MICHAEL MAHUT
Vice President, Traffic Simulation
INRO

michaelm@inrosoftware.com

A NEW ADAPTIVE, MULTI-SCALE SIMULATION

We propose a different approach to scalable dynamic traffic modelling which continuously adapts to the level of traffic congestion in order to prevent non-linear responses such as cascading queues and gridlock. In this way, model stability is ensured even when there are significant demand/supply imbalances in certain areas of the network. The new approach described here is referred to as “multi-scale” because it uses a single model with a consistent level of detail throughout the network, and an adaptive simulation approach to provide enhanced scalability for larger geographies, higher demands, and more congested conditions. The multi-scale traffic simulation was implemented in the Dynameq traffic simulation software and evaluated on several congested real-world networks, including a recent example from the Sydney area. Tests demonstrate that the multi-scale approach is effective in improving model stability in congested scenarios and in reducing run time to convergence for congested dynamic traffic assignments. Compared with hybrid approaches, the consistent level of detail throughout the network provides transparency even across wide areas.

Acknowledgement: The use of one test scenario from the Sydney area was specifically permitted by Jacobs Group (Australia). This scenario does not correspond directly to any project deliverable for which Jacobs Group was commissioned.

Introduction

Larger-scale dynamic network models based on traffic flow principles are becoming increasingly popular, from wide-area micro-simulation to sub-regional or even regional dynamic traffic assignment (DTA), which may also be a coupled with an activity-based demand model in an integrated travel demand model framework. Although the specifics of the traffic flow models may vary, from highly detailed lane-level micro-simulation to hydrodynamic or other mesoscopic models, the promise of these approaches lies largely in their ability to provide greater realism over traditional network models for investigating complex, time-dependent mechanisms, such as congestion-based high-occupancy toll (HOT) lanes, dynamic parking pricing and departure-time choice modelling.

A fundamental issue that is commonly encountered, particularly in larger-scale applications, is the response of the network model to a scenario in which the travel demand is not sufficiently commensurate with the network supply (flow capacities). A key feature shared by the types of network models mentioned above is that they strictly respect flow capacities (i.e., of links and turning movements) as well as storage (density) capacities, and more generally that they respect traffic flow behaviour as represented by the well know speed-flow-density relationship. In these models, traffic congestion is represented as physical queues which spill back upstream against the direction of traffic, and as they do they reduce the capacities of upstream links and turning movements. When this occurs, the capacity reductions upstream of a bottleneck lead to even more queueing, and thus even more spillback, which results in even more queueing, and so on… In this situation, the actual or effective capacities (throughputs) are lower than the exogenously specified capacities, sometimes drastically lower – the actual values depending on the demand itself and the degree to which it exceeds the supply. In severe cases, these cascading queues can result in severe congestion over large swaths of the network, and even total gridlock.

These severely congested conditions can have a number of negative impacts. If the assignment model fails to converge, then the results are unusable. If the model appears to converge, the results may be usable in principle but may be considered unreliable: due the nonlinear behaviour of the model, small changes to model inputs can lead to large changes in model outputs, and thus a model in this state may be overly sensitive to errors in the model inputs. In the case of an integrated travel demand model framework, even if the assignment converges, it may not exhibit much sensitivity to different levels of demand (flat line response) and thus may prevent the demand model from converging. Similarly, a matrix adjustment method used in conjunction with the network model may not function properly or converge to a solution.

From a practical standpoint, even if it is suspected that the excessive congestion is due to input errors, the fact that it is spread out over a large area makes it difficult if not impossible to identify the initial bottlenecks that triggered the large scale breakdown. It should also be noted that although the discussion here is in concerned with equilibrium dynamic traffic assignment models which use some form of traffic simulation as the network model, the problem of extreme congestion, and related questions as to the usefulness or validity of the model results, is equally relevant to one pass (non-equilibrium) applications of traffic simulation.

Proposed mechanism to improve model stability and convergence

Due to the inherent trade-off between stability and realism for these types of models, the only way to address the instability is to allow the models to be less realistic in some way. However, this must be done in a well structured way that (1) aims to maintain the overall value of the model outputs (maximize the realism that can be achieved), and (2) provides a meaningful quantitative measure of the degree to which the underlying traffic flow properties have been “relaxed” or “stretched” in order to generate a more stable solution. This paper proposes such a mechanism, describes the associated measures, and provides numerical results obtained with a detailed traffic simulation model on a relatively large real-world network.

Some attempts have been made in the past, and which are in current use today, to address this issue via built-in rules, such as removing vehicles entirely from the simulation or allowing vehicles to change paths simply in order to reduce the congestion building up behind them (in some cases this may result in the vehicle being unable to reach its intended destination). These ad-hoc approaches are unsatisfying because the outputs of the model are no longer based on the entirety of the inputs (the model violates conservation of flow) but also because the rules are arbitrary, are not parameterized, and tend to be model specific rather than based on the underlying theory that is common across all traffic flow models. As the potential applications of simulation-based network models continue to grow, along with the accessibility of the computational power that makes them feasible to use, there is a need to address fundamental methodological issues such as this in a more satisfying way.

The approach taken here is based on several observations about the behaviour of capacity-constrained traffic models (and real traffic, for that matter), and how these models differ from traditional static assignment models in which traffic speeds are calculated using volume-delay functions.

A few critical differences can be observed, which are due to the congestion spill-back effect:

(1) Volumes (or throughputs) on individual links and turns can drop well below their exogenous theoretical values.

(2) As this occurs, delay from a specific bottleneck may be incurred by vehicles that are not in fact destined for that bottleneck (and thus do not influence the demand/capacity ratio of the bottleneck).

As a consequence, travel time, and hence generalized cost, of any given path can be influenced by bottlenecks that are not on that path – which is distinct from static assignment models in which path travel time is only a function of volumes and capacities of the links (and turns) directly on the path. We propose to conceptually distinguish between these two types of delays, which we will refer to as primary and secondary delays:

(1) Primary delay: refers to delay incurred by a vehicle which has propagated from a downstream bottleneck that lies on the path that the vehicle will follow to its destination.
(2) Secondary delay: refers to additional delay incurred by a vehicle, beyond the primary delay, which has propagated from a bottleneck that is not on the path that the vehicle will follow to its destination.

In the following we propose a mechanism which reduces secondary delay through a partial relaxation of the physical constraints of the simulation model in order to mitigate the rapid, nonlinear growth of cascading queues, as a way to improve the stability and convergence of simulation-based network assignment models.

In order to only reduce secondary delay, and not primary delay, the relaxation is applied only where delay is propagated to vehicles not destined for the bottleneck from which the delay has originated, and thus, where the travel time of a path is being impacted by bottlenecks that do not lie on the path itself.

Limiting delay propagation is equivalent to a partial relaxation of the lane FIFO (first-in-first-out) property: it implies that the follower may overtake the leader at some point, since physically this is the only way that the follower can stop incurring any further delay.

In concrete, physical terms, partial lane-FIFO relaxation implies that at the end of a link, under sufficiently congested conditions, new channels of traffic flow, or virtual lanes, naturally emerge in order to permit a certain amount of overtaking that would not otherwise be physically possible.

This has two effects:

(1) Throughputs, or effective capacities, of links and turns that are impacted by congestion spill back will increase: in particular where the flow reduction was due to choke effects.
(2) Delay that is experienced in the virtual queue (emergent lane) does not propagate upstream: this delay is essentially *checked out* of the delay propagation mechanism.

The relaxation mechanism can thus be thought of as a type of driver adaptation to extreme congestion. It should be noted that in some countries, drivers are known to create more traveling lanes than are actually painted on the roadway – an effect which is not well captured by traffic simulation models if it only occurs as a response to increasing congestion. The proposed partial lane-FIFO relaxation mechanism is a way of explicitly representing this type of behaviour in a traffic simulation model.

Numerical results

The partial lane-FIFO relaxation (PLFR) mechanism was implemented in a simulation-based dynamic traffic assignment model which simulates the movement of individual vehicles on discrete lanes of the roadway using vehicle-interaction models such as car-following, lane changing and gap-acceptance (Dynameq software developed by INRO) and tested on a number of congested real-world networks using a range of congestion thresholds to trigger the relaxation mechanism. A number of model outputs were evaluated and compared:

(1) Virtual queue length measured in number of vehicles, and total delay in virtual queues, by turning movement and aggregated to links and nodes (for display purposes).

(2) Total vehicle hours of travel (VHT) and delay (VHD), as well as relative gaps to evaluate model convergence and stability.

The results discussed below are for equilibrium DTA runs using a fixed congestion threshold throughout each run. Thus, the results reflect different assignments (path choices) as well as different simulation models, but for the same demand and network.

Secondary delay

Figure 1 shows total network VHT (vehicle hours travelled), VHD (vehicle hours of delay) and VHQ (vehicle hours in virtual queues) for a single scenario for a range of relaxation levels defined by the congestion threshold parameter. The scenario has a high degree of congestion overall but the model is stable and the DTA shows acceptable convergence properties with a stable relative gap of 4% (average over all departure time intervals). The set of bars on the left represents the no relaxation case. From left to right, the bars represent an increasingly higher degree of relaxation, and thus lower congestion threshold. In all, 4 congestion thresholds were tested and will be referred to as indicated in the plot below; i.e. a higher number indicates a lower threshold and thus a higher degree of relaxation.

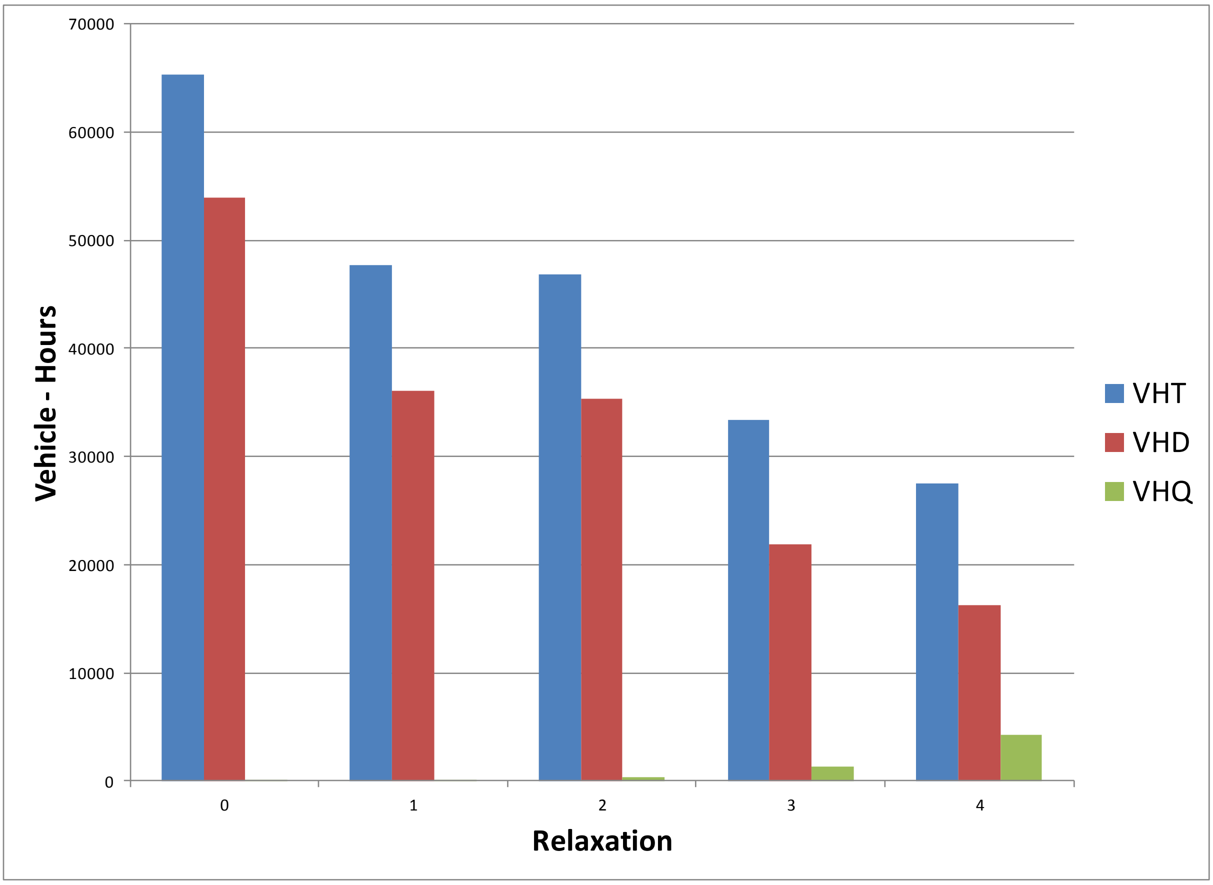

Figure 1 VHT, VHD and VHQ with increasing FIFO relaxation

This plot reveals some interesting facts about the different types of delay that occur in this network. First of all, even though each model run results in a different network assignment, there is a very consistent gap between VHT and VHD, demonstrating that total free-flow time is about the same in the different runs. Secondly, the total drop in VHT and VHD with increasing VHQ is quite remarkable: at the first level of relaxation, a very small amount of VHQ results in a significant drop in VHD and VHT. As VHQ increases with increasing relaxation, the relative drop in VHD and VHQ decreases, showing a distinctly nonlinear response of VHD and VHT to VHQ. Thirdly, the total reduction in VHD and VHT, between the no relaxation case and the maximum relaxation shown, implies that at least 60% of the total delay in the no-relaxation model run can be attributed to secondary delay.

Convergence

A second scenario was tested to investigate the relationship between relaxation and DTA convergence. This network is from the Sydney area but was tested with a somewhat hypothetical demand in order to create sufficiently severe congestion to result in a non-convergent DTA model run.

The DTA run with the second congestion threshold level, referred to as PLFR-2, converged in about 60 iterations to an average relative gap (across all departure time intervals) of 1.5%, while the DTA run with the third congestion threshold level, referred to as PLFR-3, converged in 30 iterations to an average relative gap of 1.4%. Figure 2 shows the relative gap plots, by departure time interval, for these two model runs along with the no-relaxation case. The figure shows a clear trend towards increased stability, as seen by the smoothness of the relative gap plots, as well as faster convergence (fewer iterations), with increasing relaxation.

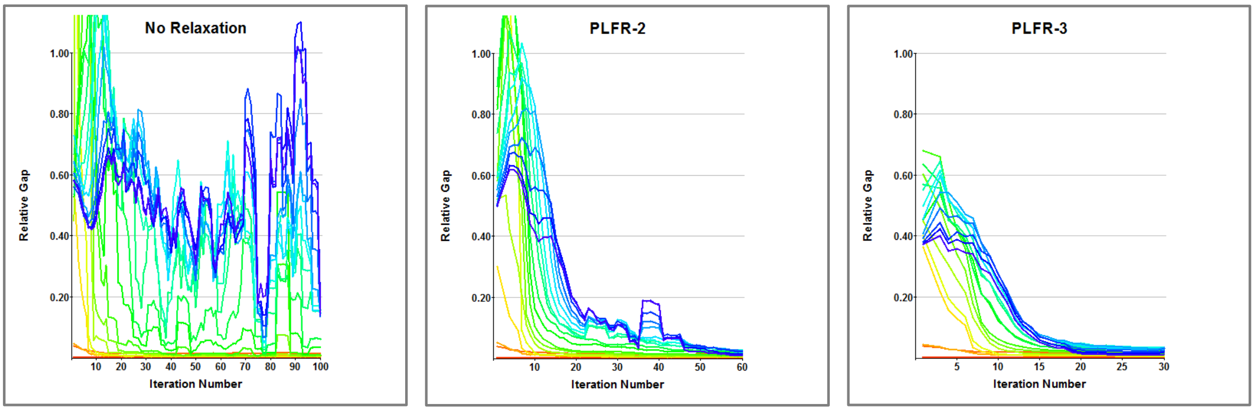

Figure 2 Relative gaps by departure time interval: No relaxation, PLFR-2 and PLFR-3.

Figure 3 shows bar widths scaled to link flows and colored by speed, and node diameter scaled to virtual queue length summed over the turning movements at each node, during the most congested 30-minute period. The virtual queues, even when summed at each node, were not nearly as prevalent as might have been expected. The largest node circle represents an average of 25 vehicles in virtual queues summed over all approaches during the most congested 30-minute period. In all, during this same 30-minute period, there were 188 nodes in the network with positive virtual queues, but only 18 of them with an average node-aggregated value of 5 or more vehicles. The PLFR-2 model, due to the higher congestion threshold, had considerably less virtual queueing than PLFR-3. During the most congested 30-minute period, there were 145 nodes with positive virtual queues, but only 4 of them had average values of 0.5 or more, with a maximum (time-averaged) value of just under 3.

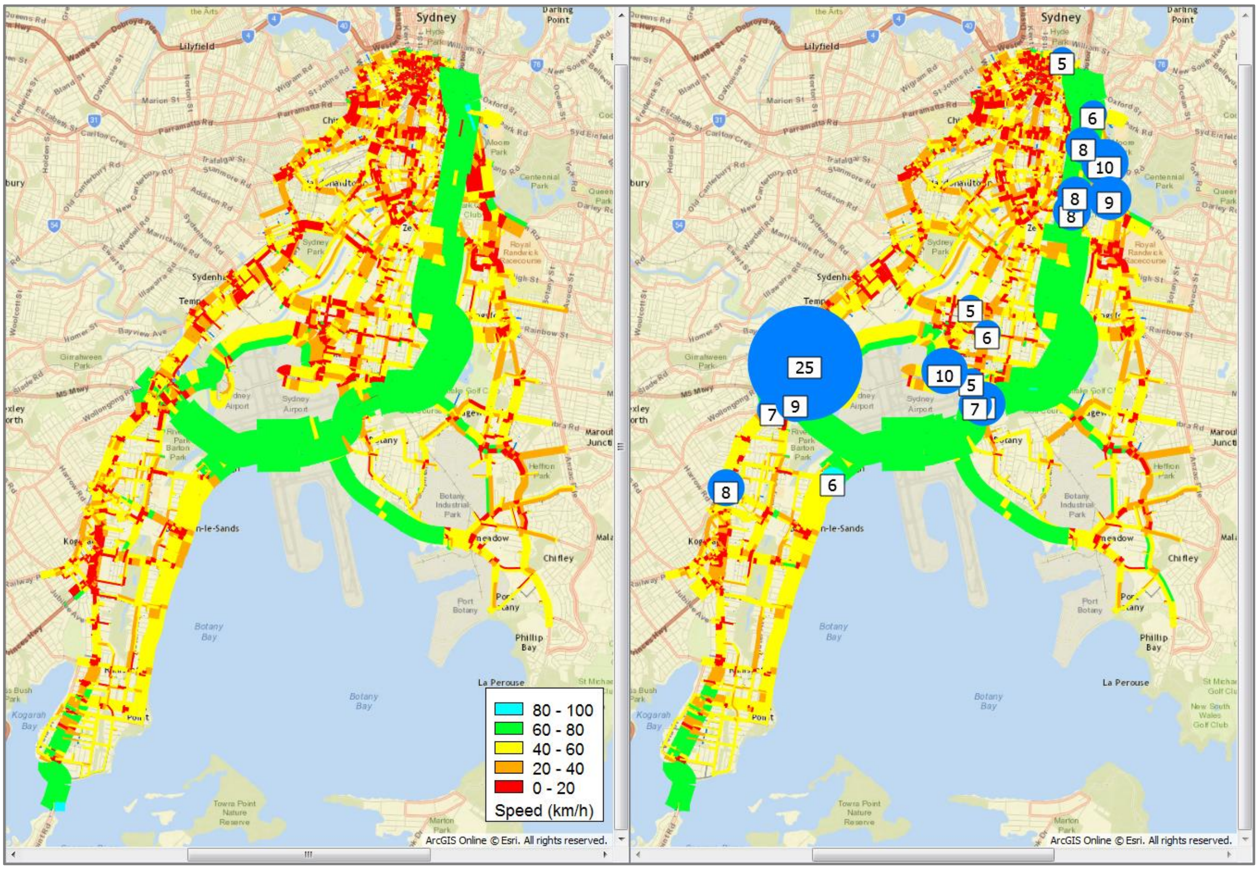

Figure 3 **Bar widths scaled to link flows and colored by speed and node diameters scaled to virtual queue length summed over all turning movements: No relaxation (left) vs. PLFR-3 (right)**

Although these results are drawn from a single network and demand scenario, the relationship between the number of vehicles in virtual queues and the resulting impact on model convergence is striking. In particular, considering the PLFR-2 results, a very low number of vehicles in the virtual queues had a major impact on model convergence.

Conclusions

The partial lane-FIFO relaxation (PLFR) mechanism provides a very effective way to address the issue of model instability due to severe congestion in simulation-based dynamic traffic assignment models. The idea behind the mechanism is to reduce the nonlinear growth of queues in severely congested conditions through a partial relaxation of lane-FIFO discipline, while respecting bottleneck capacities and turn FIFO discipline. The partial relaxation of lane-FIFO reduces the amount of secondary delay that is incurred by vehicles, which is defined as delay experienced by a vehicle that has propagated upstream from a bottleneck which that does not lie on the vehicle’s path.

The results indicated strong relationships between the amount of virtual queueing and the resulting impact on large-scale measures such as network wide vehicle-hours of delay and model convergence. In an unstable model, virtual queues of only a few vehicles (aggregated over all turning movements at each node) were enough to allow the model to achieve a convergent solution. Overall, the mechanism was found to be very effective at increasing model stability and convergence.

In a physical sense, partially relaxing lane-FIFO implies that vehicles naturally start to form new virtual lanes at the exit of a link as congestion increases. The mechanism thus reflects a form of adaptation of driver behaviour to extreme congestion. To a certain degree, this type of behaviour may in fact exist in real traffic, though the question of realism is not the primary motivation here. In the current context, the mechanism is motivated by the need to allow the simulation model to behave in a slightly unrealistic way in those situations where the model inputs are unrealistic and would otherwise result in severe congestion and effectively unusable outputs.

For cases where these conditions are due to network coding errors, the more fluid traffic flows combined with outputs quantifying the virtual queues yield far more usable results by drawing attention to the critical bottlenecks that otherwise would have resulted in widespread, severe congestion or even gridlock. For situations where the demand and supply are necessarily unbalanced, a mechanism of this kind can play a critical role in helping the model achieve convergence, without necessarily being invoked in the converged solution. Examples of this include the early iterations of an assignment run or in an integrated demand-DTA model system where the DTA is required to provide meaningful O-D impedances in response to unrealistically high levels of demand.

References

Dynamic Traffic Assignment, a Primer. Transportation Research Circular E-C153. Transportation Research Board, June, 2011. http://onlinepubs.trb.org/onlinepubs/circulars/ec153.pdf

Guidebook on the Utilization of Dynamic Traffic Assignment in Modeling, Traffic Analysis Toolbox vol. 14. Federal Highway Administration, November 2012. http://ops.fhwa.dot.gov/trafficanalysistools/

Dynameq Users Manual, Release 3.1. INRO, 2016.