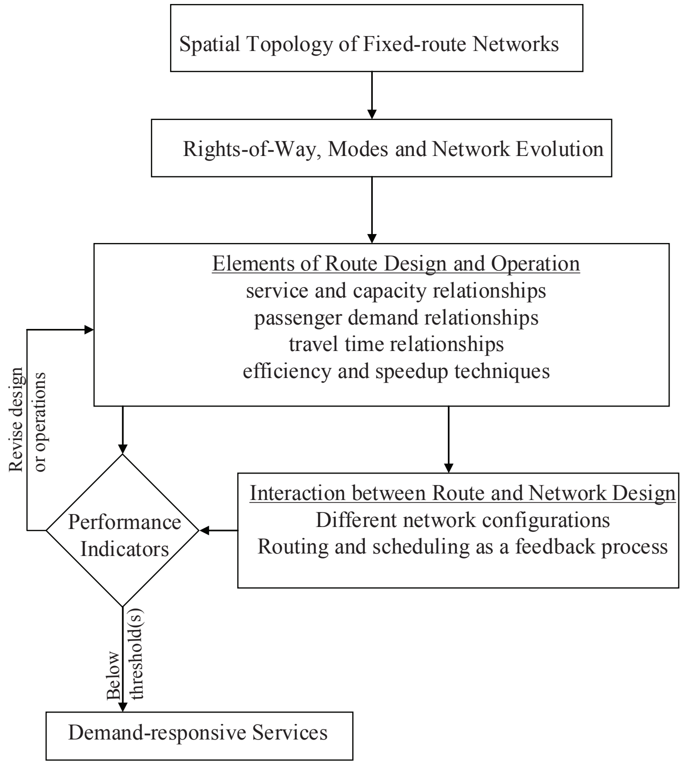
Public transportation routes are the ultimate output of public transportation agencies. But routes can't be understood in isolation from the network in which they operate. The network configuration is often central to performance and investment analysis, even of a single route. This configuration is the product, not only of conscious choice by current managers, but also of the inherited urban form—its road patterns, development densities, geographical features, previous infrastructure investment decisions, demographics, and public expectations. Thus, the existing network facilitates or hinders how well investment and performance goals can be met.

This chapter addresses several interrelated topics. The organization of it is shown on Figure 3.1. It corresponds with an evaluation process that can be followed for an individual route. The first two sections provide background. Section one presents numerous types of networks seen in actual practice. The next discusses how different levels of rights-of-way and public transport modes with different levels of performance might fit into a network, as well as how a network can be expected to evolve as it increases in size. The third section examines the fundamental relationships needed both to design routes and to plan their operations. The discussion is extended from these basic relations into speed-up and efficiency-improving techniques. The fourth section uses some recent research as a teaching tool to explain the trade-offs of various network design and operating principles.

Performance indicators are introduced throughout the chapter, as these are the means for making comparisons to other services and for informing decision making. Much of the mathematical exposition of performance indicators as well as speed-up and efficiency-increasing techniques is placed in Appendix 3.A.

The analyst might sometimes conclude that no combination of modes and services will give the desired performance at reasonable cost on a particular route, or even in an entire section of the network. Thus, the last topic in this chapter is *demand-responsive service*, also often referred to as paratransit. This is an alternative to fixed-route transit when fixed routes are deemed nonviable due to low performance measures or because passengers have special requirements. Modern transit planning requires knowledge of such alternatives.

Figure 3.1 Organization of Chapter 3


Spatial Topology of Fixed-route Networks
Rights-of-Way, Modes and Network Evolution
Elements of Route Design and Operation
service and capacity relationships
passenger demand relationships
travel time relationships
efficiency and speedup techniques
Revise design or operations
Performance Indicators
Interaction between Route and Network Design
Different network configurations
Routing and scheduling as a feedback process
Below threshold(s)
Demand-responsive Services

SPATIAL TOPOLOGY OF FIXED-ROUTE NETWORKS

There are a variety of basic network configurations. Some of the most common will be shown schematically. A convention will be used when drawing these schematics to distinguish which crossing lines have no or virtually no interchange of passengers. Figure 3.2 introduces symbology borrowed from the hydraulic engineering discipline. The semicircle indicates that the routes cross over one another but that there is no interchange (no connection between services).

Figure 3.3 shows some directly connected points, without interchange to other routes. If there are no intermediate stops on a route, only endpoints, it is a *shuttle* operation. The remaining examples of network types all have interchanges. Figure 3.4 shows a grid network. Figure 3.5 shows an elbow network, which has the property that each route crosses at least three other routes, which greatly increases the number of direct connections between lines without concentrating transfers at a center.

Each of the remaining network types is oriented relative to a center, which could be the center of the largest city in a region but could also be a subregional center, such that there is a hierarchy of networks. (It is quite possible that subregions will have different basic configurations, especially if the topographical features are different or if they were developed in different eras.) The radial network of Figure 3.6 is distinguished from the diametrical network in Figure 3.7 by the termination at the center instead of continuation through to the opposite side. Tangential additions as shown in Figure 3.8 provide routes that do not pass through a center but instead connect the arms. The composite grid-diametrical network tends to concentrate services towards the center as in Figure 3.9, yet also has a degree of parallel gridlike coverage as well. The *trunk-and-feeder network* of Figure 3.10 is like a radial network in that feeder routes converge and terminate on a point. But then passengers must transfer to another route (the trunk) that consolidates the traffic. The *trunk-and branch* network of Figure 3.11 again converges on a point, but now the routes share the same path along the trunk section. The routes are typically scheduled such that vehicles from different branches to arrive at the trunk section on an alternating basis.

Another network type not shown is the *ubiquitous network*. It essentially means one that has not developed according to a dominant pattern but has evolved by expanding to connect points that have large travel demand between each other with direct links. The eventual result is excellent coverage of the city. Paris and Tokyo are two examples.

Figure 3.2 Connectivity on Network Schematics

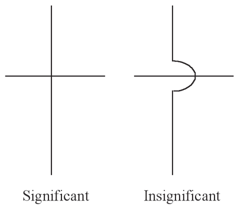


Figure 3.3 Directly Connected Points

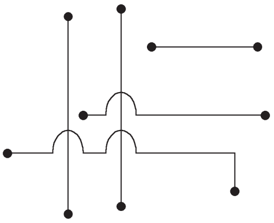


Figure 3.4 Grid Network

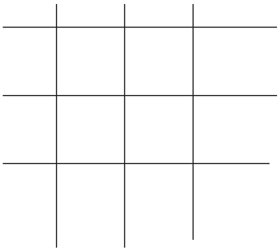


Figure 3.5 Elbow Network

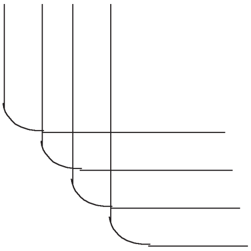


Figure 3.6 Radial Network

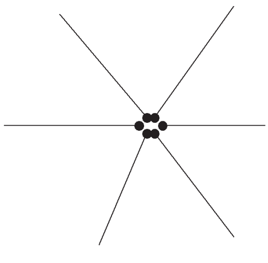


Figure 3.7 Diametrical Network

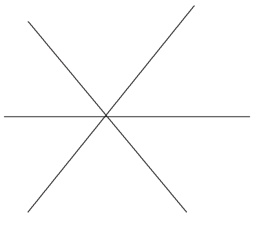


Figure 3.8 Tangential Additions to Diametrical Networks

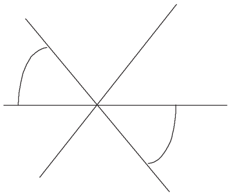


Figure 3.9 Composite Grid-Diametrical Network

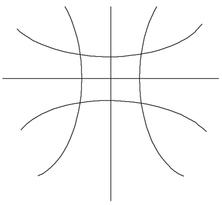


Figure 3.10 Trunk and Feeeder Network

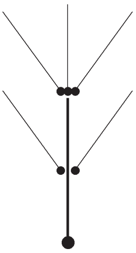
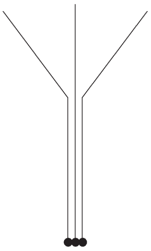


Figure 3.11 Trunk and Branch Network



It is possible to deeply analyze many aspects of a network using graph theory and other disciplines. Analysts can prepare descriptive indicators for network size, form, and topology. As examples, one can examine:

- the number of *(Origin Destination) pairs* requiring no connection;
- O-D pairs requiring X connections;
- the number of closed loops generated by the network;
- the amount of overlap of routes;
- number of paths between an O-D pair; and
- other descriptors.

These can provide some insights into the coverage by and quality of service offered. These become of serious interest when comparing prospective major changes to physical connections and to the operating plan of already highly complex networks. The interested reader is referred to Musso and Vuchic (1988) and Synn (2005). But it is not necessary to know such descriptors and indicators for a basic understanding of networks.

It is important to keep in mind that "networks" need not be comprised of only one vehicle technology. It is sometimes useful to divide networks into hierarchies, where each level is based on services that have similar speed or capacity characteristics and play similar roles. Indeed, large public transport systems typically develop several different network maps showing only one level of the network on each. As an example, a pocket map may show all rapid transit lines and some express bus routes, but not local buses and streetcars. Different maps would have these.

RIGHTS-OF-WAY, MODAL TECHNOLOGIES, AND SYSTEM EVOLUTION

The spatial relation a route has to a network is only a partial description. It also requires that the qualities of the right-of-way be considered. Table 3.1 defines different standards of rights-of-way, with examples of real-life systems provided for clarity. The scheme is consistent with that proposed by Vuchic in his textbook, *Urban Public Transportation: Systems and Technology* (1981). Right-of-way A (ROW A) is total grade separation to allow for a high-speed, highly reliable, and safe operation. Right-of-way C (ROW C) is operation in mixed traffic, with no special provision for public transportation vehicles. Right-of-way B (ROW B) is in between; it uses lateral separation, typically separate lanes or a median. But it is not full separation because of intersections. It gives speed, safety, and reliability performance somewhere in between rights-of-way A and C. As to be expected, investment costs tend to increase with the higher right-of-way standard.

TABLE 3.1 Different Standards of Right-of-Way

Definition	Examples
A grade separation	Paris Metro, Vancouver SkyTrain
B lateral separation	Gothenburg LRT, Oslo bus/taxi lanes
C mixed traffic	Most bus and streetcar (tram) systems
Combinations are also possible:	
A/B	Karlsruhe LRT – railroad and lateral street sections
A/C	San Francisco streetcars – tunnel and mixed street section

Three points need to be stressed about the vehicle technologies operated. First, some vehicle designs are committed to only one standard of right-of-way, while others can be used on more than one. As examples, rapid transit vehicles can be used only on a dedicated facility, while Light-Rail Vehicles (LRVs) can operate in the street or on a dedicated right-of-way. There are also locales where two modes share the right-of-way, even though they may have different performance characteristics that might impede one another. An example would be buses and LRVs both operating on a shared right-of-way B.

The second point about vehicle technologies is that some designs currently restricted to one standard of right-of-way can later be modified to operate on a higher or lower standard of right-of-way, while for other designs it may be impractical or impossible. As examples, LRVs can be equipped to operate on railroads, but commuter railroad trains can never operate on streets due to their length and turn radius. Indeed, the LRV has become popular precisely because it is opportunistic in its use of rights-of-way. The trade-off is that the multicapable vehicle must be more complex and might not operate equally successfully on each category of right-of-way.

The third point is that operational constraints continually evolve with vehicle technology. For example, there are now road vehicles with electronic guidance of lateral positioning; that is, the vehicle is guided along a path automatically instead of through the steering wheel (but without the positive guidance that tracks or concrete beams provide). As another example, some fuels restrict vehicles to a shorter range, which affects the ability to schedule vehicles. As yet another example, electric vehicles are increasingly equipped with auxiliary power sources allowing occasional off-route operation.

The history of transit development is one of continual replacement of vehicles, wayside, and control center components such as traffic signal controls, fare collection machines, passenger information systems, and so on, with continually more capable vehicles and systems. This, in turn, allows changes in the way that routes and networks are structured and operated. Replacement may occur as equipment begins to become unreliable and/or expensive to maintain, but may sometimes occur midlife in order to benefit from an upgrade as soon as possible.

Investment decisions should always consider whether the proposed project should be amenable to future upgrades, to mixing with operations on other standards of right-of-way, and, in general, to operational advantages or restrictions that may be implied by technological changes. Investment in nonstandard equipment, and especially modes using proprietary technology must be done with caution. Future costs can be raised beyond the point of viability. Examples of proprietary technologies that may have no alternative components suppliers include monorails, airport people movers, and optional bus guidance systems.

The history of transit development is also one of geographic expansion. As cities get larger, they tend to get denser in the center, of which skyscrapers are dramatic evidence. At the same time, distances to outlying districts would become longer. Thus, both the level of demand and the lengths of routes would increase. Congestion on roads would increase together with the increased activity. With increases in congestion and route length come increases in travel time. At some point, increase in demand no longer gets addressed simply by adding more vehicles but by the use of larger vehicles. At some point, travel time also gets addressed through increases in right-of-way standard.

Thus, there is a general evolution towards more and larger vehicles, then towards modes with higher capacity and faster speeds. The smallest town may never evolve past a minibus or taxi operating on demand. As cities get larger, buses operating on right-of-way C may grow into articulated buses and some corridors may be upgraded to right-of-way B. LRVs may be joined to become trains, the right-of-way B corridor upgraded to right-of-way A in highly congested areas, and routes might be lengthened. In the largest cities, rapid transit lines may get longer trains as cars are added, increasing frequency of service, and the network may continue to expand indefinitely.

Some cities have developed differently primarily because they have focused on accommodating the automobile, presenting new and more difficult evolutionary challenges. An urban region may increase in developed area much faster than it increases in population. Thus, the density of built-up areas may actually decrease. Metropolitan Chicago is an excellent example; it still has huge demand for travel to the Central Business District (CBD) but has depopulated in many of the surrounding districts. Over the decades since World War II, much of the population, and many of the newer employment locations have steadily migrated outwards (Sen et al. 1998). The consequently lower demand for transit through these reduced-density corridors makes the investment in higher standards of rights-of-way hard to justify. On the other hand, this improvement is needed to shorten travel times over the increasingly long distances from the CBD to where the population is shifting. In the U.S. and elsewhere, the problem is exacerbated in many post-World War II communities. As will be discussed in later chapters, street layouts that focus and collect traffic on a few arterial roads hinder both transit access to residential areas and pedestrian access to transit.

Research, experimentation, and dissemination of partial solutions to service design challenges continue. The analyst needs to stay abreast of trends.

ELEMENTS OF ROUTE DESIGN AND OPERATION

It is possible that the analyst might conclude that no current combination of modes and services will give the desired performance at reasonable cost on some routes or in entire sections of the network. One choice might then be to alter the route network anyway, accepting whatever cost consequences that brings. Another choice could be to abandon uneconomic service in some sections. Yet another could be to entirely restructure the network to better suit community goals. Regardless, understanding the network relations between routes will help to make informed and defensible decisions. Thus, there is an extended discussion of both individual routes and the network.

Important Route and Network Attributes

The distance traveled and the routing particulars of the travel path are important, but incomplete, information. The time consumed by the user to travel the complete path from origin to destination further paints the picture. Additional attributes or features of a given route and of its connecting services within the network are needed to help complete the picture. Some that are usually important are listed in Table 3.2. Together these attributes describe connectivity to the remainder of the network. Connectivity is defined as the possibilities for, and convenience of, travel between points in a network. Of these listed attributes, the route patterns that form a network have already been discussed. Most of the others will be defined more rigorously in the course of this chapter.

The number, type, and physical design details of transfer facilities for connections to and from nontransit means is also important and discussed in detail in Chapter 5.

TABLE 3.2 Some Important Route and Network Attributes

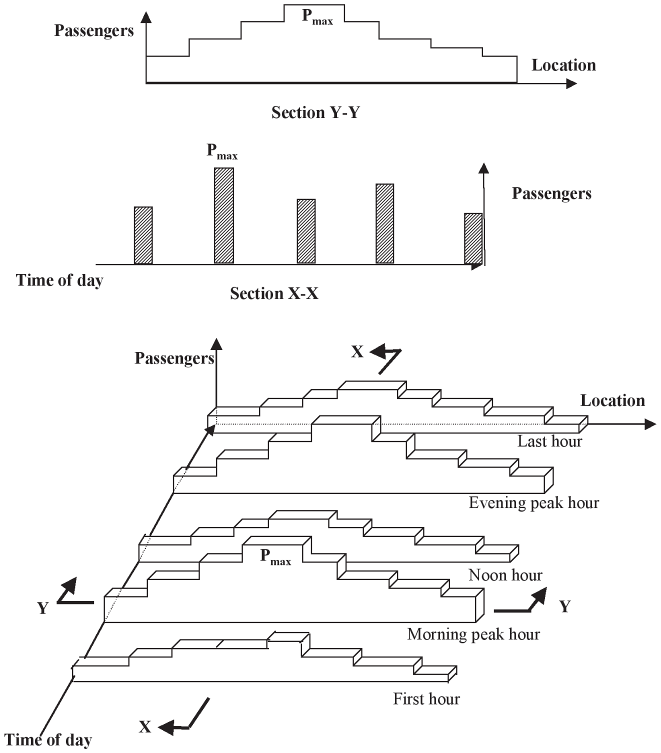
- Route pattern (radial, grid, composite, etc.)
- Temporal demand profile on each route (peaking factors by time of day)
- Spatial demand profile (distribution along routes)
- Operating speeds along routes
- Frequency along routes
- Frequency of connections
- Stop or station spacing along routes
- Area coverage
- Transfer facilities – between public transport services
- Transfer facilities – between public and other modes

Some Fundamental Level-of-Service and Capacity Relationships

Some fundamental concepts must be defined before proceeding further. The *level-of-service* is the quantity of service available as seen from the perspective of the user. One of the key measures of this is the *headway, h*. It is defined as the time separation between vehicles measured at a particular point, usually expressed in minutes, but sometimes in seconds. The *frequency* of service, *f*, is its inverse, 1/h. It has the units of vehicles per unit of time past a particular point, usually expressed in units per hour. A conversion factor of 60 minutes per hour is used when headway is expressed in minutes, making frequency 60/h instead.

Frequency, or equivalently, headway, is, in and of itself, an important performance indicator for a route from the user perspective. Ceteris paribus (that is, all else being equal), the higher the frequency, or equivalently, the shorter the headway, the more convenient is the service.

Figure 3.12 Hourly Ridership as a Function of Location and Time (Only 5 Hours are Shown for Clarity)



A distinction needs to be made between vehicles per hour and *Transit Units* (TUs) per hour. Transit Unit accommodates the fact that vehicles may actually be coupled together and move as one unit. The terminology "m-car long TU" will be used when it is necessary to specify length. (The terminology "m-car long consist" is used by many rail professionals.) The resulting *Line Capacity* is computed by multiplying vehicle capacities with frequency or inverse of headway:

$$\text{Line Capacity} = mC_v f = mC_v\,(60/h) \qquad \text{[spaces/hour]} \qquad 3\text{-}1$$

where C_v is the sum of seated and standee capacity for the particular vehicle design. Inserting the multiplication factor m allows for an m-car long TU. The factor m is always 1 for buses (except for the rare case where they pull trailers). Except when capacity is strictly limited to the number of seats present (no standees), the line capacity value is not truly fixed, but based on assumptions about the level of crowding that will be tolerated by standees. What is considered merely crowded in Japan or China, for example, would be socially unacceptable in most of North America or Europe. This can be seen by comparison of vehicle specifications. Crowding of about 4.0 persons per square meter is the upper limit in North America and Europe when manufacturers calculate available spaces. For Asian and South American vehicles, crowding of up to 6.0 persons per square meter is often assumed.

Spatial and Temporal Relationships of Passenger Demand

Demand for travel varies by time of day, and, consequently, the services offered must change as well. Thus, the temporal demand profile needs to be known. This demand profile is an aggregation of the various functions that individuals perform throughout the 24-hour day. Typically, the fraction of trips related to commuting rises substantially near major shift changes at industrial facilities and near the beginning and end of business hours for offices and other commercial facilities. Such large spikes in ridership are of great concern to transit planners. But how much and at what times of day the demand rises and falls on a given route depends on the both the particular characteristics of the area it serves and the role the route plays within the network. As temporal examples, a route serving major factories is likely to have high peaks of demand at commuting times, while one serving a shopping district and a hospital might have peaks, albeit less pronounced, in the midday and evening. As a spatial example, a tangential connector between major radial corridors could have steady demand throughout the day from persons merely passing through to reach other lines, independent of any origins or destinations along the route.

Understanding the nature of demand throughout the day is critical for selecting types of services, types of modes, indeed for basic network design. Services, even the network configuration to some extent, can be varied throughout the day. The complexity of responding to changes in demand with continual service changes, however, places practical limits on such adjustments. Apparent savings in operating costs and improvements in responsiveness to demand may well be offset by other costs and difficulties incurred in attempting to reliably manage many service changes throughout the day.

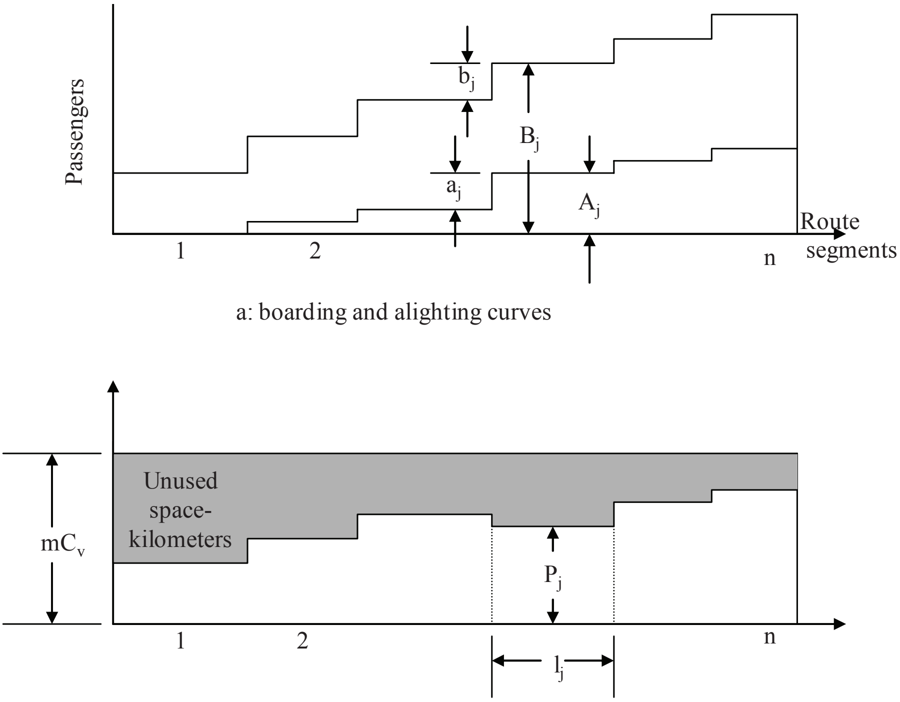
Ridership can be visualized simultaneously as a function of position along a route and by time of day through a three-axis diagram of ridership, hour, and location of the route segment. An example is shown in Figure 3.12. Only five, one-hour sections are shown for clarity. Each direction should have its own diagram. A one-hour section parallel to the position axis shows the number of passengers per hour on each route segment for this one-hour block of the day. A cross-section parallel to the time axis shows the number of passengers per hour on one route segment for each one-hour block of the day instead. Thus, section X-X on Figure 3.12 shows the number of passengers at a particular route segment for these same five one-hour periods. In this case, it is cross-sectioned at the location where the highest maximum of the day is seen. This value is labeled P_{max} on the figure. P_{max} need not be at the same location at different times of the day. The route segment at which this occurs is often referred to as the *Maximum Load Section*, or *MLS*.

A useful summary indicator to help characterize temporal distribution of ridership is the *peak-hour factor*, α, defined as the ratio of the highest maximum demand, P_{max}, to the total ridership for the day on this same segment and in the same direction:

$$\alpha = \frac{P_{\max}}{P_{\text{total}}} \qquad 0 < \alpha < 1.0 \qquad [\ -\ ] \quad . \qquad 3\text{-}2$$

An alternative definition sums both directions, but this becomes obscure when one direction has far more ridership than the other, and is not recommended. The choice of one hour is arbitrary, so a "peak of the peak" 15-minute time period, or an entire peak period of several hours is sometimes used instead. Obtaining the data contained in a three-axis ridership plot is discussed next.

Figure 3.13 Definitions Related to Passenger Boarding and Alighting



a: boarding and alighting curves

b: passenger count as the difference

Passenger Counts

Passenger count information is central to all route and network analysis. Even if it is a hypothetical design, little can be done without at least assumptions about passenger demand and its distribution along a route. In this section some definitions are provided and applied to enable continued development of further route and network performance indicators.

In general, it is highly desirable to get alighting counts in addition to boarding counts. Let a_j be alightings at stop j and b_j be boardings at stop j. Let there be n route segments. Therefore, there are n+1 stops, but there can be no alighting at the originating terminal of the route, so that a_1 equals zero. Note that b_{n+1} must also be zero since there can be no boarding at the end terminal. For any run, the difference between the accumulated boardings and accumulated alightings after any stop j gives the current passenger count on route segment j:

$$P_j = B_j - A_j = \sum_{i=1}^{j} b_i - \sum_{i=1}^{j} a_i \qquad \text{[passengers]}, \qquad 3\text{-}3$$

where P_j is accumulated passengers onboard after stop j, B_j and A_j are accumulated boardings and alightings after stop j, summed from the stop 1 to stop j. The relevant definitions of the boarding and alighting variables are shown in Figure 3.13a. Below it, Figure 3.13b shows the passenger count, which is the difference of the cumulative boarding and alighting curves. Note that the accumulated number on board on the last segment before arriving at the end terminal, B_n, must equal the total alighting there, A_{n+1}.

The difference between the current passenger count and the transit unit's capacity gives the current level of occupancy. When expressed as a ratio, it is defined as a *point load factor*, α_j, on route segment j:

$$\delta_j = \frac{P_j}{mC_v} \qquad 0 < \delta_j < 1.0 \qquad [-], \qquad 3\text{-}4$$

where, as before, m is the number of vehicles in the Transit Unit and C_v is the passenger capacity of the vehicle type. The point load factor indicates the degree of crowding on any one segment of a route.

To assess the use of capacity over the whole route, the length of each route segment, l_j, is also needed. The total passenger-distance consumed is compared to the total space-distance offered, defined as a *space-averaged load factor* or *utility ratio*. Mathematically, this is the weighted passenger-distance divided by the total space-distance:

$$\xi = \frac{\sum_{j=1}^{n} P_j l_j}{mC_v L} \qquad 0 < \xi < 1.0 \qquad [-] \;. \qquad 3\text{-}5$$

EXAMPLE 3.1

A) The following pairs of boarding and alighting data were collected on a route with 8 stops: (0,15) (2,8) (3,6) (0,5) (2,6) (4,12) (8,7) (40,0). Compute the number or persons on board on each for each route segment and the highest point load factor if the route is served with a bus having a capacity of 60 persons.

Use Equation 3-3. It is helpful to set up a table to simplify the computations, as shown below:

	a_i	b_i	A_i	B_i	P_i
Segment					
1	0	15	0	15	15
2	2	8	2	23	21
3	3	6	5	29	24
4	0	5	5	34	29
5	2	6	7	40	33
6	4	12	11	52	41
7	8	7	19	59	40
8	40	0	-	-	-

Using Equation 3-4, the highest point load factor occurs on segment 6:
δ_6 = 41/60 = .68 passengers/space.

B) Next, assume that the actual O-D matrix is available as given below for the peak direction and that the length of each segment is 0.5 mile. Compute the one-way space averaged load factor.

from/to	1	2	3	4	5	6	7	8
1	-	2	1	0	1	1	3	7
2		-	2	0	0	1	0	5
3			-	0	1	2	2	1
4				-	0	0	1	4
5					-	0	1	5
6						-	1	11
7							-	7

The computation of one-way passenger-miles is straightforward:

$$\begin{aligned}\Sigma_i L_i = {} & 2(0.5)+1(1.0)+0(1.5)+1(2.0)+1(2.5)+3(3.0)+7(3.5)\\ & +2(0.5)+0(1.0)+0(1.5)+1(2.0)+0(2.5)+5(3.0)\\ & +0(0.5)+1(1.0)+2(1.5)+2(2.0)+1(2.5)\\ & +0(0.5)+0(1.0)+1(1.5)+4(2.0)\\ & +0(0.5)+1(1.0)+5(1.5)\\ & +1(0.5)+11(1.0)\\ & +7(0.5) = 101.5 \text{ passenger-miles}\end{aligned}$$

The one-way result is:

$$\xi = \frac{\Sigma P_j l_j}{C_v L} = \frac{101.5}{(60)(3.5)} = 0.48 \text{ passenger - miles/space - mile}$$

Almost half of all available space is used.

In graphical terms, the numerator is the area under the P_j curve in Figure 3.13b and the denominator is the entire rectangular area of height mC_v and length L. It is possible to have a value greater than 1.0, in both the point load factor and space-averaged load factor. Physically, it means that crowding exists beyond the standard used to compute the nominal vehicle capacity. An alternative definition gives the two-way, space-averaged load factor, where passenger-distance is summed in both directions and 2L is used instead of L.

Boarding and alighting counts are straightforward to measure, but they do not provide information on the particular O-D pairs between which individuals are traveling. In order to obtain this information, a means of linking a boarding by one individual to an alighting by the same individual must be established. Rail systems with entry and exit gates reading a fare card can retain this information readily. "Smart Cards" and other advanced stored-value media can add this capability to rail and bus systems not having fare gates, if the media are read on exit as well as entrance. Least reliable, but often the only way to get trip O-D pair data, is to survey passengers periodically. A complete set of information about travel between all O-D pairs is referred to an *O-D Matrix*. Appendix 3.A presents another performance indicator that can be used for estimation when less complete information is available.

The Components of Travel Time

Travel on scheduled public transportation can be viewed as a series of movements and waits, where the relative importance of each varies with the length of the trip and the nature of the waits. Although travel time and its components can be understood to an extent without the aid of mathematical expressions, travel time is ultimately a quantitative concept. The relative size of each term in the series of waits and movements, as well as their relative sizes for alternative transportation choices, is central. The viability of projects can't be studied without this knowledge.

The quality of the travel experience matters as well as the quantity of time. As will be discussed in detail in Chapter 7, numerous studies show that people perceive waiting time as more onerous than in-vehicle travel time. Therefore, waiting time is equivalent to a longer in-vehicle travel time. The usual method of trying to account for quality is to add a multiplication or weighting factor to waiting times. In-vehicle travel time is rarely weighted for perceived service quality, although this too undoubtedly affects ridership in practice. In this immediate discussion, weighting factors are not included as the focus is on actual travel time, not perceived travel time.

Total user travel time using only one link of public transportation can be expressed by the sum of several terms:

$$T_{O\text{-}D} = t_a + t_{wa} + T_1 + t_e \quad , \qquad 3\text{-}6$$

where $T_{O\text{-}D}$ is the time for a user to get from origin to destination, t_a is access time, the time elapsed traveling from the passenger's origin to the boarding point of public transportation, t_{wa} is *waiting time* until departure, T_1 is the in-vehicle travel time, and t_e is *egress time*, the elapsed time traveling from the alighting point to the final destination.

The access and egress times, t_a and t_e, are not always straightforward computations. Walking speed may not be constant if there are grades, staircases, or intersections with long crossing delays. Furthermore, although access time is walking distance divided by walking speed, walking distance is not "as the crow flies". A walker must follow a rectangular grid in most cases. This decreases the number of addresses reachable from a transit stop within a given access time. See Piper (1977) for a detailed graphical explanation of access times within a grid street system. An extreme case of access restriction, perhaps even dysfunctionality, is a modern street network without pedestrian shortcuts that uses cul de sacs (that is, dead-end roads having houses laid out in a circle). Despite physical proximity of an address to a transit stop as the crow flies, the walking distance becomes three sides of a rectangular grid, creating an onerously long access distance. Another example of a serious access restriction, which, if not recognized, would cause a major miscalculation in access time, is a contiguous wall separating residences from the arterial road on which the transit stop is located.

The travel time, T_1, is equal to the travel distance on public transportation, L_1, divided by the *average operating speed*, v_o. The speed is the result of repeating cycles of an acceleration phase, a cruising phase, a deceleration phase, and a standing phase. Average operating speed over a whole line or route is computed very easily by dividing the route or line length by end-to-end travel time between terminals. Operating speed is, in and of itself, a valuable performance indicator. It corresponds with the speed experienced by customers.

Two situations will be analyzed with the help of mathematical expressions. The first is when $T_{O\text{-}D}$ is "short," the second when $T_{O\text{-}D}$ is "long." This dialectic provides insight on which components of the total travel time are important as a function of travel distance.

Short Trips An auto user typically can depart withour delay spending little time in the vehicle before arriving at a nearby destination. Since the travel time by auto is short, in order for transit to be a competitive choice, $T_{O\text{-}D}$ must be short. All time components must be short if the total, $T_{O\text{-}D}$, is to be short. But since transit can have other advantages (e.g., not needing to park,) it need not be as short. However, the shorter it is, the more trips that will be diverted from auto. One way to reduce the waiting time, t_{wa}, of course, is for the passenger to arrive just before a scheduled vehicle. But travelers going a short distance want to be able to depart their origins at random times. If the headway becomes so long that a traveler must consult a schedule, it will be far quicker to use an auto or perhaps even to walk the entire distance.

The general equation for $T_{O\text{-}D}$ can be modified for departure from an origin at random times. Traveling without regard to a schedule, the average waiting time will be equal to one-half of the headway. Thus, the average total travel time for short trips can be written as:

$$T_{O\text{-}D\,avg} = t_a + t_{wa\,avg} + T_1 + t_e = t_a + \frac{h}{2} + \frac{L_1}{v_o} + t_e \quad . \qquad 3\text{-}7$$

By studying the terms, it is apparent that headway, h, is a variable the planner can influence directly as a scheduling decision. The operating speed, v_o, can be influenced by traffic engineering, by right-of-way standard, by distances between stops and by vehicle acceleration and braking rates. But since L_1 is short by definition for short trips, an increase in the operating speed, v_o, will have little effect on total travel time since the term it affects is already small. In order for the access and egress times, t_a and t_e, to be short, the boarding and alighting stops must be nearby. This can be addressed by shortening access and egress paths, by relocating a route, or by creating additional routes. In summary, for competitive short distance travel times, operating speed is relatively unimportant, but headways must be short and walking distances limited.

EXAMPLE 3.2

A) Compute the total travel time for a trip using 3 different modes, with the bus having 3 headways of 5, 10, and 15 minutes.

The total access and egress distance to transit is 0.25 miles (0.4 kilometers). The distance on transit is 1.9 miles (3 kilometers). The other modes follow the same route for the same total distance. Assume the following average operating speeds: walk 3.1 mph (5 kph), bike 9.9 mph (16 kph), bus 12.4 mph (20 kph), and auto 18.6 mph (30 kph). A table is constructed for each travel time component of Equation 3-7:

Time Component	$t_a + t_e$	T_1	h/2	T_{O-D}	Ratio to Auto
Service Condition					
auto	0.8	6.0	0	6.8	1
bike	1.5	11.3	0	12.8	1.9
bus h=5	4.8	9.0	2.5	16.3	2.4
bus h=10	4.8	9.0	5.0	18.8	2.8
bus h=15	4.8	9.0	7.5	21.3	3.1

Note that the bicycle's slower average speed than the bus is more than offset by not needing to walk or wait for the bus. Note also how a longer headway significantly increases the ratio of bus travel time to auto.

B) Compute the total travel time for the 3 modes (bus for h = 5 only) for a range of distances. Start with 0.6 miles (1 kilometer) through 3.1 miles (5 kilometers) in even increments.

	T_{O-D} (minutes)				
Distance T_1	0.6 mi (1 k)	1.2 mi (2 k)	1.9 mi (3 k)	2.5 mi (4 k)	3.1 mi (5 k)
Service Condition					
auto	2.8	4.8	6.8	8.8	10.8
bike	5.3	9.0	12.8	16.5	20.3
bus h=5	10.3	13.3	16.3	19.3	22.3
Ratio of bus to auto	3.7	2.8	2.4	2.2	2.1

Note that ratio quickly becomes more unfavorable to buses as the distance becomes very short. As distances become shorter, persons with an auto may still choose the bus. But there must be adverse offset to the time difference, such as parking being highly inconvenient or expensive.

Long Trips At the opposite end of the spectrum of urban travel are the long-distance trips. In contrast to short-distance travel, the operating speed, v_o, becomes very important to total travel times as L_1 gets long. Within limits, travelers are more willing to consult a schedule for long trips so that t_{wa} can be minimized by shifting of departure times to arrive at the boarding point for minimal wait. They are also willing to walk longer distances or, in terms of the travel time equation, consume more access and egress times, t_a and t_e, since these are small fractions of overall travel time. This is the basis for the common North American rule-of-thumb that people are willing to walk one-quarter mile (.40 kilometer) to a bus stop and one-half mile (.80 kilometer) to a rail station. The difference arises from the generalization that rail trips tend to be longer. The actual distance varies from one individual to the next, of course. In general this distance depends on local walking conditions and is influenced by prevailing cultural attitudes. In developing countries, the distances persons will walk also tend to be longer, due to lack of an auto option and to avoid additional fares.

EXAMPLE 3.3

A) Compute the total travel time for a peak-period trip using 3 different modes, where the regional train has a headway of 30 minutes and passenger arrives either randomly or just timed to meet the schedule.

The total access and egress distance to transit is 0.5 miles (0.8 kilometers). The distance traveled on the train is 12.4 miles (20 kilometers). The other modes follow the same route for the same total distance. Assume the following peak-period average operating speeds: walk 3.1 mph (5 kph), bike 9.9 mph (16 kph), train 24.8 mph (40 kph), and auto 18.6 mph (30 kph).

Again, a table is constructed for the travel time components:

	Travel Time Component (minutes)				
	$t_a + t_e$	T_1	t_{wa}	$T_{O\text{-}D}$	Ratio to auto
Service Condition					
auto	1.6	40	-	42	1
bike	3.0	75	-	78	1.9
train h =30, random arrival	9.6	30	15	55	1.3
train h=30, timed arrival	9.6	30	0	40	0.95

Note that the bicycle is by far the slowest option. (The trip would be physically demanding as well.) The train is reasonably close to the auto even with a random arrival time at the station. The train is actually faster than the auto if one arrives just before departure. In practice, the schedule may or may not be so convenient for an individual, as it could imply wasted time at one or both ends of the trip.

B) Compute the total travel time for the 3 modes for distances ranging from 6.2 miles (10 kilometers) through 18.6 miles (30 kilometers).

	$T_{O\text{-}D}$ (minutes)		
Distance T_1	6.2 miles (10 k)	12.4 miles (20 k)	18.6miles (30 k)
Service Condition			
auto	22	42	62
bike	41	78	112
train h=30, random arrival	40	55	70
train h=30, timed arrival	25	40	55
Ratio of timed train to auto	1.1	0.95	0.89

Note that as the distance becomes longer, the higher average speed of the train more than offsets the access time and travel time becomes steadily better than the auto.

C) How will the ratio change with traffic conditions?

The train mode with right-of-way A will not be affected by congestion, while the auto mode on right-of-way C will slow down. Thus, during peak hours, the travel time advantage shifts in the favor of the rail mode. The impact on bicycles depends upon the degree to which they are separated from the auto traffic stream.

In summary, the travel time on the vehicle becomes the dominant component of travel time as L_1 increases. Therefore, v_o must increase as well in order to be competitive. Wait time can often be self-minimized by the user and persons are willing to accept longer access and egress times for long trips.

Timed-transfer Concept

When services are frequent, waiting times tend to be short, on average only one-half the already short headway of the route for which one is waiting, so that coordination is not necessary. Such is the case with grid systems in large cities, which typically run with uncoordinated services on crossing routes. When services can't be frequent, wait time between alighting from TU 1 and the boarding of TU 2 can be reduced by using the timed-transfer concept. It requires that the various routes arrive and depart at the center of radial and diametrical lines at approximately the same time, sometimes referred to as "pulsing." Further, the travel time between hubs must be slightly less than a multiple of the headway between timed transfer connections in order to provide time for passengers to alight, walk, and board another vehicle. In addition to cutting the wait time, it provides connections from origins on one route to destinations on every other route sharing a timed-transfer meet.

If there are n radial routes operating out of a center (a diametrical route is treated simply as two radial routes connected at the hub), the total number of possible connections from an origin on a particular route to a destination on another is n-1. Since there are n routes, the total number of connections is n(n-1). Thus, a transfer center with four lines has 12 pairs of permutations of one-way connections. The number of route connections, without regard to direction of travel, is n(n-1)/2. Thus, there are six different route combinations that can be paired to perform round trips.

This method also builds demand on individual routes. By collecting passengers with destinations on several different routes this method permits a higher frequency of service on each. This is the same principle most major airlines and package delivery firms use, although they tend to use the terminology *hub and spoke system* to describe the concept. The timed-transfer technique allows even the less popular routes to still meet higher frequency routes. The lower demand routes will operate at a sub-multiple of the higher frequency routes. Only a fraction of the higher frequency runs will connect to them.

Timed-transfer Network

As a region becomes larger in size but services remain infrequent, one of the few viable methods to connect O-D pairs that are not served by routes operating out of the same center, is to establish multiple centers. If there are n_1 routes at the first hub and n_2 routes at the second, counting the interconnecting route only once, there is a total of $n_1 + n_2 – 1$ routes. The number of connection possibilities for each such route is $n_1 + n_2 – 2$. The total number of combinations of route pairs then follows as $(n_1+n_2-1)(n_1+n_2-2)/2$.

If each center is connected by at least one direct route to all other hubs, a maximum of only one more wait time is introduced to any destination on any route radiating from a timed-transfer center. This second wait time can also be minimized if, once again, the TUs meet for a timed transfer. However, having a TU operating from one center meet those from another center is a more difficult proposition than meeting only those at a single center because it requires that both the arrival and departure times from each center be synchronized. Such a system is a *timed-transfer network*, or *TTN*.

As an example, a two-center TTN with four (4) routes each would have a total of 21 pairs of route connections. Connecting vehicles all have travel times of slightly less than 30, 60, or 90 minutes and arrive at both centers at approximately the same time, say three to five minutes before scheduled departure time. At the scheduled time, they all depart simultaneously. The concept is shown in Figure 3.14. Operation of TTNs can create an overriding investment goal of shortening travel time of selected routes in order to reduce them sufficiently to make timed-transfer meets. See Maxwell (1999) for a more detailed description of timed-transfer networks.

Timed transfers have certain operational advantages. As routes are shorter when feeding centers than with through routing, TUs return to the center more often. This provides opportunities for "interlining", where vehicles shift between routes make them more productive. It can also be efficient to shift vehicles of different sizes between routes during the course of the day to reflect changes in demand between them. This will be increasingly prcatical as Intellegent Transportation Systems (ITS) become commonplace. If a rail mode is used, capacity can be adjusted simply by changing the length of the TU. This is an important operating advantage that can be traded off against the higher investment cost in rail infrastructure and vehicles. There will be cost analysis examples showing this advantage in later chapters.

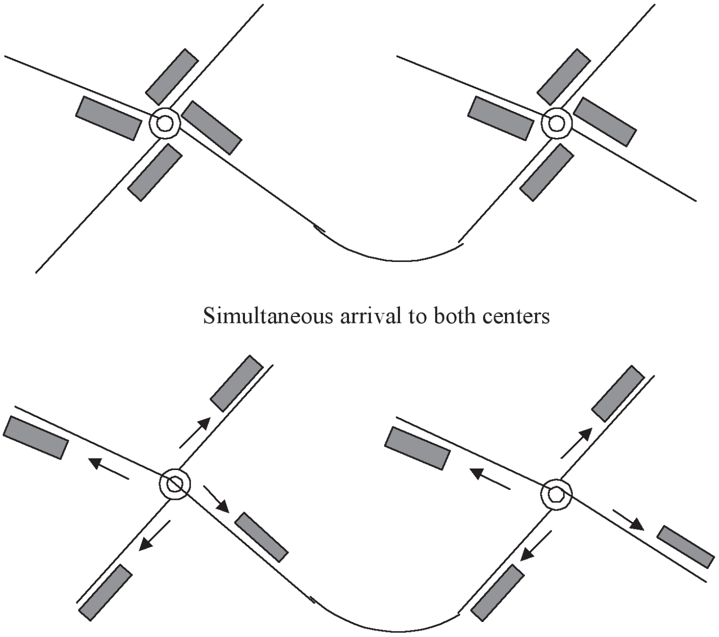
Performance Indicators and Service Improvement

Temporal aspects, such as the need for cyclic operation (a repeating schedule), form network constraints. The fact that timed-transfer services need to meet at approximately the same time in order to maintain network connections introduces constraints on scheduling. Spatial aspects, such as travel distances along routes, topographical features as well as distances to and from depots, form additional network constraints. Performance indicators that quantify some of these effects are explained in Appendix 3.A.

Since routes and the schedules under which they operate are central elements of the product delivered to the public, service must be designed primarily to meet the public's needs. The requirement to both work within the network constraints and yet also provide an effective service (that is, one responsive to the public needs) can limit the efficiency of the use of resources. Thus, a public transport system could have a dedicated work force with efficient work rules, use modern equipment that is inexpensive to operate, and still have relatively low efficiency. The converse could also be true. Nevertheless, in the interest of making the best use of resources, there are almost always measures available that can be taken to speed up service and improve efficiency. Some of these are also presented in Appendix 3.A.

The performance indicators presented in this chapter are summarized in Table 3.3. They are only suggestive. Analysts should always consider creating additional indicators that fit the peculiarities of their particular project. Only four of the indicators are of interest to the system user, while all of them are of interest to the agency planning the service. The two load indicators, operating speed, schedule efficiency, and turnover ratios all focus on an individual route and are computed for individual runs. A time-averaged indicator for each can easily be created for each by summing the individual results for each run and dividing by the number of runs.

Figure 3.14 Two Center Timed-Transfer Network with Simultaneous Arrivals and Departures



Simultaneous arrival to both centers

Simultaneous departure from both centers

INTERACTION BETWEEN ROUTE AND NETWORK DESIGN

The effect on the network from the existence of a route must also be considered. The complex interaction between variables, however, makes estimation of the effects from changes to a given route upon the network a tricky business. This holds true even with the assistance of mathematical models available to some analysts. These may overcome large computational burdens but do not overcome the often limited understanding of these interactions. Therefore, one must try to anticipate relevant interactions and consider them without the aid of models.

Different Network Configurations Having an Equal Operating Budget

A route's performance before and after a change will be evaluated differently if it is viewed as an independent entity instead of an element in a more complex mechanism. Nowhere do these differences between individual route performance and network performance come into sharper relief than in the issue of introducing transfers. The differences will be analyzed with the aid of two recent studies. The first study analyzes an isolated route as through service and the same service involving an intermediate transfer. The second study analyzes the different factors that influence route design and transfers between routes.

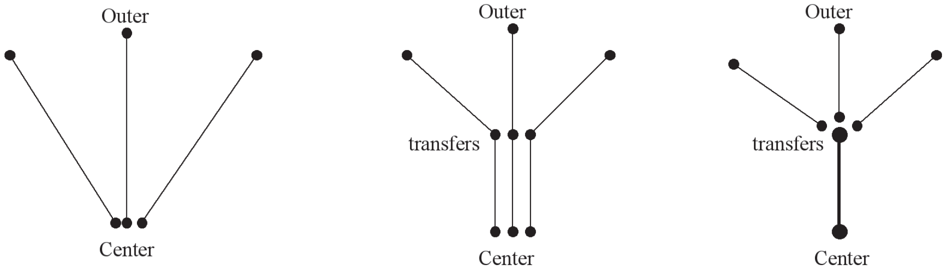
Liu, Pendyala, and Polzin (1998) provide an analysis of the effects of transfers on ridership on a single route using a mathematical simulation model from a New Jersey Transit study of the New York-New Jersey commuting corridor. It thus builds its analysis upon aggregated statistics that include potential passengers from high-income, exclusive neighborhoods of lower-density as well as urban poor living in much denser, traditional communities. Not surprisingly, ridership decreases when routes are broken into shorter sections and transfers are introduced. They find this result "discouraging" because even though steps can be taken to decrease the transfer wait time, the transfer itself is not eliminated. This is too pessimistic, since their underlying assumption is one of constant frequency of service. If more than one route can be altered at the same time to allow transfers between them at a common location, as is usually the case in real networks, a more relevant assumption is one of equal operating budgets for several alternative bundles of services within the same area.

TABLE 3.3 Some Route Performance Indicators

	Range	Boundary Conditions	Comments
Frequency (Headway = 1/f) U	f_{min} < f <f_{max}	f_{max} is function of vehicle, control system and RoW fmin is set by policy	increases with demand diminishing returns from increase to already high f operation near f_{max} is unreliable
Operating speed U,O	0 < v_o <= v_{max}	vo=v_{max} express operation	increases with higher standard of right-of-way, with signal priority, and longer station spacings
Line Capacity U, O	C>=0	C=Cmax when f=f_{max}	operation at Cmax can not be maintained due to impossibility of capacity recovery after any delay
Point load factor U,O	0 < δ_j <= 1.0	δ_j=1.0 full space usage on route segment j	δ_j>1.0 overcrowding beyond standards on route segment j increases towards center on radial and diametrical lines
Peak hour factor O	0 <α<= 1.0	α=1.0 all demand only in this one hour	commuter-oriented services tend to be higher
Space-averaged load factor O	0 < ξ<= 1.0	ξ=1.0 full space utilization ξ>1.0 crowded beyond standar	radial tends to be lower, tangential and grid tend to be higher ξ>1 only possible with overcrowding on many segments
Deadhead factor O	β>= 0	β=0 route terminus at depot	base services tend to lower factor, supplemental peak services higher, temporary storage near route terminus can reduce ratio
Schedule efficiency ratio O	γ<= 1.0	γ=1.0 using minimum terminal times	ratio tends to decrease with increasing headway ratio tends to increase with improved travel time reliability
Direction balance ratio O	η >= 0	η=1.0 balanced demand >1.0 off-peak is higher	commuter-oriented and radial services tend to be lower
Turnover ratio O	τ>= 1.0	τ=1.0 no turnover	radial tends to be lower, tangential and grid to be higher radial improves with short-distance fares on outer segments

U = of interest to user, O = of interest to operator

Figure 3.15 Comparisons of Three Service Configurations with Similar Operating Budgets



Direct radials	Radial branches sharing trunk section	Radial feeders with separate trunk section
Short travel time to/from center	Longer travel time to/from center	Travel time could be longer or shorter
Low frequency on each line	Low frequency on branches	High frequency on feeders
Connectivity only at Center	High frequency on trunk	High frequency on trunk
No transfers required to/from Center	Connectivity between Outers	Connectivity between Outers
	No transfers required to/from Center	Transfers required to/from Center

As an example, Figure 3.15 shows a commonplace situation where there are three basic types of service configuration that might serve a subregion distant from the center of a network. Assuming a similar operating budget for each one, a listing of some key attributes or features is provided for comparison. The first is direct service to the center on each route, another merges them at a common trunk section, and the third breaks the routes into segments terminating between outer and trunk sections, mandating a transfer to continue towards the center.

The leftmost or "direct radial" configuration in Figure 3.15 provides direct routing towards the center. It provides the shortest travel time and direct service to the center. But it requires travelers going to destinations on other routes to transfer in the center. Thus, except for destinations on the same route, routing is circuitous and travel time is long.

The middle or "radial branches sharing trunk section" configuration reroutes travelers to and from the outer branches to a common trunk section. It improves connectivity because it opens up the possibility of transfers between outer points on the branches without going all of the way to the center first. Focusing on only one route highlights the negatives of a more circuitous trip for those going to the center and the need to transfer for those going between outer points, while neglecting the positive features. One of these positive points is the very existence of connections between outer origins and destinations without going all of the way to the center. In reality, there may never be a more direct route between them when there is low demand. If there is, it will likely be of low frequency.

There is another potential positive for the network as well from the "radial branches sharing trunk section" configuration. It increases the frequency of service and capacity along the corridor into which all of the routes have been funneled. Thus, this reconfiguration would be very attractive in situations where one corridor needs more service and the inner areas through which the other routes would no longer operate still have sufficient coverage from other routes.

So far, the middle configuration in Figure 3.15 has not enabled increased frequency on the branches. For a route operating with a given vehicle size, replacement with smaller vehicles with lower operating costs is always a possibility in order to increase frequency along the entire route. But this is often impractical. Operating costs don't go down in direct proportion with vehicle size, so total capacity would be reduced. Moreover, if the trunk route is already congested, more vehicles only aggravate the situation.

Now the logic of the rightmost configuration in Figure 3.15, the "radial feeders with separate trunk section," becomes clear. If connectivity between outer points is important, traffic needs to meet at a common point closer to the outer ends. If this has to be done in any case, a cooperative use of modes can improve frequency on the branches as well as the trunk. By operating fewer but higher-capacity TUs on the trunk section, the saved vehicle-hours from the smaller vehicles that otherwise would have operated on this section can be reinvested into increasing the frequency along the branches. Thus, even better connectivity can be offered to all points. The Liu, Pendyala and Polzin model did not consider this possibility but did recognize the negative attribute of the separate trunk section configuration: the fact that transfers are now mandatory for all travelers between the trunk and branches. Nevertheless, on balance, this is often a good trade-off when viewed from a network perspective.

There is empirical evidence to support this contention. Thompson and Matoff (2003) compared nine U.S. transit systems. They were categorized as:

- traditional all-bus radial systems serving the downtown;
- mixed bus and rail radial systems serving the downtown;
- express bus services superimposed on the radial services; and,
- what they refer to as "postmodern" systems, which are essentially trunk-feeder systems.

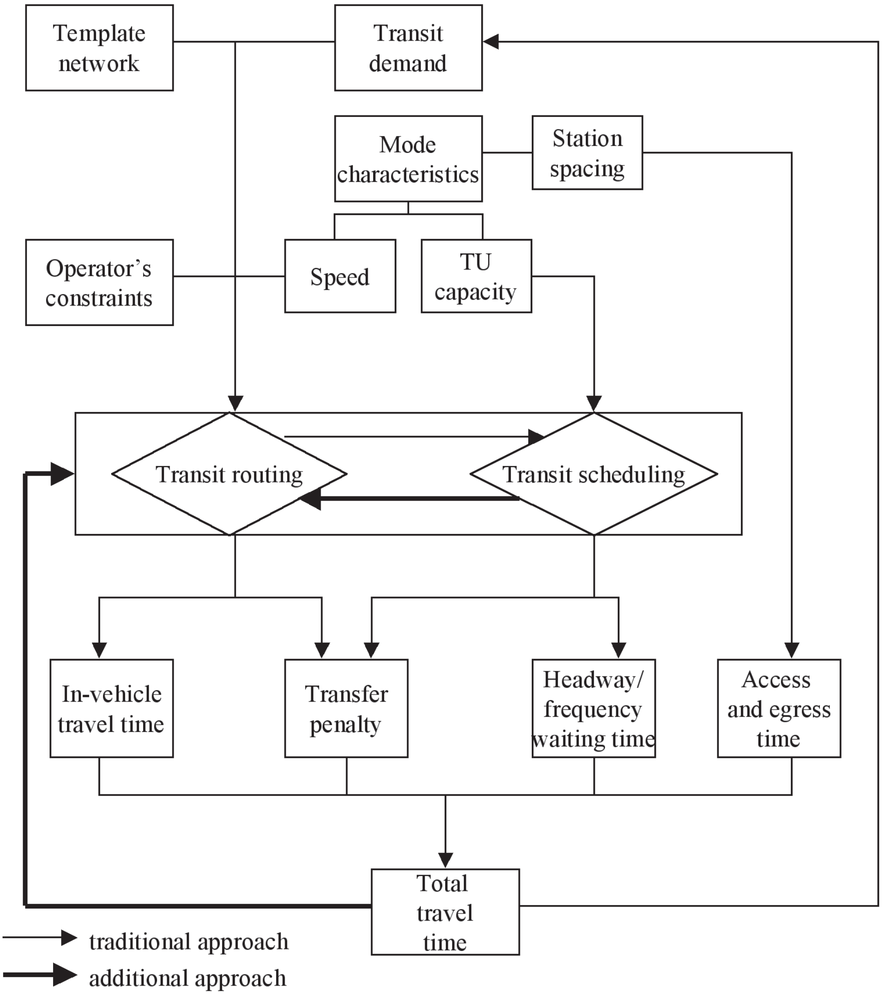
They found that the postmodern systems had lower costs per passenger and that they had more off-peak trips. This is as expected because they had a larger number of O-D pairs available throughout the day. Furthermore, the traditional radial systems had declined in ridership while the express bus services showed modest ridership gains, but the postmodern systems showed the largest ridership gains.

The need to accommodate an even larger accumulation of passengers with time can also be an argument for the use of a separate trunk configuration. At some point, an upgrade of the trunk to the ROW B or A standard becomes justified. Many large public transport networks with very high ridership use the radial feeders with separate trunk configuration and witness a large volume of transfers. When rail technology is used, another advantage of this configuration is that peak demands on the trunk can be accommodated at low marginal cost, through the simple lengthening of TUs. Some examples will be given in the chapters on cost estimation.

Routing and Scheduling as a Feedback Process

The traditional method of analyzing networks is to separate the routing phase from the scheduling phase. Routes are first created that reflect the designers' knowledge about travel patterns and existing available infrastructure. Scheduling then follows as a procedure to match supply with demand. But ideally there should be feedback. The total user trip time of passengers should be considered and fed back to the route generation stage. In actuality, access and egress times usually treated as constants can instead be treated as variables to be influenced by the location and number of routes. The waiting times are also influenced by frequency of service, which again depends upon route network design. The in-vehicle travel time, too, is influenced by the *circuity* of routing, defined as the ratio of actual path traveled to the shortest path. Once again, this is a function of network design. Figure 3.16 shows scheduling that incorporates all of these aspects in the routing analysis in an iterative process.

Figure 3.16 Relationship Between Routing and Scheduling



Source: Lee 1998

In practice, routing and scheduling have rarely been done simultaneously except in the most theoretical manner to gain general insights. The problem is too complex mathematically; the complete form is an optimizing mathematical program of astronomical computational size and fearsome complexity, which includes integer variables and nonlinear constraints. Approaches have been used historically that reduce the problem size by discarding unlikely routing possibilities early in the process and through mathematical simplifications that exclude a few of the variables affecting total user travel times. Optimization is then aimed at one of three objectives:

1) Minimization of the sum of all user travel times (not including access and egress times)
2) Minimization of operating costs as approximated by some combination of the total number of vehicle operating hours and total distance operated
3) Minimization of total social cost, which is, in effect, some combination of low user travel times and low operating costs

In the third optimization objective, the one most applicable to real network designs, the results are typically candidate route network designs that provide a paired set of a good service solution to the public with a low-cost solution to the operator. This requires a sophisticated analyst who understands the model limitations, can pre-reject poor candidate solutions, and can then select among results that require multiobjective trade-off. These stringent requirements have resulted in its limited applicability in the real world to date.

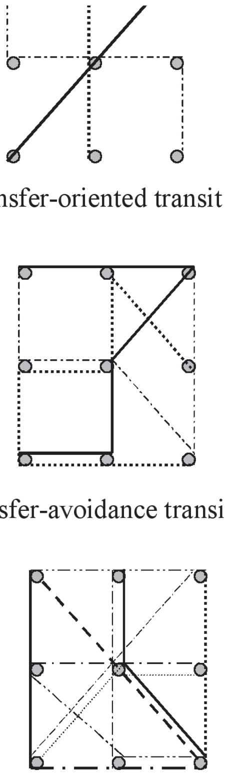
Recent research by Lee (1998) fully incorporates all elements of user total travel time. It avoids the massive optimization problem and resulting implications for advanced analyst ability as well as undesirable simplifications that partially defeat the purpose of the analysis. Instead, starting with a given O-D matrix, it initially provides direct connections where passenger flows are high, and frequent connecting routes from areas where passenger flows are lower. It then iteratively adjusts routes and frequency of service on routes (scheduling) until the minimum sum total of all user travel time is achieved. This is not a true optimization procedure but provides a “near optimum.” Its results are insightful and the approach can even be implemented for real-life project analyses.

In Lee’s model, there are three critical input variables the analyst can adjust: level of passenger demand, travel time on links, and the transfer penalty. This penalty takes the form of weighting factor applied to waiting times. The first two types of inputs are clearly grounded in physical reality, whereas the transfer penalty represents an average that could change with time, change with the types of passengers, and change with quality of transfer facilities. It must be seen for its relative rather than absolute influence on results.

Lee describes three distinctly different types of networks generated according to the relative values of the input variables: demand level, travel speed, and the transfer penalty. Figure 3.17 is an aid to help to explain the differences. It has a matrix of nine points, each of which must be served as both an origin and destination. The demand between the various O-D pairs and the operating budget are both fixed. Each different route connecting some of these points uses a different line type to distinguish it.

The *transfer-oriented network* is Figure 3.17a. It consists of relatively few, short routes

Figure 3.17 Three Types of Transit Networks



a. transfer-oriented transit network

b. transfer-avoidance transit network

c: directly-connected transit network

Note: different line styles are used to indicate individual routes

Source: Lee 1998

with relatively high frequencies, and tends to provide moderate in-vehicle travel times. Two of the routes are short, connecting only three points, the third extending between five points. Transfers take place at the middle of the network between three diametrical routes. Despite transfers, the waiting times are relatively short, due to the relatively high frequencies.

The *transfer-avoidance network* is shown in Figure 3.17b. There are still only three routes, but they meander. Two of them connect six points, the third connecting seven. It has somewhat lower frequencies because longer routes spread the limited number of vehicles the operating budget can support farther apart. It also has longer in-vehicle travel times on average for its passengers because of longer, more circuitous routing. There may still be attractive transfer points, but the average waiting times will be somewhat longer since frequencies are somewhat lower.

The *directly connected network* is shown in Figure 3.17c. There are now 10 routes, none of which connects more than three points. Although the routes are short, because they are so numerous, average frequency must be substantially lowered. It has the shortest travel times underway and relatively few transfers because of direct connections for the majority of popular O-D pairs. But because frequency is lowered, there is likely to be a longer wait for a vehicle. Furthermore, for those trips that do require a transfer, a long wait time is again likely.

Lee also studied a much larger hypothetical simultaneous routing and scheduling network. It was chosen for its use by other modelers as well. Thus, the interested reader can begin with Lee's solution and compare it with other approaches. Lee found solutions that could be characterized as each the three aforementioned types of networks. He subjected the network to "high" or "low" input values for demand, travel speed, and transfer penalty, in various combinations. Some interesting relationships between network type and the levels of these inputs were identified.

Each network type is evaluated as being incompatible/inferior performance, adequate performance, or good performance after being subjected to these different combinations of input values. The transfer-oriented network can give good or adequate performance at any level of demand. High travel speeds actually favor it because speed can compensate for the long waiting times for connections riders might expect under low demand conditions. Not surprisingly, and indeed by definition, a high transfer penalty discourages use of a transfer-oriented network. The transfer-avoided network, also as to be expected, has good performance with a high transfer penalty, but only in combination with a high travel speed. This is because circuitous routing combines with low travel speed to cause excessively long user travel times. Nor does the transfer-avoidance network perform well when demand is high. If demand is high, frequent service can be justified; the circuitous routing causes more delay than would be caused by transfer wait times. Lastly, the directly connected network seems to work adequately with most combinations of input values, except for the important and common case of low demand levels. This is not surprising because frequencies must then also be low and, consequently, wait times must be long. These results are summarized in Table 3.4.

TABLE 3.4 Conceptual Relationship Among Network Types and Critical Inputs

	Demand		Travel Speed		Transfer Penalty	
	High	Low	High	Low	High	Low
Transfer-oriented network	O	O	O	-	X	O
Transfer-avoidance network	X	O	-	X	O	-
Directly connected network	O	X	O	O	O	O

"O" good or adequate
"X" inferior or incompatible
" - " no strong relationship

Source: Lee 1998

Liu, Pendyala, and Polzin's conclusion that not properly including the transfer effect will overestimate ridership can be elaborated upon and somewhat disputed with these additional insights. Looking at the parameters they used, it represents a composite of potential users, so it is not known what the specific transfer penalty would be for subclasses of potential users. The penalty would be lower for captive riders having few alternatives, the young and fit, leisure travelers, and others with trips of a nonwork nature. It would be higher when and where the network caters to auto owners, higher-income persons, and so on. Since the transfer penalty will vary from high to low given this wide range of user types, it would be more accurate to use a prediction model where the high transfer penalty is applied to only part of the population. Moreover, the effect on ridership might be offset, perhaps even more than offset, by the increased number of destinations and frequency enabled by the existence of a transfer point.

Many small networks in the developed countries would likely be classified according to Lee's scheme as having low demand and low travel speed inputs, with transfer penalty ranging on the low side. the low demand and low speed correspond to pedestrian unfriendly, congested arterials. The low penalty follows from the type of ridership that can be expected. Most persons with a high penalty would use autos almost exclusively. Under these circumstances, referring again to Table 3.4, the direct-connected network and transfer-avoided network would both be ill-advised. The former because the service frequency would be very low, the latter because the trip times underway would be very long. This leaves the transfer-oriented network, as it is the only one that can give overall adequate service under these circumstances. This probably explains the widespread use of timed-transfer centers in smaller or minimally funded networks.

Network Economies

The current discussion, even though it refers to research using mathematical models, really does not express controversial conclusions. It is well known that there are *economies of density*, meaning that high density of demand makes route and networks more efficient. It is also well known that there are *economies of scale*, meaning that large systems can collect passengers from more routes who are interested in connecting to any other particular route, again raising efficiency. The small systems, minimally funded systems, and low-density systems that evolved towards transfers did so because it is the best option they have to the challenge of operating in far-from-ideal circumstances.

Network operations research models raise one more interesting issue. Depending upon the objectives set for them, a different right-of-way in the same corridor could have been selected for a high-capacity, high-frequency operation with parallel routes receiving much less service or finding themselves redesigned to feed the higher capacity route. Conversely, several parallel routes of similar capacity and frequency might have been created. This throws into question the concept of *cross-subsidization.* Economists arguing for deregulation of buses in the UK argued that better performing routes should not have to subsidize inferior ones. But when viewed from a network perspective, individual routes do not exist only to serve particular communities, but to serve larger objectives like the need to concentrate passenger flows for the sake of high-frequency service. Allowing higher fares to be charged or service to be removed because a route is not "profitable" might disrupt system design and arbitrarily penalize those using the inferior route. The empirical evidence from the early years following the UK deregulation did indeed show network performance degradation through large ridership losses after reduction of service on some unprofitable routes (Pickup et al. 1991; Fawkner 1995).

DEMAND-RESPONSIVE SERVICES

When performance indicators reveal that a service is below an acceptable productivity threshold, alternatives to fixed routes can be considered. They can serve low-demand areas, perhaps in the range of 1 to 10 persons per vehicle-hour. Since they usually use smaller vehicles, they can go where larger vehicles can't fit or are not welcome. Of increasing importance is the accommodation of persons with disabilities that deter or prevent them from using fixed routes. Demand-responsive services can also serve infrequently made trips between O-D pairs that fixed routes can't serve effectively. In sum, they can extend the network into fringe areas and to people who otherwise could not be accommodated.

The key characteristic of all demand-responsive services is that they depend upon the specific requests they receive. In addition, such services may have no preset route and offer nonexclusive rides, such that pickups and drop-offs overlap. The cost of providing service and the quality of the service are highly dependent upon the rules used to assign trips to vehicles. Taxis provide the best service by having no schedule, by responding immediately, by having no route of any kind, and, in most cases, by carrying only one party. Accordingly, they usually cost the most to provide. To control costs, publicly funded services usually offer a less exclusive service.

Selecting a demand-responsive service design and its operating rules is a large topic all by itself. There is also controversy over what the best solution is given the specifics of the operating environment and the potential ridership. This is because it can be very hard to estimate what the costs of an alternative service would be in order to make a comparison. Furthermore, services are often regulated as to the objectives that they must meet in order to prioritize use of a tight budget. These constraints can make both analytical studies and experimentation problematic. As a result, planners rely heavily on peer examples. They also depend upon software assistance to design services, even if the underlying assumptions imbedded in this software and the implications for this proposed service design may not be fully understood.

The design principle for basic taxi service is quite simple. First, distribute taxis to areas where experience has shown that demand is likely to be present. Some demand is then matched simply by hailing from the street. The remaining trip requests are accepted by telecommunications and assigned to vehicles nearby. In some applications, each vehicle's location and status is continually reported to the dispatcher to improve the assignment process. In smaller towns, the vehicles may just be waiting at a stand since the response time is short under all circumstances.

The "on-demand" fixed-route service design is also quite simple. The vehicle operates on a fixed route, but it is only dispatched if at least a service request is made, there- by saving resources when no one will be riding it.

Other service possibilities quickly become more complicated. There are basic design principles that are typically used, however, whether through manual effort or computer algorithms. The first step is to build "skeleton routes," or "quasi-routes" around *subscribers* (that is, persons who make recurring requests to be picked up and dropped off at the same times and same locations). The next step is to insert other trip requests into the closest quasi-route having enough available time to make the additional pickup and drop-off. After all trips have been initially accommodated, then revisit all routes to see if some trips could be swapped between routes or could be reinserted into other routes, given the changes that have occurred since the initial insertion.

The same quantitative productivity and efficiency indicators used for fixed routes, such as passengers per hour, cost per passenger, and cost per unit passenger-distance, can still be used to assess whether proposed changes are improvements. In some operations, it is permissible to use taxis and other occasional providers to accommodate trips that don't fit well into the quasi-routes. Although it raises coordination issues, this can be less expensive than using a larger vehicle carrying only one person or one that must deviate far from all other pickups and drop-offs.

There are further complications to developing schedules. Callers making requests must either be assigned a trip immediately in a real-time scheduling process or must be called back after many are scheduled simultaneously in a *batch scheduling* process. If one assigns a trip immediately to obviate the need for a callback, opportunities for swapping and reinserting trips are diminished. On the other hand, there are costs of having to call back. This increases the time and complexity of making reservations for the call taker, while also forcing the trip requestor to give longer advance notice.

Another complication is that there are pickup and drop-off time "windows." A requestor can often not be given the exact time they requested but is given a range of time when the pickup and drop-off can be expected. The wider the window, the easier it is to develop efficient schedules. On the other hand, this detracts from the quality of the service the requestor receives.

Yet another complication is the total time any rider can spend on board. It is not reasonable to hold someone captive as the vehicles meanders around until, eventually, the vehicle comes to the vicinity of their drop-off point. Thus, there is a constraint of maximum time on-board. In the U.S., this is typically set at about twice the amount of travel time it would take to go from the origin (pickup point) to the destination (drop-off point) using the fixed-route network.

The complexity of developing schedules and returning calls to trip requestors is such that, at some point, a scheduler can't manage all of the constraints effectively. Even if they can be met, the solution is not likely to be very efficient. Thus, computer scheduling/dispatching packages are used in all larger operations. Based on a survey of the industry, some researchers set a threshold for fleet size where computer assistance becomes necessary at about 30 vehicles (Lave, Teal, and Piras 1996). On the other hand, such software also takes time and skill to set up properly. It requires adequate and current information about travel times along arterial roads and map coordinates that link to street addresses. The results are also very sensitive to the parameters that reflect the time windows allowed, the maximum time allowed on board, and to rules about connecting passengers with other services.

There may also be eligibility requirements for the persons requesting trips. If eligibility is strict, at first glance it seems that demand can be reduced and costs contained. On the other hand, it may actually lower productivity and waste resources as vehicles travel through neighborhoods where latent trips by noneligible persons are denied. Technological improvements are continually influencing such trade-offs in favor of more complicated service designs that might combine previously separate passenger-market niches. These can blur traditional boundaries between operating domains of public transport agency fixed and demand-responsive services, human service agencies, and private transportation. This will be discussed in more detail in the next chapter.

SUMMARY

This chapter began with a description of spatial network types. There may be subnetworks of different types within a region that reflect road-building patterns of different eras, different topographies, and different modes. Networks tend to evolve as cities and demand grow and as travel distances become longer to include higher-capacity Transit Units and faster modes.

Three primary right-of-way standards for individual routes were introduced. ROW A is total separation from all other traffic, and usually requires tunnels, elevated sections, and other measures requiring substantial investment. ROW C is simply operation in mixed traffic, with little distinction from other vehicles. ROW B provides lateral separation from other traffic but, because of intersections, not full separation. This gives performance in terms of speed and reliability somewhere in between A and C.

Passenger demand varies by time of day and location. Thus temporal and spatial distributions of demand are needed to form a picture of the travel needs to be satisfied. The three-axis diagram was presented as an insightful method for displaying this information. Indicators were introduced to characterize passenger demand and usage of available capacity.

The methods that can be used for data collection and the completeness of the information collected depend upon the installed technologies. Manual methods are tedious and expensive and are therefore done infrequently. One of the most important is the fare collection technology, but there are also devices made specifically for automatic passenger counting.

The various components of travel time (access time, waiting time, in-vehicle travel time, transfer waiting time, and egress time) were analyzed as to their relative importance as a function of travel distance. Key points are that access and egress time must be shorter and service frequency must increase for shorter-distance trips and speed must increase with longer-distance trips in order to be competitive with the automobile. Further, the timed-transfer concept is often used to reduce transfer wait times and to connect O-D pairs on different routes.

The need to run cyclic schedules and the inherent properties of infrastructure alignments introduce constraints that limit efficiency of time usage. Some suggested route performance indicators were defined that help to characterize this efficiency both in comparison to other routes in the same network and against peer routes elsewhere.

Route interaction with the network, especially the issue of transfer time being weighted more heavily by users than in-vehicle travel time, was analyzed with the aid of two research studies. The study by Lui, Rendyala and Polzin et al. looked at a route mostly in isolation. It studied the effect of a break in journey versus a single ride from origin to destination. It found that ridership loss could be significant but did not include the possibility of an offsetting effect from the additional destinations available at the transfer point. The study by Lee was done in a network context. Briefly stated, Lee's network analysis classified three network types: transfer-oriented, transfer-avoidance, and directly connected. They were described as to how well they performed as a function of "high" and "low" levels of demand, travel speeds, and transfer penalties. The results were summarized in Table 3.4. Similar to the isolated route study, Lee found that when demand is low and speeds are low (the most difficult operating environment), a high transfer penalty argues against a transfer-oriented network. But the additional insight from Lee's model is that the other types of networks perform even worse in this same situation. In smaller networks, high transfer penalty travelers probably use autos and can't be attracted anyway. Instead, most potential riders are those who have a low transfer penalty. Thus, transfer-oriented networks are commonly used in smaller and minimully funded networks.

There are economies of scale and economies of density in transit networks. But the concept of cross-subsidization of routes used by some economists was challenged on the grounds that there is some latitude in network design. Which route is chosen to become a high-capacity trunk line can be somewhat arbitrary. Furthermore, routes often serve a network purpose beyond serving origins and destinations only along its own length.

Fixed routes are supplemented by demand-responsive services. These are characterized by a lack of a preset route, a preset schedule, or both. These services are applied to very-low-demand areas or to accommodate riders with special needs or disabilities that prevent them from using fixed routes. The service design principle usually involves building ad-hoc, daily, quasi-routes that build around subscription users. There are complications such as pickup and drop-off windows, and circuity restrictions. As demand-responsive operations get larger in scale, efficiency and productivity can be greatly enhanced by scheduling software.

Methods to speed-up operations and increase efficiency are described in the Appendix 3.A. The least expensive is usually just to increase transit stop or station spacing. Vehicle acceleration and braking rates can be increased. Public policy can also be changed such that merging transit vehicles have the right-of-way. Traffic Signal Priority that favors transit is becoming easier to implement as it becomes localized at only one intersection at a time, although queue bypasses are sometimes necessary if it is to be effective. Reducing dwell times at stops can be accomplished through faster fare collection techniques, public education, and the use of vehicles with more door channels. Short-turn versions of routes can sometimes be created on long routes. Extending routes can sometime be done as layover time permits.

APPENDIX 3.A

Indicators of Efficient Use of Resources

In order to quantify the effects of scheduling constraints, the basis scheduling equations need to be analyzed. The fundamental relationship for a simple route where vehicles cyclically repeat is:

$$T = Nh \qquad 3\text{-}8$$

where T is the cycle time, or time for a vehicle to return to its initial position, and N is the number of Transit Units required to maintain a constant headway, h. N must be integer, while h is usually divisible into 60 minutes (e.g. 2,2.5,3,4,5,6,7.5,10,12,15,20,30,60) so that the schedule can repeat hourly ifor convenience in memorization. As a consequence, T can take on only discrete values.

If a route is of length L, having an operating speed v_o in both directions, then T is composed of the two travel times underway plus two terminal times, one for each route end:

$$T = 2\frac{L}{v_o} + tt_1 + tt_2 \qquad tt_{\min} \leq tt_1,\ tt_2\ . \qquad 3\text{-}9$$

The value tt_{min} is the minimum time that a vehicle can be scheduled between arriving at a terminal and then departing in the opposite direction again. This is set by the minimum time specified or contractually required for vehicle operator breaks and by the need to reposition TUs when the passenger alighting and boarding locations are not the same. In practice, there often is a need to add extra time, or *slack*, to schedules to allow for delays from congestion, heavy passenger loads, and other randomly recurring events. Also, in practice, in order to meet the constraints on the cycle time that arise from the integer restriction on N and clock headway restriction on h, either or both of the terminal times must usually be extended anyway. Thus, a TU may have to stand at terminals for periods longer than tt_{min}. In so doing, it stands at a terminal instead of doing productive work. In this way, scheduling inefficiency stems from the requirements to not run behind schedule repeatedly and/or to maintain cyclic operations. It can be useful to define a ratio of the minimum possible cycle time to the schedule-constrained cycle time, or schedule efficiency ratio:

$$\gamma = \frac{T_{\min}}{T} = \frac{2\frac{L}{v_o} + tt_{1\min} + tt_{2\min}}{2\frac{L}{v_o} + tt_1 + tt_2} \qquad \gamma \leq 1.0 \qquad [\text{-}]\ . \qquad 3\text{-}10$$

Scheduling relationships will be developed and used further in connection with the estimation of the cost of operating routes in a later chapter.

EXAMPLE 3.4

A) A route has the following properties: operating speed of 11.16 mph (18 kph), a one-way length of 6.82 miles (11 kilometers), and a headway of 15 minutes. The drivers' union agreement states that total terminal time for each round trip should be at least 15 percent of the total travel time. Find the number of vehicles required, the minimum cycle time, the actual total terminal time, and the schedule efficiency ratio.

$$2L/v_o = 2(6.82\ \text{miles})/(11.16\ \text{mph}\ /60) = 73.3\ \text{minutes}$$
$$tt_{min1}+tt_{min2}= 0.15(73.3\ \text{min}) = 11\ \text{minutes}$$
$$T_{min} = 2L/v_o + tt_{min1}+tt_{min2} = 84.3\ \text{minutes}$$

T = N(15) is also a requirement, where N is integer. To not violate the union constraint, T_{min} must be rounded up, not down, to find the actual cycle time. The fleet size must be 6, as it is the first integer value giving an actual value higher than 84.3 minutes:

$$T = 6\ (15) = 90\ \text{minutes}$$

The actual total terminal time must then be:

$$tt_1+tt_2 = 90 - 73.3 = 16.7\ \text{minutes}$$

The schedule efficiency ratio follows directly from substitution of values into Equation 3-10:

$$\gamma = \frac{T_{\min}}{T} = \frac{84.3}{90} = 0.94$$

APPENDIX 3.A (CONTINUED)

Another source of unproductive time is deadheading, the repositioning movements of vehicles when they are not in scheduled revenue service. These stem from travel between the depot and the starting and ending terminals for a day's work, and from any repositioning between routes during the day. It can be useful to define a ratio of total vehicle-hours to revenue vehicle-hours, or *deadhead factor:*

$$\beta = \frac{\text{Total vehicle - hours - Revenue vehicle - hours}}{\text{Revenue vehicle - hours}} \qquad \beta \geq 0 \qquad [-]\ .$$

Some care is needed when comparing routes, especially routes from external peer systems. What constitutes revenue vehicle-hours can be somewhat arbitrary to define, particularly when demand is unbalanced in opposite directions of a route. It is in fact very common during certain hours for demand to be much higher in one direction. Even when accepting passengers in the off-direction, a Transit Unit's primary task may be to reposition to the peak direction. In some cases, it may use an express path like a parallel highway instead, in order to return to the peak direction more quickly.

EXAMPLE 3.1 (CONTINUED)

[Please note: This example is an extension of Example 3.1 found on page 56.]

C) Compute the two-way space-averaged load factor, if the same vehicle deadheads on the return trip.

The two-way factor is found by doubling the length and adding zero passengers for the second direction:

$$\xi = \frac{\sum P_j l_j + 0}{C_v 2L} = \frac{101.5 + 0}{(60)2(3.5)} = 0.24 \text{ passenger - miles/space - mile} \quad .$$

Note the dramatic 50 percent reduction in space efficiency.

APPENDIX 3.A (CONTINUED)

In the interest of a better comparison, one can create a threshold of ridership in the off-direction, below which the movement is declared a deadhead. The total boarding counts for the peak and off-peak directions during the peak period are compared as a *direction balance ratio*:

$$\eta = \frac{\text{Total boardings in off - peak direction}}{\text{Total in peak direction}} \qquad \eta \geq 0 \qquad [\text{-}] \; .$$

For example, if an agency chose a threshold of less than or equal to .10 and the calculated value was less, most off-peak direction runs would be considered deadhead runs as would those at its prospective peers. Their associated service hours would be subtracted from the revenue vehicle-hours total. The exceptions would be the few runs made to meet any minimum frequency standard, as some base service is always offered regardless of demand. In this way, comparisons can be made using a common definition.

Absent the complete picture that O-D pair information provides, there is another indicator to characterize demand that can be of practical value. Even without knowing the lengths of segments between stops, the ratio of total boardings to accumulated passengers arriving at the end terminal, or *turnover ratio*, roughly implies the length of trips. Mathematically, this ratio is:

$$\tau = \frac{\sum_{j=1}^{n} b_j}{P_n} \qquad \tau \geq 1.0 \qquad [\text{-}] \, . \qquad \text{3-11}$$

The word "turnover" alludes to the concept of space being occupied by more than one passenger over the course of a run. A ratio of exactly 1.0 would mean that no passengers alighted before the terminal. This is approximately the case with many radial commuter services, particularly in the peak period. Other types of alignments, particularly tangential or cross-town grid routes, or perhaps even the same radial route at a different time of day, would have a higher ratio. Since turnover is proportional to fare revenue collected, this is a particularly important indicator of financial performance when a complete O-D matrix is not available. In general, services that accumulate passengers have a lower turnover ratio and collect less revenue per unit passenger-distance than those with higher turnover ratios. If the fare structure is flat (that is, if fare is the same regardless of distance) this is strictly true.

EXAMPLE 3.1 (CONTINUED)

[Please note: This example is an extension of Example 3.1 found on page 56.]

D) Compute the turnover ratio for this example. Use Equation 3.11:

$$\tau = \frac{\sum_{j=1}^{n} b_j}{P_7} = \frac{15+8+6+5+6+12+7}{40} = 1.48\, persons / space$$

Speedup and Efficiency Increasing Techniques

Beyond its use for determining which trips are to be declared deadheading, the direction balance ratio is important for characterizing the efficiency with which demand can be served. The closer to 1.0, the more equal the demand in opposite directions. Generally, a route with a large traffic generator at only one end will have a lower direction balance ratio than one with more distributed origins and destinations. Extending a route that begins at one major traffic generator to reach a second major traffic generator would greatly improve the balance ratio. An example would be an extension of a route that has a CBD at one end to a major airport at the other end.

Up until this point it has been assumed that the number of TUs, N, is fixed for any given headway, h. But N is not truly fixed, rather it is just treated as fixed since partial vehicles can not be removed from service while in operation. In reality, as cycle time is reduced, at some point, N can be reduced by 1 to N-1. The difference in cycle time from T to the reduced cycle time, T′, that is required is exactly h, as can be shown by the following derivation based upon Equation 3-8 on page 76:

$$T' = (N-1)h = Nh - h = T - h \qquad \text{3-12}$$

To reduce T, each of the variables in Equation 3-9 is a candidate for change. Changing v_o might require some investment or route modification. Increasing v_o is the most advantageous way to reduce T to T′. It not only reduces operating costs by requiring one fewer TU while providing the same headway, it can also reduce the capital cost since a smaller fleet need be supplied for this service. Furthermore, noticeably higher operating speed increases ridership on long trips and consequently the fares collected, which in turn lowers the operating subsidy required.

The least expensive way to increase average operating speed is to increase stop or station spacing for new designs and to remove existing stops for existing routes. Often stops are very close together, ostensibly for the convenience of the passengers. But it can be easily shown that increased access and egress distances are usually more than offset by decreases in time onboard the vehicle. If operating speeds are slow because there are consistently heavy passenger loads or gradients along a route, a vehicle design with a higher output propulsion plant and better brakes may be needed. These will restore acceleration rates and braking rates to the normal range.

In some cases, operating speed is inconsistent. Success at reducing inconsistency can often permit the reduction of terminal times, tt_1 and tt_2, when these were originally set long to accommodate a wide variation in travel times. Public policy changes will sometimes alleviate random delays and thereby reduce variability of travel times. For example, in some regions, there are laws requiring merging buses to receive the right-of-way in conjunction with clearly marked bus pullout locations. This saves them the random time loss associated with the wait for a gap to open in traffic.

Transit can be favored and thereby be both sped up and reduced in inconsistency through signal re-timing and by altering signal phases when a bus is detected. Traditional methods of providing *Transit Signal Priority (TSP)* involved expensive equipment both wayside and onboard the vehicle, and perhaps revision to a centralized, computerized traffic signal control network. As control technologies have improved over time, decentralized control of signals one intersection at a time requires less time and effort to implement. Queue bypasses at busy intersections in conjunction with TSP can also sometimes be used when auto traffic otherwise would trap transit vehicles in the queue. Another measure is the "bus bulb," where the sidewalk is extended to the traffic lane. This allows bus passengers to board and alight without pulling to the side and risking a long wait to remerge with traffic.

Efforts can also be made to cut the *dwell time* at stops, defined as the time actually spent with the doors open. Fare collection can be moved off of the vehicle for the largest impact. Vehicles can be selected which have more door channels so that boarding and alighting queues are shorter. Fare collection procedures can be sped up. For example, prepayment and contactless fare deducting media (which can be read from a distance) both reduce the time each individual spends in doorways. In addition, the public can be educated in proper boarding and alighting procedures.

Shortening the route length, L, is also a possibility. One strategy is to redesign a meandering route to follow a shorter course in order to reduce length sufficiently to save one headway, h. As another strategy, if passenger demand is significantly higher on one portion of a route, some of the vehicles can operate over a shorter length in a *short-turning* operation. In this case, there can actually be savings of more than one vehicle. This can be easily modeled by dividing the basic route into two subroutes, a short one plus a long one. This strategy can be applied only to the extent that the lower frequency of service on the nonoverlapping section of the longer route remains high enough to meet any minimum frequency standards.

Use of nonproductive time can sometimes be improved by extending route length, L, instead to put to use time otherwise spent standing in the terminal. The extra distance, ΔL, that it might be possible to extend a route with existing total terminal time, $tt_1 + tt_2$, is given by:

$$2\Delta L = v_o\,(tt_1 + tt_2 - tt_{\min 1} - tt_{\min 2}) \quad . \qquad \text{3-13}$$

In practice, whether the route should be extended would also depend on the potential ridership and the availability of a new terminus. Moving a terminal for an extension that carries little ridership would introduce a new form of inefficiency in the form of poor use of vehicle capacity.

EXAMPLE 3.4 (CONTINUED)

[Please note: This example is an extension of Example 3.4 found on page 77]

B) How much would the route have to be shortened to save one TU?

The revised cycle time must meet the fleet integer constraint of Equation 3-12:

$$T' = (N\text{-}1)\ h = (6\text{-}1)\ 15 = 75 \text{ minutes}$$

But the revised cycle time must also meet the travel time requirement of Equation 3-9:

$$T = 2L'/(v_o'/60) + tt_{min1} + tt_{min2} = 2L'/(v_o/60) + 0.15(2L'/(v_o/60))$$

Inserting values gives:

$$75 = 2L'/(11.2/60) + 0.15(2L'/(11.2/60))$$

Solving the expression for L' gives 6.1 miles (9.8 kilometers). Thus, the route must be 0.7 miles (1.2 kilometers) shorter.

C) How much could the route be extended without requiring an additional vehicle?

Use Equation 3-13:

$$\Delta L = v_o\ (tt_1 + tt_2 - tt_{\min 1} - tt_{\min 2})/2 = (11.2/60)(16.7 - 11)/2 = 0.53 \text{ miles} \quad (0.86 \text{ kms})$$