Rail and road capacity are not such different things

Capacity of rail systems is a key factor in many current debates. The case for high speed lines largely depends on the likelihood that the current network will run out of capacity, whilst investment in new signalling technology claims to postpone that day. Likely limitations on investment capital mean we must be very clear what "capacity" we already have and may get for our money. Too often, assessments of capacity based purely on technical factors are misleadingly optimistic. Maximising utilisation of theoretical rail capacity will often require some compromise against objectives for punctuality of operation and minimum journey times.

Technical Line Capacity is simply the frequency with which trains travelling at full speed can follow each other without seeing caution signals. The distance separating trains reflects the maximum permitted speed for which the signalling was designed, known as the Line Speed, as signals must be spaced to provide braking distance between a caution and the subsequent danger signal. The time-separation is the "inter-green" period for each signal, and is the time taken for one train to pass through the separation distance so that the next can be given a clear aspect. To confuse things, by "Headway" the rail operator means this minimum time-separation between trains, rather than the actual timetabled interval.

For this ideal situation, Speed : Flow curves for road and rail are very similar, except that the rail operator plots headway (rather than frequency) against speed, and so the curve is inverted. The key difference is that, whilst road drivers adjust their separation as speed varies, railway signals are permanent features, so that separation distance remains the same whatever the actual speed of trains. To destroy line capacity, design the signalling for one speed and then run slower - trains will receive caution aspects at the same distance separation as for full speed, but passing slowly through that distance does the time-separation no good at all.

On UK Multiple Aspect Signalling, the optimum Line Speed for maximum flow is 45 mph and allows headways of less than 1 minute - so long as signals can be located ideally, which, to give drivers a chance of seeing them despite bridges and tunnels, is not very often. For stopping trains, Headway is also increased by station dwell times, and time lost in accelerating and braking. However, Line Capacity is fundamental to behaviour of a rail system only in the way that oxygen is to human behaviour – given enough of it, it has no influence on what we then choose to do. True capacity depends less on technical features of the assets as on how those assets are used.

Some trains have to stop at stations whilst others don't. When the schedule or "path" for a train is represented as a line on a Distance : Time graph, a stopping path will be seen to fall steadily behind a preceding through path. Without opportunities for following through trains to overtake, a wedge of unusable capacity builds up. The timetable aims to allow all trains to run on clear aspects – in effect, "green waves" for all. - which is not always possible, so "pathing time" added to schedules trades journey time for capacity.

To maximise use of capacity, run trains of similar speeds consecutively, in "flights". This may not match commercial aspirations, so clever planning will aim to use that wedge of dead capacity in other ways – perhaps fitting in some short workings to intermediate termini, or sneaking another train across the route at a junction.

Junctions raise issues that are directly comparable with roads, in that using one route may block others. Paths that occupy the same time and space are said to "conflict" (signals will prevent anything worse than delay in practice). The trick is to exploit "parallel moves", that is, to path trains on independent routes to pass simultaneously through a junction or station throat. For conflicting paths, the signalling sets minimum time-separations. With a bit added to allow for slight deviations from the planned timings, headways and junction margins are the essence of the "Rules of the Plan". Given typical values of 2 or 3 minutes, unlike roads, the train delaying you may not even be in sight.

And routes lead to termini, and can't feed trains in any faster than they can be started back again. On simple systems, very short turnrounds can be achieved by "stepping back" of drivers; on a diverse suburban network it is a good idea for whoever drives a train in also to take it away, minimising the risk of being stuck in the station for a crew change that hasn't happened, but this means that the time taken to change ends becomes critical.

Finally, the more trains that run, and the more complex the junction layouts, the greater is the risk to punctuality through knock-on delays should something go wrong.
The Institution of Railway Operators defines capacity as:

"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 within agreed performance targets in the face of prevailing levels of Primary Delay".

The many useful parallels with highways, albeit given a number of vital differences, should make the subject accessible to transport planners.