What TransCAD Users Should Know about New Static Traffic Assignment Methods

Recent research by Caliper and others has led to a variety of new and improved methods for calculating user equilibrium traffic assignments. New options provide much faster computing and also the achievement of much tighter convergence, resulting in more accurate impact assessments and select link analysis.

Even before any of the improvements discussed below, TransCAD was significantly faster in computing user equilibrium than any other commercial software. We believe that the further improvements have widened TransCAD’s lead in computing speed considerably.

Greater speed is needed because of the desire to do larger and more complex problems, and greater accuracy is needed because, without it, there can be large errors in estimates of the impacts of plans and projects. The discussion below purposely limits technical detail so as to be accessible to a wide audience of users and managers, but there are numerous references that provide a full discussion of prior research and the topics discussed. Also, we at Caliper would be happy to discuss any questions that you might have about these matters.

Multi-threading of the MMA Assignment

Most large regional models use the MMA (multi-mode, multi-class assignment) routine in TransCAD. This model accommodates HOV lanes, toll roads including those with entrance-to-exit tolls, multiple user classes and class prohibitions, trucks of different sizes, and varying values of time, and is appropriate for problems that need one or more of these features.

For the conventional UE assignment, based upon a well-tuned Caliper implementation of the Frank-Wolfe (FW) algorithm, multi-threading in TransCAD results in a nearly proportional speedup in computation that is a function of the number of physical cores in a computer’s CPU. Figure 1 illustrates the speed of convergence for a rather large and congested network assignment problem for the metropolitan Washington, D.C. region with 2500 zones, over 36,000 links, and 5 user classes. We use this particular traffic assignment problem as a test case because it is from a real, deployed model and matches ground counts closely. It is much more challenging from a computational point of view than all the test networks used in the research literature. The presence of 5 user classes results in much longer running times than would be experienced for a single class assignment.

We use the relative gap (RG) measure of convergence, which is a common and reasonable figure of merit, for the presentation of assignment convergence results in the discussion here. Most models in the U.S. have traditionally used a relative gap of .01, but that is insufficient for impact assessment, and as this has become more widely known, increasingly tighter assignments are being computed. Nevertheless, gaps below .001 are rarely encountered for large models due to concerns about computing time.
The computer used to generate these results has two quad-core Xeon CPUs that run at 2.93 GHz. It was purchased more than a year ago and does not have the hyper-threading featured on newer chips from Intel. At that time, it was one of the faster options available. For some of the tests, we disabled one of the CPUs and/or some of the cores.

As one can see from Figure 1 below, there is a significant improvement due to multi-threading, yielding a nearly proportional speedup in the FW MMA assignment as a function of the number of CPU cores. Doubling the number of cores generally halves the amount of time it takes to reach a given level of convergence.

With the new 6 and 12 core processors coming on the market, even further gains will be easily achievable, and there is always the possibility of using computers with 2 to 4 or more CPUs. Also, newer chips from Intel have more effective hyper-threading, which adds additional improvement, although not as much as additional CPUs or cores.

About five years ago, we implemented distributed processing for the MMA assignment. We have since dropped it, as it is much less efficient than multi-threading due to data communication overhead.

Figure 1: Convergence graph for the MMA FW assignment with 2, 4, and 8 computer cores
Irrespective of the number of cores or the speed of the assignment, you can also see that the FW assignment method’s rate of convergence tails at a certain point, an observation well-known to both theoreticians and practitioners. This means that this method will not be able to achieve orders of magnitude lower convergence. You may find it interesting to know that it takes the FW method 650 and 4145 iterations to reach relative gaps of .0001 and .00001, respectively.

**Conjugate Descent Options Added to MMA**

Proposed by Daneva and Lindberg (2003), the bi-conjugate descent FW (BFW) method uses a little more memory than the conventional FW assignment, but not so much that it would typically be a concern today. FW holds two link flow vectors in memory whereas conjugate descent methods keep 3 or more link flow vectors in memory, which are used in choosing a more effective search direction than FW. Conjugate descent methods are easily multi-threaded, so there is no tradeoff in using them. These options were added in Release 3 of TransCAD 5. Figure 2 shows the running times for the test network with the bi-conjugate FW method with 1, 2, 4 and 8 computer cores.

![Figure 2: Convergence graph for the Bi-conjugate traffic assignment method with 1, 2, 4 and 8 computer cores](image-url)
The results indicate a very significant improvement in efficiency using the bi-conjugate Frank-Wolfe method. Two things can be noted. First, the bi-conjugate method, as implemented in TransCAD, cuts the running time by a factor of 2 or more at low convergence and by a factor of 6 or more in terms of the time taken to reach a gap of $10^{-5}$. It makes computing to that level realistic for this assignment problem. It can also be observed that multi-threading is very effective for this method, too, so further improvements in running time can be obtained by using more powerful computers.

If you are using TransCAD interactively, you can use the new bi-conjugate traffic assignment option by choosing it from the traffic assignment method pull-down list in the Traffic Assignment dialog box and setting the N-conjugate value to 2 as shown below.

If you want to run the new bi-conjugate assignment method in batch mode, change the traffic assignment macro code from

```
Opts.Global.[Load Method] = "UE"
```

to

```
Opts.Global.[Load Method] = "NCFW"
```

```
Opts.Global.[N Conjugate] = 2
```

In our experiments, we found the bi-conjugate option just as effective as the k-conjugate method for k>2. However, because each network assignment problem can be different, some users may wish to experiment with these options.
A New Algorithm—OUE Assignment

Another way to solve an MMA UE problem is by using the OUE assignment option that was introduced in TransCAD 5 and has been evolved to handle all of the MMA procedure features and options. This method is based upon an algorithm developed by Robert Dial who worked with Caliper on its implementation. Dial’s Algorithm B does not tail like a conventional assignment and, as a result, it can achieve unprecedented levels of convergence and do so quickly. The OUE assignment generates a solution for each origin’s link flows and therefore requires more memory than FW and BFW methods. We chose to feature this method in TransCAD after evaluating and testing various other new UE assignment methods for which favorable claims were made in the literature. Specifically, we believe that it is significantly faster than route-based, projected gradient methods and other origin or bush-based methods and has other advantages as well. Importantly, it is not only faster, but unlike the FW or conjugate descent method, it drops inappropriate routes as it generates successive iterative solutions, yielding a cleaner assignment solution. In Figure 3, we show how OUE based upon algorithm B performs on the Washington Regional Network. You can observe that OUE benefits from multiple cores, but not nearly so much as the other assignment methods. The reason is that OUE is not fully multi-threaded at the present time.

![Figure 3: OUE Cold Start Convergence Rates PM Assignment](image-url)
Figure 4 compares OUE with the FW and BFW methods. It illustrates the superior rate of convergence of OUE; however, it is not faster when started from scratch than BFW because of the multi-threading benefit afforded by the 8 cores until relative gaps are sought that are below $10^{-5}$. For lower relative gaps, BFW tails but OUE does not.

When started without a prior solution or, in other words, from a cold start, OUE reaches a relative gap of .00001 in 1.5 hours on the machine with 4 cores and is only somewhat faster on the machine with 8 cores. This is acceptable for practitioners in comparison to the 12.5 hours that FW takes to get to the same level of convergence on the 4 core computer or the 6 hours for the 8 core computer. When compared to the BFW method, OUE is faster on this problem to $10^{-5}$ on a quad core computer, takes a similar amount of time on an 8 core computer, and can achieve orders of magnitude tighter convergence. In addition, OUE has further advantages to which we now turn.
Immediate Availability of Select Link Analysis as a Post Process

When the results are saved after the computation of an OUE assignment, select link analysis can be performed for any query. This means you do not have to specify the query before running the assignment, and you can save a great deal of time by not having to redo traffic assignments when another query is required.

Warm Start Speedups with OUE

A particularly important aspect of OUE (and Algorithm B) is that it can re-compute a new equilibrium from a prior solution, even if the trip table or network attributes have changed somewhat. Prior solutions are nearly always available because traffic assignments are computed over and over again in the course of model development. Of course, this requires that a previous solution has been saved, and there is some computational overhead associated with writing and reading a prior solution.

To illustrate this warm start advantage, we computed the test assignment to a relative gap of $10^{-5}$ with a quad-core computer and then investigated how long it takes to compute a new solution to the same level of convergence for a similar, but not identical trip table. As shown below in Table 1, when we randomly perturbed the trip tables for the Washington regional model by $+ \text{ or } -5\%$, we were able to calculate a new equilibrium to a gap of $0.00001$ in less than 9 minutes compared to the 1.5 hours it takes from a cold start. With a $10\%$ random perturbation, the time grows but only to about 11 minutes.

<table>
<thead>
<tr>
<th></th>
<th>Time to converge</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cold Start</td>
<td>01:28:02</td>
</tr>
<tr>
<td>+/-5% perturbation run 1</td>
<td>00:08:51</td>
</tr>
<tr>
<td>+/-5% perturbation run 2</td>
<td>00:08:53</td>
</tr>
<tr>
<td>+/-5% perturbation run 3</td>
<td>00:08:45</td>
</tr>
<tr>
<td>+/-10% perturbation run 1</td>
<td>00:11:10</td>
</tr>
<tr>
<td>+/-10% perturbation run 2</td>
<td>00:11:18</td>
</tr>
<tr>
<td>+/-10% perturbation run 3</td>
<td>00:10:00</td>
</tr>
</tbody>
</table>

Table 1: OUE Warm Start Convergence to $10^{-5}$ with Perturbed Trip Tables

From these results, it is clear that there is a massive efficiency gain from using the warm start which, for this model, yields roughly an order of magnitude improvement in computing time. With the warm start, the calculation of a tight equilibrium takes essentially no more time than calculating FW to a 1% relative gap.

In a model with feedback loops, trip tables change at each iteration while the network remains fixed. The warm start is particularly effective in this situation. In Table 2, we show the calculation times for a model run with 4 loops and traffic assignment convergence to a relative gap of $0.0001$ and to a travel time skim RMSE of $0.1\%$.
As is evident, the time needed to compute the traffic assignment steps declines significantly after the first loop and declines further although less so with successive loops. Loop 4 takes only one-fifth of the time taken for the first loop. Of course, one could use a warm start for the first loop using a prior solution, which would further reduce the overall running time necessary to achieve feedback convergence.

The warm start can also be used when the network changes somewhat. Generally speaking, if the prior solution is feasible for the modified network, the warm start will be effective. If the prior solution is not feasible, it may not work or, if it does, it may be much less efficient.

Inexplicable Link Flow Changes Need Not Be a Feature of your Model

Research that we have performed has confirmed other findings that poor convergence leads to large errors in traffic assignments and, thus, in all aspects of planning models and impact assessments. (Slavin et al, 2009, Boyce et al, 2004). In particular, poor convergence accounts for the implausible changes in link flows that are often observed in traffic assignments at locations that are remote from the specific projects being evaluated.

As an example, we edited the test network southeast of D.C. to reflect the opening of a new flyover ramp from the Capital Beltway to southbound Route 5 in Prince George’s County. Route 5 is heavily traveled and is used by more than 122,000 vehicles per day. The new ramp was intended to reduce congestion and ease a difficult merge at the interchange.

In Figure 5, we show the differences between the base case flows and the flows with the new interchange for different relative gaps. At a 1% RG, there is significant convergence error, and the convergence error dominates the project forecast. The green links represent those that had a gain in flow and red links are those that lost flow of more than 200 vehicles. While the flows on the new links are similar, there are spurious effects far away from the project.
Tighter convergence cleans up and localizes the projected impact. The spurious link changes are mostly gone at a RG of .0001, but you can also see that the solution is different and cleaner at a RG of .00001.

**Importance of Convergence for Impact Analysis**

The calculated VHT impacts show that poor convergence can lead to erroneous conclusions about project benefits. At a gap of 0.01, the model predicts that the VHT will actually increase by 408 vehicle hours after the modification of the interchange. At a gap of 0.001, the model still indicates a VHT increase of 94 vehicle hours. At a smaller gap of 0.0001, the model predicts a VHT saving of 20 vehicle hours. At an even tighter RG, the VHT savings are negligible, which may be due to replacing older ramps with longer ones.
Table 3 indicates the number of links in the scenario that have an absolute flow change of more than 200 vehicles from the base case, VHT for the base case, VHT for the scenario, and the VHT differences.

<table>
<thead>
<tr>
<th>Gap</th>
<th>Number of Links with Abs Flow Diff &gt; 200</th>
<th>VHT Base Case</th>
<th>VHT Scenario</th>
<th>VHT Saving (Veh-Hrs)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1e-2</td>
<td>162</td>
<td>1,105,726</td>
<td>1,106,134</td>
<td>-408</td>
</tr>
<tr>
<td>1e-3</td>
<td>56</td>
<td>1,091,153</td>
<td>1,091,247</td>
<td>-94</td>
</tr>
<tr>
<td>1e-4</td>
<td>44</td>
<td>1,090,136</td>
<td>1,090,116</td>
<td>+20</td>
</tr>
<tr>
<td>1e-5</td>
<td>45</td>
<td>1,090,090</td>
<td>1,090,088</td>
<td>+2</td>
</tr>
<tr>
<td>1e-6</td>
<td>47</td>
<td>1,090,084</td>
<td>1,090,079</td>
<td>+5</td>
</tr>
<tr>
<td>1e-7</td>
<td>45</td>
<td>1,090,087</td>
<td>1,090,085</td>
<td>+2</td>
</tr>
<tr>
<td>1e-8</td>
<td>45</td>
<td>1,090,089</td>
<td>1,090,086</td>
<td>+3</td>
</tr>
</tbody>
</table>

Table 3: Comparison of Base and Scenario – Interchange Project

Not only are estimates of project impacts affected by convergence levels, but so are overall VHT estimates. This means that it is important to run both base cases and forecasts to the same level of convergence, so as to avoid biased calculations.

While more research will be needed to establish general guidelines, evidence accumulated to date by us suggests that a relative gap of .0001 or better should be used for accurate results. This can be tested for specific models rather easily using TransCAD by computing OUE to high convergence and then assessing the link flow errors associated with lesser convergence.

The tests that we have conducted illustrate clearly that estimates of project benefits vary substantially with the convergence level utilized in the traffic assignment. The link flow errors are not randomly distributed, but will typically reflect a particular geographic bias associated with forecast use of some routes instead of others.

The bottom line is that achieving good convergence in the traffic assignment is essential for obtaining valid estimates of project impacts.

Why does Poor Convergence Plague Models?

One reason that poor convergence plagues most deployed models is that the disturbing and hard-to-explain spurious impacts are not very noticeable when evaluating an entire future plan scenario. It is much easier to identify problems when one can check for a logical connection between a specific project and its forecast impacts. But that is not the only reason why poor convergence has often been masked.

In particular, a variety of common mistakes have led to poorly converged models. Some of these problems, in other software, arose from use of a poor convergence measure, inaccurate calculations, and poorly conceived assignment options. For many years some older commercial packages offered equilibrium “closure” based upon VHT differences between successive iterations and did not have a valid gap measure. This caused modelers to think that their
assignments were converged when, in fact, they were far from it. Use of integer or limited
precision link costs may also artificially inflate estimates of the convergence gap. Limiting the
V/C ratio either explicitly or internally in the executable code has the same effect as does
“dampening” of costs or flows.

Other practices are also problematic. Using a fixed number of iterations for an assignment is
still common, but it is not a good practice. Often it results in arbitrary choice of the gap
achieved, and it can easily lead to different levels of convergence for the base case and the
forecast scenario thus introducing another source of error in any forecast.

Obviously execution time considerations have had a dominant impact. This is no longer
necessary in most cases due to the various types of speedups available. Once again, testing the
appropriate level of convergence is easy to do with TransCAD and can help you avoid
convergence error in your model and its forecasts.

**Most Likely Route Flows, Proportionality, and Select Link Analysis**

The user equilibrium assignment computes a set of unique link flows if it is carried out properly
and to a small enough relative gap. The route flows associated with any particular equilibrium
assignment are not unique, despite the fact that estimates of these route flows are used when
performing critical link analysis. Similarly, the class flows on each link derived from a multi-
class assignment are not uniquely determined.

In a recent study, several university researchers examined the select link analysis results
produced by various commercial packages with several different UE methods. (Boyce et al.,
2010). This research, while not necessarily definitive due to limitations of several types, led to
some interesting results. Perhaps the most important of these results was that unadjusted
route-based and origin-based assignments lead to biased select link analysis, and conventional
methods may also do so, but to a lesser extent. The reason is that iterative cost updates are
made by origin for origin-based methods, and by origin or by origin-destination pair for route-
based methods that enumerate routes. This leads to order dependence of the critical link
assignment results because the specific new paths added are influenced by selection of prior
ones.

Select link analysis based upon the FW or conjugate descent methods is more democratic in
that no origin or route is given priority in terms of order of computation. Its defect is that in
congested situations, the solutions reflect some unreasonable routes that are added in early
iterations after the first set of shortest paths is computed and that are not dropped. This is
mitigated, but only somewhat, by achieving good convergence. (With these methods, higher
convergence also comes with a price in that there are more and more utilized paths with tiny
fractional flows-- a result that is not realistic).

If there are multiple possible route flow solutions for a given set of equilibrium link flows, it
becomes natural to try to identify the most likely route utilization. The maximum entropy or
most likely set of route flows provides an attractive solution to this problem. This methodology has not been commercially available until recently in TransCAD 5, Release 3.

In his Master’s thesis, which still makes good reading today, Tom Rossi suggested the maximum entropy solution as a means of achieving a consistent method of allocating traffic to development projects (Rossi, 1987). Unfortunately his method of calculating the solution required route enumeration and use of a separate optimization package and might not be workable for the large networks used in current practice. Bar-Gera proposed the TAPAS assignment method (Bar-Gera, 2010) for achieving proportional route flows across paired alternative links in the solution for each origin, arguing that this very closely approached the maximum entropy solution for the route flows. This addresses a defect in origin-based and route-based traffic assignment with respect to select link analysis.

The idea of proportionality can be understood with a simple example, such as the one that follows. The network below shows the equilibrium flows resulting from a highly converged traffic assignment solution. This network has a demand of 1800 vehicles from origin node 1 to destination node 7 and of 1200 from origin node 3 to destination node 7. Such a case, in which different origins compete for limited infrastructure, is commonly found in real networks.

![Figure 6: Proportionality Example](image)

The answer to the question of how much flow is contributed to the total equilibrium flow on the highlighted link by each origin is that it should be in the same proportion as the flows on the two alternatives routes, regardless of the point of origin. However, due to the order dependence of origin and path-based assignment methods, the select link results will be distorted. For OUE, we recognized that we could remedy this problem, and we immediately came up with a solution that achieves proportionality and does so with rather little additional computation. Table 4 below shows the select link analysis for highlighted red link with and without this correction.
Without proportionality, the flow on the critical link exclusively comes from the OD pair 1-7 and almost no flow comes from the OD pair 3-7. Note that this extreme solution still satisfies the link flows. If proportionality is used, the flows contributed by the OD pairs 1-7 and 3-7 are 714 and 476, which is proportional to the OD flows of 1800 and 1200 from node 1 and 3 respectively.

<table>
<thead>
<tr>
<th>Origin</th>
<th>Destination</th>
<th>Flows without Proportionality Correction</th>
<th>Flows with Proportionality Correction</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>7</td>
<td>1189</td>
<td>714</td>
</tr>
<tr>
<td>3</td>
<td>7</td>
<td>1</td>
<td>476</td>
</tr>
</tbody>
</table>

Table 4: Select Link Flows With and Without Proportionality

In a forthcoming project report, Boyce et al. present findings demonstrating proportionality for select link analysis from TAPAS converged to $10^{-12}$ but compared the results with other methods converged only to $10^{-4}$, presumably because not all methods in other software could go lower. The report suggests that extreme convergence is necessary to achieve route flow proportionality, but that is not so. Unlike Bar-Gera’s TAPAS method, OUE can achieve proportionality at the convergence levels that would normally be sought in large planning applications.

At present, TransCAD is the only commercial package that has this more robust select link analysis. Nevertheless, we should point out that the proportional route flow solution is a mathematical construct, so it is not a guarantee of behavioral realism.

A Note on Speed Comparisons

Research publications and sales literature are rife with claims from researchers and vendors stating that they have the fastest assignment method. We have made a practice of testing these claims to the extent possible and evaluating most of the new methods proposed, either from examination of published running times or through our own implementations. Based upon our investigations and published data, we believe that TransCAD is significantly faster in achieving convergence than other software. This statement applies to both our FW and to OUE in comparison to the new route-based methods from INRO and Citilabs and the LUCE algorithm in PTV software.

Even if performed with the same test problems, speed comparisons are fraught with difficulty due to variations in computer hardware, varying numerical precision of the software utilized, and use of differing convergence measures. Processor performance varies with the CPU architecture, on-chip memory caches and, of course, as a function of clock speed and the number of physical cores. As mentioned before, convergence behavior is strongly influenced by the precision of the calculations, with limited precision inflating convergence speed in some instances. There are many different metrics that are used to calculate the convergence gap in general and the relative gap in particular, so comparisons that use different measures are not
likely to be valid. We have also surmised that some results ignore the time it takes to actually test for convergence which, in of itself, takes significant computation time for large networks.

We have developed some comparative data using the well-traveled Chicago example popularized by David Boyce and Hillel Bar-Gera, and we have also had others make some runs for us with other commercial software. Based upon the evidence that we have gathered to date, the TransCAD OUE is significantly faster than the LUCE method in VISSUM and the projected gradient method in EMME/3. All of our assignments are many times faster than those in Cube Voyager. We plan on doing some further verification on different test problems before we publish these results. It remains to be seen how the performance of different algorithms and software implementations vary with problem size and other dimensions of difficulty. Also, there is always room for improvement, so don’t be surprised if our assignment methods get faster in the future.

A Note on Multi-threading

You might wonder why the OUE and some other procedures in TransCAD are not yet fully multi-threaded. Multi-threading is a natural for computations that can be done in parallel, but brings another level of complexity to software development. If not done properly, multi-threading can actually slow down sequential calculations or result in a situation where the same computation yields different results on computers with differing numbers of cores. Slowdowns can occur when one execution thread has to wait for another to finish before proceeding. These so-called “race conditions” can also make the order of calculations somewhat dynamic. Changes in the order of calculation will often lead to small numerical differences in calculations, and, for iterative processes like UE traffic assignment, these small differences will be compounded into larger ones later in the calculation of results. Multi-threading can also greatly increase the amount of memory required leading to either insufficiencies or memory paging to disk, which has substantial performance penalties. So while multi-threading brings great benefits, it must be done with care.

We thought we should mention that a key component of our TransModeler traffic simulation software is multi-threaded, and that we recently did very well in a competition held by AMD to suggest the best use for their new 48 core offering. A You Tube video, for which we won third place, can be viewed at http://blogs.amd.com/work/2010/04/15/winner-announced-what-would-you-do-with-48-cores/. We are continuing to multi-thread more and more of both TransCAD and TransModeler, as large numbers of cores are clearly the wave of the future.

Deployment of OUE

OUE is deployed in several regional models for production use. Unlike some of the other competing new methods, it is feature complete and handles multiple user classes, entrance-to-exit tolls, and turn penalties, so there is no impediment to its applicability. However, for some very large problems with multiple user classes, OUE requires more than 2GB of memory. Thus a 64-bit solution is needed, and TransCAD 6, which will soon be in beta, will provide it.
References


Caliper Corporation (2008) Travel Demand Modeling with TransCAD, Version 5, Newton, MA

Daneva, M. and Lindberg, P.O. (2003) Improved Frank-Wolfe Directions through Conjugation with Applications to the Traffic Assignment Problem, Linkoping University, Dept. of Mathematics


May 3, 2010 ©2010 Caliper Corporation All Rights Reserved