Application of computer simulation to rail capacity planning

W.M. Barter
Comreco Rail Ltd., York, England - www.comreco-rail.co.uk

Abstract

Developing arrangements for “open access” to rail networks, together with the establishment of service reliability targets and penalty charges, requires a new understanding of both system capacity and the effect on reliability of additional services. The separation of trains on any one line depends upon an interaction between the traction, route alignment and signalling. The capacity of the whole system however, is also determined by junctions, and the structure of the intended timetable, reflecting commercial aspirations. However, the practical capacity limit must also take account of service reliability in the face inevitable disruptions. Service reliability is principally a function of intensity of operation, increasing the chance of “delay by reaction” as each late train potentially delays others. Conventional planning techniques are adequate to plan timetables, and suggest whether changes to the traction, route, signalling or timetable will enhance or jeopardise capacity and reliability. However, decisions involving capital investment or the balance of advantage between competing factors require quantified data on reliability. Computer simulation can be used initially to refine the headways and junction clearance times on which the service is based. Then it is possible to merge data on the traction, route and signalling with a timetable plan to simulate operation of a complete service. Further, input data for the probability and magnitude of delays to trains on handover to the simulation area can be added, testing the risk that these delays will escalate “by reaction”. The Comreco Rail Ltd. RailPlan™ simulation system offers detailed output on the operation of individual trains and of complete services. Comparison of initial and final delays to trains quantifies the “delay by reaction” within an area. This information can then form an input to patronage and revenue forecasting, and to investment appraisal.
Line capacity - what is it?

The capacity of a rail system, typically measured in terms of Trains per Hour (tph), is considerably easier to discuss than to quantify. The more precise a calculation of capacity appears to be, then typically the more difficult it will be to realise the calculated capacity in a useful train service.

Whatever type of signalling is employed, it will be the principal determinant of capacity. Signal sections, whether defined by lineside signals or by cab signalling, are located so that the distance between the first caution signal and the danger signal to which it applies allows a train to come to a stand from a nominal maximum speed. This maximum speed, which is not necessarily related either to the capabilities of the traction or civil speed limits, is termed the Line Speed.

A figure for the minimum interval between trains - the Headway - can then be arrived at from calculations based principally on the Line Speed and the length of the signal sections. This allows a simple calculation of the capacity of the track to be made, in terms of tph, by dividing the Headway in minutes into 60. Inevitably, some upward variation of section lengths is inevitable, because of restrictions on the location of signals. Such restrictions may arise from approach visibility, or from the need to locate signals in relation to station platforms or turnouts. The longest section or series of sections thus tends to determine the headway for the route.

Such calculations are standard practice. However, they overlook a number of issues which tend to inhibit operation of trains at the theoretical frequency. Some of these issues are resolved by factors other than objective technical criteria.

For instance, conventional practice considers the headway to have been infringed unless the second train of a pair can run on green aspects. Approaching a station, however, caution aspects may be less restrictive on train speed than braking for the station stop. In complex areas such as terminal approaches with local speed restrictions well below the Line Speed, a caution aspect may be seen well before any braking action is actually required. The fact that a train running with regard to its capabilities and the civil speed restrictions may not have to respond to that caution may lead to the caution being disregarded in the headway assessment.

Moreover, it may be possible for some trains with either lower speed or better braking capabilities than the norm to run at significantly closer intervals than the calculated headway, without having to brake other than for local speed restrictions. However, planning a service which will
require trains to run on caution aspects is a dubious practice, in view of the risk of habituation to cautions on the part of the driver, and so this opportunity to run a more intensive service than is theoretically possible may not be exploited.

Even between stations, the Line Speed may not actually be achievable because of gradients or local civil speed restrictions, or by trains with restricted maximum speeds. Once a line has been signalled for a given speed, the separation in time required to allow trains to run on clear aspects increases dramatically as the actual speed of trains drops below the Line Speed. This effect will be severe should the slow running coincide with a stretch with long signal sections. Conversely if sections in the restricted area happen to be close to the minimum length, no detriment may be apparent compared with long sections elsewhere.

The criterion for identifying the headway may therefore be more a matter for a management decision than an objective feature of the signalling. Even once the headway criterion has been decided, the calculated capacity will be a maximum, and many factors inherent in operation of complex systems will combine to prevent operation of a service of the theoretical frequency.

**Turning capacity into a train service**

Usage of the theoretical capacity tends to be limited by other factors, mainly those of benefit to passengers, such as stopping at stations. For a metro-style line, the principal determinants of line capacity are likely to be the junction arrangements at terminal stations, and platform reoccupation and dwell times at intermediate stations.

The platform re-occupation time is defined as the interval between departure of one train and earliest possible time for arrival of the next in the platform. This time is made up of two components:

- the time from departure of the first train to clearing the critical Track Circuit that allows the station approach signal to clear for a second train to enter the platform;
- the time then taken for the second train to run into the station from the last clear signal approaching the station;

The overall platform cycle time, i.e. from departure of one train to departure of the next, is then the reoccupation time plus whatever station dwell time is required for passenger purposes. Unless terminal station arrangements are unduly restrictive, then the platform cycle time for intermediate stations is likely to be the binding constraint on line capacity.
Whilst the platform dwell time is dependent upon passenger numbers and design of the rolling stock and station accesses, the reoccupation time is essentially determined by:

- acceleration of trains away from the platform;
- location of critical Track Circuits ahead of the platform, such as the overlap of the platform starting signal;
- location of the platform entry signals;
- the braking profile of trains running into the platform.

The first and last of these factors will in turn be influenced by features such as the gradients and speed limits in the station area.

If the platform cycle time is indeed the critical factor determining line capacity, then signalling between stations may be well short of ideal without imposing a constraint. Both in the case of plain line headways and platform reoccupation times, the true constraints on line capacity are potentially subject to a complex interaction of features of the route alignment, the traction capabilities and the detail of the signalling.

The situation is further complicated by the presence of junctions, characteristic of a surface suburban operation. The individual lines approaching a diamond crossing may well be signalled so as to offer very short theoretical headways. What becomes critical in such a case is the junction clearance time, that is, the minimum interval between two trains on the conflicting routes that still allows the second train to receive clear aspects. The major factor in this is the length of signal sections approaching the junction, so that the junction clearance time will be very similar to the headway.

This implies that, given that trains approach the diamond alternately from each direction, each line can then only be used at half its theoretical capacity. However, the effect is less severe if the service on one of the lines is greater than on the other. Moreover, if the diamond lies near to a platform, it may be possible to plan the conflicting move to take place whilst a train for the main route is running in and waiting for station duties to be completed.

A well-known problem in the case of multi-purpose railways is that of running a mix of fast and slow or stopping trains through a “corridor”, that is, as section of route between two regulating locations. If a fast train is to follow a slow train without being delayed, the interval between departures at the entrance to the corridor must be greater than the feasible minimum.

Similarly, if a stopping train enters the corridor just after a fast, the time separation between the two trains will steadily widen, so that at the next regulating point the interval between the trains is much greater than the minimum allowed by the signalling. However, no subsequent train is
able to use that interval, as between regulating points it has no chance to pass the stopping train.

However, depending upon the precise circumstances, the unused capacity is by no means unusable. For instance, a further train may start at an intermediate station, following a fast train but without interfering with the following slow train. Or perhaps the growing interval between the two allows a conflicting move to take place across the line in question, without detriment to the flow of trains on the main line.

The classic planning response to running trains of varying speeds is “flighting”, that is, running a group of fast trains followed by a group of slow trains. Unused capacity then occurs only between the flights of trains rather than between individual trains. This is exemplified by Eurostar services, which run in pairs over Railtrack lines and through the Channel Tunnel, to minimise loss of capacity for slower domestic and Shuttle services. However, the consequence is a service which may not be commercially optimal. The more fast trains that are flighted together, then the longer intermediate passengers will have to go without a service at all, and the greater the gap that will eventually arise in the fast service between flights.

Most intensively worked railways, such as the London suburban lines at peak times, achieve their service frequencies by standardising the schedules in the timetable, for instance, restricting stops at intermediate stations, and banning freight services at peak times. Metro-type operations of course largely avoid mixtures of train speed and stopping pattern, and minimise interactions between lines at junctions.

Typically on a multi-purpose railway, by the time the number of trains in the plan reaches half the number that might be run on any one line in isolation, sensible paths for further trains are very hard to find. The reasons for this are also difficult to explain to an observer at a point on the system, who sees only a flow of trains at what might seem to be excessive intervals. So in assessing capacity, account needs to be taken not only of the interactions between traction, route and signalling, but also of the pattern of the train service.

**Planning for reliability**

One further complication needs to be considered, and is the most difficult to quantify and manage - reliability. An intensive service might be feasible, in that paths for all trains can be found. In other words, if all trains start on time, and suffer no incident on route, they will encounter no conflicts with other trains at junctions, and will finish their journeys on time. Such an ideal situation of course very rarely occurs, and the
question left unanswered by the pure timetable planning process is how well the service will stand up to a degree of disruption.

Should a train run late on the approach to a junction, two possible consequences exist. It is possible that a low intensity of service allows the late-running train to proceed through corridor sections and over junctions in spare paths, avoiding any impact on other trains. However, it is more likely that the late-running train will encounter a conflict with another train that is avoided in the plan. Either the punctual train will be delayed, or the late train will incur further delay. If the punctual train is delayed, it may then itself cause delay to other trains in a chain reaction, and the total delay escalates. Then, if the traction units are capable of improving on the planned running times, delay may be recovered before arrival at the terminus. In practice, the behaviour of a complete system will be a balance between the effects. Clearly, escalation of delays is most likely to be significant where the system is both tightly-timed and intensively used.

Simulation - quantifying reliability

Service reliability, then, is determined by interaction between the traction, the infrastructure and the timetable. This makes reliability effectively impossible to forecast by conventional planning techniques, when a change to one or more of the three is being considered.

Experience may very well indicate the direction of an impact on reliability, that is, whether the consequence of a change to the system will be good or bad. However, putting a figure on the effect, as an input to a business decision-making process, is quite another matter. For instance, increasing the intensity of the timetable will present a risk of increased delay, but the operator needs numerate data to judge whether this risk is acceptable.

Moreover, most real scenarios will involve a balance between competing effects. For instance, a restrictive signalling layout might be mitigated by improved traction performance. A balance needs to be struck between the benefits for reliability of allowing time in the schedules for recovery purposes, and the ongoing adverse effect on revenue of advertising extended journey times. For an infrastructure owner, the benefit of selling more of the potential capacity has to be set against increased penalty payments if the reliability of the system is jeopardised. The qualitative nature of conventional techniques leaves them powerless in assessing the balance of advantage.

Simulation is an established tool for assessing the building blocks of the timetable - Sectional Running Times (SRT’s), line headways and
junction clearance times. However, to ensure that the potential capacity of rail networks is utilised to the greatest possible extent, simulation is increasingly being used to minimise uncertainty when assessing reliability of complete services, as determined by the combination of the traction, infrastructure and timetable.

Unlike a standard traction performance simulator, a network simulator will combine a calculation of the traction performance in the face of the physical characteristics of their routes with interactions between trains caused by the signalling and interlocking. Reports will be offered detailing the simulated running and punctuality of trains. Headway assessments can then be conducted by modelling the operation of flights of trains, and iteratively reducing the interval between trains until the event chosen to indicate infringement of the headway is observed. The procedure for assessing junction clearance times is similar.

Together with SRT’s derived from a simple traction performance simulation, the data for timetable compilation has now been generated. Once a timetable plan has been compiled, its operation with the planned traction and infrastructure can be simulated, to verify the accuracy of the compilation process. The full power of simulation, however, comes into play in studying the behaviour of the train service when subject to typical daily variations from planned operation.

The question which cannot be answered by conventional planning techniques is whether initial disruption imposed on a system can be absorbed, or whether delays to trains will escalate “by reaction” as the service operates, so that the ultimate delays become more severe than the original problem. A comprehensive simulation system will have the capacity to randomly allocate start or handover delays, or excess dwell time to trains according to user-defined probabilities. Output data from such a simulation will indicate whether the imposed delays escalate because of interactions between trains, or are absorbed by buffer capacity in the timetable.

Any operator can specify infrastructure to minimise escalation of delays, such as multiple-tracking or grade separation, but the additional revenue attracted by the change in timetable or reliability may not justify the capital outlay. Conversely, investment purely to improve reliability may prove to be justified, if only the effect can be quantified in terms which can translate into a patronage or revenue forecast.

To determine the balance between recovery and delay by reaction, the End Delay, that is, the total delay to trains at the end of their journeys or on leaving the simulation area, is compared with the total delay to trains at the start of their journeys or on entering the simulation area, termed the Start Delay. If End delay is less than Start Delay, then overall recovery is
taking place. However, if End Delay exceeds Start Delay, then the balance of effects is dominated by delay by reaction. When comparing two scenarios, the key indicator is the difference in delay by reaction between the two.

Once a figure for delay by reaction is available, different traction and infrastructure options can be studied to observe their effects on punctuality of trains, and timetable or infrastructure plans can be modified before commitment. Techniques exist for incorporating statistics on reliability into patronage forecasting, allowing future cash flows to be compared with capital outlay in an investment appraisal. Decisions can be made on such factors as:

- the number of trains that can be operated within an acceptable delay criterion;
- the level of Start Delay that must be achieved in order for a planned service to operate within an acceptable delay criterion;
- the level of capital or revenue expenditure justified by consequential reduction in delay.

**The RailPlan™ simulation system**

The Comreco Rail Ltd. RailPlan™ system models the operation of trains on a rail network, allowing the user to investigate the effects of changes to an existing system, and the operation of new systems. The system combines a calculation of the traction performance in the face of the physical characteristics of their routes with interactions between trains caused by the signalling and interlocking. Reports are offered detailing the simulated running and punctuality of trains.

Data is entered to RailPlan™ either directly through the front end input windows, or by importing from Excel spreadsheets. The latter method has been favoured by users as being an extremely fast method of entering large amounts of data, especially where use of a spreadsheet is appropriate in data preparation tasks such as conversion of units. The input windows are however valuable for entry of or alteration to individual data items, and for ordered display of the database.

Input data for a RailPlan™ model covers: Geography - station locations, and physical characteristics of the routes that link them; Signalling - the routing options through the network, and associated interlocking; Traction and trains - the capabilities of the traction units and characteristics of the trains.

Simulation models will vary depending upon the issue in question. The simplest models will be designed to study the performance of individual trains, and will therefore require only traction and route data. As
separation of trains is not an issue at that level, signalling detail can be omitted, or entered just in a nominal form as needed to allow the simulation to run, and to track the progress of trains.

Such a single-train model will deliver SRT’s as the input to the process of train scheduling. The effect of changes to the route or traction will be apparent, possibly revealing locations where speed restrictions constrain train speeds disproportionately, or, conversely, where investment in raising the speed profile of the line will show minimal benefits because of limited traction capabilities. As journey time is a direct input to patronage forecasting, such a model allows straightforward investment proposals to be assessed.

Multi-train models will require full detail of signalling and interlocking so as to capture correctly the interactions between trains. Such models will, however, differ in the level of timetable detail entered, again depending upon the purpose of the model. For instance, the train planning margins such as on-route headways, junction clearance times and platform reoccupation times, required to allow individual train schedules to be blended into a timetable plan, can be identified by setting up specimen blocks of trains and adjusting their separation until clear aspects and free-flow are observed.

The geographical extent of the model is initially outlined by entering the stations to be modelled within simulation. The system to be modelled is then defined in terms of sequences of stations between junctions, known as Line Branches.

The Track Profile, or physical characteristics of the system such as gradients, speed restrictions and curvature, is entered for each Line Branch. Where parallel tracks have different characteristics, separate Line Branches must be defined even though the sequence of stations may be the same.

The precise paths through the network available to trains, at the level of station platforms and tracks between stations, are known as Train Routes. Each Train Route identified is given a number, which is carried by individual trips in the timetable, and read by the simulator to control the sequence of signal sections taken at junctions.

To relate route physical characteristics to trains in the simulation, each Train Route is related to a defined sequence of Line Branches.
**RailPlan™** models the signalling of the network in terms of Track Sections. Track Sections represent a portion of line which can be occupied by only one train at a time, and are occupied sequentially as a train moves between stations. The possible choices of "Next Section" offered by the track layout, and the conditions under which trains can move between sections, are defined in the Route Selections windows associated with each Track Section. These conditions may include any interlocking activated by the move, or speed restrictions applicable to turnouts.

Where a choice of "Next Section" is offered, the selection is determined by the Route Number of the trip in question. It is possible to enter more than one choice for a given Route Number, allowing the "Next Section" to be selected on the basis of traffic conditions. To model blockages of Track Sections other than those caused by movement of trains, such as level crossing closures or signal failures, the user can enter probabilities and magnitudes of such blockages, to be triggered when a Track Section is approached by a train.

Each category of train, as defined by common traction units and features such as braking rate, length, and mass, is identified and allocated a Train Type number. If required to reflect variations, train mass can be entered as a user-defined probability distribution. For each train simulated, a mass is then selected by the simulator from within the
specified range. Each timetabled trip is tagged with a Train Type number and Train Route number. The trip schedule is then defined by entering the departure time from the origin and arrival time at the destination, supplemented by arrival and departure times at intermediate stations. Probability distributions can be entered for late starts from the origin station, and excess dwell time at intermediate stations. These distributions can be made specific to each Train Type, or even to individual timetabled trips.

**RailPlan™** generates each trip at its timetabled origin station, in the appropriate platform for its Train Route number. The trip will then depart at the timetabled time so long as Track Sections forming the route out of the station are free. If use of Departure delay probability distributions has been selected, a delay may also be allocated from the distribution.

Movement of trains is modelled by deriving acceleration from the Tractive Effort and mass of each train. As Tractive Effort varies with speed, typically reducing as speed rises, the acceleration is constantly recalculated throughout each train’s journey. The speed of trains is adjusted to allow for gradients, and rolling and aerodynamic resistance. Acceleration continues until the train reaches the most restrictive of: The maximum permitted speed for the Train Type; The applicable route speed restriction; The maximum speed compatible with the braking distance available to the end of the next free Track Section.
As a train enters one Track Section, the next section ahead is requested, and if available, locked against other train movements, and included in the braking distance seen by the train. The Route Selections window is used to define any further sections on conflicting routes that are required to be locked, and the distance to be travelled before these sections are released. Should any of these conflicting sections already be held by another train, then the move being requested cannot be made.

If the next section is not free, braking is initiated so as to bring the train to a stand at the end of its current section, and a “Caution” is registered in the output statistics. The caution is then classified either as “Caution-stop” or “Caution-clear”, depending upon whether the train actually has to come to a stand. Once braking has been initiated, in normal operation trains will be released from the braking curve immediately the subsequent section becomes free. This functionality is modified when use of the various transmission-based signalling systems is selected.

The mode of operation described above is equivalent to 3-aspect fixed block signalling. Where greater braking distance is required in order for permitted speeds to be achieved, for instance, with 4-aspect signalling, input data can be amended accordingly. It is also open to the user to select use of a variety of transmission-based signalling systems, or speed signalling with or without speed monitoring.

The run to each station is accomplished in accordance with the train’s capabilities, the distance and route characteristics between stations, and any conflicts encountered with other trains. At intermediate stations, punctuality is then reported with respect to the timetabled arrival time. If early, the train will wait until the timetabled departure time, but if late, it
will wait for a dwell time equal to the difference between the arrival and departure times. If use of Dwell delay probability distributions has been selected, a delay may also be allocated from the distribution. At the destination station, delay or early arrival is assessed against the timetabled arrival time, and incorporated in the summary output data for the simulation run.

**RailPlan™** offers a wide range of outputs to allow the user to analyse the operation of the system being modelled. These outputs comprise: An animation of train movements, similar to a signalbox mimic diagram; Graphical displays; A chronological list of all events in the simulation; Tabular statistics for both individual trains and complete services.

Graphical displays of output available are the Train Graph, and Speed:Distance plots for selected trips. In addition, plots of certain input factors can be retrieved, for traction characteristics (tractive effort v. speed curve) of each train type, and the track profile applicable to different sections of the network. The chronological Events Listing details for every train the times of events such as station arrivals and departures, section-to-section moves and release of signal sections and track circuits, as well as instances of conflicts between trains. A search facility is incorporated, allowing the user to extract lines of entry for selected events, such as: All events connected with one train; Departures from a selected station; Release of a particular track circuit; Movements past a given signal.

Processed tabular output includes: Station arrival and departure times for each train, together with delay compared with the timetable; Numbers of caution aspects presented to trains, broken down by individual signal section; Summaries of each train’s journey, including running time, overall journey time, energy consumed by train movement; number of signal checks experienced, and overall delay; Totals of train movements through each signal section; Overall total, and summary for each train type, of actual delay and percentages of trains in different delay-bands, Overall total, and summary for each train type, of energy consumed by train movement.

In addition, the user can select before each run up to five trains to be subject to a detailed report, giving their speed, position and time at specified distance intervals, as low as ten metres. These reports allow close study of the acceleration and speed of trains. **RailPlan™** output files are designed to be opened via a spreadsheet package such as Microsoft Excel, so that the data can be manipulated as necessary, and charting facilities can be used to prepare graphical displays according to the user’s needs.