Scheduling Techniques for Maximizing Urban Passenger Rail Service While Minimizing Vehicle Requirements

COLIN F. WHEELER

A variety of techniques for maximizing the utilization of rapid transit vehicles exist. The use of one or more of these techniques would be appropriate for operators of light rail, commuter rail, or heavy rail transportation systems attempting to minimize fleet size and by properties faced with a rail car shortage. Methods include the use of fallback scheduling, skip-stop scheduling, zonal scheduling, reverse-direction deadheading, Dutch switching, and shortening headways while operating more low-capacity trains. Examples of agencies currently using each technique are provided.

WITH THE INCREASINGLY HIGH acquisition costs of new rapid transit vehicles and continuing pressure from the federal government to reduce capital and operating expenditures, it is important that operators of urban passenger rail services minimize the size of their rail vehicle fleets. Providing the greatest amount of rail service with the least amount of equipment can be accomplished with various scheduling techniques, whether used as permanent measures or as short-term remedies for dealing with a car shortage caused by an increase in demand beyond existing capacity or a high rate of vehicle malfunction.

Because total vehicle requirements are determined by peak vehicle requirements, most of the following techniques reduce the round-trip cycle times of vehicles on a rail line to a level at which some equipment can be circulated through the line more than once during the peak, thus reducing the number of

3508 S.E. Carlton Street, Portland, Oreg. 97202.

pull-outs required and thereby saving vehicles. This is generally accomplished by shortening each train's recovery time or by reducing the running time along the entire line or a portion of it. The last two techniques in this paper, however, allocate high-capacity trains only to those segments of a line where the capacity is needed (Dutch switching), and only during those times when this greater capacity is absolutely needed (shorter headways plus more low-capacity trains).

FALLBACK SCHEDULING

Fallback, or dropback, scheduling is designed to make intensive use of vehicles. Each train is scheduled for a minimum amount of recovery time. Because operator break time is not only desirable but often mandatory due to labor contracts or local laws, operators are required to back-trade with (or "fall back" to) either their immediate follower or a subsequent vehicle after each trip so that they can take their breaks. Consequently, the number of operators assigned to a line at any one time when fallback scheduling is employed exceeds the number of trains in service on the line.

As an example, consider a hypothetical line on which 10-min headways are operated in both directions with 10-car trains that have a 120-min round-trip running time (excluding layovers). If trains are given layovers of 10 percent of the round-trip running time plus 5 min (a commonly used formula for calculating required operator break time), 14 trains, 14 operators, and 140 cars would be required on the line at any one time. However, if fallback scheduling is used and trains are given only 5-min layovers at each terminal, 13 trains, 14 operators, and only 130 cars would be required.

With fallback scheduling, operator break time is determined by the line's headway in relation to the number of departures operators are required to fall back to (e.g., a 5-min headway would result in a 15-min break if operators were instructed to fall back three trains). Although the back-trading of operators can occur at any point along the line, the time required to change operators dictates that it occur at only one, or both, of the line's terminals rather than at intermediate stations. Although it is generally desirable to schedule operator train assignments in advance, it is sometimes appropriate to instruct operators to join a pool of other operators taking their breaks at relief points. This is particularly true when changes are frequently made to a line's operating schedule by supervisory personnel to match capacity to demand as closely as possible on a day-to-day basis. In such a situation, a supervisor assigns operators to specific departures, and operators do not know in advance which trains they will operate during the course of their shift. To minimize operator overtime, it is important that the supervisors

assigning operators to departures be aware of how much time each operator has worked before the end of the shift.

An additional benefit of fallback scheduling is that, by reducing the number of vehicles operating on a line, the likelihood that terminals will be scheduled beyond their capacity is reduced. Not only does fallback scheduling minimize vehicle requirements, larger properties operating long trains with short layovers also use this technique because of the lengthy amount of time it takes for operators to walk from one end of a multicar train to the other.

An inherent disadvantage of fallback scheduling is that, because of reduced vehicle recovery time, it can lead to a reduction in schedule reliability. This method is most successful; therefore, when it is used on lines on which, by virtue of an exclusive right-of-way, transit preferential traffic signals, etc., short layovers will not severely affect the quality of the line's on-time performance. To provide at least a minimal amount of recovery time, most properties currently using the fallback technique schedule their trains for layovers of at least 3 min.

One common way of dealing with the reduced ability to recover from service delays of extended duration created by fallback scheduling is the deployment of strategically located "gap" trains. In the event of a major disruption in service, these trains are dispatched to cover the trip or trips missed by the train caught in the delay. Upon its eventual arrival at one of the terminals from which gap trains are dispatched, the delayed train becomes the new gap train. Although the use of gap trains reduces the amount of vehicle savings that fallback scheduling makes possible, the ability of such trains to minimize the negative impact on schedule reliability justifies their use.

A second disadvantage of fallback scheduling is that it increases labor costs beyond the absolute minimum level that would exist if this technique were not used. A final drawback of fallback scheduling is that some operators who prefer to stay with the same vehicle for most or all of their shift may object to changing trains for each trip.

Fallback scheduling is commonly used on both light rail and heavy rail systems throughout North America. Properties making extensive use of this technique include New York's Metropolitan Transportation Authority (MTA); the Washington (D.C.) Metropolitan Area Transit Authority (WMATA); Pittsburgh's Port Authority of Allegheny County (PAT); San Diego's Metropolitan Transportation Development Board (MTDB); and the Toronto Transit Commission (TTC).

SKIP-STOP SCHEDULING

Unlike fallback scheduling, skip-stop scheduling reduces vehicle cycle times not by shortening layovers but by increasing the speed with which trains

operate over a line. Skip-stop scheduling is essentially the overlapping of two or more different versions of limited trains. Trains make different sequences of stops along the same route. By bypassing a portion of a line's stations, total vehicle dwell times can be reduced significantly. To identify the different stop sequences, trains are usually identified as "A" trains, "B" trains, or, sometimes, "C" trains. Most commonly, only "A" and "B" trains are used, in which case each rail station is designated an "A" station, a "B" station, or an "AB" station. "A" and "B" trains are scheduled to alternate with one another, with "A" trains stopping only at "A" and "AB" stations, and "B" trains stopping only at "B" and "AB" stations. Because "A" and "B" trains do not pass each other, they can use the same tracks. Ideally, there should be a high degree of travel between the stations served by each type of train. Although this may be hard to accomplish, origin-destination surveys may aid in determining which stations should be linked together under the same stop category. Common stations, served by both "A" and "B" trains, are usually major focal points of activity, such as timed-transfer centers, major downtown stations, or park-and-ride lots.

The following example illustrates the ability of skip-stop scheduling to save cars. If the running time from one terminal of a line to the other is the same as that used in the example illustrating the use of fallback scheduling (i.e., 60 min when trains stop at all stations), the round-trip running times of local, or all-stop, trains would be 136 min (assuming that trains are given 8-min layovers at each terminal). If peak demand levels are such that 10-car trains operating on 10-min headways are required, 14 trains and 140 cars must be operated. However, if skip-stop scheduling is employed and the running time from one terminal to the other can be reduced to 50 min, the round-trip running time of skip-stopping trains would be 114 min (assuming that trains are given 7-min layovers at each terminal). By operating 5-car trains on 5-min headways (10-min service to each type of station) the operator would be able to provide an amount of capacity equal to that provided by local service only, but would be able to realize a savings of 25 cars (23 trains and 115 cars would be required with the operation of skip-stop service). As this example shows, skip-stop scheduling generally has a greater potential for reducing vehicle requirements than does fallback scheduling.

There are two variations of the skip-stop scheduling technique. Under one variation, "A" trains are scheduled to travel locally from the extremity of a line to a point midway along the line, after which they begin skip-stopping until they reach the line's other terminal (usually a region's central business district). Under this strategy, "B" trains are scheduled to begin service at the point where "A" trains start skip-stopping, and stop at the stations bypassed by "A" trains. Another variation is to schedule both "A" and "B" trains to begin service at the outer terminal of a line and to stop at alternating stations on their way to the other terminal of the line.

Aside from the reduction in vehicle requirements made possible with this technique, the decrease in travel times provided by skip-stop service can be a valuable marketing tool for increasing ridership. One drawback of skip-stop scheduling, however, is that passengers wishing to ride from an "A" station to an "B" station, or vice versa, must change trains at an "AB" station. Another disadvantage is that skip-stop scheduling often leads to a deterioration in the frequency of service provided to stations served by only one type of train. A third drawback is that it results in an inconsistency of service, which can be confusing to passengers. A final disadvantage of this technique is that it can antagonize passengers if they are frequently passed by trains not scheduled to stop at their station. To identify the stop designation of trains, it is imperative that cars display the proper signage, especially on outbound trips.

Although this technique is most commonly used on heavy rail and commuter rail systems, there is no reason that it could not be used on light rail systems as well. The Chicago Transit Authority (CTA) is the major user of skip-stop scheduling in North America. Other users include Boston's Massachusetts Bay Transportation Authority (MBTA) commuter-rail service (the Attleboro Line); Philadelphia's Southeastern Pennsylvania Transportation Authority (SEPTA) (the Market-Frankford subway-elevated line); and New Jersey's NJ Transit (the Morris and Essex line east of Summit, the North Jersey Coast line west of Matawan, and the Boonton line). Skip-stop scheduling was recently reintroduced on the New York MTA's "D" and "Q" lines. Visitors to Expo '86 in Vancouver, British Columbia, will recall that the monorail used at the fair used skip-stop scheduling.

ZONAL SCHEDULING

Zonal scheduling is similar to skip-stop scheduling in that vehicle cycle time is reduced by increasing average train speed. But this technique involves the operation of limited or express service between one terminal of a line and different sections along the line. Rather than operating limited trains over the entire length of a line, as with skip-stop scheduling, trains operate locally within designated zones and then travel express to the major terminal of the line. Because of the ability of zonal scheduling to reduce the travel time between the extremity of a line and the line's major terminal, this technique is best suited to lines on which passenger demand is oriented primarily toward a single station or group of stations instead of being evenly distributed along the line.

Similarly, zonal scheduling is best suited to longer lines on which the operation of local service over the entire length of the line would make for a very slow (and therefore unattractive) trip from one end to the other. This

technique is also appropriate on lines on which demand for arrival times in, and departure times from, a specific terminal is heavily peaked. In North America, zonal scheduling is used primarily by operators of commuter rail lines providing long-distance, highly peaked service to and from large cities.

As with skip-stop scheduling, there are a variety of substrategies for zonal scheduling. Under true zonal scheduling, a line is divided into a series of zones and trains are scheduled to operate only between their assigned zones and a common terminal with no intermediate stops. The line is, in effect, segmented into several different services. Although each zone is provided with a high level of service to and from the common terminal, little or no service is provided to and from the other zones on the line. Each zone should be situated so that trains are approximately at their capacity as they pass the zone boundary departing for, or arriving from, the major terminal. Because this strategy involves a great deal of short-lining, train volumes on a line's inner portion(s) are much heavier than they are on its outer portion(s). Examples of this variation include Chicago's Metra commuter rail service (the Chicago-Aurora line operated by the Burlington Northern Railroad); and NJ Transit (the Northeast Corridor service).

Closely related to the previous strategy is a variation of zonal scheduling in which zones are not specifically designated, but long-line express trains are operated in combination with short-line local trains. A drawback of this method is that passengers wishing to travel between a station along the long-line portion of a line and a station along the short-line portion must transfer at a common station served by both local and express trains. Examples of this strategy include Philadelphia's Port Authority Transit Corporation (PATCO) (the Lindenwold line); and Boston's MBTA commuter rail service (the Stoughton and Franklin branches).

Another variation, which differs from the previous two in that short-line trains are not used, operates all trains over the entire length of a line with each train traveling a different distance on an express basis. This strategy is best suited to lines on which there is a large amount of travel between the two terminals and on which demand is sharply peaked in one direction. An example of this variation is California's CalTrain commuter rail service between San Francisco and San Jose.

These three variations of zonal scheduling technique have a number of characteristics in common. First, they each involve the operation of express trains. Whenever express service is operated on a line, there must either be a third track for peak direction express service, or headways must be wide enough to provide a "window" through which express service can operate. The operation of express service works best when separate tracks are available exclusively for express trains. The use of express tracks also maximizes time savings for express trains and thus provides the greatest potential for

minimizing vehicle requirements. Obviously, the additional capital cost of laying such tracks must be weighed against the various benefits of doing so (including the ability of express service to attract ridership).

With careful scheduling it is possible, however, to operate express service on lines with passing sidings at appropriate locations along the line. This method can also be employed on lines without separate tracks or passing sidings if headways are long enough and if certain scheduling precautions are taken to prevent slow local trains from getting in the way of fast express trains. When the latter condition applies, express trains should be scheduled during the morning peak to arrive at the line's major terminal just behind local trains and, during the afternoon peak, be scheduled to depart from the line's major terminal just ahead of local trains. This, incidentally, is just the opposite of bus operations in which, to equalize passenger loads, local buses are generally scheduled in the morning to arrive at a line's major terminal just behind express buses and to depart in the afternoon from a line's major terminal just ahead of express buses. Because the ability of express trains to achieve a time savings over local trains is directly related to the amount of time express trains are operating in that capacity, the operation of express and local trains over the same tracks is most successful on relatively short lines.

The scheduling of CalTrain's peninsula service between San Francisco and San Jose provides an example of the scheduling of express and local service over the same tracks. During the morning peak, trains are scheduled to arrive in San Francisco approximately 5 min apart, with the first train having traveled local for most of its trip, and the second having operated express from a station somewhat farther away from San Francisco than the first train. This sequence is continued for three more trains before all-stop service resumes. In the afternoon, trains are scheduled to depart from San Francisco approximately 4 min apart, with the first train traveling express for most of its trip, and the second operating express to a station somewhat closer to San Francisco than the first train. As in the morning, this sequence continues for three more trains before all-stop service resumes. Although an automatic block signaling system is used, the scheduling of express trains in this way helps to spread trains out and thus minimizes the likelihood that one train will overtake another.

Another inherent feature in nearly all variations of zonal scheduling is the use of short-lining. Whenever short-lining is used, the line must have one or more midroute turnbacks, and short-lining trains must be able to remain in a pocket track without fouling the blocks of either of the mainline tracks until the schedule dictates that they are needed for a trip in the return direction.

Although the short-lining of trains can occur at more than one point along a line, the greater the amount of short-lining, the more difficult it is for passengers to travel between a line's inner and outer segments. Consequently,

the main drawback of short-lining is that it involves a deterioration in the quantity of service provided to a line's extremity. Another disadvantage of short-lining is that it results in an inconsistency of service, which can be confusing to passengers—especially new users of a system. To prevent passengers wishing to make a long-line trip from boarding a short-line train, outbound trains must display their destinations. To allow passengers to travel between a line's inner and outer segments without having to wait for more than one headway, it is recommended that long-line trains be preceded by no more than one short-line train.

Because the adoption of zonal scheduling can lead to passenger animosity if passengers at inner stations are regularly passed by trains operating express from outer zones, this technique works best on lines on which express trains reach their maximum capacity approximately at the point where they begin operating express (i.e., where passengers at inner stations realize that there is no room for them on board the express trains). As express service is very desirable to most passengers, a side benefit of this method is that its adoption can lead to an increase in patronage along the extremity of a line. As previously mentioned, the major disadvantage with zonal scheduling is that it does not allow passengers to travel between two zones without having to transfer at a common station served by all trains.

REVERSE-DIRECTION DEADHEADING

As with bus operations, one way of reducing the number of required pull-outs is to deadhead equipment back to either the beginning of a line or a point midway along the line. This technique can also be used on rail lines with either passing sidings or, ideally, separate tracks dedicated exclusively to use by express and deadheading trains. Because the deadheading of trains on lines with only passing sidings can be difficult to schedule and can present safety hazards, it is recommended that this technique be employed only when additional tracks are available. Reverse-direction deadheading is particularly appropriate on lines on which demand is strongly peaked in one direction, such as on many commuter rail lines. This technique can also be used, however, on both light rail and heavy rail systems, and is most commonly used in combination with the various forms of zonal scheduling. Although trains do not generate revenue while they are deadheading, this drawback is offset by the fact that their repositioning makes it possible for them to pull one or more additional high-revenue peak-direction trip(s).

A side benefit of reverse-direction deadheading is that by reducing the number of stops and accelerations trains are required to make, power consumption can be lowered somewhat. The main disadvantage with this method is the deterioration in the quantity of reverse-direction service, which discourages much back-haul activity from being made on the line. A second

drawback is that it can antagonize passengers wanting to make reverse-direction trips if they are regularly passed by empty trains traveling in the direction they wish to go. A variation of this technique, designed to respond to the previous two disadvantages, is to have deadheading trains travel instead as limited trains, stopping at only the stations with the heaviest demand on their way back to the end of the line.

Examples of properties making use of this technique include Newark's Port Authority Trans-Hudson Corporation (PATH), which deadheads every second train in the reverse-peak direction between the World Trade Center and Newark; NJ Transit, which uses deadheading extensively on its multiple-track lines; Philadelphia's PATCO; and Toronto's GO Transit commuter rail service.

DUTCH SWITCHING

A somewhat obscure technique for minimizing vehicle requirements, Dutch switching is essentially the short-lining of cars rather than trains. At one time Dutch switching was used extensively by operators of interurban rail systems throughout North America. Now used primarily by intercity railroads, this technique (also referred to as car dropping) is designed to match capacity as closely as possible to demand along each segment of a line. Trains originating at a line's major terminal and passing through the peak-load point are composed of enough cars to provide adequate capacity through the portion of the line with the heaviest demand. As trains proceed toward the extremity of the line and demand drops off, cars are detached from each train at appropriate locations and temporarily stored on pocket tracks. These dropped cars are attached shortly thereafter to the front of trains traveling in the opposite direction along the line. Dutch switching is similar to the "changing gauge" practice used in the commercial aviation industry (i.e., multistage flights are scheduled to make one or more changes in aircraft size).

To illustrate the use of Dutch switching, if one portion of a line requires trains with 10-car consists, and the other portion requires trains with 5-car consists, trains passing from the 10-car section to the 5-car section must drop their last 5 cars before proceeding. These dropped cars are then coupled to the front of the next train traveling from the 5-car section into the 10-car section.

The principal advantage of this method is that it provides frequent trains to all sections of a line without requiring heavy use of cars. Dutch switching requires careful scheduling to ensure that the window of time between the dropping and adding of cars is long enough to prevent trains traveling from the heavier demand section into the lighter-demand section from missing their connections due to late arrivals. To allow for the time it takes trains to add and drop cars, it is also important that an adequate amount of dwell time be built into the schedule at the point(s) where Dutch switching occurs.

Because of this required dwell time and the fact that most heavy rail lines operate with headways approaching the amount of time it takes to couple and uncouple cars, this technique is best suited to light rail and commuter rail systems. Dutch switching can be employed anywhere along a line where a pocket track of adequate length is available, and it can, if necessary, be employed at more than one location. Ideally, a third, center track should be available at the point(s) where Dutch switching occurs. It is recommended that a supervisor be stationed at the point(s) where Dutch switching takes place so that he or she can assist in the coupling and uncoupling of cars. To speed operations, it is also recommended that inbound passengers be allowed to board dropped cars prior to the arrival of the next inbound train.

Dutch switching is best suited to longer lines on which long stretches of the line require significantly less capacity than other sections, but over which it is still desirable (e.g., for political reasons) to operate a relatively high level of service. This technique is not to be confused with the practice employed by San Francisco's Municipal Railway (Muni) of dividing multicar trains at a point midway along the line and operating each car to different branches as a separate train. Although similar to Dutch switching, this technique requires additional operators to run the branch-line trains.

Dutch switching is also well suited to lines on which headways cannot be shortened further, meaning that long trains must be operated through those portions of the line with the heaviest demand. Instead of operating long trains along the entire length of the line, train length is reduced at one or more stations along the line. A side benefit of Dutch switching is that, because trains of shorter length (and therefore less weight) are operated over portions of a line, power consumption is lower than it would be if long trains were operated over the entire length of the line. So that passengers will be segregated into the correct cars, it is imperative that all cars in a train, not just the head car, display their destinations. It may also be advisable for operators to announce over the public address system the vehicle numbers or locations of the cars that will be dropped at some point along the line.

Disadvantages of this technique include the difficulty some passengers may have with understanding that although the train they are on will traverse the entire length of the line, the car they are in will not necessarily do so as well. Other disadvantages include a possible increased rate of coupler fatigue as a result of the frequent joining and cutting of trains, and the safety issues involved with the coupling of cars with passengers on board.

A good example of the use of this technique is Chicago's South Shore and South Bend commuter rail line on which trains destined for or arriving from South Bend, Indiana, drop and add cars in Gary and Michigan City. Dutch switching was also used at one time on the 90-mi-long Chicago North Shore and Milwaukee interurban line; cars were dropped from northbound trains

and added to southbound trains in Waukegan, Illinois. Variations of this technique were also used in New York and in New Jersey.

SHORTER HEADWAYS PLUS MORE LOW-CAPACITY TRAINS

The technique described in this section is similar to Dutch switching in that it is designed to match train length with demand as closely as possible. Unlike Dutch switching, however, this method involves a temporal, rather than a spatial, matching of capacity to demand. Under this technique, the number of cars on trains passing through the peak-load point during the shoulders of the peak is reduced to an absolute minimum and, to maintain adequate capacity, headways are shortened. High-capacity consists are operated only on those trains passing through the peak-load point in the peak direction at the peak of the peak. For example, rather than operating a line with 11 two-car peak trains (22 peak cars) providing a 7.5-min peak headway, 14 peak trains could be operated. Four of these would be two-car trains scheduled to pass through the peak-load point during the peak of the peak. Ten would be one-car trains scheduled to pass through the peak-load point during the shoulders of the peak. The latter schedule would provide a 5-min peak headway but would require only 18 cars, a savings of four cars.

Because this technique involves the operation of a single headway throughout the peak and the assignment of high-capacity consists only to specific trains, it is most appropriate on lines with sharp peaking characteristics. Although this technique can be attractive to the public because of the increased frequency of service, its main drawback is that it can be expensive to operate because of additional manpower requirements. This may be difficult for an agency to justify, in light of the fact that reduced labor costs are supposed to be one of the main justifications for the construction of a rail line. The increased revenue resulting from the appeal of high-frequency service and the decreased maintenance and power consumption needs resulting from the reduced peak vehicle requirements may act to offset this disadvantage, however.

Although this technique can be used on any type of rail system, it would be inappropriate on a line with a peak demand period of a long duration (i.e., most of the trains on the line pass through the peak-load point during the peak of the peak), and very short lines, where it would be impossible to schedule low-capacity trains to "miss the peak." Although it perhaps goes without saying, if this method is adopted and very short headways are operated, it is important that the line have a very good automatic block signaling/automatic train stop system.

CONCLUDING REMARKS

An additional technique for reducing round-trip vehicle cycle times that has not been addressed because it does not pertain directly to scheduling is that of simply increasing the maximum authorized speed along sections of a line where it is possible to do so without compromising high levels of safety. Segments of a line with an exclusive right-of-way or through which station spacing is long are particularly well suited to this method. Agencies attempting to increase the speed limit along all or a portion of a line should be aware that, in general, the faster trains are expected to operate, the longer the block lengths should be.

Another way of dealing with a car shortage that has not been discussed is that of arbitrarily canceling trains or operating shorter consists on those days when not enough rail vehicles are available. Because both of these practices are likely to elicit a great deal of passenger criticism as a result of missed connections or overcrowded trains, it is recommended that they be avoided if at all possible.

Although the scheduling techniques described in this paper are relatively low-cost ways of dealing with a car shortage problem, other more expensive measures for addressing this problem exist, such as the implementation of a self-service fare collection system and the construction of high-level loading platforms (both of which would increase the speed of operation). Other measures include changing the vehicle seating configurations to increase passenger capacity (an action that would allow fewer vehicles to provide the same capacity as that provided with the old seating configuration), and increasing the peak/off-peak fare differential (an action that would shift some demand away from the peaks, thus enabling peak capacity to be reduced).

Because peak vehicle requirements determine peak spare ratios, it is conceivable that an agency could choose to employ any of the above methods only during the peaks. It is also conceivable that more than one of these methods could be employed at the same time (e.g., zonal scheduling used in combination with reverse-direction deadheading). Although the operation of one or more of these techniques on a routine basis would allow an agency to minimize vehicle acquisition costs and maximize the number of maintenance hours available per rail vehicle, agencies could also choose to implement one or more of these methods on a contingency, as-needed basis, substituting them for the regular rail schedule only on those days when not enough rail cars are available.