ERTMS Level 2: effect on capacity compared with "best practice" conventional signalling

W. A. M. Barter
First Class Partnerships Ltd, UK

Abstract

This paper reviews potential capacity benefits attributable to ERTMS Level 2 compared with UK Multiple Aspect Signalling (MAS), tests scope for their exploitation against the practicalities of preparing a comprehensive timetable for a suburban rail network, and proposes simulation experiments to confirm the benefits.

The UK is preparing for adoption of ERTMS Level 2, System D, as its standard signalling system. At the same time growth in passenger demand deriving from privatisation and socio-economic factors is continuing at levels beyond those forecast by conventional planning models, and major expenditure to cater for demand is becoming necessary. It is widely hoped that ERTMS will offer capacity benefits to help cater for demand in the medium term. A variety of claims for the potential capacity increases deriving from ERTMS Level 2 have been made. Many are felt to be simplistic or optimistic.

Effects of ERTMS Level 2 System D are seen to arise principally in the context of reduction of line headways. Compared with UK MAS, headways are expected to be reduced by cab signalling normalising block boundaries once lineside signals are eliminated, and from division of the train safety separation into shorter signalling blocks. However, line headways are only one factor in determining the capacity of a network, and other crucial factors are largely unaffected by the chosen signalling system.

The practical potential of the likely benefits is then tested for plausibility against the example of the commuter operation serving London's Charing Cross station. A 10% increase in capacity is found to be plausible, but only so long as outputs from unrelated projects can be assumed, and track circuit arrangements are redesigned for the purpose. Some methods of operation, and public expectations of the type of service, must also be modified.

Keywords: signalling, operations, capacity, ERTMS.

WIT Transactions on The Built Environment, Vol 103, © 2008 WIT Press
www.witpress.com, ISSN 1743-3509 (on-line)
doi:10.2495/CR080221

1 Introduction

Pressure on capacity of the UK rail network continues to increase. Passenger journeys in 2007 amounted to 1.2 billion, an increase of 7.8% on the previous year, whilst passenger miles exceeded 30 billion for the first time since 1946. The point appears to have been reached at which major expenditure to increase capacity is becoming inevitable, initially through lengthening of trains and through infrastructure work at bottlenecks. Work to increase the capacity of the Thameslink cross-London North-South route with improved signalling and additional tracks at London Bridge is in hand, and a project to create the new London East-West Crossrail route has been authorised.

In the longer term there is growing pressure to consider new capacity in the form of high speed lines for long-distance services, although double-deck solutions for existing lines are probably ruled out by infrastructure constraints and the relatively small capacity benefit achievable within UK vehicle dimensions.

Against this background, it is essential that capacity benefits attributable to ERTMS are on the one hand exploited to the full as an option for comparison with other major infrastructure solutions, and on the other hand are soundly based, to ensure that theoretical benefits can be realised in practice.

ERTMS Level 2, System D, is emerging as the preferred UK option. Although the principal benefits on which its adoption is predicated are safety and the reduced cost of equipment compared with conventional signalling, it is likely that some capacity benefits will need to be identified in order to formulate a positive business case for the adoption of ERTMS.

A variety of estimates have been made for the potential capacity benefits of ERTMS. However, in simply assuming that benefits claimed will transfer in practice to the UK context, considerable uncertainty is encountered. For instance, Invensys [1] suggests that ERTMS Level 2 on the High Speed Line Córdoba-Málaga will enable 24 trains per hour, compared to the current Spanish national system capacity of 7.5 trains per hour. However, the figure suggested for ERTMS seems to be a theoretical maximum, whilst the comparison appears to be made with a historic actual figure, rather than with best practice conventional signalling if applied to the new line. In the UK, the Strategic Rail Authority and Railway Safety and Standards Board [2] describe the potential capacity benefits of System D as “significant”, offering an “increase by potentially up to 1 in 10 train paths”.

First, the basis for comparison of many claims needs to be clarified. UK 4-aspect signalling has been in use since 1925. Since then, standards for the system have evolved to find a sophisticated balance between safety and capacity. In intensively-worked areas such as the South London suburban lines, the signal engineers have become extremely skilled in exploiting the system to best advantage. So long as the signals are located exactly as required to provide the braking distance for the intended maximum speed, and trains actually run at that speed, theoretical headways are remarkably low, around 90 seconds on 4-aspect signalling for 160 kph trains, and little over a minute at half that speed [3]. In

claiming benefits for ERTMS, comparison needs to be made with this highly-evolved best practice.

Then, contrasting with many networks, that of the UK retains a widespread mixed traffic capability operating over complex track layouts. The South London suburban system sees significant, and growing, use by freight trains, serving both Channel Tunnel and seaborne container flows, and domestic traffic such as aggregates for distribution in the London area, or dredged from the Thames Estuary for use outside London. Many routes are limited to double track, but still have to carry both fast and stopping passenger trains, and frequent junctions with only limited grade-separation are a legacy of the evolution of the network.

2 What is "capacity"?

Many assessments of capacity are simplistic, using the technical headway to calculate line capacity glibly in terms of "trains per hour", and are inadequate in the face of the realities of a complex, multi-purpose network.

For each line in a network, the signalling system sets the "headway" - the minimum possible interval between trains that avoids restrictive signal aspects. The headway is constrained by the realities of lineside signals, which must be clearly visible to drivers of approaching trains, not just the wrong side of bridges or tunnels, or out of sight round curves in cuttings. We tend not to place signals in the middle of station platforms so as not to stop trains frustratingly half in and half out of stations. Access for maintenance may militate against placing them in tunnels or on viaducts, which also avoids the risk of trains being stopped at locations that passengers might find unnerving. As signal sections cannot be shorter than is necessary to give braking distance, all these problems can only lead to longer sections and thus longer headways, and the worst group of sections sets the headway for the route.

All in all, once the signalled headway has been rounded off for the convenience of timetable planners, and some allowance made for robustness in practice, a 200 kph line will probably end up with a planning headway of 3 minutes, and line on a suburban route, 2 minutes.

That is all well and good for one line in isolation, and for a continuous flow of trains running at the full permitted speed, but hardly describes any real railway system. In practice, trains stop at stations, so that their dwell time, which is completely independent of the signalling, adds to the separation. And some trains stop at stations while others don't, so that a wedge of unusable capacity builds up between a through train and a following stopping train.

This loss of capacity can be mitigated by "flighting" - running trains of the same speed in pairs or batches. However, intermediate stations may then find their stops concentrated into short periods, and a more passenger-friendly pattern may be laid down in franchise specifications at the expense of capacity.

Other factors combine to reduce the calculated capacity further. Flat junctions destroy opportunities to run trains simultaneously on conflicting routes. At each end of the line, trains need to turn back at terminal stations - the rate at which this can be done, determined largely by the turnround time and the number of

platforms, is normally much less than the rate at which each approach line might feed trains in or out. Finally, reality suggests it is unwise to work continuously to the limits of capacity.

So line capacity measured in "trains per hour" is really a technical abstraction, useful for comparing some details of signalling schemes, but for little else. In fact, terminal capacity is probably the binding constraint on usage of much of the UK national network.

The UK Institution of Railway Operators' definition of network capacity, adopted in the Department for Transport's Rail Technical Strategy [4] is:

"The number of trains that can be incorporated into a timetable that is conflict-free, commercially attractive, compliant with regulatory requirements, and can be operated in the face of anticipated levels of primary delay whilst meeting agreed performance targets".

3 Charing Cross – a practical example

London's Charing Cross station caters for inner suburban trains from South-East London and outer suburban trains from the county of Kent. The intensity of operations was recognised as long ago as 1922, when the South Eastern & Chatham Railway introduced its "parallel working" timetable, optimising the train service around critical junctions approaching Charing Cross where trains diverged to serve the "City" terminus at Cannon Street. This style of working persisted until 1975 when extensive track and signalling alterations, with a limited application of grade separation, allowed trains for Charing Cross and Cannon Street to be allocated to separate tracks 9.6km out, at Parks Bridge Junction.

Key features of the infrastructure approaching Charing Cross are:

- **Charing Cross station**: 6 platforms, worked as two groups of three, each group served by a pair of approach tracks, known as the "Fast" and the "Slow" lines, although the permitted speed on both pairs is 40 kph;
- The two pairs of lines continue through **Waterloo East station**, 1km from Charing Cross, with one platform per track. This is a major interchange location with trains at Waterloo Main Line station as well as the London Underground, and also serving directly growing employment areas in Southwark, and the areas North of the River Thames accessed by way of Westminster and Waterloo bridges. Passenger usage demands a typical dwell time of 1 minute;
- At **Metropolitan Junction** (2.2km), the two pairs of tracks converge into one at an at-grade double junction, and run as such for 0.26 km to just short of London Bridge station;
- Approaching **London Bridge station** (3km) the two tracks fan out into four "paired by direction". The station provides interchange with Cannon Street services, and with the London Underground to access the growing employment areas in the former London Docklands. Passengers transferring off inner suburban and outer suburban trains from South London and Sussex also transfer onto Charing Cross and

Cannon Street trains. For commercial reasons, the objective is to maximise the number of trains that stop at London Bridge, although one of the two tracks in the "Up" (to London) direction has no platform face and is only used by through trains. Again, passenger usage including interchange is such that a dwell time of $1^1/_2$ minutes is called for at peak times;

- Between London Bridge and **Parks Bridge Junction** (9.7km from Charing Cross) the Charing Cross lines reduce to a single pair once more, and run adjacent to a pair of tracks for Cannon Street trains. At Parks Bridge Junction itself, at-grade connections allow exchange of trains between the Charing Cross and Cannon Street lines, with some very limited grade-separation to access branches of the suburban network;
- Between Parks Bridge Junction and **Orpington** (22.2km from Charing Cross) the four lines run in pairs segregated "by use", with the extension of the Charing Cross lines catering for through trains and stopping trains allocated to the extension of the Cannon Street lines. After Orpington, where many inner suburban trains terminate, the four lines converge into two;
- This double track continues to **Sevenoaks** (35.6km from Charing Cross), carrying outer suburban trains and remaining inner suburban trains serving intermediate stations. The section features two long tunnels and three intermediate stations. Sevenoaks is the limit for inner suburban services.

The net effect is a complex network with many at-grade junctions, carrying a mix of fast and stopping trains through two major interchange locations to a relatively small terminus.

Today, 30 trains arrive at Charing Cross in the busiest 60-minute period of the morning peak, even though, based on planning headway alone, the Fast and Slow lines immediately outside the station could feed in 48 between them (the situation is of course complicated by the short stretch of double-track between London Bridge and Metropolitan Junction, offering just one line for Up trains).

As Figure 2 shows, the three Slow line platforms work continuously through the peak of the peak at the minimum turnround of 7 minutes. With 3 minutes between occupations of each platform, this comes to 18 trains. This is just 75% of the theoretical capacity of the Up Slow line, which is set by the station stop at Waterloo East. Meanwhile, the three Fast line platforms handle only 12 trains, largely as many trains work back in service according to a clockface timetable rather than just at planned arrival plus 7 minutes. Even so, an average of 4 trains per platform per hour is fully comparable with other London terminals such as Victoria and Waterloo, and free time in the busiest hour equates to less than 5 minutes per platform.

Amongst the figures that have been suggested for the potential capacity benefits of ERTMS is 10%, a nice round number. So for 30 train-per-hour Charing Cross, that means three more, or 33 trains per hour on 6 platforms. Is that possible?

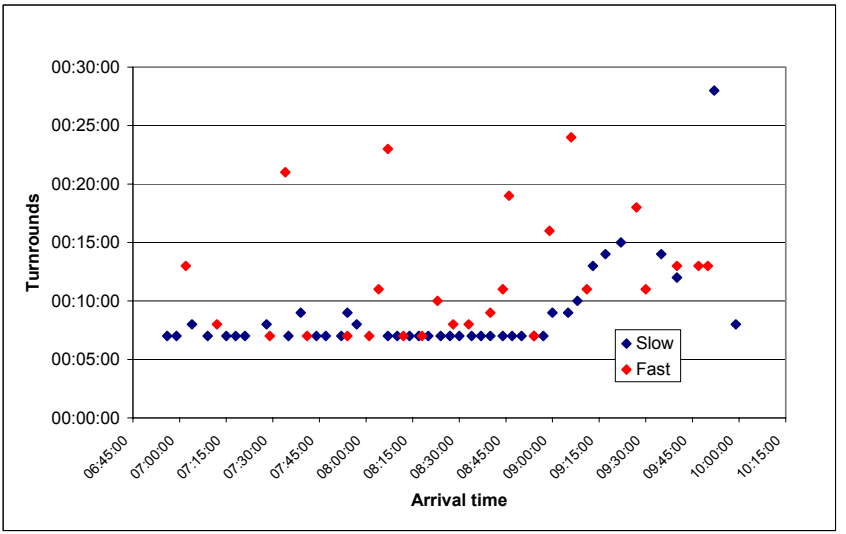

Figure 1: Charing cross morning peak turnrounds.

4 How might ERTMS help?

The features of ERTMS relevant to capacity derive essentially from cab signalling and Automatic Train Protection.

If lineside signals are done away with, the message to drivers becomes simply a safe speed at which to drive, calculated by the on-board computer and displayed in-cab. So a practical system could have shorter blocks, and more of them between trains, without the need to display different aspects and expect a driver to comprehend them - perhaps the equivalent of 10-aspect signalling. Some things follow immediately from this:

- Any fixed block system puts one more section between free-flowing trains than is actually required for braking distance. With a given separation provided by a large number of short blocks, this extra section adds less to the total separation – in effect, the benefit of 4-aspect signalling compared with 3-aspect, taken to extremes.
- By decoupling block boundaries from the constraints of sighting lineside signals, block lengths can be closer to the theoretical minimum, minimising excess separation. We might, however, still be reluctant to split tunnels into more than one section, but is this really valid in these days of central door locking, good lighting, open stock and public address systems?
- With lineside signalling, trains running on greens are separated by the full braking distance for the maximum permitted speed, even if their own permitted speed is lower and their required braking distance shorter. ERTMS can give an unrestricted “movement authority” to a slow train on the basis only of the braking distance it actually needs,

rather than the worst case (probably the fastest) train, so a flight of slow trains can run with less time-separation than fast trains.

- Given the Automatic Train Protection functionality of ERTMS, the risk of misjudged braking is virtually eliminated, so signal overlaps might be reduced significantly or even abandoned, further reducing separation.

So ERTMS potentially reduces headways, if track circuit arrangements and block boundaries for ERTMS Level 2 are redesigned specifically, rather than simply being ported over from the previous conventional schemes. All things considered, a 3-minute planning headway on 4-aspect signalling might become 2 minutes under ERTMS.

That sounds excellent - line capacity goes up from 20 trains per hour to 30. The problem is that very few lines with 3-minute headways now actually carry anything like 20 trains per hour, for all the reasons of junctions, differing speeds, and terminal capabilities outlined above. ERTMS will do very little for those problems.

With regard to the mix of train speeds, the underlying issue is one of differing running times, not of headway. True, at the point where trains enter a "corridor", a slow train might follow a fast a bit more closely to start with, but the lost capacity on route will not change. Perhaps once the fractions of minutes mount up, another complete train might be run, but which sort of train – another fast, another stopper, or what? The benefit of improving headways is only felt when trains of the same speed and stopping pattern follow each other.

At junctions, some benefit might be found. Without signal overlaps, the last block boundary before a junction can be closer to the point of conflict than a fixed signal would be. With route set only as far as necessary for braking distance, slow trains could approach the point of conflict more closely before the interlocking needs to "deny" it to other trains. "Advisory speeds" may allow regulation of trains short of the junction so as to coincide with a free path at the junction rather than stopping clear of the junction to wait for a path - particularly beneficial for freight trains with low rates of acceleration, and also offering environmental benefits by mitigating fuel consumption for restarting after a stop. But using one route over a point of conflict still prevents trains running on all conflicting routes.

And the ability of terminals to accept, turn and despatch trains will not change. In a suburban operation, the limiting factor is the time taken for drivers to change ends (crew changes at the terminus in the peak are not a good idea). For long distance trains, other factors come into play, such as servicing, as well as a greater robustness allowance.

5 Can these benefits be exploited in practice?

Now consider the actual constraints on capacity in the example of Charing Cross and the lines that feed it. ERTMS may well improve line headways, but can this show a benefit given other constraints such as terminal capacity, junctions, and the mix of train speeds?

In terms of platform capacity at Charing Cross, the answer is easy – the Fast platforms will have to work as hard as the Slow platforms. The clockface timetable will be at risk, and outer-suburban trains will have turnrounds as short as inner-suburban. The extra trains will have to be formed of rolling stock that can inter-work with the outer suburbs. In theory that gives us six more trains per hour, but let's not overdo things.

Work back from Charing Cross itself, and see what other constraints arise. Waterloo East comes next, where all trains stop. Again the Fast Lines could do what the Slow lines already do, in terms if frequency of service.

The problem comes at Metropolitan Junction, where the two-track section from London Bridge changes to four "paired by use". That means a diamond crossing, where the 18 trains up the Slow line have to cross the Down workings on the Fast line which, on the basis that what goes Up must come Down, now total 15. 33 trains per hour over a diamond crossing is a lot, even though the current Rules of the Plan allow 2 minutes separation, with $1^1/_2$ minutes "not for successive moves". Exploiting that to the full with 33 trains would lead to the diamond being locked out for 58 minutes out of 60 – too high for a reliable service. But if overlaps short of junctions can be eliminated, and speed of trains controlled to keep them moving, maybe ERTMS brings enough to make it realistic.

But first the 33 trains have to use the one Up line from London Bridge, and we must assume the headway benefits of ERTMS will allow this.

Once the Thameslink Project is implemented, at London Bridge there will be two platforms for Up Charing Cross trains, needing to handle 16 or 17 each per hour. This is less than the current single Up Charing Cross platform does now off-peak when almost all trains call, albeit with off-peak dwell times. However, we can hope that peak dwell times will reduce - there will be 10% more trains to carry the passengers, and the project will improve station accesses, distributing passengers better. And as signal locations are currently heavily constrained on this complex layout, we can also hope that ERTMS will reduce platform reoccupation times.

Below London Bridge, headways effectively set capacity, as the intermediate station platforms are on the Cannon Street lines. We can reasonably hope for success, at least until we get to Parks Bridge Junction. Here trains are transferred between the Fast and Slow lines so as to sort them out for Cannon Street and Charing Cross. The numbers of trains making these crossing moves will increase, as the 10% increase we are aiming for should apply to Cannon Street as well, and much will depend upon how well the pattern of service exploits scope for parallel moves. This is a big unknown, especially as, to make the London terminal work, clockface patterns are jeopardised.

From Orpington to Parks Bridge, headway is again the main constraint on the Fast lines, although usage is lower as the network fans out into branches. But trouble starts again below Orpington, where the line via Sevenoaks is only double track, and capacity is limited by the speed differential between fast and stopping trains. The fast headway is already 2 minutes, and there are long tunnels

in which we may still want to limit the number of trains. The likelihood of being able to run three extra trains over this section in one hour is low.

But with the Thameslink Project to help at London Bridge, some doubts at Parks Bridge Junction, and some heroic assumptions about signal overlaps, a 10% increase in trains can be made to sound plausible - in the inner suburban area.

But our starting point was that the extra trains at Charing Cross would have to be capable of working round with outer suburban trains. So what limits the potential of ERTMS in this thought-experiment is a commercial desire to have trains that suit the passengers they carry, just the sort of trap in the realities of preparing a timetable that is overlooked by glib talk of "trains per hour".

6 How to refine this analysis

First and foremost, some decisions need to be reached in respect of safety standards. Will elimination of signal overlaps be permitted given the Automatic Train Protection functionality of ERTMS? Will we feel able to place block boundaries in tunnels or on viaducts with the risk of trains being stopped in such places? Decisions in this respect have not yet been made.

Given these decisions, however, it is quite a simple application of a simulation package to derive new Rules of the Plan – line headways, junction margins, and (crucially) platform reoccupation times.

Then comes the essentially human task of timetable planning. It is all very well identifying a bit of capacity here and a bit there – but to put a train in a timetable, these bits have to link up into a conflict-free path, with platform slots at origin and destination, and a return path out of the terminus back to the origin or a stabling point. And unless such a path can actually be incorporated into a timetable, "capacity" cannot be said to exist.

The key role for simulation returns of course in analysis of the robustness of the resulting timetable. Additional trains will be operating over sections of the network that are fundamentally unchanged, eroding spare capacity. Even where ERTMS can influence the capacity, the fact that additional trains are running will exacerbate the effect of line blockages. Appropriate functionality will also capture the impact of system response times and probabilities of communication breakdown - "dropped calls".

7 Conclusion

In the complex and intensively-worked area that was the subject of this "thought experiment", an increase in operational terms of 10% in the number of trains run given ERTMS Level 2 is found to be plausible, albeit it at the upper extreme of plausibility.

However, this conclusion depends upon intensification of terminal workings to accommodate additional trains, requiring standardisation of the rolling stock fleet between inner and outer suburban trains, which may not be commercially

acceptable. Completion of planned infrastructure changes at London Bridge also needs to be assumed.

The conclusion also requires adjustment of standards, principally the effective elimination of signal overlaps, which has not yet been accepted.

Robustness of the intensified service is appropriately tested by simulation, once a timetable has been prepared.

References

[1] Invensys, Transport Capacity Research Paper, Credo, November 2007
[2] Railway Safety and Standards Board, Towards a Better, Safer Railway.
[3] Nock, O. S., Railway Signalling, A & C Black, 1980
[4] Department for Transport, Rail Technical Strategy, The Stationery Office, July 2007