A note on section scheduling on the Indian Railways

Narayan Rangaraj
IEOR, IIT Bombay
400076

1 Introduction

The investment on track and associated capital equipment forms a major chunk of the resources used by the railways. The effective utilisation of this resource is of crucial importance. Although the modes of resource utilisation by non-homogenous traffic and varying track configurations makes it hard to quantify, a serious attempt must be made to define, measure and improve the effectiveness of this resource.

It is hoped that this note will go some way towards defining such measures of performance and discusses how far current operating practices go towards maximising these. The objective of such an exercise is to improve long term performance of services and key assets, while leaving operating policies flexible enough to respond to local requirements.

Specifically, this note has tried to achieve the following:

1) Outlining a methodology for section scheduling in a general setting

2) Identification of some major algorithmic issues in simulation of section scheduling

3) Proposing a definite set of measures for monitoring section performance

4) Constructing links of the section scheduling decision with other operational issues in the railways.

The approach is user oriented, and is meant to be qualitatively from the point of view of section control as visualised by the Indian Railways today and in the near future.

2 Operating policies

The Indian railways have over the years evolved certain sound operating practices of section movement, beginning with the all-important conditions of safe operation. The goal of prioritised movement of traffic to achieve goals of timely performance is translated into a hierarchical set of objectives.

For example, major objectives of railway operations are punctual running of high priority trains and quick turnaround of freight trains. These are translated into sectional objectives of efficient transfer of trains to other sections, or other zones.

A few of the most important operating practices relevant to the issue at hand are:

1) A block section may be occupied by only one train at a time

2) A train should not be delayed because of a train of lower priority train (note that a train can certainly wait - if early). At the simplest level, this is interpreted to mean that when competing resources have to be allotted (even ahead of time), the higher priority one wins. But if we consider a situation in which there are cascading delays, it is not easy to say if some delay is because of some other train. This is discussed in more detail, the section on cascading effects of bottlenecks in Section III. The situation is further complicated if priorities themselves change over time. For example, suburban trains acquire highest priority during the peak commuting hours.

3) A first come first served discipline is followed among trains of the same priority asking for the same section resource. However, see Section III, 3, where the scheduling strategy may permit this discipline to be overridden to permit late running passenger trains to recover time.

3 Objectives

If we view the section scheduling problem as one of dynamically allocating resources to activities, a number of standard concepts from scheduling theory are relevant [8].

3.1 Due dates and tardiness:

Every passenger train has a got a scheduled time of arrival at the end of a section and indeed at all stations in the section. The time specified in the timetable typically allows for some slack, to permit trains to make up time lost due to unforeseen reasons, such as occasional speed restrictions and unscheduled halts. The relevant measure of performance as far as each train is concerned is tardiness. (Tardiness is nothing but lateness when it is positive and zero otherwise). An aggregate measure using tardiness will not give credit to early arrivals, since they do not contribute positively to efficient movements later on (i.e. the train has to wait for the scheduled time path anyway).

3.2 Aggregate lateness and throughput

In the case of freight trains, the situation is more complex. Minimising the aggregate delay is more relevant, since this ensures quicker delivery to customers, faster wagon turnaround and asset utilisation. If we consider the aggregate performance of a whole bunch of freight trains of the same priority, this goal can be translated into throughput related performance of certain key resources on the section.

3.3 Utilization

In certain conditions, the utilization or minimum idle time usage of a certain bottleneck resource can be taken as a surrogate for the performance of the entire section. This bottleneck resource is equivalent to the ruling block section under normal operating conditions. What complicates the situation is that the bottleneck section keeps shifting because of the particular traffic pattern and the way it uses the track configuration.

For example, if a single loop is known to be occupied by a (low priority) train for a long time, effectively the two sections on either side behave as one long block section, which may become the ruling section.

4 Scheduling strategies

Well known strategies in scheduling theory can be combined to yield reasonable operating policies. A possible framework is the following:

1) Classify traffic at a point of time as "free" or "congested".

   Free traffic refers to normal running of trains, where the number of trains in the system is comfortably below some nominal capacity of a section (a simple estimate of the capacity is the maximum throughput past the ruling section, or longest time section). Free traffic also means that there are sufficient resources (mainly loop resources) to permit the required passing or crossing of trains of different priorities. Congested traffic is where passing of trains is difficult and the idle time of a certain resource (the bottleneck resource) directly translates into lost time for the overall objective.

2) For free traffic, a rule based on due dates, which will minimise tardiness is appropriate for trains with fixed schedules. A version of an Earliest Due Date (EDD) rule will ensure that late running trains will be allowed first passage. Although this violates the First Come First Served (FCFS) operating rule, it tries to restore the correct sequence of trains on the section. For other trains, without fixed schedules, the FCFS rule is optimal under free traffic conditions.

   The outstanding issue here is that there are trains of different priorities and some policy will have to appropriately penalise the lateness of each train or at least each class of trains. Two options are possible:

   a) A weighted measure is formulated (weights refer to commercial or other importance of the particular class of train)

b) Thresholds are specified and priorities are redefined if the thresholds are crossed. For example, a low priority train which is more than K hours later will be reclassified to have a higher priority is a possible rule.

The latter rule is easier to implement.

The cost of implementing non FCFS strategies is that they may involve significant passing or crossing of trains, and traffic is slowed down (a low priority train which is ahead will have to decelerate, come to a halt on a loop, and after some block working time, a passing or crossing will take place).

All this presupposes that there are loops available for this purpose. It also assumes that the overall throughput of the section is not affected. If the traffic density increases, or the traffic pattern and track configuration force too many delays of the type above, then the traffic condition moves to "congested".

3) In congested traffic, a bottleneck is first identified. This may be done directly by calculating the current ruling section, but since the section is something like a flow shop, the real reason for waiting trains may be a section a few stations ahead. So queues before (and after, in single line sections), a section may also be a good indicator of the cascading impact of a bottleneck section. Also, since waiting trains may create further bottleneck sections, by preventing passing, the detection of bottlenecks is a dynamic and subtle issue.

Cascading

The true test of a bottleneck is when a unit of capacity lost at that facility implies a unit lost on the overall performance parameter of throughput or cumulative delay or makespan. Since this is difficult to compute dynamically, we look for indirect evidence of a bottleneck such as queues waiting for the use of a facility. Since waiting space may be finite, queues may be shifted to neighbouring sections and so it becomes important to allocate waiting to a particular section.

We can take the algorithmic analog of trying to force as much as possible and observing where material waits. In any algorithm for section scheduling, there will be some waiting component as sections are busy. In addition, there will be some backtracking in the algorithm where trains are not scheduled on sections even when free because there is no track (waiting room) available at the other end. We can refer to this as cascading. This is recursively applied to ensure that there is track resource kept available for passing (high priority) trains, by looking ahead in time.

In any case, algorithmically speaking, any waiting beyond a scheduled halt can be unambiguously allocated to a particular section, from the origination of the associated backtracking procedure. This might yield a more accurate picture of bottleneck conditions together with resource constraints.

4) Once it is decided that a certain section is a bottleneck and is (likely to be) so for some time, an operating rule based on bottleneck utilisation can be proposed. Some rules for single machine scheduling rules are potentially applicable here. No version of the SPT (Shortest processing time) rule is really relevant, because all the jobs are not available at one time. The simplest rule which minimises idle time of the section is one which minimises the block working time between successive trains. This may lead to some platooning (or flighting) of trains, where trains in one direction scheduled first, without passing and then in the other.

   Two issues emerge here

   a) In case an entire yard is viewed as the bottleneck resource (which is often the case from the point of view of the entire railway network), this issue may require some consideration of "the sequence dependent set up" problem in scheduling. In addition, there are significant routing decisions in yards and the analysis of this is not discussed here.
   b) In the section scheduling context, any strategy of bottleneck utilization has to periodically verify the bottleneck status of the resource. This has been touched upon earlier, but needs reiterating, because any strategy such as platooning will lead to certain movements and loop occupancies, which may create other bottlenecks.

5 Summary of scheduling strategies

In the above section, the following major issues have been discussed:

1) Classification of traffic as ”free” or ”congested” and periodically reviewing this status,

2) A suitable version of a due date based rule for scheduling in free traffic conditions, and an appropriate mechanism for handling multiple priorities,

3) A periodic and definite rule for identification of bottlenecks in congested traffic conditions,

4) A utilisation based rule for scheduling the bottleneck resource.

   In each of 1)-4) above, the best procedure in a given case is open to research in the given context.

6 Simulation

The basic tool for computer support for section scheduling is simulation of the train paths, ahead of time, based on operating conditions and decision rules. A number of studies haver reported sophisticated models and implementation of simulation based models in practice. For example, the results of Petersen et al [3,4] give some early experience of simulation and several works by Carey, Hachemane and others, for example [1,7,10], summarize more recent efforts, based on a mixture of mathematical programming and simulation.

Simulation, or look ahead is by itself fairly complex, since in practice,

1) a number of events are simultaneously scheduled

2) a number of decisions (such as loop preferences) have to be taken at the time of scheduling.

Note: Routing decisions of a complicated nature (other than loop preferences) are not considered here.

With these factors in mind, the section scheduling procedure can be simulated to capture as much of the operating rules (implicit and explicitly stated) and computed as efficiently as possible. There are two important issues in any procedure for scheduling trains. In what follows, we use the term resource allocation to refer to either section or loop allocation to a particular train.

6.1 Physical deadlock

This happens when, in a physical sense, there is no passage possible unless one train is reversed to at least the preceding passing station. This is obviously extremely undesirable (however, refer to the section on accidents and unusual situations) and has to be avoided in the simulation outcome as well. Simple examples will show that following the basic rules in above can easily lead to such physical deadlocks.

Therefore, some amount of look ahead is required in any resource allocation decision. The more the look ahead, the more is the computational time spent in each decision and therefore the entire simulation. This may very quickly reach levels which make it unacceptable as an on-line tool.

Also, for this reason, some other models [9,11] have resorted to scheduling trains through to the end of a section before taking up another train. This has some major advantages:

a) It is guaranteed to terminate in a result which is not a physical deadlock

b) It is in fact done for higher priority trains, as a class, before taking up lower priority trains and if we have a continuum of priorities (e.g. super fast over fast over ordinary mail express over passenger, etc., within each class, earliest due date first, and for freight trains according to classification of goods), this policy can be interpreted very reasonably

c) In a sense, the FCFS policy implemented section wise yields this policy

d) The coding of such an algorithm is considerably easier.

6.2 Back tracking and look ahead

In the algorithmic sense, certain conflicts of resources later in time are detected only after simulating the event. This may require algorithmic (but not physical) backtracking or de- reserving of certain resources.

In practice, section schedulers who do manual scheduling do not do algorithmic backtracking as it is tedious. They learn to estimate the possibilities of backtracking by a very sophisticated but often informal set of rules. This may lead to overly conservative behaviour during scheduling, and result in less efficient operations. This is one definite advantage of a computerised scheduling system, since look ahead computations can be done ahead of time and fairly efficiently.

7 Section performance

We now turn to defining a reasonable measure of section performance in the medium/long term. If physical signalling and block section infrastructure is in place, the first consideration in estimating section performance is the capacity, as determined by the ruling section. As a first cut rule, there is a case for having equally spaced (more precisely, equal traversal time) block sections, as far as possible.

Capacity:

Unlike in road networks, there are some difficulties in defining the capacity of a rail section. Even if we standardise the concept of capacity to mean the number of trains that a given section can handle without suffering unmanageable delays, the following need to be taken into account:

1) The traffic is not of a homogenous variety and there is no equivalent of the Passenger Car Unit (PCU) homogenisation of flow units, that researchers have defined for road networks [5].

2) The interaction between traffic of different types and section geometries is quite complicated and makes it very difficult to develop any analytical estimate of capacity.

3) Like in many resource constrained system, capacity is a nominal concept and is heavily influenced by what sort of operating policies are in place. The section scheduler has to consider a number of constraints apart from the objectives discussed in the earlier sections, and it is difficult to come up with a very clear measure for each section.

The problem of capacity estimation on rail networks is discussed, for example, in [2,9,11].

Simulation and measures of performance

Simulation and measures of performance:

One proposal could be based on section schedulers themselves simulating the section under traffic profiles drawn from previous data and under some standard operating policies. Sufficient time could be added for exceptions and contingencies (which would anyway be recorded separately in actual operation and monitoring), and an actual section performance can thus be derived.

Speed flow diagrams - of number of trains in a section against average waiting times - can be derived which indicate that the carrying capacity of sections is limited by some density of traffic after which section waiting times shoot up. This indicates an upper limit on the number of trains which can be comfortably handled in a section. It also gives some expected average times of waiting for a given level of traffic, which can be used to monitor operating performance.

Similarly in the case of passenger movements, due date based scheduling rules can be easily tested, which are likely to be in force in normal free movement conditions. If the section condition is diagnosed as being free of congestion, then the measures of tardiness translate directly into the railway goals of punctuality for passenger train operation.

To validate the relevance of such measures, it is necessary for operating personnel to use the tools of section simulation and actually build up of rules, so that the norms are based on operating reality. Perhaps the use of such simulators in the general training of section schedulers (rather than implementing automatic section scheduling) is a first step towards this validation.

8 Section scheduling vis a vis other operational decisions

The section scheduling decision is only part of the set of operational decisions involving resource allocation. We discuss some interconnections of this decision with some other important decisions.

Crew scheduling:

In the normal run of things, once train scheduling decisions are taken, then the other operating decisions, including crew allocation for the services are decided accordingly. Crew is a particularly important resource (although it may not be the most expensive resource). The human element in this resource means that there are a number of conditions which have to be followed for safe and comfortable operation of trains by crew.

These conditions (in the Indian Railways), have to do with

1) Length of duty,

2) Relief at appropriate times

3) Additional duty

4) Locational constraints for providing relief

5) Appropriate advance notice for performing duty, etc.

We outline some important issues, mainly in the case of freight trains, where timings are much more uncertain. The basic argument is that a basic simulation of section scheduling will provide inputs to the other resource allocation decisions.

1) A clear picture of section scheduling performance in the next few hours, perhaps by a simple simulation of expected performance is of practical use to schedule crew at appropriate places. The times at which crew replacements will be required can be accurately estimated. This will avoid undue excess provision of stand-by duties.

This is a somewhat delicate issue, as one operational goal of efficiency of section scheduling (which may even have some incentives) has to be done in tandem with a cost control exercise (where there may be some disincentives, for not having stand by crew, for example). There is a temptation to play safe by allotting extra crew (perhaps with additional - even overtime - costs), because of the inherent uncertainty in section performance. Therefor, unless these decisions and trade offs of cost control and efficiency are under the same control, it is difficult to implement this.

2) If section scheduling forecasts, through simulation, reveal unduly long halts of low priority freight trains, the priorities of trains can be reworked in the light of the following requirement. There is a period of discretionary duty by crew (where they can choose to accept or not, if sufficient notice is given) and a period beyond which they cannot be asked to perform any more duty without proper rest. If this time happens to be in the middle of a long section, then the crew is entitled to proceed to the nearest relief facility (running room), with the loco of the freight train to transport them, if necessary.

This could mean considerable additional delay for the freight train and also loop occupancy for a long duration. An overriding priority can be considered for such freight trains, failing which alternate arrangements have to be made. This analysis, of course, requires detailed data about the crew on each train.

Loco allocation:

The use of automated section scheduling for providing clear time estimates of arrival of freight trains at end of sections is useful in the following way also. At points where traction changes, most commonly from diesel to electric and vice versa, locos need to be allocated efficiently. The operating policy in such cases is either that the loco should not be kept waiting for the rake or that the rake should not suffer detention for want of a loco, depending on which is the locally felt scarce resource or need. In either case, the section scheduling procedure can be adapted to provide useful inputs to this decision.

Case 1: If the loco use is important, then the desired time of arrival of freight rakes can be calculated and by backward scheduling, the appropriate time paths of freight trains can be calculated. This will in all probability make the scheduler's job easier, by providing a slack on some activities and allowing freedom in prioritising others appropriately.

Case 2: If the rake movement is of primary importance, loco allocation based on estimated times of arrival as determined by the simulation should yield better results.

Rake information to customers and to the commercial department

Along the same lines, customer information about rake arrivals and availability can be accurately made available, in those cases where rakes terminate along the section in question. If the larger umbrella of a Freight Operations Information System is in place, the appropriate input to that MIS can be made. Accurate customer information ahead of time has got short term implications in how demurrage and other costs to do with service and delivery are concerned, and long term implications on what services and service guarantees to offer to customers.

Accidents and unusual situations:

The use of automated systems for section scheduling is of obvious utility in accident situations, where unusual traffic conditions are in force and there is a lot of associated decision making, such as rerouting of trains, cancelling some trains, regulating some trains and so on. Similarly, it is of use, during planned maintenance of any aspect of track, or overhead equipment or signalling, even when running times are significantly different from normal conditions, but especially when temporary diversion of traffic has to be planned.

Algorithmically speaking, the following issues emerge:

1) It should be possible to specify track access and operating policies as per the new situation in an effective manner.

2) The issue of moving between different regimes for scheduling (free to congested) is particularly sharp here and one would have to build in fail safe rules to ensure no physical deadlocks. Manual control of the section may have to be restored to, perhaps to facilitate interrelated decision making about revised timetabling and other decisions.

3) These conditions are likely to push the effectiveness of, for example, look ahead in scheduling, to the limit and it may well be the case that the formal rule base constructed from the section schedulers experience may not be algorithmically sufficient (after all, humans have many other informal mechanisms of decision making). The section scheduling algorithm needs to have a contingency search procedure for generating at least some reasonable schedules in such cases. This is quite a challenging problem, since the best known mathematical representation of the problem can solve very small problems and it is demonstrated even with powerful commercial software that general search procedures fail in this case, unless supported by well thought out rules to guide the search. This is a challenging field for research.

Of course, the construction of optimal or even efficient rules for section scheduling in a given regime of operation is the area which will see most fruitful work being done, both theoretically and in implementation.

9 Acknowledgements

I would like to acknowledge the contributions by B.C.S.Krishnamohan, Priya Mahadevan, Julie Shah and Manish Garg, who have worked on different aspects of this problem with me, Kejitan Dontas for several useful discussions and pioneering efforts in simulation and G.Raghuram for introducing me to this problem. I also thank the several railway officials who have patiently explained the intricacies of this most fascinating issue to me.

10 References

1) Patrick Hachemane (1996) Evaluation de la capacite de reseaux ferroviaires (Evaluation of capacity of rail networks), proceedings of the 7th World Conference on Transport Research, Sydney, 1995, Pergamon-Elsevier Science Ltd., Oxford, U.K.

2) Florio, L., and Mussone, L., (1996) A method of capacity computation for complex railways systems, proceedings of the 7th World Conference on Transport Research, Sydney, 1995, Pergamon-Elsevier Science Ltd., Oxford, U.K.

3) Petersen, E.R. (1977) Railyard Modelling: Part I Prediction of throughput time, Transportation Science 2.

4) Petersen, E.R. (1977) Railyard Modelling: Part II The effect of yard facilities on congestion, Transportation Science 2.

5) Transportation Research Board (1985) Highway Capacity Manual, Special Report 209.

6) Assad, A.A. (1980) Models for Rail Transportation, Transportation Research, 205-220.

7) Carey, M. and Lockwood, D., (1995) A Model, Algorithm and Strategy for Train Pathing, Journal of the Operational Research Society, 46, pp. 988-1005.

8) Baker, Kenneth R., (1974) Introduction to sequencing and scheduling, Wiley, New York.

9) Krishnamohan, B.C.S. (1994), Capacity estimation and enhancement on transportation networks", M.Tech Project Report, IEOR group, Indian Institute of Technology, Bombay.

10) Rudd, D.A., Storry, A.J. (1976) Single track Railway Simulation - new models and old, Rail International, p. 335.

11) Raghuram G., and Venkat Rao, V. (1991) A decision support system for improving railway line capacity, Public Enterprise, pp. 64-73.