

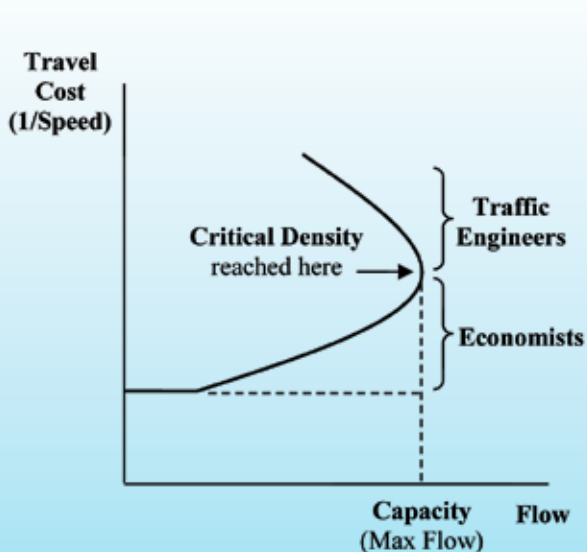


RESIDENTIAL AND  
CIVIL  
CONSTRUCTION  
ALLIANCE OF  
ONTARIO

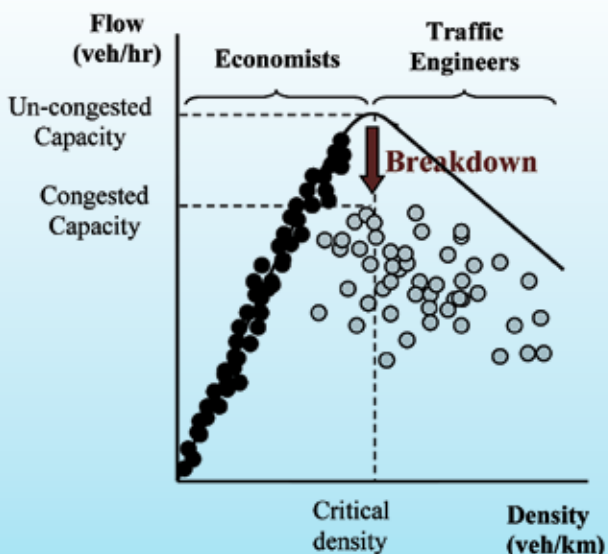
An Independent Study Commissioned by

RCCAO

Constructing Ontario's Future



a) Travel cost vs. traffic flow



b) Traffic flow vs. traffic density



## Congestion Management in the GTHA: Balancing the Inverted Pendulum

April 2013



## RCCAO

25 North Rivermede Road, Unit 13

Vaughan, Ontario L4K 5V4

Andy Manahan, executive director

e [manahan@rccao.com](mailto:manahan@rccao.com) p 905-760-7777

w [rccao.com](http://rccao.com)

The Residential and Civil Construction Alliance of Ontario (RCCAO) is composed of management and labour groups that represents a wide spectrum of the Ontario construction industry. The RCCAO's goal is to work in cooperation with governments and related stakeholders to offer realistic solutions to a variety of challenges facing the construction industry and which also have wider societal benefits. For more information on the RCCAO or to view copies of other studies and submissions, please visit the RCCAO website at [www.rccao.com](http://www.rccao.com)

RCCAO members include: Carpenters' Union • Greater Toronto Sewer and Watermain Contractors Association • Heavy Construction Association of Toronto • International Union of Operating Engineers, Local 793 • International Union of Painters and Allied Trades, District Council 46 • Joint Residential Construction Council • LIUNA Local 183 • Residential Carpentry Contractors Association • Toronto and Area Road Builders Association

---

# **Congestion Management in the GTHA: Balancing the Inverted Pendulum**

An Independent Study  
Commissioned by the  
RESIDENTIAL AND CIVIL CONSTRUCTION  
ALLIANCE OF ONTARIO

April 2013

Baher Abdulhai, Ph.D.  
Professor, University of Toronto  
Director, The Toronto Intelligent  
Transportation Systems Centre

---

## Table of Contents

---

<b>Executive Summary</b>	<b>8</b>
Motivation	10
Essential Background	10
Congestion: The Nature of the Beast	10
Categories of Solutions	12
Capacity Expansion	12
Demand Reduction (Management)	13
The Third Alternative: Intelligence	14
The Silver Bullet: Capacity Expansion, Demand Reduction or ITS?	15
Report Objectives	16
High Potential ITS Technologies for the GTHA	17
<b>Chapter I: MARLIN</b>	<b>18</b>
Traffic Lights: Source of Delay or Efficiency?	19
Introduction to Adaptive Traffic Signal Control	20
In Lay Terms: Stakeholder Benefits	23
From Single-Agent to Multi-Agent Reinforcement Learning	24
Reinforcement Learning	24
Multi-Agent Reinforcement Learning	24
Multi-Agent Reinforcement Learning for an Integrated Network of Adaptive Traffic Signal Controllers (MARLIN-ATSC) Platform	25
Agent	26
The Learning Environment	27
Interface	28
Experimental Results	28
Testbed Network	28
Benchmarks	29
Toronto Results and Discussion	30
Burlington Results and Discussion	36

---

• Base Case Model with Actuated Signal Control	37
• Adaptive Traffic Control Strategies	37
• Supply Management Strategies	37
Observations and Analytical Results	38
Conclusions and Future Work	38
Chapter I References	40
<b>Chapter II: OSTE</b>	<b>42</b>
Severe Congestion: Just Add Chaos!	43
Evacuation 101	44
State-of-the-art Scan	45
Evacuation Scheduling	46
Destination Choice	46
Traffic Routing and Control Strategies	47
OSTE Motivation	48
OSTE: Framework for Optimizing Multimodal Evacuation	49
Optimal Spatio-Temporal Evacuation Model	49
Optimal Routing and Scheduling of Mass Transit Vehicles	50
Demand Estimation by Mode	51
Data Source: The Transportation Tomorrow Survey (TTS)	51
Model Estimation	51
The Overall Framework: Putting the Pieces Together	52
Large-scale Application: Evacuation of the City of Toronto	54
Supply Modeling	54
The Network Simulation Model	54
Transit Infrastructure: The Toronto Transit Commission (TTC) Fleet	55
Evacuation Demand Estimation	55
Subway Demand	55
Shuttle Bus Demand	56

---

Representation of Noncompliant Traffic	56
Experimental Results and Analysis	56
OSTE: Genetic Optimization Process	56
OSTE: Traffic Assignment Outputs	57
Optimal Routing and Scheduling of Transit Vehicles	60
Shuttle Buses	61
Summary, Conclusions, and Future Research	62
Chapter II References	64
<b>Chapter III: Win-Win Congestion Pricing</b>	<b>66</b>
Introduction	67
What is Congestion Pricing?	67
To Toll or Not to Toll: Tragedy of the Commons	68
Brief History of Road Pricing Studies and Practices	69
Congestion? Microeconomic and Traffic Engineering Perspectives	71
Getting Technical with Tolls: Static vs. Dynamic, Revenue vs.	
Social Welfare Maximization, and Single Roads vs. a Network	73
Static Pricing	74
Dynamic Pricing	77
Turning Prices into Control Actions: How to Control Hypercongestion	
using Dynamic Pricing	79
Open-Loop versus Closed-Loop Dynamic Congestion Pricing	80
Traffic Network Optimization versus Regulation Using Pricing	81
Revisiting Recent Advances in Congestion Pricing Methods	82
General Congestion Pricing Framework	82
Facility Pricing	82
Network Pricing	83
User Responses to Congestion Pricing	83
Putting the Pieces Together: Aspects to Consider in a	
Comprehensive Congestion Pricing Strategy	84

---

The Roadmap: An Integrated Congestion Pricing Strategy for Large Cities and the GTHA	87
Summary and Future Work	90
Chapter III References	91
<b>Chapter IV: Open Transport Innovation</b>	<b>93</b>
The Point	94
Introduction	94
Background and Motivation	96
The Proactive Transportation Management Challenge	96
The Sustainability of Transportation Innovation Challenge	96
Connected Vehicles and Aspirations for the Connected Traveller	97
The Tsunami of Data Opportunities and Challenges	98
Theme I: Open Transport Service Innovation Platform	98
The Transport Open Service Innovation Framework	100
Theme II: A Sensing Platform and Gateway for Traffic, Transit and Freight Monitoring	102
A Gateway for Multi-Protocol Data Sensing and Delivery	103
Theme III: Multi-Sensor Data Fusion	104
Data Fusion Example: Emerging Technologies for Traffic Data Collection	105
Enabling Open Intelligent Transportation Systems: the Tools	108
Online Network Enabled-Intelligent Transportation Systems (ONE-ITS)	108
A Sensing Platform using Smartphone and On Board Diagnostics Device Application	112
Data Fusion System Application for Speed and Travel Time Estimation	114
Summary	116
Chapter IV References	117
<b>Closing</b>	<b>118</b>
Looking Ahead: Autonomous-Vehicles – The Next Traffic Revolution	118
References	119

## EXECUTIVE SUMMARY

---

**A**s congestion levels soar to unprecedented levels, solutions become more complex, creating a strong potential for technology and an important role for advanced research.

Congestion mitigation strategies typically focus on infrastructure capacity expansion or demand reduction. An important, but less evident, category of solutions is enhancing the efficiency of the existing system, in other words, how do we use existing infrastructure more intelligently using technology?

Intelligent Transportation Systems (ITS) help increase the effective capacity of infrastructure, manage demand, and maximize efficiency. It is important to realize though that there is no single silver bullet solution to congestion problems. A comprehensive solution consists of a three-pronged approach: (1) capacity expansion where warranted, (2) demand management to rationalize use, and (3) intelligent systems to dynamically enhance efficiency of the existing system, before building more or imposing harsh restrictions on users.

This report offers an in-depth look at promising technology-oriented congestion mitigation approaches for large cities, particularly the Greater Toronto and Hamilton Area (GTHA), as part of any comprehensive portfolio of sustainable transportation solutions. The report consolidates and recommends ITS solutions for the GTHA, based on state-of-the-art research at the University of Toronto's ITS Centre and Testbed over the past 14 years and knowledge of the state-of-the-art worldwide.

More specifically, this report presents and discusses in detail, in a practical manner, four specific ITS solutions that have a common objective of easing congestion during typical daily peak congestion periods and under atypical emergency situations that cause rare but very significant and volatile congestion for which we have to prudently anticipate and plan. These technologies are:

1. Smart self-learning traffic lights to cut down intersection delay by half.
2. Optimization of large-scale emergency evacuation for managing the transportation system under evacuation demand and cutting down evacuation time by as much as 75%.
3. Win-win dynamic road pricing for effective but socially conscious transportation systems management.
4. Open innovation business model and technical platform for accelerating transport innovation in Canada.

---

The intent is to bring to practice the most pragmatic solutions from academic research to achieve the following:

1. Reduce congestion levels in the GTHA.
2. Close the gap between state-of-the-art and state-of-the-practice in the GTHA in particular and in Canada in general.
3. Promote the most pragmatic Canadian-grown ITS solutions for local consumption and exportation.
4. Invite public sector agencies to have a closer look at the presented technologies and assess the benefits to the public through pilot deployments and further testing.
5. Invite the government(s) to put into place special procurement processes to try new solutions that emerge from innovation-based organizations such as, but not limited to, universities.
6. Invite the private sector to adopt the new technologies and offer them to public agencies, the public at large, and other countries.

*RCCAO wishes to thank Professor Robin Lindsey, Professor, Sauder School of Business, University of British Columbia, who provided an independent review of the draft report. His fast turnaround and insightful comments were most helpful to the authors in finalizing the report. Most recently, Professor Lindsey co-authored a study for RCCAO with Professor Emeritus Harry Kitchen on 'Financing Roads and Public Transit in the Greater Toronto and Hamilton Area' (January 2013).*

---

## Motivation

This report offers an in-depth look at promising technology-oriented congestion mitigation approaches for large cities, particularly the Greater Toronto and Hamilton Area (GTHA), as part of any comprehensive portfolio of sustainable transportation solutions. The report consolidates and recommends Intelligent Transportation Systems (ITS) solutions for the GTHA, based on state-of-the-art research at the University of Toronto's ITS Centre and Testbed over the past 14 years and knowledge of the state-of-the-art worldwide.

The technologies offered are not only cutting-edge, but also Canadian, which is equally important. Canadian congestion mitigation technologies have the potential to respond to local conditions at reduced cost, and they offer opportunities for exporting the solutions internationally. The research and development presented in this report target closing the gap between the state-of-the-art and the state-of-the-practice in transportation in Canada, bringing to practice the most pragmatic solutions from academic research.

As congestion levels soar to unprecedented levels, solutions become more complex, creating a strong potential for technology and an important role for advanced research. This report is also an invitation to both the public and private sectors to adopt promising solutions that can have a strong impact on congestion in large Canadian cities and the overall quality of life.

## Essential Background

### Congestion: The Nature of the Beast

Congestion is simple to explain, but extremely complex to solve.

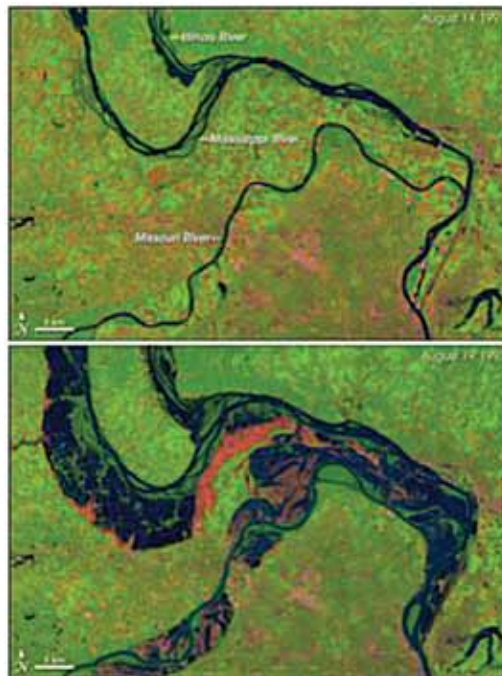
When and where the number of users (cars, travellers, transit riders, etc.) exceeds the capacity of the infrastructure (road, bus line, etc.), we have congestion (or hyper congestion, which will be defined in detail later). More formally, when demand exceeds available capacity, the system starts to break down exhibiting instability and excessive delays. This happens somewhere and sometime in any town or city, small or large. However, the larger and more populous the city or a metropolis, the more congestion spreads over space (more and more locations become congested, contiguous or not) and over time (longer periods of the days are congested). The overall effect of widespread congestion, spatially and temporally, is lower average speeds to get from anywhere to anywhere; for instance, longer commute times and high impedance to the movement of goods. Such spread of congestion has ramifications and impacts on quality of life, the economy, and the environment. When

is serious intervention needed? Unfortunately, the decision is subjective and depends largely on how much pain we are willing to endure. It also depends on the available resources to solve the problem and keep the economy moving.

Unlike the collapse of a building, it is important to realize that transportation system failure is neither as sudden nor as catastrophic. Transportation failure is gradual, and it becomes worse with time at an increasing rate (exponential deterioration). More demand and/or less capacity (due to accidents or construction) means more spread of congestion over wider space and longer time, resulting in lower overall service quality.

The exponential nature of deterioration of the transportation system performance (i.e. congestion) means that, although it is gradual, it can accelerate and get out of hand fairly quickly. Deterioration can indeed reach the near-catastrophic stage when it becomes unbearable for citizens and businesses. Understanding the spatio-temporal nature of congestion, however, helps us understand that congestion is not a binary thing; i.e. now you see it, now you don't. Rather, it is a continuous phenomenon. In other words, we can work to make it better, but there is no such a thing as a world without congestion, unless you live in the middle of nowhere. Actually, while it has negative consequences, congestion can be a sign of vibrancy.

Another important observation worth highlighting is the “flooding” nature of congestion spread. Much like an overflowing river, vast areas of land can be covered very quickly. As shown in Figure 1, a river, and similarly, a major freeway, can carry significant flow under normal circumstances in a well-confined stream. If it is allowed to overflow, however, significant areas will be covered with excess water very quickly. When a major freeway overflows, a fairly small percentage of diversion onto the streets can cause widespread congestion. This is the reason it is important to properly maintain high-capacity major routes throughout the city.



**Figure 1: Overflowing River**

---

## Categories of Solutions

The good news is that there are plenty of approaches to address congestion: traditional and non-traditional, technical and non-technical (e.g. habitual, social etc). In general, we agree that congestion occurs when demand volume ( $v$ ) exceeds capacity ( $c$ ), somewhere, sometime, i.e.  $v/c$  ratio exceeds 1. The logically straightforward solutions are:

1. Add capacity such as new lanes, new roads, new transit lines, and/or
2. Reduce demand (voluntarily or using disincentives such as road pricing), such as:
  - a. Don't travel, e.g. work from home
  - b. Travel in less congested times, e.g. off peak
  - c. Travel in less congested places, e.g. away from downtown cores (area) or away from congested roads (routes)
  - d. Change mode and use transit or bike
  - e. Carpool, etc.

These solutions are easier said than done.

## Capacity Expansion

Building capacity is infrastructure-intensive and costly, both in the financial and environmental senses. We can never build enough lanes to completely eliminate congestion at all times and everywhere. Many people see building more roads as an unsustainable alternative. If we all opt to drive cars alone (single occupant), it will eventually be prohibitive to serve the resulting demand at a reasonable level of service, not to mention the environmental impact. Therefore, rushing to building more roads should not be the first alternative. However, there is a danger in suggesting that we can avoid building roads altogether. If we do not build high- capacity lines, we risk flooding vast areas of city streets.

Building transit makes more sense where demand is available. If we replace 50 cars with one bus, the benefit is evident. The danger though, again, is inefficiency. We all see empty buses running in the evening in the suburbs and even in downtown cores. Running empty transit is cost-prohibitive, financially impractical, and environmentally mad. Too much transit can also be unsustainable. However, an empty bus is often justified on the basis of social equity. Sustaining public jobs sometimes further complicates the matter. Cutting off-peak transit service may mean cutting jobs for transit workers. Automating subways or joining streetcars together may also mean cutting further jobs for drivers. In such cases, it is no longer a technical or service quality problem, but rather a social and political one.

---

## **Demand Reduction (Management)**

Reducing demand is another option to bring down the  $v/c$  ratio. It is harder to do because it is prohibitive—it involves telling people what not to do. It is often called demand management because “reduction” is too harsh a term and can agitate the public. Travellers rarely volunteer to reduce travel to help the system, and disincentives such as congestion pricing are often fiercely opposed.

As we will later explain, congestion pricing is not as bad as it sounds. If executed correctly, it can be beneficial for all. Regardless of the political aspects of congestion pricing, disincentives in general “push” travellers away from congestion by shifting modes, departure time, and/or location of travel. Remember that congestion is a continuous phenomenon in both space and time, i.e. wherever and whenever there is congestion, there are other places and other times that are less congested. Therefore, the majority of demand management techniques target the redistribution of demand over space (modes, roads, etc.) and time (peak vs. off-peak).

Also remember that transportation demand is “derived” from socioeconomic activity. People travel not only to perform activities, but those activities have time constraints. It is impossible to balance demand over 24 hours, so there will always be peak periods. We just try to flatten the peak when and/or where demand exceeds capacity by spreading demand more evenly over time and space, to the extent feasible and in a controllable manner.

It is important, early in the discussion, to briefly explain a few important points about the often widely opposed topic of congestion pricing. If any commodity is subsidized, we will tend to over consume it. More than a century ago, the British discovered that when cattle are allowed to feed freely on public land, they tend to overgraze and grass depletes, to the detriment of all. This phenomenon is known as the tragedy of the commons. To the contrary, imposing a fee rationalizes demand and prevents the system from collapsing. This option is well established in economics (see details in chapter 3).

By the same principle, drivers should fully pay the cost of what they consume: road travel. While our taxes partially pay for the road, and we pay for the car, for gas, etc., we do not pay for the congestion we impose on all others. Just by popping up on the freeway, you slow down everyone else a tiny tad, but this tiny tad affects thousands of others. This “marginal” delay is the social subsidy for your travel. Paying for this external cost of congestion rationalizes demand, reduces congestion, and maximizes social welfare. Moreover, when you show up on the freeway at 5:00 PM, the external cost of congestion or your marginal impact can be substantial, i.e. not just a tiny tad. A little more demand may cause the freeway to slip into a stop-and-go pattern and lose some 25% of its capacity. Preventing this situation by imposing a dynamically varying fee means gaining back 25% capacity at the time we are desperate for it. It is important to know that the price you need to pay to achieve this situation is far less than monopoly

---

pricing on privately owned roads, where a road owner is trying to maximize profit. In social road pricing, the target is to benefit the public at large, i.e. maximum social welfare. The impact is higher throughput and the toll is mild. We treat these issues in detail in chapter 3.

### **The Third Alternative: Intelligence**

When we focus on v/c ratio, an important, but less evident, category of solutions, is enhancing efficiency of the existing system, i.e. use the existing infrastructure more intelligently. For instance:

1. If traffic lights are designed to learn to be agile and switch back and forth in accordance with demand fluctuations, we can cut delay at intersections by as much as 60%. This is a technology-based solution that does not require physical infrastructure expansion or demand reduction. It is simply efficiency enhancing technology.
2. If a disaster requires the evacuation of a large city with millions of people, running to the road will not help anyone because we will all sit on the road and move nowhere. A computerized evacuation optimization system can guide travellers through the process and inform them when to evacuate or mobilize, where to go (safe destination), and how to get there (mode and route). It does so by optimizing the spread of demand over time and space in a manner that reduces congestion and breakdowns. This system can reduce evacuation time of a city like Toronto by 75%, a whopping reduction without building a single road.
3. A computer system that monitors a freeway and dynamically sets varying congestion pricing fee in real time, such that demand never exceeds capacity, will accommodate 25% more demand.
4. If you wonder why buses come in threes, it is because once a bus is delayed in traffic, it reaches the stop late, stays longer to fill up and leaves full, while the next one and the next one are running emptier and faster until they catch up and bunch together. A smart traffic light would sense the formation of this bunching and accelerate the first but hold the second until the schedule is restored.

Such technological solutions fall under the umbrella of Intelligent Transportation Systems (ITS), a field that has been growing very rapidly since 1990s. As congestion escalates, especially where infrastructure expansion does not keep up with increased demand, ITS is growing more important. A common key characteristic of such solutions is real time operation that allows decisions to be made in seconds or minutes based on current (or even predicted) congestion conditions. Hence, ITS require heavy use of Information and Communication Technologies (ICT). The cost of ITS and ICT is far less than the cost of infrastructure expansion.

---

## **The Silver Bullet: Capacity Expansion, Demand Reduction or ITS?**

Based on the introduction, here is a summary of the key points.

1. Congestion is a continuum in space and time, not an on-off event.
2. Congestion can be reduced in space and time, but it cannot be feasibly eliminated at all times and everywhere.
3. Solutions or categories of solutions are plenty, with varying degrees of cost, ease of implementation, and impact.
4. Congestion is dynamic, happens here or there, at this time or that time, and needs to be dealt with using dynamic systems control.

The logical question to follow is that if there is a silver bullet solution to congestion, is it building capacity, reducing demand or ITS?

There is no single silver bullet solution. The solution is a combination of several things: (1) capacity expansion where warranted, (2) demand management to rationalize use, and (3) intelligent systems to dynamically enhance efficiency of the existing system, before building more or imposing harsh restrictions on users. Also, ITS can be a capacity expansion mechanism. Consider the smart traffic lights example. We managed to achieve in the lab the same benefits of a multi-million dollar intersection expansion using smart traffic lights, with anticipated deployment cost of approximately 5% of the expansion cost. ITS offers capacity expansion at a fraction of the cost of infrastructure expansion. It can also be used for demand management. A real-time information system can provide travellers with congestion conditions, routing options, or congestion prices on their smart phones, so they can plan their departure time, mode choice and route choice accordingly, in daily commute situations or in emergency situations.

Moreover, ITS enhances multi-modality options for travellers. In the past, we either drove or took transit. It is becoming more and more common that travellers drive to transit, or choose between the two modes dynamically depending on their time of travel. For instance, driving is perhaps the fastest option for a trip to the downtown core at 1:00 PM. On the other hand, if the same trip has to happen at 8:00 AM, transit might be a more sensible choice, especially for single travellers. Having real-time information on the status of roads and transit services, therefore, expands travel options and greatly helps travellers to choose. Last but not least, ITS enables dealing with congestion in a dynamic fashion, providing dynamic solutions when and where congestion is, in real time, to keep the system moving while being fully utilized. Managing congestion in real time is much like balancing an inverted pendulum, a common textbook control problem. To illustrate the challenge, put a stick on the palm of your hand, try to balance it upward, and observe the dynamics.

---

In conclusion, intelligence matters in dealing with complex large-scale systems such as transportation systems, especially under soaring congestion. Although ITS is only one of the alternative categories of solutions, it should be examined before, and examined in conjunction with the other categories. Non-wasteful transit should be expanded, but ITS should also be used to make transit more efficient. Roads, whether we like them or not, will have to be expanded and ITS will make them efficient, reducing the need for more. Demand must be managed and ITS will help us manage it with precision.

### **Report Objectives**

This report makes recommendations for ITS solutions for large congested cities, with a specific eye on the GTHA. The recommendations are based on state-of-the-art research at the University of Toronto's ITS Centre and Testbed over the past 14 years and intimate knowledge of the state-of-the-art internationally. The intent is to bring to practice the most pragmatic solutions from academic research to achieve the following:

1. Reduce congestion levels in the GTHA.
2. Close the gap between state-of-the-art and state-of-the-practice in the GTHA in particular and in Canada in general.
3. Promote the most pragmatic Canadian-grown ITS solutions for local consumption and exportation.
4. Invite public sector agencies to have a closer look at the presented technologies and assess the benefits to the public through pilot deployments and further testing.
5. Invite the government(s) to put into place special procurement processes to try new solutions that emerge from innovation-based organizations such as, but not limited to, universities.
6. Invite the private sector to adopt the new technologies and offer them to public agencies, the public at large, and other countries.

---

## High Potential ITS Technologies for the GTHA

ITS solutions are numerous and are the subject of very aggressive research and development worldwide. In this report, we focus on technologies that have been developed and/or tested in a Canadian context, more specifically in Toronto, without loss of generality in terms of transferability to other cities. The rest of the report presents and discusses in detail, in a practical manner, four specific ITS solutions that have a common objective of easing congestion during typical daily peak congestion periods and under atypical emergency situations that cause rare but very significant and volatile congestion that has to be prudently anticipated and planned for. These technologies are:

1. Smart self-learning traffic lights to cut down intersection delay by half.
2. Optimization of large-scale emergency evacuation for managing the transportation system under evacuation demand and cutting down evacuation time by as much as 75%.
3. Win-win dynamic road pricing for effective but socially conscious transportation systems management.
4. Open innovation business model and technical platform for accelerating transport innovation in Canada.

# Chapter I: MARLIN

**Multi-agent Reinforcement Learning  
for Integrated Network of Adaptive Traffic  
Signal Controllers (MARLIN-ATSC):  
Methodology and Large-Scale  
Application in the GTHA**

Abdulhai, B., El-Tantawy S. and Abdelgawad H.

---

## Traffic Lights: Source of Delay or Efficiency?

Stopping at a traffic light, or several traffic lights in a row, one sometimes wonders about the purpose of traffic lights. They seem like a source of delay. If I were not stopped, I would have been moving; the difference is pure delay, right? Not really. Traffic lights are intended to resolve right-of-way conflicts at intersections. If all cars try to go through an intersection at the same time, they will come to a halt, or even crash. Therefore, traffic lights are intended to assign the right-of-way to a group of non-conflicting movements at a time, enhancing both safety and efficiency. Imagine downtown Toronto at 5:00 PM under a power outage with no functioning traffic lights. It would be gridlock, plain and simple.

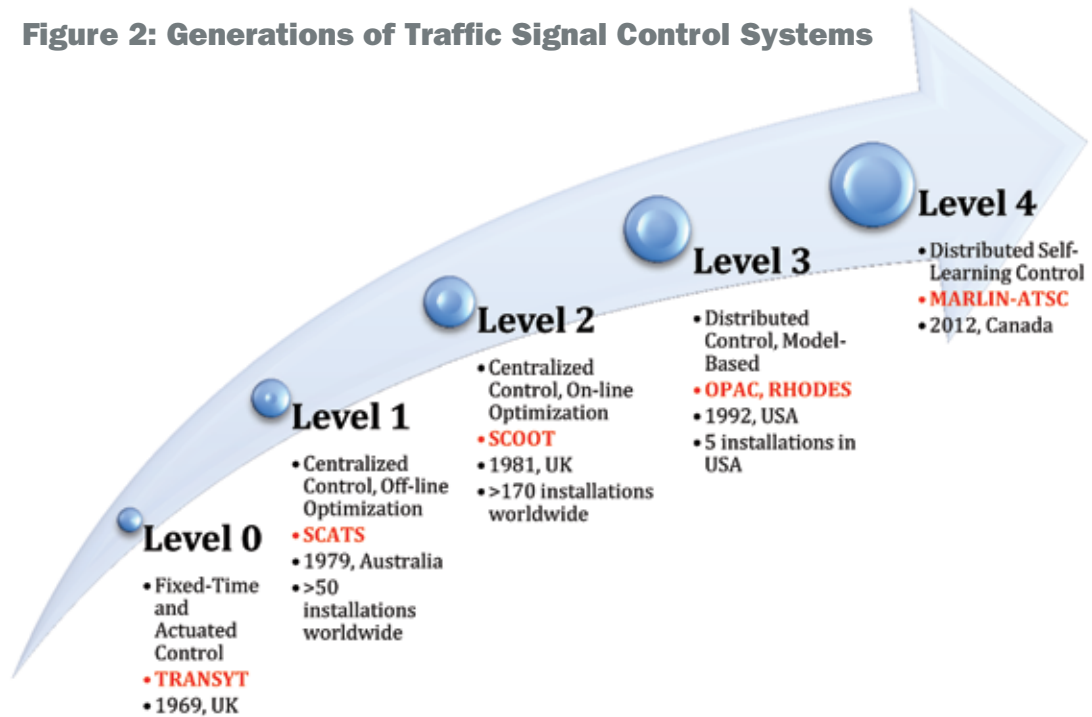
Having said that, we admit it is not entirely convincing. We all stop at a red traffic light while no traffic goes through the green, and that causes unnecessary delays. We also stop, seemingly endlessly, at congested traffic lights that do not cope with demand. The solution, one may quickly say, is to make traffic lights change in an agile manner in accordance with demand. The devil is in the details. It is also important to make a clear distinction between two completely different reasons that traffic lights can be inefficient.

1. Inefficiency in light traffic. When one direction is stopped in absence of demand from opposing directions.
2. Inefficiency in heavy traffic. When traffic is heavy and heavily fluctuating from all directions and the traffic light is not able to meet the heavy and oscillating demand with precision.

The first case has a relatively easy fix. For one, if traffic is light, the problem is not as pressing and the number of affected travellers, overall, is not high. In such cases, actuated traffic lights do a fine job. The control system assigns the green signal only after a vehicle(s) is detected. No vehicles, no green.

The second case is tougher. The traffic light is faced with heavy and variable demand and long queues from all directions. The question is how to optimize green allocations for maximum efficiency. Doing such optimization on a second-by-second basis is not a trivial task, especially if this optimization needs to be done over a large number of intersections simultaneously. The control system would need traffic sensing (detection), communication networks, heavy computing, and, most importantly, a smart control logic: the brain of the system that runs the controller and switches the traffic lights in real time. Depending on the desired efficiency, the system may become more complex and intractable. Solving such complex problems has been in the making for decades with several generations of traffic control methods, briefly summarized in Figure 2. In this report, we focus primarily on the latest generation, level 4, and the MARLIN-ATSC, developed at the University of Toronto. We highlight the technological, methodological, and operational differences compared to the previous generations and the state-of-the-art. We also present case studies showing the performance of the system in virtualized GTHA scenarios, and compare them to state-of-the-practice, i.e. current conditions.

**Figure 2: Generations of Traffic Signal Control Systems**



## Introduction to Adaptive Traffic Signal Control

Population is steadily increasing worldwide and the GTHA is no exception. Consequently, the demand for mobility is rapidly increasing and congestion is turning into a daily chore, hampering not only quality of life but also economic competitiveness. When the growth in social and economic activities outpaces the cash-strapped growth of transportation infrastructure, congestion is inevitable. Among the myriad of demand and supply management possibilities to combat congestion, Adaptive Traffic Signal Control (ATSC) targets enhancing infrastructure efficiency by adjusting the traffic signal timings in real time in response to traffic fluctuations to achieve a chosen objective (e.g., minimize delay).

ATSC, in general, has great potential to outperform both pre-timed and actuated control [1]. In Toronto, for instance, almost one-quarter of the traffic lights are controlled by a system of British origin named SCOOT. Existing ATSC systems, however, have non-trivial limitations that make them relatively inefficient, expensive, and difficult to maintain, ultimately limiting their potential benefits. For instance, they treat intersections as isolated nodes independent of neighbouring intersections, which limits the efficiency gains of such technology.

Therefore, optimally controlling the operation of multiple intersections simultaneously can be synergetic and beneficial. Such integration certainly adds more complexity to the system that science has not been able to resolve until very recently. Coordination has been typically approached in a centralized way (e.g., SCOOT [2], TUC[3]) which is only feasible if communication channels amongst all intersections and the central control

---

location are available, which is resource demanding and prone to communication failure. SCATS [4] is another example of an adaptive signal control system that is hierarchical and distributed system in which an area is divided into smaller subsystems (in the range of 1–10 intersections) that perform independently. PRODYN [5], OPAC [6], RHODES [7] are also examples of adaptive systems that are decentralized, but their relatively complex computation schemes make their implementation costly [8].

The coordination mechanism in the systems in practice is employed along an arterial (where the major demand is). It is desirable to efficiently operate traffic signals along arterials where the major demand is. Despite the importance of maintaining good progression along an arterial, it is also important to consider the network-wide effect of how we operate the traffic lights in two-dimensional networks, especially when major east-west arterials for instance pour traffic demand onto north-south arterials, as is the case in downtown Toronto. In a signalized urban network setting, considering a network-wide objective has the potential to improve overall network performance and mobility, and reduce emissions. It is important to not confuse the widely common offset coordination with the newer concept of agent coordination in adaptive control. In the traffic signal control practice, offset coordination (also known as synchronization or progression) is concerned with adjusting the time offset between successive intersections such that vehicles can pass through several intersections without stopping. On the other hand, action coordination mechanism between control agents (signalized intersections) has a much broader meaning as agents coordinate their policies such that a certain objective is achieved for the entire traffic network.

Based on scientific advances in the last decade, coordination can be plausibly achieved using reinforcement learning and game-theoretic approaches [8]. Reinforcement Learning (RL), from Artificial Intelligence (AI), has shown good potential for self-learning closed-loop optimal traffic signal control in the stochastic traffic environment ([9],[10]). RL has the added advantage of being able to perpetually learn and improve service over time. In RL, a traffic signal represents a control agent that interacts with the traffic environment in a closed-loop system to achieve the optimal mapping between the environment's measured traffic state and the corresponding optimal control action, offering an optimal control law. The mapping from states to actions is also referred to as the control policy. The agent iteratively receives a feedback reward for the actions taken and adjusts the policy until it converges to the optimal control policy. The University of Toronto successfully demonstrated the value of RL systems in traffic signal control almost a decade ago. The challenge, however, has been how to scale the system up and apply RL to a transportation network of multiple signalized intersections. Control agents typically react to changes in the environment at the individual level, but the overall behaviour of all agents may not be optimal. Each agent is faced with a moving-target learning problem in which the agent's optimal policy changes as the other agents' policies

---

change over time[8]. The University of Toronto has investigated Game Theory, which provides the tools to cast the multi-agent systems as a multiplayer game and provide the rational strategy to each player in a game. Multi-Agent Reinforcement Learning (MARL) is an extension of RL to multiple agents in a stochastic game (SG) (i.e. multiple players in a stochastic environment), a well-suited approach for dynamics and stochastic traffic systems ([8], [11]).

MARLIN is designed and developed to directly address the challenges faced by MARL-based systems. First is the exponential growth in the state-action space with the increase in the number of agents. Second is that the majority of the MARL-based ATSC in the literature assumes that agents learn independently, in which case each agent acts individually in its local environment without explicit coordination with other agents in the environment. Although independent operation simplifies the problem, it limits efficiency in case of a network of agents. For example, in over-saturated traffic conditions, queues could easily propagate from a downstream intersection (agent) and spill back to the upstream intersections (agents) in a network-wide cascading fashion. Such cases require a network-wide multi-agent coordination as discussed earlier. Thus, flexible and computationally efficient ATSC systems are becoming instrumental in controlling a network of agents; plausibly by employing heuristics and approximate approaches based on modifying the existing MARL techniques [8].

To address these limitations, The University of Toronto presents a novel Multi-Agent Reinforcement Learning for Integrated Network of Adaptive Traffic Signal Controllers (MARLIN-ATSC) that offers the following features and characteristics:

1. Decentralized design and operation, typically less expensive compared to the centralized system.
2. Scalable to accommodate any network size.
3. Robust with no single point of failure.
4. Model-free - does not require a model of the traffic system that is challenging to obtain
5. Self-learning - reduces human intervention in the operation phase after deployment (the most costly component of operating existing ATSCs)
6. Coordinated – by implementing mode 2 (integrated mode), which coordinates the operation of intersections in two-dimensional road networks (e.g. grid network), a new feature that is unprecedented in ATSC state-of-the-art and practice.

In addition, MARLIN-ATSC is tested on a large-scale simulated network of 59 intersections in downtown Toronto using the input data (e.g. traffic counts, signal timings, etc.) provided by City of Toronto. It is also simulation-tested onto two of the major hotspots in the City of Burlington, a suburb west of Toronto, using local traffic information provided by the City of Burlington.

---

## **In Lay Terms: Stakeholder Benefits**

As in the case with any knowledge-intensive application, technical details can be overwhelming and take away from the bottom line: the benefits. In this section, we attempt to clearly state the anticipated benefits of MARLIN to stakeholders, municipal operators, and travellers.

Developed at the University of Toronto, MARLIN is a state-of-the-art traffic signal control system. It is AI-based control software that enables traffic lights to self-learn and self-collaborate with neighbouring traffic lights to cut down motorists' delay, fuel consumption, and the negative environmental effects of congestion. The target user sector is municipal traffic departments in medium to large cities experiencing chronic congestion. The ultimate beneficiaries are drivers and commuters who are suffering in escalating congestion in major urban areas.

MARLIN offers unprecedented value to both municipal operators and motorists. The system decreases delay at intersections by an average of 40% and up to 75% in some areas. It improves travel times on major corridors like Toronto's Lake Shore Boulevard by 25% and decreases emissions by 30%. These values enable motorists to enjoy improved mobility, save time and money, lower unpredictable delay risk, and enhance their travel convenience and overall quality of life.

In addition to enabling municipal operators to better serve the public and fulfill their mandates, MARLIN cuts down implementation cost and it is easier to use than existing knowledge-intensive products in the market. MARLIN achieves these benefits due to its novel decentralized design that puts intelligence in the traffic light in the field. It does not need expensive communication infrastructure to link individual intersection to a central control room, which means it can offer major financial saving and reliability enhancements compared to competitive products. MARLIN is also self-learning and hence relieves municipalities of the burden of hiring and maintaining highly skilled operators, which is a major challenge even for large cities like Toronto. MARLIN also uses non-intrusive detection, relieving municipalities of the need to use common pavement-embedded detectors that often break or fail, and are hard to repair in heavy traffic corridors and harsh winter weather.

Overall, the value of MARLIN to motorists and municipal operators presents a new generation of intelligent traffic control, made in Canada.

---

## From Single-Agent to Multi-Agent Reinforcement Learning

### Reinforcement Learning

Typically, RL is concerned with a single smart agent operating in an environment so as to maximize its cumulative long-run reward, inspired by how humans learn and function. The environment's state depends on the agent's actions. The most common single agent RL algorithm is Q-learning [12]. The Q-Learning agent learns the optimal mapping between the environment's state  $s$  and the corresponding optimal control action  $a$  based on accumulating rewards  $r(s,a)$ . Each state-action pair  $s,a$  has a value called Q-Factor that represents the expected long-run cumulative reward for the state-action pair  $(s,a)$ . In each iteration,  $k$ , the agent observes the current state  $s$ , chooses and executes an action  $a$  that belongs to the available set of actions  $A$ , and then the Q-Factor is updated according to the obtained reward  $r(s,a)$  and the state transition to state  $s'$ .

The sequence  $Q^k$  is proven to converge to the optimal value after sufficient training [12]. The agent can be trained on the actual environment or in virtual replica of the environment. During training, the agent explores possible actions and policies and increasingly learns to exploit the best-known actions. To balance the exploration and exploitation in Q-Learning, algorithms such as  $\epsilon$ -greedy and softmax are typically used [13]. After the completion of training, the agent is deployed in the actual environment to control it, exploiting its knowledge from the training phase. The agent, once deployed, can cease to learn or, if desired, can continue to fine tune its knowledge with experience while operating.

### Multi-Agent Reinforcement Learning

MARL is an extension of RL to multiple agents (signalized intersections). The simplest way to extend RL to the MARL is to consider the local state and local action for each agent assuming a stationary environment and that the agent's policy is the prime factor affecting the environment. However, the traffic environment is non-stationary since it includes multiple agents learning concurrently; i.e., the effect of any agent's action on the environment depends on the actions taken by the other agents. The local behaviour of traffic at one intersection affects and is affected by traffic at the neighbouring intersections. Each agent is, therefore, faced with a moving-target learning problem because the best policy changes as the other agents' policies change, which accentuates the need for coordination among agents. Coordination can be achieved by considering the joint-state and joint-action for the other agents in the learning process. Moreover, given that all agents are acting simultaneously, the agents' choices of actions must be mutually consistent to achieve their common goal of optimizing the signal control problem. Therefore, the agents require a coordination mechanism to make the optimal decision from the possible joint actions (i.e., agents have to coordinate their choices/

actions so as to reach a unique equilibrium policy). Agent coordination in this context is not to be confused with conventional traffic signal coordination that maximizes green bands, offsets, etc.

Game theory offers the theoretical framework of MARL. Agents act as players in a game setting. The agent's objective is to find a joint policy (known as equilibrium) in which each individual policy is a best response to the others [14]. Further in depth technical details can be found in [15] -[27].

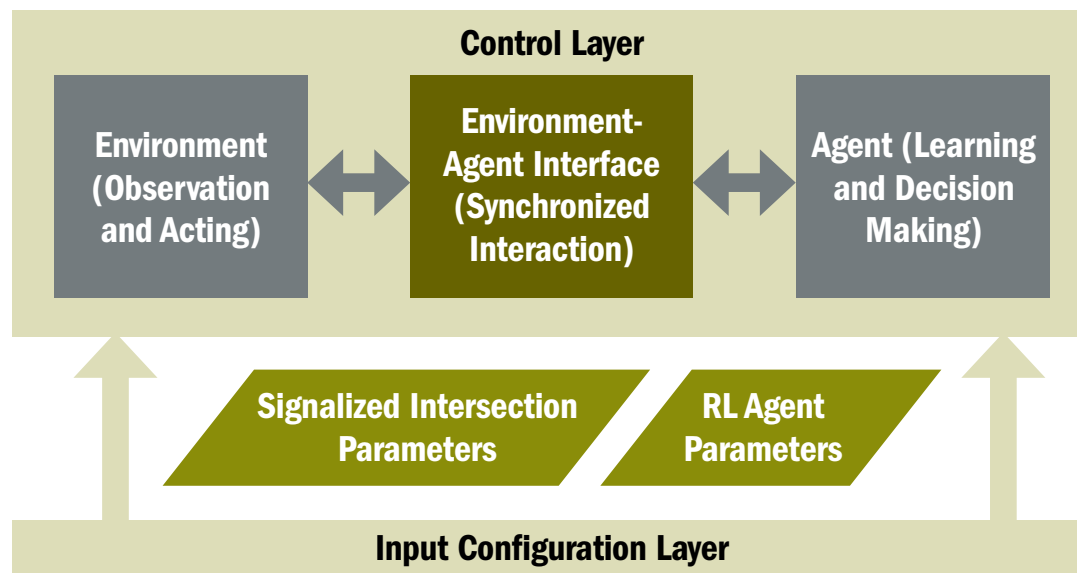
### **Multi-Agent Reinforcement Learning for an Integrated Network of Adaptive Traffic Signal Controllers (MARLIN-ATSC) Platform**

The MARLIN-ATSC platform is illustrated in Figure 3. The platform consists of two main layers. The first is an input configuration layer that is responsible for configuring and providing the necessary input to the second layer.

The configuration layer has two main roles. 1) It configures the training / learning environment using actual traffic data, and 2) it configures the RL-design parameters.

The second layer is a control layer that includes three interacting components (as shown in Figure 3).

**Figure 3: MARLIN-ATSC Platform**



---

## Agent

The agent component implements the control algorithm. The agent is the learner and the decision-maker that interacts with the environment by first receiving the system's state and the reward and then selecting an action accordingly. A Generic agent model is developed using Java programming language such that different levels of coordination, learning methods, state representations, phasing sequence, reward definition, and action selection strategies can be tested for any control task. In MARLIN-ATSC, agents can implement one of the following two control modes:

- **Independent Mode:** In this mode, each controller has an RL agent working independently of other agents using Multi-Agent Reinforcement Learning for Independent controllers [*MARL-I*] in which each agent implements a Q-Learning algorithm [12].
- **Integrated Mode:** In this mode, each controller coordinates the signal control actions with the neighbouring controllers by implementing MARLIN learning algorithm.

## MARLIN Learning Approach

MARLIN presents a new control system that maintains an explicit coordination mechanism while addressing the curse of dimensionality for a large-scale network of connected agents by:

- Exploiting the principle of the locality of interaction [31] among agents. The principle of locality of interaction endeavours to estimate a local neighbourhood utility that maps the effect of an agent to the global value function while only considering the interaction with its neighbours. Hence, it is sufficient to consider the neighbours' policies to find the best policy for the agent.
- Utilizing the modular Q-learning technique [32]. Modular Q-learning partitions the state-space to partial state spaces that consist of two agents. As a consequence, the size of the partial state space is always  $|S|^2$  regardless of the number of agents, and therefore results in a reasonable state space.

In MARLIN, each signalized intersection (agent) plays a game with all its adjacent intersections in its neighbourhood. The agent has a number of learning modules; each corresponds to one game. The state-space and the action-space are distributed such that the agent learns the joint policy with one of the neighbours at a time following the principle of modular Q-learning.

---

## The Learning Environment

The learning environment component models the traffic environment. At the University of Toronto, Paramics, a microscopic traffic simulator, is used to model traffic environment [33]. Paramics models stochastic vehicle flow by employing speed regulations, car-following, gap acceptance, and overtaking rules. Paramics provides three methods of traffic assignment that could be employed at different levels: “all-or-nothing” assignment, stochastic assignment, and dynamic feedback assignment. In this application a dynamic stochastic traffic assignment was used. Paramics Application Programming Interface (API) functions were used to construct the state, execute the action, and calculate the reward for each signalized intersection.

Parts of the training process of MARLIN are the design of the state definition, action definition, and reward definition. In [34], a comprehensive investigation of these key issues in RL-based signal control for isolated intersections is conducted. The state definition, action definition, and reward definition recommended in [34] and [35] are adopted as follows (for more details on the definitions, please refer to [34]):

### State Definition: Queue length

The agent’s state is represented by a vector of  $2+P$  components, where  $P$  is the number of phases. The first two components are: 1) index of the current green phase, and 2) elapsed time of the current phase. The remaining  $P$  components are the maximum queue lengths associated with each phase.

### Action Definition: Variable Phasing Sequence

The agent is designed to account for variable phasing sequence in which the control action is either to extend the current phase or to switch to any other phase according to the fluctuations in traffic, possibly skipping unnecessary phases. Therefore, this algorithm is an acyclic timing scheme with variable phasing sequence in which not only the cycle length is variable but also the phasing sequence is not predetermined. Hence, the action is the phase that should be in effect next. MARLIN can also implement fixed phase sequence if preferred by the traffic operators.

### Reward Definition: The Reduction in the Total Cumulative Delay

The immediate reward for certain agent is defined as the reduction (saving) in the total cumulative delay associated with that agent, i.e., the difference between the total cumulative delays of two successive decision points. The total cumulative delay at time  $k$  is the summation of the cumulative delay, up to time  $k$ , of all the vehicles that are currently on the intersections’ approaches. If the reward has a positive value, this means that the delay is reduced by this value after executing the selected action. However, a negative reward value indicates that the action results in an increase in the total cumulative delay.

---

## Interface

Interface component manages the interactions between the agent and the simulation environment by exchanging the state, reward, and action. The interaction between the agent and the environment is associated with the following design elements:

- A synchronized interaction between the agent and the environment was designed to ensure that the simulation environment is held while the agent is performing the learning and the decision-making processes and finally produces the action that should be executed by simulation environment. At the same time, the agent should be put on hold until the action is executed in the environment and the resultant state and the reward are measured.
- The system was designed such that the interaction frequency is variable for each agent. The interaction occurs each specified time interval (1 second) as long as the current green for a signalized intersection that is associated with an agent  $i$  exceeds the minimum green time. Otherwise, the interaction starts after the minimum green.

The agent was designed to learn offline through a simulation environment (such as the micro-simulation model employed in the experiments) before field implementation. After convergence to the optimal policy, the agent is ready to be deployed in the field by mapping the measured state of the system to optimal control actions directly using the learned policy, or it can continue learning in the field by starting from the learned policy.

## Experimental Results

### Testbed Network

MARLIN-ATSC has been tested on several simulated networks including a fairly large simulation of the lower downtown Toronto network and two smaller networks in the City of Burlington. The lower downtown of Toronto is the core of the City of Toronto. The network in this study is bounded to the south by the Queens Quay corridor, to the west by the Bathurst Street, to the east by the Don Valley Parkway (DVP) and to the north by Front Street. Toronto is the oldest, densest, most diverse area in the region and its downtown core contains one of the highest concentrations of economic activity in the country.

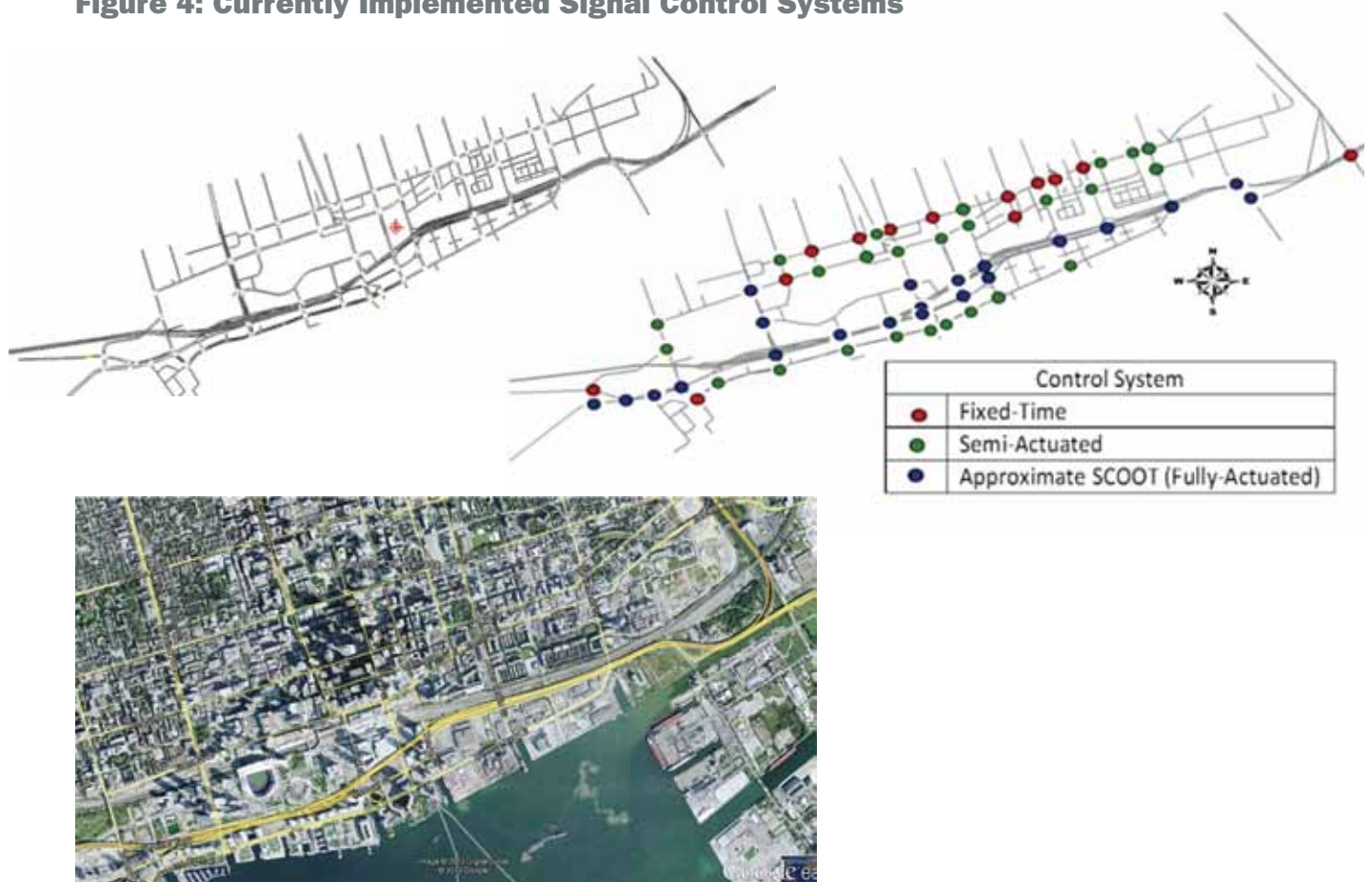
A base-case (BC) simulation model for the lower downtown core was originally developed using Paramics, a microscopic traffic simulator, in the ITS Centre and Testbed at the University of Toronto. In this application, the model is further refined to reflect the signal timing sheets provided by the City of Toronto. The analysis period considered in this application is the AM peak hour, which has around 25,000 vehicular trips. After the very positive results from this test were obtained, the model was further tested on two smaller networks in the City of Burlington, Ontario, around the Walker's Line and Harvester intersection and Guelph Line and Harvester intersection. Both areas

are immediately to the south of the Queen Elizabeth Way (QEW). The analysis period covered the PM peak, which is characterized by heavy traffic leaving office buildings on South Service Road and Harvester, heading towards and interacting with the freeway.

## Benchmarks

It is typically difficult to find a benchmark for large-scale traffic signal control problems, given that the operational details of most traffic control systems are not easily available due to obvious commercial reasons. The performance of MARLIN-ATSC approach is compared to the BC scenario in which traffic signals, as defined and operated by the City of Toronto, are a mix of fixed-time control, semi-actuated control, and SCOOT control, as shown in Figure 4. It is worth noting that due to the limited technical details about the operation of SCOOT, it is approximated in this testing as an enhanced fully actuated control in which loop detectors are placed on all approaches and the extension times are conducted second-by-second.

**Figure 4: Currently Implemented Signal Control Systems**



---

## Toronto Results and Discussion

The results are reported for BC control systems (existing conditions), MARL-I (represents MARLIN-ATSC Independent Mode with no communication between agents), and MARLIN (represents MARLIN-ATSC Integrated Mode with coordination between agents).

The performance of each control system is evaluated based on the following measures of effectiveness (MOEs):

- Average Delay Per Vehicle (sec/veh),
- Average Max. Queue Length Per Intersection (veh), representing traffic backups,
- Average Standard Deviation of Queue Lengths Across Approaches (veh), representing queue balance,
- Number of Completed trips, representing throughput,
- Average CO<sub>2</sub> emissions factors (gm/km), representing environmental impact,
- Average Travel Time for Selected Routes (min), representing a focused look at major corridors as opposed to network wide performance.

Table 1 compares the performance of the BC against the MARLIN-ATSC system with and without communication among agents, i.e, MARLIN and MARL-I, respectively.

The analysis of the results shown in Table 1 leads to the following findings:

- The two MARLIN-ATSC algorithms result in lower average delay, higher throughput, shorter queue length, and stop time compared to those from the base case. The most notable improvements are the average delay (38% MARLIN vs BC), standard deviation of average queue length (31% MARLIN vs BC), CO<sub>2</sub> emission factors (30% MARLIN vs BC).
- These substantial improvements are not only due to the intelligence of the RL algorithm, but also as a result of the coordination mechanism between the agents to reach a network-wide set of actions that minimize delay. This automatic coordination results in the well-known “metering” effect from the upstream intersection to the downstream intersection while accounting for the queues and delays at the downstream intersection. Such metering in practice is often done manually by operators who are experts in the local traffic conditions, who can identify which intersections to meter and why. In the MARLIN case, the system not only automatically discovers the best candidate intersections for metering but also optimizes the extent of metering.
- In fact, the tangible savings in the standard deviation in the queue length is interesting because this means balanced queue among all intersection approaches.

- MARL-I outperforms the BC in all the MOEs, most notably for the average intersection delay (27%) and the CO<sub>2</sub> emission factor (28%). However, comparing MARLIN to MARL-I, it is found that the latter experience relatively higher delays because in MARL-I the actions are based only on locally collected data and thereby results in more vehicles retained in the network at the end of the simulation (6% throughput improvement in MARLIN vs 2.8% throughput improvement in MARL-I).

**Table 1: Network-Wide MOE in the Normal Scenario**

System MOE	BC	MARL-I	MARLIN	% Improvement MARL-I Vs. BC	% Improvement MARLIN Vs. BC	% Improvement MARLIN Vs. MARL-I
Average Intersection Delay (sec/veh)	35.27	25.72	22.02	27.06%	37.57%	14.41%
Throughput (veh)	23084	23732	24482	2.81%	6.06%	3.16%
Avg Queue Length (veh)	8.66	6.60	5.88	23.77%	32.07%	10.88%
Std. Avg. Queue Length (veh)	2.12	1.62	1.47	23.37%	30.74%	9.61%
Avg. Link Delay (sec)	9.45	8.50	5.04	10.07%	46.73%	40.76%
Avg. Link Stop Time (sec)	2.74	2.57	2.02	5.95%	26.06%	21.38%
Avg. Link Travel Time (sec)	16.81	15.81	12.32	5.97%	26.70%	22.05%
CO <sub>2</sub> Emission Factor (gm/km)	587.28	421.34	412.21	28.26%	29.81%	2.17%

To further understand which intersections contribute the most to the above noted savings the spatial distribution of delay of the BC – normal scenario – is plotted in Figure 5. It is interesting to note that some intersections encounter delay in the range of 0–10 sec/veh while others encounter 70–110 sec/veh in the BC scenario.

**Figure 5: Spatial Distribution of Average Delay for the Base Case Normal Scenario**

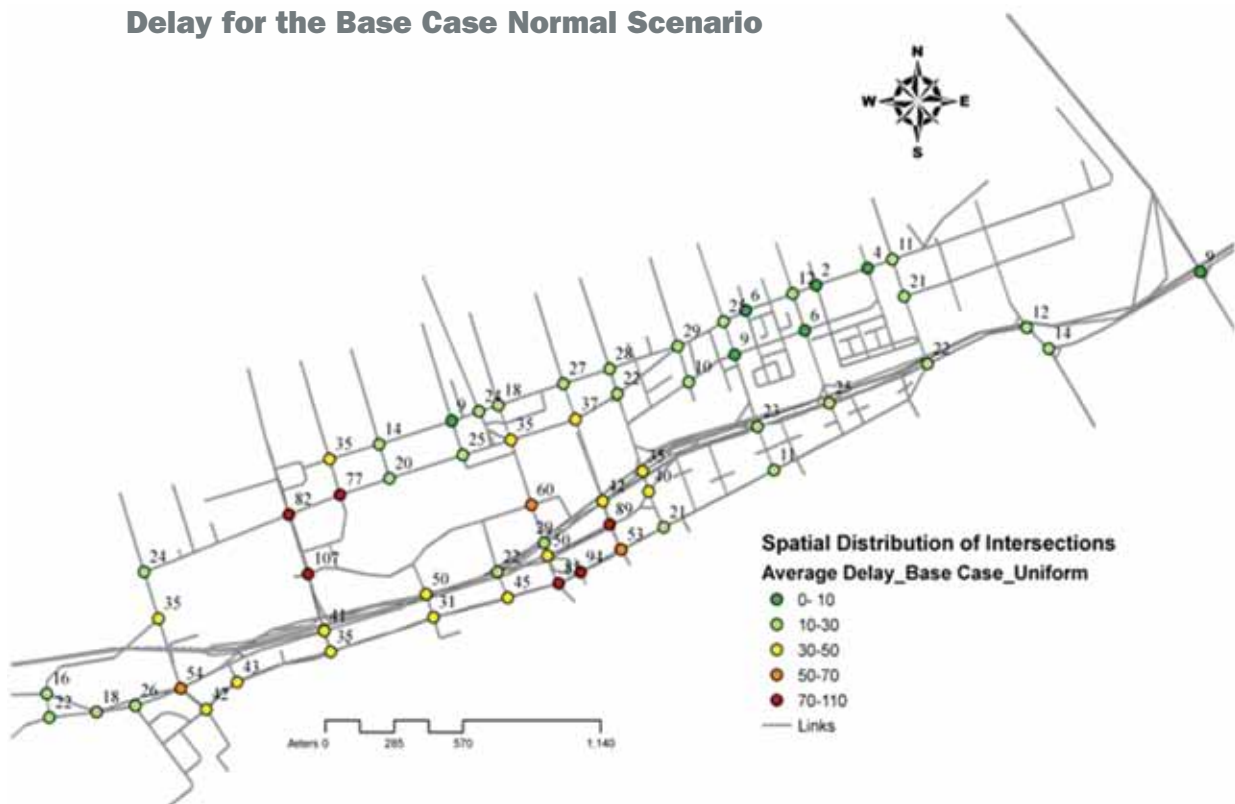


Table 1 shows a very promising overall performance of MARLIN. However, as shown in the wide range of average delays among intersections, the improvements at some intersections are much higher than the network averages. Therefore, the spatial distribution of percentage improvement is presented in Figure 6.

**Figure 6: Spatial Distribution of Average Delay Improvements for MARLIN vs. BC**



---

The analysis of Figure 6 leads to the following conclusions:

- In MARLIN, most of the savings are found at the busiest downtown intersections, but what is interesting to note is the formation of a “corridor of savings” in the MARLIN-IC case as shown in the dotted shaded areas in Figure 4. To illustrate, the intersections along York St. (from Queens Quay St. to Bremner St., shown as Area 1) exhibit substantial savings. This is primarily due to coordination among agents. This observation is extremely important as this is the methodical way to determine which intersections warrant MARLIN-ATSC with communication between agents as opposed to MARLIN-ATSC without communication. A similar coordination is warranted along Yonge St. and to a lesser extent along Spadina Ave. and Jarvis St.
- It is even more interesting to observe the formation of an “area control” effect in some chronic areas in the downtown core. The off-ramps at Spadina St, Yonge St., and York St. form major access points to the City’s core, therefore if traffic is heavily congested at these locations (and their downstream intersections) the entire network could easily deteriorate into a gridlock condition. The MARLIN algorithm automatically identifies and captures such a pattern while controlling these critical intersections, hence the major savings compared to the BC. As a result, the noted Area 1 in Figure 4 warrants coordination among all the intersections within this area. Similarly, Area 2 is chronically congested during the morning peak due to the dense business and economic activities near Front St., Wellington St., York St., and Simcoe St. In addition, in the west end of the network, LS (LakeShore) and Fort York St. are two alternative routes for a high number of trips destined to Bathurst St., which is also observed to be congested in the morning peak. This area (Area 3) exhibits considerable savings that warrant “area control” using MARLIN.
- It is worth noting that there are a few intersections that exhibit substantial savings (81%), such as the case of Front St. and Princess St.; however this saving is not as important because the BC delay is observed to be minimal (2 sec/veh).

It is important to study the effect of various control systems on the travel time and travel time variability for selected key routes in the lower downtown core of Toronto. Selected routes are defined and analyzed (Table 2 and Figure 7).

Route travel times and standard deviation in travel time for the BC, MARL-I, and MARLIN scenarios are presented in Table 2. The routes in Table 2 are arranged in descending order from the worst to the best in terms of % improvements in average route time for MARLIN vs. BC. To further study the route travel times within the simulation hour, the travel times for selected routes are plotted in Figure 5. The analysis of Table 2 and Figure 5 leads to the following conclusions:

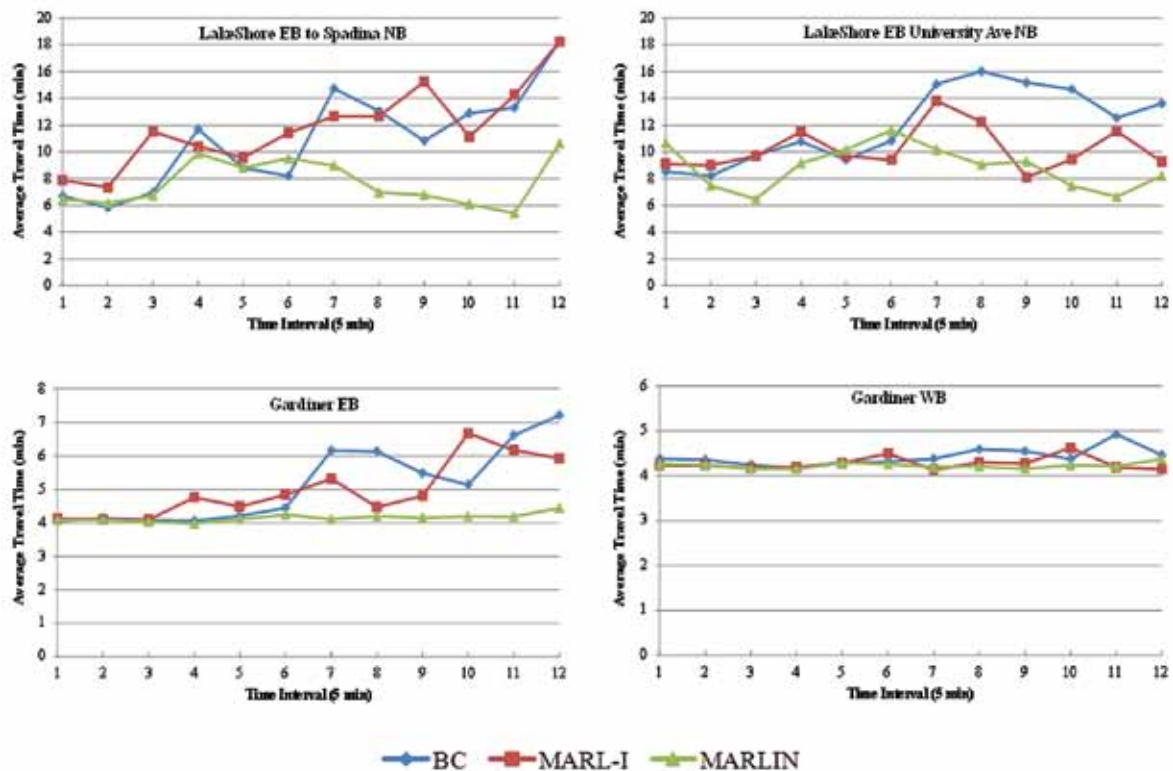
- It is clear that MARLIN outperforms MARL-I and BC in all routes. The % improvements range from 4% in route 1 to 30% in route 8. MARL-I outperforms BC in almost all cases; the % improvements range from 3% in route 5 to 15% in route 6.

- It is interesting to find that the Gardiner Expressway eastbound traffic (inbound) travel time improves by 19% in the MARLIN scenario. Alleviating the congestion on Spadina St. and York St off-ramps contributes the most to these savings. This clearly shows the effect of the downstream capacity on the freeway performance. For the Gardiner westbound direction traffic was not as congested as the eastbound but MARLIN still attains 4% improvement in average route travel times.
- The most congested routes appear to be routes 7 and 8; through which traffic originated at the west end of the study area and destined in the downtown core (Spadina St and University Ave). MARLIN achieves 26% and 30% improvements in route 7 and 8, respectively, which reflects the superior effect of the two-dimensional coordination between agents.
- From observing the temporal distribution of route travel time across the simulation hour, it is generally found that MARLIN is stable and exhibits less variation compared to the BC and MARL-I scenarios. While the BC scenario exhibits the highest variability in travel time (as shown in the standard deviation values in Table 2), MARL-I still shows some variations, most notably in the most two congested routes (route 7, 8). MARLIN shows stable route travel times in all routes.

**Table 2: Route Travel Times for BC, MARL-I, and MARLIN**

System Route	BC	MARL-I	MARLIN	% Improvement MARL-I Vs. BC	% Improvement MARLIN Vs. BC	% Improvement MARLIN Vs. MARL-I
1- Gardiner EB	5.14	4.98	4.15	3.18%	19.30%	16.65%
St Dev	1.15	0.86	0.12	24.59%	89.70%	86.34%
2- Gardiner WB	4.42	4.27	4.23	3.35%	4.35%	1.04%
St Dev	0.20	0.15	0.06	26.52%	68.03%	56.49%
3- Front EB	10.65	9.13	7.88	14.28%	13.69%	13.69%
St Dev	2.15	1.22	0.60	43.26%	72.27%	51.13%
4- Front WB	5.55	5.34	5.10	3.81%	8.15%	4.51%
St Dev	0.92	0.79	0.49	13.39%	47.10%	38.93%
5- LakeShore EB	16.31	13.28	12.10	18.60%	25.77%	8.82%
St Dev	3.74	1.37	1.37	63.38%	63.49%	0.31%
6- LakeShore WB	10.31	9.07	8.46	12.02%	17.91%	6.70%
St Dev	1.03	0.69	0.50	33.30%	51.09%	26.67%
7- LakeShore EB to Spadina NB	10.94	11.86	7.70	-8.40%	29.59%	35.05%
St Dev	3.75	3.07	1.75	18.25%	53.44%	43.05%
8- LakeShore EB University to Ave NB	12.05	10.24	8.87	15.04%	26.38%	13.36%
St Dev	2.81	1.66	1.64	40.80%	41.75%	1.62%

**Figure 7: Average Route Travel Time for Selected Routes**



## Burlington Results and Discussion

The Queen Elizabeth Way (QEW) freeway runs through the City of Burlington resulting in significant traffic flow destined to the City's arterial streets (see Figure 8). Arterial streets and service roads in the City are very close to the freeway on and off-ramps creating excessive long queues during the peak periods because of the high traffic volume approaching the freeway. The most chronic intersection in the City is Harvester Road and Walkers Line as it experiences excessive delays and queue spillback that often block the upstream intersections.

The subject study area of this test is therefore the Walkers Line and Harvester Intersection and the neighbouring intersections. Modeling only one intersection could lead to misleading results, as the operation of one intersection is affected by flow at upstream intersections and congestion at downstream intersections. Therefore, a wider footprint is considered as shown in Figure 8 to include six signalized intersections. The queue spillback described above is more vivid during the afternoon peak period (4 PM to 5 PM) as traffic from most businesses and office buildings in the area is accessing the freeway eastbound / westbound directions at that time.

**Figure 8: Study Area**



Five scenarios were selected for an in-depth analysis. These scenarios were intended to evaluate the performance of various traffic signal control and supply management strategies. These scenarios were compared with the base case model – that is based on actuated traffic signal control – for the PM peak period.

#### **Base Case Model with Actuated Signal Control**

The base case model involved the 2012 existing conditions for the PM peak period. The model developed consisted of six signalized intersections. All intersections operate under a standard NEMA actuated controller with stopline and extension loop detectors located at the intersection approaches.

#### **Adaptive Traffic Control Strategies**

As mentioned earlier, two ATSC strategies were considered in this research:

- 1) Independent Mode (MARL-I);
- 2) Integrated Mode (MARLIN).

#### **Supply Management Strategies**

One of the congestion mitigation options considered by the City of Burlington is the expansion of the Harvester-Walkers Line intersection by adding left turn and right turn lanes. This expansion is costly to implement and therefore it is important to assess the potential improvements before costly deployment. In this case, two expansions were modeled:

- 1) Addition of one left turn lane to the existing eastbound left turn lane (hereafter called the double left turn strategy);
- 2) Full intersection expansion by adding a lane in each approach of the intersection in anticipation of traffic growth in the future.

---

Due to the stochastic nature of traffic, each simulation run can be regarded as a random experiment, i.e. a random day in real life. Therefore, 10 simulations were completed for each scenario, including the base case model. The final results, averaged over the multiple runs, were reported.

### **Observations and Analytical Results**

The findings of the experimental setup are summarized here.

- The base case model experienced significant delays and queues at the intersection of Walkers Line and Harvester resulting in queue spill back that blocks the upstream intersections. The average delay at that intersection was estimated to be 90 sec / veh. This observation was confirmed through a field visit. The City is therefore considering a multi-million dollar expansion project to alleviate the congestion and delays at that particular intersection.
- MARLIN outperformed the base case model by an average of 76% in travel time, 93% in average speed, and 30% in remaining number of vehicles in the simulation model at the end of the hour.
- MARLIN outperformed MARL-I by a range of 4% in average travel time, 5% in average speed, and 4% in the remaining number of vehicles, which indicates that network coordination could potentially result in more savings.
- The results showed that the overall network performance under MARLIN cases outperformed the supply management strategies by an average of 3% in travel time, 3% in average travel speed, and 4% in remaining number of vehicles in the simulation model.
- In the route-level results, MARLIN was found more stable as it minimized the variation in travel time when compared to the full expansion case.

In conclusion, the new self-learning MARLIN system was found a viable alternative to costly expansion. MARLIN achieves better overall network performance at a fraction of the cost of the full expansion.

### **Conclusions and Future Work**

In this chapter, we highlighted the state-of-the-art and state-of-the-practice in Adaptive Traffic Signal Control (ATSC) and its limitations. We also introduced the latest ATSC system from the University of Toronto's ITS Centre and Testbed and demonstrated the performance of MARLIN-ATSC on a large-scale urban network of 59 intersections in downtown Toronto and smaller networks in the City of Burlington. We reported results for base case control systems (represented existing field conditions using signal timing sheets provided by municipalities, MARL-I (represented MARLIN-ATSC Independent Mode with no communication between agents), and MARLIN (represented MARLIN-ATSC Integrated Mode with coordination between agents). Results showed that MARL-I and MARLIN

significantly outperformed the BC in all the MOEs. MARLIN, due to explicit coordination amongst neighbouring intersections, outperformed MARL-I. In terms of route travel time, it was generally found that MARLIN exhibited less average route travel time and less variation of the temporal distribution across the simulation hour compared to the BC and MARL-I scenarios. The daily economic benefits (i.e., travel time savings) were estimated to be around \$53,000. MARLIN-ATSC would cost approximately \$1.2 million to implement across a network of 59 intersections. Consequently, the payback period is 23 days.

It is logical to ask how to take the promising MARLIN-ATSC system from a perfectly controlled lab environment to the real world. There are two essential steps to be taken before widespread deployment. The first step is to integrate the MARLIN control software with physical field controllers (hardware) available in the market today and commonly used by most municipalities. The second step is to conduct a field operational test at one or a few intersections.

To prepare MARLIN for field implementation, the University of Toronto is currently integrating the MARLIN control software system with one of the most advanced controllers (hardware) in the market, PEEK ATC1000, which is widely used in North America and in Toronto in particular. Both the MARLIN software and the PEEK hardware are being tested using hardware-in-the-loop simulation (HILS) methodologies. The HILS mimics field operation by integrating MARLIN with the actual field controller using common (NEMA and NTCIP) standards and protocols. The integrated software/hardware system is connected to a high fidelity traffic simulation platform (e.g., Paramics) for testing under varying traffic conditions and network configurations, in a realistic field-like setup as shown in Figure 9. It is noteworthy that this HILS testing does not change the findings presented above. It is only intended to assure seamless integration of the control system components using industry standards and off-the-shelf hardware. The final step is to conduct a field operational test at one or a few intersections in collaboration with a willing municipality.

**Figure 9: MARLIN-HILS Architecture**



## Chapter I References

---

- [1] W. R. McShane, R. P. Roess, and E. S. Prassas, *Traffic engineering*: Prentice Hall, 1998.
- [2] P. B. Hunt, D. I. Robertson, R. D. Bretherton, and R. I. Winton, "SCOOT-a traffic responsive method of coordinating signals," Technical Report, Transport and Road Research Laboratory, Crowthorne, England, 1981.
- [3] C. Diakaki, M. Papageorgiou, and K. Aboudolas, "A multivariable regulator approach to traffic responsive network-wide signal control," *Control Engineering Practice*, vol. 10, pp. 183-195, 2002.
- [4] A. G. Sims and K. W. Dobinson, "SCAT-The Sydney Co-ordinated Adaptive Traffic System-Philosophy and Benefits," presented at International Symposium on Traffic Control Systems, 1979.
- [5] J. L. Farges, J. J. Henry, and J. Tufal, "The PRODYN real-time traffic algorithm," presented at The 4th IFAC/IFIP/IFORS Symposium on Control in Transportation Systems, Baden-Baden, Germany, 1983.
- [6] N. H. Gartner, "OPAC: A demand-responsive strategy for traffic signal control," *Transportation Research Record: Journal of the Transportation Research Board*, vol. 906, pp. 75-81, 1983.
- [7] K. L. Head, P. B. Mirchandani, and D. Sheppard, "Hierarchical framework for real-time traffic control," *Transportation Research Record*, vol. 1360, pp. 82-88, 1992.
- [8] A. L. C. Bazzan, "Opportunities for multiagent systems and multiagent reinforcement learning in traffic control," *Autonomous Agents and Multi-Agent Systems*, vol. 3, pp. 342-375, 2009.
- [9] B. Abdulhai and L. Kattan, "Reinforcement learning: Introduction to theory and potential for transport applications," *Canadian Journal of Civil Engineering*, vol. 30, pp. 981-991, 2003.
- [10] S. El-Tantawy and B. Abdulhai, "An agent-based learning towards decentralized and coordinated traffic signal control," presented at 13th International IEEE Conference on Intelligent Transportation Systems (ITSC), 2010.
- [11] S. El-Tantawy and B. Abdulhai, "Towards multi-agent reinforcement learning for integrated network of optimal traffic controllers (MARLIN-OTC)," *Transportation Letters: The International Journal of Transportation Research*, vol. 2, pp. 89-110, 2010.
- [12] C. Watkins and P. Dayan, "Q-learning," *Machine learning*, vol. 8, pp. 279-292, 1992.
- [13] R. S. Sutton and A. G. Barto, *Introduction to reinforcement learning*. Cambridge Mass.: MIT Press, 1998.
- [14] T. Basar and G. J. Olsder, *Dynamic Noncooperative Game Theory*, 2nd ed. London, U.K: Classics in Applied Mathematics, 1999.
- [15] L. Busoniu, R. Babuska, and B. De Schutter, "A comprehensive survey of multiagent reinforcement learning," *IEEE Transactions on Systems, Man and Cybernetics*, vol. 38, pp. 156-172, 2008.
- [16] C. Claus and C. Boutilier, "The dynamics of reinforcement learning in cooperative multiagent systems," presented at The 15th National Conference on Artificial Intelligence and 10th Conference on Innovative Applications of Artificial Intelligence, Madison, US, 1998.
- [17] M. Weinberg and J. S. Rosenschein, "Best-response multiagent learning in non-stationary environments," presented at The 3rd International Joint Conference on Autonomous Agents and Multiagent Systems, 2004.
- [18] T. Thorpe, "Vehicle traffic light control using sarsa," Master's Project Report, Computer Science Department, Colorado State University, Fort Collins, Colorado, 1997.

- 
- [19] M. Wiering, "Multi-agent reinforcement learning for traffic light control," presented at The 17th International Conference on Machine Learning 2000.
- [20] B. Abdulhai, R. Pringle, and G. J. Karakoulas, "Reinforcement Learning for True Adaptive Traffic Signal Control," *Journal of Transportation Engineering*, vol. 129, pp. 278-285, 2003.
- [21] E. Camponogara and W. Kraus Jr, "Distributed learning agents in urban traffic control," presented at The 11th Portuguese Conference on Artificial Intelligence, 2003.
- [22] D. De Oliveira, A. L. C. Bazzan, B. C. da Silva, E. W. Basso, L. Nunes, R. Rossetti, E. de Oliveira, R. da Silva, and L. Lamb, "Reinforcement Learning-based Control of Traffic Lights in Non-stationary Environments: A Case Study in a Microscopic Simulator," presented at EUMAS06, 2006.
- [23] S. Richter, D. Aberdeen, and J. Yu, "Natural actor-critic for road traffic optimisation," in *Advances in Neural Information Processing Systems*, vol. 19. Cambridge: MIT Press, 2007.
- [24] I. Arel, C. Liu, T. Urbanik, and A. G. Kohls, "Reinforcement learning-based multi-agent system for network traffic signal control," *IET Intelligent Transport Systems*, vol. 4, pp. 128-135, 2010.
- [25] T. Li, D. B. Zhao, J. Q. Yi, and Ieee, "Adaptive Dynamic Programming for Multi-intersections Traffic Signal Intelligent Control," presented at 11th International IEEE Conference on Intelligent Transportation Systems, 2008.
- [26] A. Salkham, R. Cunningham, A. Garg, and V. Cahill, "A collaborative reinforcement learning approach to urban traffic control optimization," presented at IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology, 2008.
- [27] J. C. Medina and R. F. Benekohal, "Q-learning and Approximate Dynamic Programming for Traffic Control – A Case Study for an Oversaturated Network," presented at Transportation Research Board Annual Meeting, 2012.
- [28] L. Shoufeng, L. Ximin, and D. Shiqiang, "Q-Learning for Adaptive Traffic Signal Control Based on Delay Minimization Strategy," presented at IEEE International Conference on Networking, Sensing and Control, 2008.
- [29] L. Kuyer, S. Whiteson, B. Bakker, and N. Vlassis, "Multiagent reinforcement learning for urban traffic control using coordination graph," presented at The 19th European Conference on Machine Learning, 2008.
- [30] A. L. C. Bazzan, "A distributed approach for coordination of traffic signal agents," *Autonomous Agents and Multi-Agent Systems*, vol. 10, pp. 131-164, 2005.
- [31] R. Nair, P. Varakantham, M. Tambe, and M. Yokoo, "Networked distributed POMDPs: A synthesis of Distributed Constraint Optimization and POMDPs," presented at The 20th National Conference on Artificial Intelligence, 2005.
- [32] N. Ono and K. Fukumoto, "Multi-agent reinforcement learning: A modular approach," presented at The Second International Conference in Multi-Agent Systems, 1996.
- [33] Quadstone Paramics, "Paramics Microscopic Traffic Simulation Software ": <http://www.paramics-online.com>, 2012.
- [34] S. El-Tantawy and B. Abdulhai, "Comprehensive Analysis of Reinforcement Learning Methods and Parameters for Adaptive Traffic Signal Control," presented at Transportation Research Board, Washington D.C., 2011.
- [35] S. El-Tantawy and B. Abdulhai, "Neighborhood Coordination-based Multi-Agent Reinforcement Learning for Coordinated Adaptive Traffic Signal Control," presented at Transportation Research Board Washington D.C., 2012.

# Chapter II: OSTE

**Optimal Spatio-Temporal  
Evacuation of Large Cities**

Abdulhai, B., and Abdelgawad, H.

---

## **Severe Congestion: Just Add Chaos!**

Large cities in Canada, as all large cities in the world, are badly congested. In the event of major natural and/or man-made disaster, high levels of congestion could mean the difference between life and death. Planning for disaster includes finding ways to move large masses of people to safety.

Thankfully, Canada has not been subject to numerous severe natural and man-made disasters, especially compared to the United States, for instance, and more recently, Japan. However, disasters have occurred that required the consideration of mass evacuation such as the flooding of the Red River in Manitoba, and the 1979 train derailment in Mississauga that caused the release of chlorine gas. In addition, Canada increasingly relies upon nuclear energy and there are several nuclear plants in close proximity to populated areas. While nuclear mishaps are rare, the impact has the potential to be deadly and disaster management strategies require rigorous planning.

Transportation systems evolve with city growth, coping with the daily activity and travel patterns of people. In large-scale emergency situations, the picture is completely different. There is sudden and extreme surge in demand, infrastructure and transportation supply reductions, extreme time pressure, and, of course, panic. Moreover, the transportation demand patterns that arise during emergencies are not just the extreme versions of the daily patterns; they are totally different. The prime goal is to get people to safety, away from the core of the emergency event. In these instances, demand patterns and the transportation system suddenly do not match, further aggravating the risk.

On one hand, immediate simultaneous evacuation (the run-for-your-life scenario) results in wide-scale gridlock. On the other hand, if we knew of the event way in advance, we can take our time moving evacuees slowly but surely to safe alternatives, with reduced congestion and no panic, but obviously this is very rarely the case in emergencies. Between the extremes of immediate simultaneous evacuation and the rare luxury of very lengthy and relaxed evacuation, there is an optimum. Science and technology, as well as stakeholder coordination, could provide the possibility to move the largest number of people, as quickly as possible, to safe destinations, in a controlled manner.

Many large Canadian cities are currently rethinking their transportation systems. With a lack of transportation infrastructure investment, governments at all levels are in the process of figuring out how to catch up, often taking into consideration daily congestion. In the process of planning, it is important to keep emergency evacuation in mind as well. In Toronto, for example, the future of the Gardiner Expressway is unsettled. Whether it stays, gets replaced, gets buried or is simply removed, the options will be subject to rigorous planning, modeling and assessment before the final solution is chosen. Stakeholders should also think of emergency evacuation. Do we have enough arteries leading in and out of Toronto?

---

This chapter presents a novel framework that optimizes the evacuation of large cities using multiple modes, including vehicular traffic and mass transit shuttle buses.

A large-scale evacuation system is developed and tested at the University of Toronto on a mock evacuation of the entire City of Toronto under a hypothetical emergency city-wide evacuation scenario. A demand estimation process is first designed to accurately quantify the evacuation demand by mode (drivers vs. transit users), over time of the day when the crisis begins and over space (location). The output of the demand estimation process is then fed into two optimization platforms: (1) an Optimal Spatio-Temporal Evacuation (OSTE) system that synergizes evacuation scheduling, route choice, and destination choice for vehicular traffic and (2) a mass-transit management system to optimize the routing and scheduling of all available mass-transit vehicles.

The testing concluded that the OSTE system can clear the City of Toronto four times faster than having no optimized management. The OSTE average automobile evacuation time for the 1.21 million people using their cars is close to two hours. The optimization of the routing and scheduling of the readily available Toronto Transit Commission (TTC) fleet (four Rapid Transit lines and 1,320 transit buses used as shuttles) can efficiently evacuate the transit-dependent population (1.34 million) within two hours. The optimized evacuation system can get the last person out of Toronto in eight hours, compared to 30 hours without optimization. These figures are based on the assumption that 80% of the population will cooperate (comply) with recommendations, while 20% will ignore the evacuation orders. The level of compliance, although important, is not our main point, however. Public education increases the odds of compliance and compliance also increases with the public's confidence in the provided guidance. Rather, the main point is that the value of rigorous and scientifically sound guidance cannot be overstated.

## **Evacuation 101**

Planning for emergency evacuation in densely populated cities is challenging. The core objective is to move people from potential hazard zones to safe destinations in the quickest and most efficient way. Given the drastic consequences of large-scale emergencies, proper development of emergency evacuation plans is paramount. Also, given the typical diverse demographic characteristics of most communities, an efficient evacuation strategy should integrate multiple modes to particularly aid transit-dependent people who have no access to automobiles at the time of evacuation or at all.

Numerous notable emergency evacuation planning models have been developed over the past few decades. They propose and investigate the effect of one or more strategies that have the potential to improve the performance of the evacuation process. Approaches in the literature use various modeling and optimization techniques. However, most of these systems are typically focused on automobile-based evacuation using a certain strategy (e.g.

---

evacuation scheduling) without considering other modes of transportation or attempting to simultaneously synergize several other possible strategies such as destination choice optimization, route selection optimization, etc. In addition, the absence of accurate representation of the spatio-temporal distribution of the population makes estimating evacuation demand difficult and hence renders the interpretation and the conclusions of the methods themselves questionable. Furthermore, the lack of quantifying mode-specific populations (e.g. transit-dependent) magnifies the vulnerability of those people to threats in cases of large-scale emergencies, a major drawback of existing planning models.

In this chapter we present the development of a large-scale multimodal evacuation planning system that is fast enough to be usable in real time. We test the system using the evacuation of the entire City of Toronto under hypothetical large-scale emergencies as a case study. Of particular importance in this study: 1) optimization of multimodal evacuation using global and efficient optimization techniques while assessing multiple objectives, and 2) accurate estimation of the spatial and temporal distribution of the population by mode of travel to better estimate evacuation demand.

The remainder of this chapter is structured in three main parts. The first part discusses the relevant state-of-the-art, background and the motivation of developing our system. In the second part, we present the details of the multimodal evacuation optimization methods. In the third part, we apply the framework to the evacuation of the City of Toronto. Finally, we present a discussion and analysis of the results alongside the major findings.

## **State-of-the-art Scan**

Transportation networks in urban areas evolve over long time spans. The transportation infrastructure is typically designed to cope with recurrent daily traffic and is not designed to cope with unexpected menaces that may cause sudden surges in travel patterns. Due to the frequent occurrence of natural and man-made disasters, the performance of the transportation infrastructure under such extreme events received significant attention over the past decade. Typically, major cities and urban centres are congested and operating at or near capacity; therefore network performance can severely deteriorate if drastic changes in Origin-Destination (O-D) demand patterns and/or significant loss of capacity occur during or after a disaster (Tuydes and Ziliaskopoulos, 2004). In case of a disaster or hazardous event, the ultimate goal is to optimize the utilization of the existing infrastructure through traffic management centres and a myriad of emergency response resources. Emergency operation centres face multi-faceted challenges to anticipate traffic flows under emergencies, to provide proactive actions to coordinate the efforts of relevant stakeholder, and to guide the public to safety (Chiu et al., 2006).

Designing a transportation network for extreme rare events and related demand patterns is indeed financially infeasible. A better alternative is to utilize the available capacity and

---

resources more efficiently using systems approaches (Theodoulou and Wolshon, 2004; Wolshon et al., 2005; Dixit and Radwan, 2009). Numerous investigations examined several evacuation management strategies that could potentially expedite the evacuation process. The following sections summarize the most relevant strategies in the planning for emergency evacuation.

### **Evacuation Scheduling**

Scheduled evacuation is a widely used control strategy to guide evacuation flows. Evacuation scheduling aims to better distribute the evacuation demand over the evacuation horizon (time). In simultaneous evacuation, evacuees are advised to evacuate immediately to their destinations; whereas in staged evacuation, evacuees are advised when to evacuate so as to minimize the network clearance time (Sbayti and Mahmassani, 2006). It has been proven that controlling the release of evacuees to the network can expedite the evacuation process (Abdelgawad and Abdulhai, 2009). However, effective approaches to obtaining such starting times during a staged evacuation have not been adequately addressed in the literature.

Investigating the effectiveness of staging evacuation, Chen and Zhan (2006) concluded that staging the evacuation is essential in communities where the street networks have a “Manhattan structure” and the population density is high. Mitchell and Radwan (2006) proposed a heuristic prioritization of emergency evacuation to study the effect of some zone-based parameters (population density, road exit locations/capacity, and major evacuation routes) that might affect the staging decision. Chiu et al. (2006) applied a system optimal dynamic traffic Cell Transmission Model (CTM) to a simple evacuation event to solve the evacuation destination-route-flow-staging problem for non-notice emergency events. Although the demand scheduling problem incorporates many pressing issues such as how and what to optimize, how to perform dynamic traffic assignment, and how, when and what to advise evacuees, little research has been conducted to simultaneously combine and analyze these issues. The investigations by Sbayti and Mahmassani (2006) and Abdelgawad and Abdulhai (2009) are steps in this direction.

### **Destination Choice**

In typical evacuation planning models, evacuees are assigned to pre-determined destinations that are based primarily on the geographical context and their daily activities. However, this may result in sub-optimal solution due to congestion, road blockage, chaos, incidents, hazards associated with the emergency situation, and limited destination/shelter capacity. One promising concept is to relax the constraint of assigning evacuees to pre-fixed destinations. In other words, instead of assigning the evacuation demand to pre-fixed destinations, evacuees are directed to the nearest and

---

fastest-to-reach safe destinations outside of the impacted area. This can be achieved, from a modeling perspective, by directing evacuees to one amalgamated super destination beyond the existing destinations and let the optimization approach find the shortest route to this super destination. Chiu et al. (2006) and Yuan et al. (2006) proposed the One-Destination evacuation model in which the traditional road network with  $m$  origins to  $n$  destinations has been modified to a network with  $m$  origins to one destination. Yuan et al. (2006) reported that a reduction of approx. 60% in the overall evacuation time can be achieved in their regional evacuation case study of a nuclear power plant mishap, and a reduction of 80% in the overall evacuation time when modeling traffic routing and en route information accompanied by the one destination framework.

### **Traffic Routing and Control Strategies**

Traffic routing, as one of the main control efforts, aims to identify the best set of routing decisions so as to fully utilize the available capacity of an evacuation network. User equilibrium (UE) conditions mean that drivers follow time-dependent least travel time paths, while system optimum (SO) conditions result from drivers following time-dependent least marginal travel time paths (least total travel time increase caused by an extra trip). Simulation-based approaches are used to perform UE, SO, or multiple user classes assignment, though the equilibrium conditions are only heuristically approximated (Peeta and Ziliaskopoulos, 2001).

From this perspective, there are two schools of thought in dynamic traffic assignment (DTA) during evacuation. One group argues that in emergency evacuations, the major concern of planners is the overall system performance; therefore, it is more plausible to use system-optimal traffic assignment that minimizes total evacuation time. Among these studies is the work proposed by Sbayti and Mahmassani (2006) on evacuation scheduling and by Han and Yuan (2005) who proposed a one destination approach for emergency evacuation assignment using SO DTA. Conversely, the other group argues that reaching SO in actual evacuation operations is neither practical nor equitable. Travellers act individually and selfishly to minimize their own travel times (Brown et al., 2009). During emergency situations it is expected that selfish behaviour will increase among evacuees and lead towards user-optimal decisions, a concern that motivated the work conducted by Chiu and Mirchandani (2008) on dynamic traffic management for evacuation. They quantified the system performance in case of emergency evacuation. The study compared the results from the route choice decisions made by evacuees and the SO real-time pre-trip route guidance scheme and, based on feedback from observed traffic, a controller was designed to influence the system performance towards an optimum level. The results showed degradation in the system performance, which was postulated to occur due to evacuees' deviation from optimal paths.

---

## OSTE Motivation

Despite the numerous approaches (e.g. evacuation scheduling, destination choice, route choice, etc.) that have significantly contributed to expediting the evacuation process, we are still lacking an integrated optimal evacuation planning tool. More effort is needed to synergistically combine all or some of these promising strategies to further improve the efficiency of the evacuation process. Traffic route guidance, evacuation staging, and destination optimization can be combined into a single comprehensive solution. Also, an accurate description of the spatial distribution of population, by time of day and mode of travel is essential to realistically address major population evacuation. Unlike day-to-day travel patterns, planning for emergency evacuation has unique demand distribution that should be carefully examined to produce accurate evacuation performance measures (Murray-Tuite and Mahmassani, 2003; Wilmot and Mei, 2004).

Furthermore, automobile evacuation has received the most attention; consequently multimodal evacuation is still largely missing in most emergency evacuation studies (TRB, 2008). A significant portion of the population in cities like Toronto use public transit particularly within, towards, and out of the downtown core. This portion of the population does not have access to automobiles during the day or at all. Utilizing the readily available transit capacity is therefore essential to not only serve the transit captives but to also improve the evacuation process and reduce network clearance time by moving people in masses. For example, a single bus-only highway lane can carry up to six times the passengers as that of a passenger car-only highway lane (Litman, 2006). In addition, standard buses, LRT, and Rapid Transit (subway or metro) can carry up to 5,400, 28,800, and 72,000 passengers/hour respectively (Vuchic, 2005). Therefore, transit services afford huge capacity that can significantly reduce the clearance time in cases of evacuation. New schedules and routes need to be optimized however in the case of emergency evacuation.

In summary, the following elements are found essential to realistically plan for emergency evacuation:

- Accurate assessment and representation of the transportation infrastructure, most notably the roadway and the public transit networks (transportation supply).
- Accurate estimation of the spatial and temporal distribution of population (transportation demand).
- Accurate identification of available modes and captive population to certain modes.
- Integrated framework that accounts for various evacuation strategies such as evacuation scheduling, route choice, and destination choice.
- Multimodal evacuation strategies that synergize the effect of multiple modes.

- 
- Accounting for background traffic or noncompliant evacuees (the percentage of travellers not following the evacuation plan).
  - Robust and extensible optimization and solution algorithms that can tackle such multi-dimensional non-deterministic problem.

While the aforementioned elements constitute separate modeling, analysis, optimization and operational tasks, they are closely interrelated. Each is indispensable for the design and implementation of an effective emergency evacuation plan. The University of Toronto OSTE is geared to amalgamate all these elements into one framework that is designed to plan for realistic large-scale multimodal emergency evacuations. It is worthwhile to emphasize that the major point in our approach is not which evacuation tool works best. The point is how to employ all of them together in a methodical, scientifically sound, and computationally feasible way to maximize their combined effect.

The following section details the approaches to model and solve such complex and multi-dimensional problem.

## **OSTE: Framework for Optimizing Multimodal Evacuation**

### **Optimal Spatio-Temporal Evacuation Model**

OSTE can be achieved through optimizing the evacuation scheduling and the destination choice simultaneously (Abdelgawad and Abdulhai, 2009). The concept of synergizing evacuation scheduling and destination choice aims to spread the evacuation demand over time and space respectively in an optimal manner, i.e. optimize the spatio-temporal demand pattern. While condensing demand release in time and space causes gridlock, spreading demand over a too-long time period unnecessarily extends the evacuation process, which defeats the purpose of expedient evacuation. Not only representing the temporal patterns of mobilization and evacuation, i.e. the mobilization/loading curve,  $L(t)$  and evacuation curve,  $E(t)$  is paramount to this approach, but also captures the dynamic interaction between both curves. Therefore, an optimal spatio-temporal plan is sought. The output of the OSTE system is simple yet effective traveller information and guidance for evacuees regarding 1) when to evacuate (schedule), 2) where to go (destination choice) and, 3) how to get there (route choice).

The OSTE platform is built on the interaction between Dynamic Traffic Assignment (DTA) and Evolutionary Algorithms (EA). The decision variables are the staging percentages in the vector  $\mu=(\dots, \mu_t, \dots)$  with a modified super-destination network representation to achieve optimal destination choice as a by-product of the DTA process. In this approach, the problem is formulated as a multi-objective optimization problem in which both the waiting time before entering the network (mobilizing) and travel time within the network are simultaneously minimized. This dual objective is different from

---

the typical minimization of vehicle travel time or maximizing the number of evacuees reaching safe destination within a preset evacuation time horizon as in Abdelgawad and Abdulhai (2009).

OSTE utilizes Genetic Algorithms (GA) as the core optimization method. GA overcome many of the problems coupled with the traditional deterministic optimization methods (Bethke, 1976; Kruchten et al., 2004). They start the search from a population of initial solutions, and not from a single point. Therefore, the odds of finding the global optima without entrapping in local minima are higher than in most conventional approaches, and they do not require differentiation of the objective function. Moreover, GA are inherently parallelizable, allowing the power of several computers or CPUs to be harnessed using High Performance Computing (HPC) clusters.

OSTE uses and expands on GENOTRANS (Generic Parallel Genetic Algorithms Framework for Optimizing Intelligent Transportation Systems) developed at the University of Toronto (Mohamed, 2007). GENOTRANS is a GA platform in which the objective function is evaluated and constraints are satisfied through a simulation model. The simulation model replicates the transportation network and performs DTA. Vehicles are loaded onto the network according to a temporal profile set by the GA, and navigate through the network towards their destination(s). The details of the parallel distributed GA approach utilized in OSTE are reported in Abdelgawad and Abdulhai (2009).

### **Optimal Routing and Scheduling of Mass Transit Vehicles**

The aim of the transit module is to cast and solve the mass transit evacuation problem with multi-objective optimization and constraints techniques. It demonstrates how mass-transit systems can be utilized to evacuate carless (transit-dependent) population in no-notice evacuation events—an attempt to fill the gap in the state-of-the-art of multimodal evacuation. The problem is formulated as a variant of the well-known Vehicle Routing Problem (VRP) to include (i) Multiple Depots (MD) to account for the dispersed presence of transit vehicles and to account for the availability of different types of buses at different depots such as municipal transit bus depots, commuter bus depot, school buses, etc., (ii) Time Constraints (TC) to ensure optimal use of the transit vehicles within the evacuation time window, and (iii) Multiple Pick-up and Delivery (PD) locations for evacuees to allow for picking up evacuees from dispersed stops to avoid excessive walk distances. Abdelgawad et al. (2010) describes the technical details of the transit evacuation system.

In our approach, we explicitly minimize the (1) in-vehicle travel time and (2) the waiting time for evacuees. Minimizing transit time only may cause excessive wait times before evacuees are picked up and hence the importance of explicitly minimizing the wait time together with the transit travel time.

---

## **Demand Estimation by Mode**

The objective of the demand estimation module is to quantify where people are at the time crisis hits and what mode they used (auto driver, auto passenger, transit user, etc). This determines the auto and non-auto evacuation demands, which are the inputs to the evacuation optimization system. The process described below can be conducted for any time of the day. In our experimental analysis, we use the peak evacuation demand as a worst-case scenario (described below).

### **Data Source: The Transportation Tomorrow Survey (TTS)**

The Transportation Tomorrow Survey (TTS) is the largest and most comprehensive travel survey in Canada and is conducted once every five years. The TTS covers 5% of all households in the Greater Toronto and Hamilton Area (GTHA) and surrounding areas, selected at random. Data reported by the TTS include two sets of data: demographic characteristics such as age, gender, household size, dwelling type, etc. and travel patterns such as trip purpose, time of the day, mode of travel, etc. (DMG, 2003). The demand estimation model includes the entire GTHA, which is divided into six regions: Toronto, Durham, York, Peel, Halton, and Hamilton.

### **Model Estimation**

The detailed records of each person in each household are tracked during the course of a 24-hour period (DMG, 2003). The following attributes are used to construct a query to extract the demand data each half hour during the entire day: Household sample number, Person number within household, Start time of trip (24 hour clock), Primary mode of trip<sup>1</sup>, GTHA zone of trip destination, and GTHA zone of household.

The estimation process includes the following steps for each time interval:

- Group people according to the start time of their trip.
- Identify people who drive and identify their home location (zone). This results in an OD matrix in which origins are the current location of people who drive and their default destinations in case of evacuation are their homes (unless another safe destination is suggested by the destination choice module).
- Identify people who do not drive and identify their home location (zone). This results in an OD matrix in which origins are the current locations of non-automobile people and their default destinations in the cases of evacuation are their homes.
- Identify people who returned home.
- Identify people who have not yet made a trip.
- Identify people who are at their homes by combining the previous two steps.

<sup>1</sup> In this application, modes are categorized to Drive (auto driver) and NonDrive modes (auto passenger, local transit, GO train, walk & cycle, other).

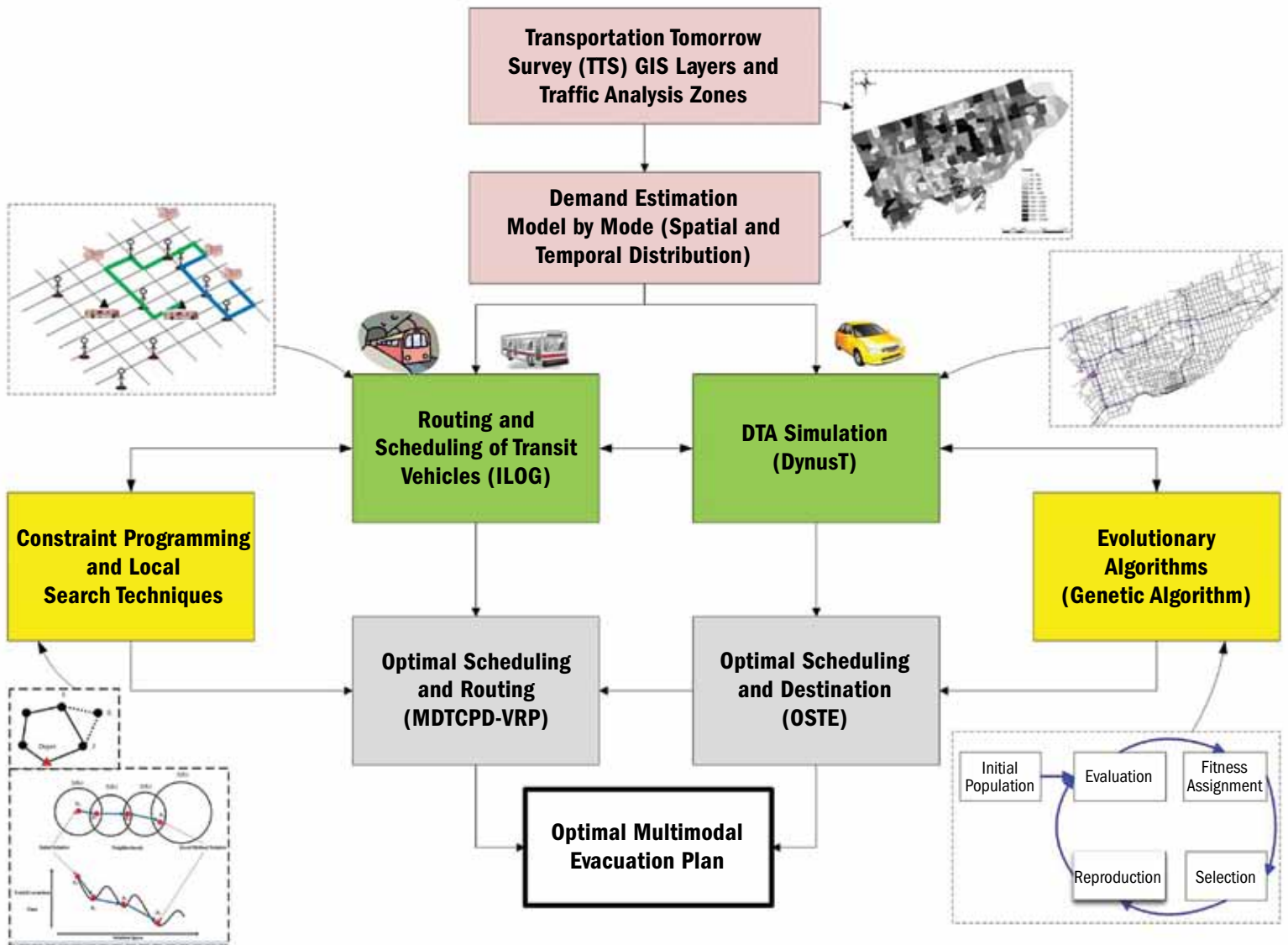
---

The outputs of the demand estimation process are the spatial and temporal characteristics of the trips that are made by each of the three classes of people (travellers using automobiles, travellers using other modes, and travellers who are still at home or returned home). For those who are traveling when the crisis hits, their home locations are known and assumed to be their default destinations in the absence of a better destination choice. Ultimately, this method accurately identifies where people are located by time of day and by mode of travel.

### **The Overall Framework: Putting the Pieces Together**

The overall system attempts to optimize the use of multiple modes during emergency evacuation as shown in Figure 10. We start by estimating the evacuation demand using a regional demand survey (e.g. TTS) and a representation for the traffic analysis zones, the output of which is an accurate representation of the spatial and temporal distribution of population and their modes of travel. We then generate OSTE plans for the vehicular demand using Genetic Algorithms (GA) as a global optimization technique and a dynamic traffic assignment tool. OSTE also produces link travel times that are used as input for the optimal routing and scheduling of transit vehicles. The routing and scheduling problem of transit vehicles is then solved using constraint programming. The automobile OSTE plan and the transit optimal routing and scheduling plan are finally combined for dissemination to evacuees. The next section describes a large-scale implementation of the proposed approach to evacuate the entire City of Toronto.

**Figure 10: Framework for Optimization of Multimodal Evacuation**



## Large-scale Application: Evacuation of the City of Toronto

Our ultimate goal is to produce a system for emergency evacuation planning and optimization that is applicable to large cities and hence smaller regions as well. In this section we demonstrate the application of the system to optimally evacuate the entire City of Toronto in cases of emergency. The City of Toronto is a typical example of fairly large North American cities with a population of 2.37 million. The City of Toronto is located in the centre of the Greater Toronto and Hamilton Area (GTHA). It is the oldest, densest, and most diverse area in the region.

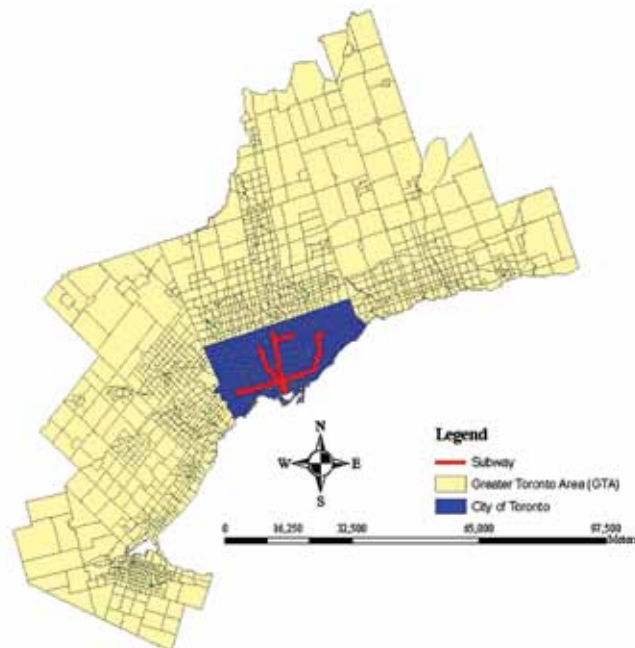
### Supply Modeling

#### The Network Simulation Model

In this work, the GTHA and City of Toronto road networks are developed in DynusT© (Dynamic Urban System in Transportation), a mesoscopic DTA model that is well suited for dynamic traffic simulation and assignment on a regional scale.

Since the evacuation area includes the entire City of Toronto the simulation model covered the whole GTHA at the centre of which is the City of Toronto (see Figure 11).

**Figure 11: City of Toronto and the Greater Toronto and Hamilton Area (GTHA) Study Area**



---

## **Transit Infrastructure: The Toronto Transit Commission (TTC) Fleet**

The TTC fleet consists of bus, light rail transit (streetcar), and rapid rail transit services. The fleet includes about 1,500 vehicles (1,320 buses plus 180 streetcars) during the peak AM period. The Rapid Transit service includes four subway lines (Bloor-Danforth, Yonge-University-Spadina, Sheppard, and Scarborough RT), which cover a large area of the City of Toronto. We assume that, in cases of emergency, the entire bus fleet will be available for our system to reschedule and reroute based on the needs of the evacuation process. Regular transit services will no longer be in effect. Streetcars are not used in this application.

## **Evacuation Demand Estimation**

The output of the demand estimation method is the temporal distribution for the total number of people in the City of Toronto plotted for a 24-hour clock that starts at 4 AM on the trip day to 4 AM the next day. In total, the number of people in Toronto peaks at 108% of the City's population. This increase is attributed to the high concentration of economic activities in the City of Toronto and particularly the business and financial district. Results also show a wide peak activity period that starts from 7:00 AM in the morning and ends at 6:00 PM in the afternoon. The peak demand is found to be 2.56 million people and occurs at the 11:30 AM- Noon interval, which constitutes the worst-case scenario for evacuating the City of Toronto. It is assumed that evacuees will start the evacuation at the time of order (noon time).

In the case of evacuation, it may be harder to persuade drivers to abandon their cars and take any other mode. Also transit users and carless population are captive to transit modes because their choices are limited. Therefore, it seems practical to assume that evacuees will use the same mode of transport when commuting to the City of Toronto, i.e. the Not-At-Home drivers will evacuate using their cars and the Not-At-Home Non-Drive users will evacuate using transit modes. Available transit modes in this implementation include the rapid transit system (subway lines) and surface street buses used as shuttles. For At-Home evacuees, trips are assigned to modes based on the mode split reported by the TTS for trips made by residents of the City of Toronto (DMG, 2003). The overall results are 1,216,886 evacuation trips by automobile and 1,344,942 evacuation trips by transit. The automobile trips form the input origin-destination matrix to the simulation-based DTA model, while the transit trips are assigned to transit shuttles.

## **Subway Demand**

An access distance buffer zone (1,000 metres) is defined around subway lines and we assume that evacuees within the accessible buffer zone are carried by subway to safe destinations. Safe destinations for subway travellers are subway terminals (e.g. Kipling, Finch Subway stations). Evacuees are then spatially assigned to the nearest subway station. The average walking distance to the closest station is found to be around 460 metres.

---

## Shuttle Bus Demand

Major bus stops are extracted from a planning model and the TTC bus routes are provided by the University of Toronto Map Library<sup>2</sup>. Then evacuees are spatially assigned to the nearest bus stop, which resulted in average walking distance to the nearest bus stop of around 330 metres.

## Representation of Noncompliant Traffic

In notice emergency evacuation such as hurricane evacuations, a considerable percentage of evacuees seek out their homes and relatives first (Wilmot and Mei, 2004). In the case of evacuating the City of Toronto, evacuees are advised to head to the optimal shelter and are assumed to start the evacuation at the time of order. However, deviation from the evacuation plan is inevitable. Therefore, a certain percentage of evacuees are assumed not to comply (Noncompliant Evacuees) with the optimal plan and seek their home. The percentage of noncompliant evacuees is treated in this case study as a sensitivity parameter, i.e. exogenously set. We do not attempt to estimate this percentage on the basis of evacuees' behaviour, which is beyond our scope. During the demand estimation process, some of the evacuees who are not at home are directed towards their homes based on the specified percentage, i.e. two components form the total demand in the evacuation planning model: 1) the compliant evacuation demand (superzone demand), and 2) noncompliant demand, as follows:

Total Demand (D) = Evacuation Demand (super-zone demand) + Noncompliant Demand

To the best of the authors' knowledge and based on the literature, evacuee compliance to guidance is rarely modeled or reported. Although it is challenging to model evacuee behaviour, stated-preference surveys and post-analysis surveys of certain evacuation scenarios may be plausible avenues for modeling such behaviour. In the absence of past evacuation surveys in Ontario and Toronto, we assumed that 25% of evacuees will not comply with the provided guidance and will seek their homes first.

## Experimental Results and Analysis

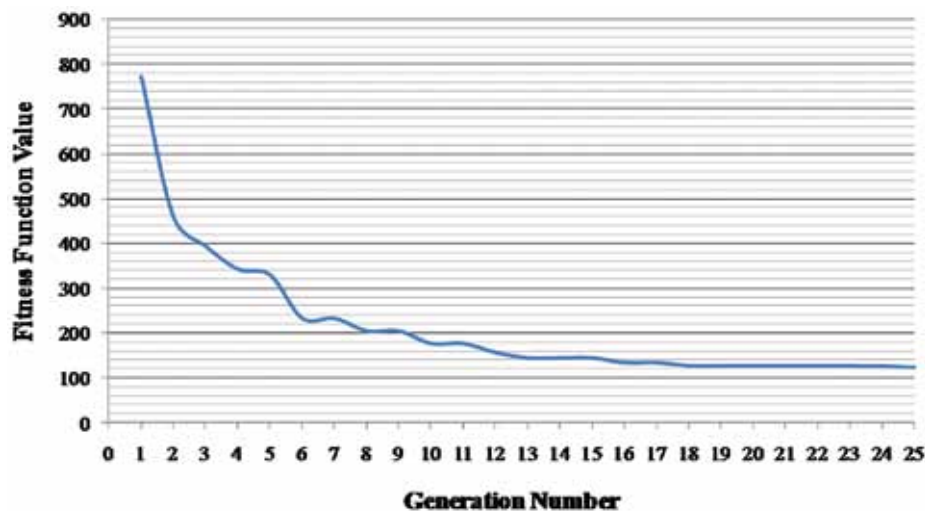
### OSTE: Genetic Optimization Process

The output of the genetic-based optimization process is a scheduling vector ( $\mu$ ) that minimizes total system evacuation time (travel time and waiting time) with a modified network representation to model the destination choice problem. The OSTE framework, integrated with GENOTRANS, outputs the optimized scheduling vector and the detailed routing plan for the evacuation scenario.

<sup>2</sup> The University of Toronto Map Library (online source, [www.main.library.utoronto.ca](http://www.main.library.utoronto.ca))

The objective function (fitness function) values for various scheduling vectors and destination choices are reported. Conceptually, the fitness value in this application is a function of the scheduling vector and the destination choice; however, there is no closed-form equation that represents this relationship. The optimization-by-simulation approach applied in this implementation enables such a relationship to be modeled and thereof optimized. Once the optimization process is completed, the optimal set of parameters is the one (or one of those) with the least (in case of minimization problems) fitness function value. Figure 12 illustrates the evolution of the minimum fitness function value with the number of generations.

**Figure 12: Fitness Function Value with the Number of Generations**



### **OSTE: Traffic Assignment Outputs**

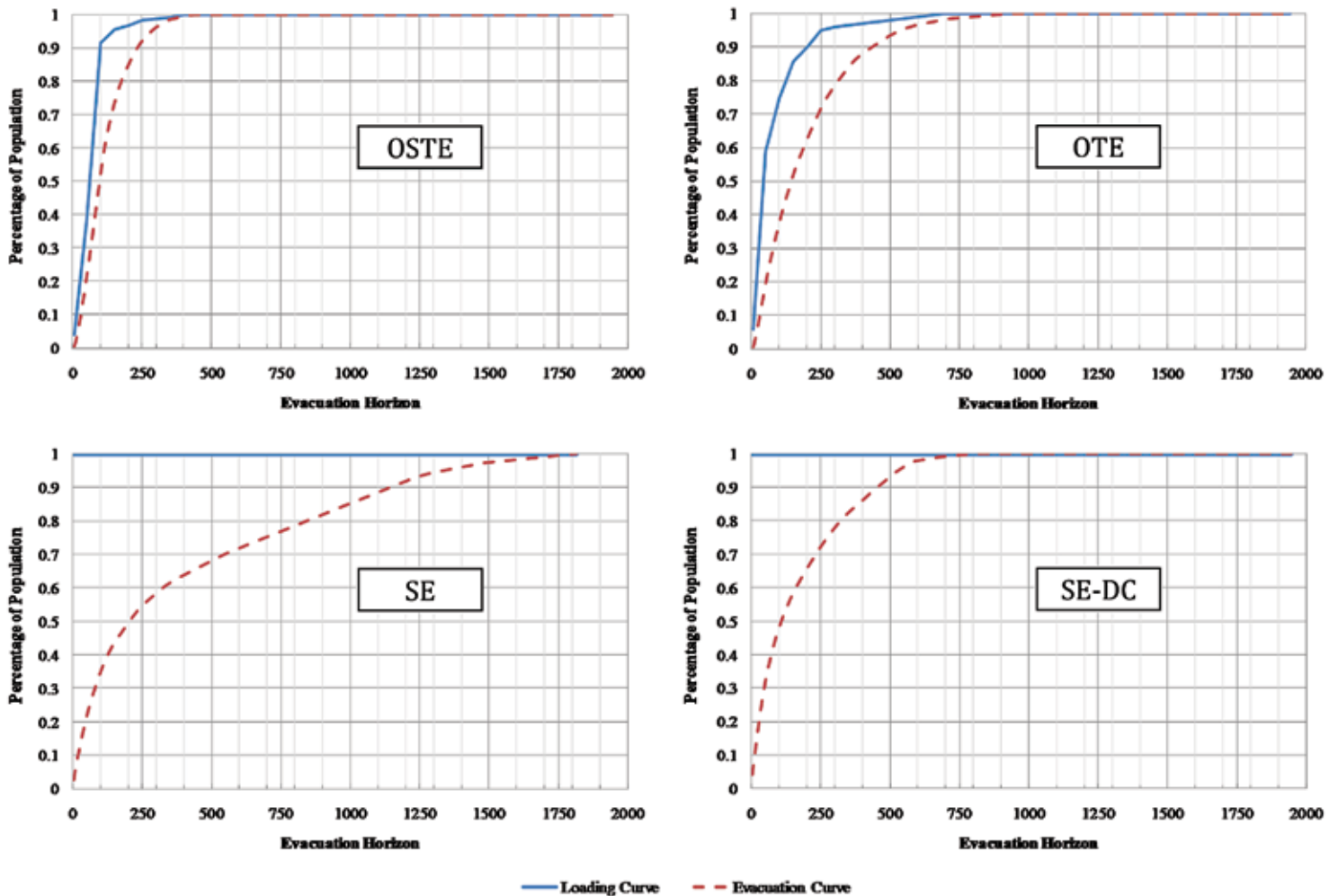
The output of the genetic optimization process (the optimal scheduling vector) is evaluated using the DynusT traffic assignment model. The mesoscopic representation of the traffic simulation model provides sufficient details for the analysis of departure time, destination choice (shelters), and routing plan for each vehicle in Toronto. An important factor in planning for emergency evacuations is the status of evacuees in/out of the City with the evolution of the evacuation plan. In this implementation, special attention is paid to studying the dynamic interaction between the mobilization and evacuation curves and the area between them.

We evaluate three scenarios: OSTE (optimal spatio-temporal evacuation), OTE (optimal temporal evacuation), and SE (simultaneous evacuation). Each scenario examines certain level of integration for different evacuation strategies. While OSTE integrates evacuation scheduling and destination choice optimization, OTE examines

the effectiveness of evacuation scheduling only. SE, on the other hand, presents the do-nothing scenario where evacuees are rushed to the transportation network without any pre-announced schedule. In fact, the SE evacuation scenario is evaluated twice: once with destination choice optimization (SE-DC) and a second time without destination choice optimization (SE). The latter replicates the worst-case scenario where evacuees immediately seek their preferred destination, which is not necessarily the optimal one.

The optimal loading and evacuation curves are shown for the four scenarios (OSTE, OTE, SE, and SE-DC) in Figure 13. The following measures of effectiveness are reported to judge the efficiency of the evacuation strategies when compared to the baseline scenario: average waiting time, average travel time, average total system evacuation time, average trip distance, network clearance time, and average stop time.

**Figure 13: Loading and Evacuation Curves for Four Strategies**



It is clearly shown in Table 3 and Figure 13 that the SE strategy performs the worst since it results in the longest network clearance time (NCT, end of evacuation curve), and the most congested travel times (area between the loading and evacuation curves).

**Table 3: Comparative Analysis of Evacuation Strategies**

MOE Scenario	Average Waiting Time (min)	Average In-Vehicle Travel Time (min)	Average Total System Evacuation Time (min)	Average Trip Distance (km)	Network Clearance Time- NCT (min)	Average Stop Time (min)
SE	0	412	412	11	1815	380
SE-DC	0	175	175	11	800	148
OTE	82	112	194	14	940	88
OSTE	65	50	115	12	445	33

Given the surge in demand in emergency evacuation, this result is not surprising. SE-DC can reduce in-vehicle travel time, but not to the level that any scheduling strategy can achieve, i.e., OTE and OSTE always result in less in-vehicle travel times compared to SE and SE-DC. Average stop time (time when vehicles are caught in congestion) is the longest in the case of SE and SE-DC. In terms of NCT, OSTE performs the best; however, SE-DC performs slightly better than OTE; as OTE explores the optimal scheduling curve so as to minimize the total system evacuation time, this typically results in evacuees being held back from being rushed to the transportation network (in this case, by up to 82 minutes). Also, SE-DC cuts down travel distances significantly by selecting more accessible destinations and avoiding congested / gridlock routes. On the other hand, it should be noted that network stability is another performance factor to consider. SE and SE-DC may result in a network (infrastructure) that has no further room for manoeuvre and could come into gridlock in the case of further panic and/or secondary events (i.e. possibly less stable). Stability however is not explicitly tested in this work. We hypothesize this conclusion based on the observation that SE and SE-DC have the longest stop time, i.e. traffic is more often in a stop-and-go condition.

It is clear that OSTE outperforms all other strategies since it synergizes the scheduling and destination choice in a one shot optimization. The improvements are clear especially in terms of NCT and average stop time. This reflects the essence of a wise evacuation strategy that holds evacuees at the origins up to a point where if they are released into the network and seek the dynamic optimal destination, they will clear the network promptly

---

with the minimum stopping time, encounter less congestion, and contribute to a shorter overall network clearance time. An order of magnitude savings in NCT (75%), total system evacuation time (72%), average stop time (92%), and in-vehicle travel time (89%) are reported. This means that in OSTE, the network is cleared four times faster than the do-nothing strategy (SE), evacuees travel eight times less than in the do-nothing strategy, evacuees stop eleven times less than in the do-nothing strategy, and finally the total system evacuation time can be reduced to one quarter.

### **Optimal Routing and Scheduling of Transit Vehicles**

In general, transit services are designed based on the seating and standing capacity of transit vehicles; however, in emergency evacuation evacuees are expected to tolerate more crowding and make the best use of any available space in the transit vehicle. Therefore, it is plausible to consider the crush capacity of a transit vehicle when planning for emergency situations. It is assumed that one transit bus can carry at most 90 passengers, a subway vehicle is assumed to carry at most 330 passengers, and a subway train with six vehicles can carry up to 2,000 passengers. Dwell times<sup>3</sup> that reflect the physical and operating characteristics of transit units are calculated at transit stops (Vuchic, 2005). It should be noted that dwell times are mode-dependent; therefore, dwell times for buses are different from those for subway trains.

Two scenarios that represent two objective functions are evaluated in this case study of Toronto; the first objective is to minimize the in-vehicle travel time (routing time) for transit vehicles (TT), while the second minimizes the travel time of transit vehicles and the waiting time of evacuees at stops (TT.WT) simultaneously—that is, minimizing the total system evacuation time. Subway trains and shuttle buses are evaluated based on the following measures: average number of runs per transit vehicle, average total in-vehicle travel time, average in-vehicle travel time, average travel distance, and average waiting time. Table 4 shows the MOEs for each subway line/shuttle bus in each scenario. The following sections discuss the relevant findings for both modes.

<sup>3</sup> Dwell time consists of the time lost prior to opening and after closing the transit vehicle doors, and the time required for boarding /alighting of passengers at most heavily used doors. Factors affecting the calculation of dwell times include: vehicle floor height and platform height, number of boarding/alighting channels (doors), and fare type and fare collection.

**Table 4: Measures of Effectiveness of Rapid Transit Lines and Shuttle Buses \***

Line/ Mode  MOE	Bloor Line		Yonge Line		Scarborough Line		Sheppard Line		Shuttle Buses	
	TT	TT.WT	TT	TT.WT	TT	TT.WT	TT	TT.WT	TT	TT.WT
Average No of Runs**	2.1 (1.2)	2.1 (0.8)	4 (3)	4 (0.6)	1.6 (0.9)	1.6 (0.4)	1.2 (0.5)	1.2 (0.5)	5.59 (30)	5.59 (3)
Average Total In-Vehicle Travel Time (min)	96 (70)	90 (35)	228 (117)	216 (35)	47 (29)	47 (10)	19 (8)	19 (8)	114 (425)	122 (44)
Average In-Vehicle Travel Time (min)***	42	43	55	56	27	29	16	16	23	24
Average No of Runs Average Travel Distance (km)**	20.5	21	27.4	27.8	15.17	16.61	6.9	6.6	21	22
Average Waiting Time (min)	72 (63)	40.6 (36)	178 (157)	98 (70)	27 (27)	23 (15)	12 (9)	9 (8)	914 (1844)	53 (43)

\* Numbers between parentheses represent the standard deviation of the MOE across the number of vehicles.

\*\* Average number of runs: The average number of runs each transit vehicle travels to serve all the evacuees.

\*\*\* In-vehicle travel time: the total in-vehicle travel time divided by the number of runs per vehicle.

## Shuttle Buses

The model generates the optimal routing and timetable for each bus as it shuttles between pick-up points and nearest shelters. It should be noted that buses are initially assigned to pickup points according to their location at the onset of the evacuation. After the initial pickup, buses shuttle evacuees to the nearest shelter then go back into the system, but not necessarily to the same pickup points. This means that each bus seeks the best pick-up point in order to achieve a certain objective function. Shuttle buses loop between the optimal pick-up points and shelters until all the demand (total number of carless population) is exhausted and finally head back to safe shelters.

---

Examination of the results leads to the following conclusions:

- On average, transit shuttle buses make around six runs to transport evacuees to safe destinations.
- Including the waiting time in the objective function has evened out the average in-vehicle travel time and the average number of runs/vehicle. This is clearly shown by the significant drop (44 vs 425) in the standard deviation of the total in-vehicle travel time averaged over transit units.
- The average in-vehicle travel time is found to be 23 minutes in the TT scenario and 24 minutes in the TT.WT scenario; although close to each other, the pattern across transit units is significantly different. Including the waiting time in the objective function has increased the in-vehicle travel time and total travel distance. However, it is worth noting that the average total in-vehicle travel time increased by only 7%, yet the average waiting time of evacuees dropped by an order of magnitude (53 vs 914).
- The standard deviation of the average waiting time for evacuees dropped by nearly two orders of magnitude (43 vs 1844) when including the waiting time in the objective function. This demonstrates that the TT.WT scenario provides a more reliable service for evacuees in emergency situations.

### **Summary, Conclusions, and Future Research**

This chapter covered optimizing multimodal evacuation for large-scale emergency evacuations. A framework is presented that optimizes automobile-based evacuation and mass transit-based evacuation. The integrated platform is tested on the evacuation of the City of Toronto using rapid transit (subway), shuttle buses and automobiles. The OSTE platform utilizes a parallel genetic algorithm optimization to guarantee a near global optimal solution. OSTE is designed to model multi-objective optimization evacuation of large cities (minimization of vehicle travel time and evacuees' waiting time). Moreover, a new transit module that provides an optimal scheduling and routing plan for transit vehicles is presented. To the best of the authors' knowledge, this is the first attempt to model transit-vehicles routing and scheduling from a multi-objective perspective with real-life constraints for large-scale evacuation, in parallel with optimized automobile evacuation.

We conclude that an accurate demand-estimating model has the potential to identify the evacuation demand by mode over time and space. These attributes are essential to realistically plan for regional evacuation scenarios. The demand estimation model concluded that the noon period is the time where the maximum number of people is present in the City of Toronto totalling 2.56 million people (1.22 million automobile evacuees plus 1.34 million transit evacuees)

---

We also conclude that considering only the travel time in emergency evacuation inappropriately ignores the waiting time of evacuees, especially in case of non-notice evacuation. On the other hand, minimizing only the waiting time will obviously lead to simultaneous evacuation, which results in early gridlock and severe congestion. A good compromise is to account for both the waiting time of evacuees at the origins and their travel times through the transportation network.

When applying the multi-objective structure of the model to the City of Toronto, it is found that OSTE clears the network four times faster than the do-nothing strategy (SE), evacuees travel eight times less than in the do-nothing strategy, and finally evacuees stop eleven times less than in the do-nothing strategy; in which case the average automobile-based evacuation time across 1.22 million people is about two hours and the NCT is about eight hours. Furthermore, transit systems (rapid transit and buses) can substantially improve the evacuation process due to the readily available capacity of transit vehicles. The Toronto Transit Commission (TTC) fleet is capable of evacuating the transit-dependent population (1.34 million) within two hours on average.

These findings are indeed encouraging in terms of system performance. One potentially critical dimension that needs further investigation is user equity, which may also have a direct bearing on compliance. Since evacuation scheduling, particularly for automobile-based evacuation, implies holding back some evacuees and releasing others, the question is who to let go first. If evacuees perceive their holding up as risky or unfair, they may simply ignore the provided guidance, which defies the purpose of scheduling. One can envision the administration of scheduling on the basis of priority and exposure to risk, which requires further investigation of the nature of the risk (e.g. exposure to nuclear radiations for instance). Finally, further research is required to assess evacuee's potential compliance to the given recommendations. Behaviour can be potentially assessed by analyzing evacuees' actions under past real evacuation scenarios, possibly in other countries, such as the United States and Japan.

## Chapter II References

---

- Abdelgawad, H. and B. Abdulhai (2009). "Optimal Spatio-Temporal Evacuation Demand Management: Methodology and Case Study in Toronto." *Proceedings, the 88th Annual Meeting of the Transportation Research Board, Washington, D.C.*
- Abdelgawad, H., B. Abdulhai and M. Wahba (2010). "Optimizing Mass Transit Utilization in Emergency Evacuation of Congested Urban Areas." *Proceedings, the 89th Annual Meeting of The Transportation Research Board, Washington D.C., 2010.*
- Bethke, A. D. (1976). "Comparison of Genetic Algorithms and Gradient-based Optimizers on Parallel Processors: Efficiency of use of Processing Capacity." *Logic Computing Group, Univ. Michigan, Ann Arbor, MI, Tech. Rep 197.*
- Brown, C., W. White, C. v. Slyke and J. D. Benson (2009). "Development of a Strategic Hurricane Evacuation-Dynamic Traffic Assignment Model for the Houston, Texas, Region." *Transportation Research Record: Journal of the Transportation Research Board* 2137(1): 46-53.
- Chen, X. and F. Zhan (2006). "Agent-Based Modelling and Simulation of Urban Evacuation: Relative Effectiveness of Simultaneous and Staged Evacuation Strategies." *Journal of the Operational Research Society* 59(1): 25-33.
- Chiu, Y. and P. Mirchandani (2008). "Online Behaviour-Robust Feedback Information Routing Strategy for Mass Evacuation." *IEEE Transactions on Intelligent Transportation Systems* 9(2): 264-274.
- Chiu, Y., J. Villalobos, B. Gautam and H. Zheng (2006). Modeling and Solving the Optimal Evacuation Destination-Route-Flow-Staging Problem for No-Notice Extreme Events. *Proceedings of the 85th Transportation Research Board, Washington, DC.*
- Dixit, V. and E. Radwan (2009). "Hurricane Evacuation: Origin, Route, and Destination." *Journal of Transportation Safety & Security* 1(1): 74 - 84.
- DMG (2003). Transportation Tomorrow Survey 2001: Design and Conduct of The Survey. *Data Management Group, University of Toronto, Joint Program in Transportation (January).*
- DMG (2003). Transportation Tomorrow Survey Summaries 2001. *Data Management Group, University of Toronto, Joint Program in Transportation.*
- Han, L. and F. Yuan (2005). Evacuation Modeling and Operations Using Dynamic Traffic Assignment and Most Desirable Destination Approaches. *Proceedings of the 84th Transportation Research Board, Washington, DC.*
- Kruchten, N., B. Abdulhai, L. Kattan and D. de Koning (2004). Galapagos: A Generic Distributed Parallel Genetic Algorithm Development Platform for Computationally Demanding ITS Optimization Problems. *Proceedings of the 83rd Transportation Research Board, Washington, DC.*

- 
- Litman, T. (2006). "Lessons from Katrina and Rita: What Major Disasters Can Teach Transportation Planners." *Journal of Transportation Engineering* 132(1): 11-18.
- Mitchell, S. and E. Radwan (2006). Heuristic Prioritization of Emergency Evacuation Staging to Reduce Clearance Time *Proceedings of the 85th Transportation Research Board, Washington, DC*.
- Mohamed, M. (2007). Generic Parallel Genetic Algorithms Framework for Optimizing Intelligent Transportation Systems (GENOTRANS). *Master Thesis, University of Toronto*.
- Murray-Tuite, P. and H. Mahmassani (2003). "Model of Household Trip-Chain Sequencing in Emergency Evacuation." *Transportation Research Record: Journal of the Transportation Research Board* 1831(1): 21-29.
- Peeta, S. and A. Ziliaskopoulos (2001). "Foundations of Dynamic Traffic Assignment: The Past, The Present And The Future." *Networks and Spatial Economics* 1(3): 233-265.
- Sbayti, H. and H. Mahmassani (2006). "Optimal Scheduling of Evacuation Operations." *Transportation Research Record: Journal of the Transportation Research Board* 1964(1): 238-246.
- TRB (2008). "The Role of Transit in Emergency Evacuation." *Speicla Report* 294.
- Theodoulou, G. and B. Wolshon (2004). Modeling and Analyses of Freeway Contraflow to Improve Future Evacuations. *Proceedings of the 83rd Transportation Research Board, Washington, DC*
- Tuydes, H. and A. Ziliaskopoulos (2004). Network Re-design to Optimize Evacuation Contraflow. *Proceedings of the 83rd Transportation Research Board, Washington, DC*.
- Vuchic, V. (2005). *Urban Transit: Operations, Planning and Economics*, John Wiley & Sons.
- Wilmot, C. G. and B. Mei (2004). "Comparison of Alternative Trip Generation Models for Hurricane Evacuation." *Natural Hazards Review* 5: 170-178.
- Wolshon, B., E. Urbina, C. Wilmot and M. Levitan (2005). "Review of Policies and Practices for Hurricane Evacuation. I: Transportation Planning, Preparedness, and Response." *Natural Hazards Review* 6(3): 129-142.
- Yuan, F., L. Han, S. Chin and H. Hwang (2006). "Proposed Framework for Simultaneous Optimization of Evacuation Traffic Destination and Route Assignment." *Transportation Research Record: Journal of the Transportation Research Board* 1964(1): 50-58.

# Chapter III: Win-Win Congestion Pricing

**A Roadmap for an Effective and  
Socially-Conscious Congestion Pricing Strategy**

Abdulhai, B. and Aboudina, A.

---

## Introduction

### What is Congestion Pricing?

Congestion pricing, also loosely referred to as road pricing, congestion charges, and road tolling, is a collection of methods and systems for surcharging road users to reduce congestion due to excessive demand.

The grand objective of congestion pricing is to manage the transportation infrastructure. Freeways are meant to be free of interruptions and hoped to be free of congestion, but not necessarily toll-free. However, the devil is always in the details. Different stakeholders have different views and objectives. For instance, traffic engineers may think of congestion charges as a tool to control traffic much like the use of traffic lights and speed control. Transportation planners may opt to use charges to influence land use and combat sprawl or influence mode choice. Economists view congestion charging as a means to rationalize demand for road usage via removing social subsidy and making road users pay the full cost of using the road. Governments may view congestion charges as a new revenue source, pure revenue, period, to meet budget deficits, or revenue to be earmarked for spending on enhancing the transportation infrastructure. Some road users may accept road charges in return for improved service (less congestion, better road and transit infrastructure, etc.). Other users may perceive it as new taxation that should be avoided like a plague. Investors approach toll roads as a business, and hence legitimately target profit maximization. The implications of the varying views are public policy confusion as well as variations in the extent (severity) of pricing, its impact on system improvement, the economy, and public reaction. Last, but not least, is how to allocate the raised money.

It is already complicated. However, in our view, the different stakeholders can either battle their way through conflicting views and objectives or try to achieve a win-win approach, which is a sensible target.

We have several possibilities with congestion pricing, simplified as follows:

1. Charge nothing leading to the status quo of road over-consumption and rapidly escalating congestion.
2. Charge a lot to maximize revenue generation, risking negative impacts on public opinion, social equity (pricing out the poor), and the economy at large.
3. Seek an optimal compromise that helps us achieve significant road efficiency benefits, moderate road consumption, increase capacity (explained further later), moderate revenue generation for system enhancement, and manage public perception when people see tangible benefits and maximum social welfare.

---

In the rest of this chapter, we lay the foundation and rationale for such a win-win approach. The approach pursued at the University of Toronto is based on dynamic (varying with actual traffic conditions) road pricing. This approach can eliminate queuing (congestion) delay while boosting capacity (on freeways for instance) in the order of 25%, all while moderately raising funds for infrastructure enhancements including, but not limited to, transit expansion.

### **To Toll or Not to Toll: The Tragedy of the Commons**

To toll or not to toll is no longer the question. The concept of full-cost pricing has been established for decades. The “tragedy of the commons” concept was established centuries ago and has been widely discussed by Garrett Hardin (Journal Science, 1968) and many others since then. The tragedy of the commons is “a dilemma arising from the situation in which multiple individuals, acting independently and rationally consulting their own self-interest, will ultimately deplete a shared limited resource even when it is clear that it is not in anyone’s long-term interest for this to happen” (Wikipedia, 2011). A famous example involves herders who are given free access to open grassland for their cows to graze. Most times, the cows will tend to overgraze and deplete their source of sustenance to the detriment of everyone.

The parallel to the tragedy of the commons in traffic could not be more direct. While transportation authorities and society at large would like to optimize travel and minimize overall cost of travel, travellers act very differently. Travellers act independently and rationally, consulting only their self-interest, i.e. minimizing their direct cost while not paying attention to the societal cost and the detriment to others. In doing so, for instance, travellers seek their shortest route until all routes have more or less the same travel time, a condition known as user equilibrium, as per the well-known Wardrop’s principle.

A traffic network in this state is stable in a non-optimal state, i.e. stuck. The result is congestion, lack of mobility, and limited accessibility to activities, frustration, and pollution. The costs of congestion encountered by commuters each day are partly the result of the inefficiency in the way the independent and self-centred travel decisions of individuals interact and influence the transportation system.

From an economic perspective, and for most goods, prices send the correct signals to coordinate supply and demand and facilitate the decision-making of firms and individuals, hence efficiently allocate resources. However, this is not the case for the transportation market. When individuals consider whether or not to make a trip, they consider their direct benefits and costs that they will face in making the trip and ignore the delay and related externalities that their presence on the road causes other motorists. The level of congestion that would occur if road users took proper account of societal costs would be less than what we are currently facing and would be the economically efficient level.

---

Accordingly, the purpose of congestion pricing is to ensure a more rational use of road resources. This is accomplished by charging fees for the use of certain roads in order to reduce traffic demand or distribute traffic demand more evenly over time and space across the network and throughout peak periods. A toll is designed to relieve congestion by influencing and changing travellers' behaviour, especially choice of departure time (when to travel), mode choice (drive or use transit for instance) and route choice (which routes to use).

### **A Brief History of Road Pricing Studies and Practices**

Road pricing has a long history, with turnpikes dating back at least to the seventeenth century in Great Britain and the eighteenth century in the United States (Small and Verhoef, 2007). Road pricing for congestion management is more recent. The earliest modern application is Singapore's Area License scheme, established in 1975. Since then, other applications have appeared, varying from single facilities such as bridges or toll roads to tolled express lanes as in the United States, toll cordons as in Norway, and area-wide pricing as in London, England.

A number of cities have implemented or are in the process of implementing road pricing. Highway 407 Express Toll Route (ETR) in Toronto, which was opened to traffic in 1997, is the world's first all-electronic, barrier-free toll highway, in which tolls are charged based on vehicle type, distance driven, time of day, and day of week (Lindsey, 2008). The toll on the Autoroute 25 expressway linking Laval and Île de Montréal in Quebec also varies by time of day. Except for the Highway 407 ETR and Autoroute 25, tolls in Canada do not vary over time, and no area-based road-pricing scheme has been implemented in Canada, which lags behind the United States and a number of countries in Europe and Asia with respect to pricing practices. In the GTHA, congestion pricing has been contemplated for years. However, without a thorough understanding of the enabling methods and their pros and cons including public acceptance, it is neither a simple technical task nor a wise political move.

Numerous studies have investigated the potential of road pricing schemes in reducing the vehicular demand subject to travel and behavioural characteristics. In a study conducted at University Drive (Burnaby, BC), single-occupant vehicle (SOV) commuters completed a discrete choice experiment in which they chose between driving alone, carpooling or taking a hypothetical express bus service when choices varied in terms of time and cost attributes.

The results of this study indicate that a potential increase in drive alone costs brings greater reductions in single-occupant vehicle (SOV) demand than an increase in SOV travel time or improvements in the times and costs of alternatives (i.e. carpooling and bus express service) beyond a base level of service (Washbrook et al., 2006). Another study conducted by economists at University of Toronto assessed the potential of congestion

---

pricing against capacity expansions and extensions to public transit as policies to combat traffic congestion. The study concludes that vehicle kilometres travelled (VKT) is quite responsive to price (Duranton and Turner, 2011). Moreover, Sasic and Habib (2012) showed that the recommended strategy to lighten peak period demand while maintaining transit mode share in the GTHA involves imposing a modest toll (around \$1) for all auto trips in addition to a 30% flat peak transit fare increase applied for the majority of the morning peak period. Furthermore, the results reported suggest that such a policy would have a larger effect on shifting travel demand over time than any other policies, not including a road toll. These findings suggest that policy makers interested in reducing demand for auto travel should place as much emphasis on financial disincentives for auto use as they do on improving the supply of alternative travel modes.

Consequently, the purpose of congestion pricing is to ensure a more rational use of road resources. This is accomplished by charging fees for the use of certain roads (i.e., customers should pay the full cost of what they consume) in order to reduce traffic demand or distribute it more evenly over time (away from the peak) and space (away from congested facilities) across the network and throughout the day.

If we agree that congestion pricing has significant merit, the next step is to determine the specifics of how much, when, and where congestion pricing should be considered.

Numerous studies have investigated a variety of congestion pricing models and what is the best pricing structure to be used in congested urban areas (see de Palma and Lindsey, 2011). The scope of these studies ranges from applying a flat or simple pricing structure (e.g. Lightstone, 2011; and Sasic and Habib 2012) on a small or sometimes hypothetical network, (e.g. Gragera and Sauri, 2012; and Guo and Yang, 2012), to a rigorous network-wide pricing scheme (e.g., Verhoef, 2002; and Morgul and Ozbay, 2010).

Although these studies contribute considerably to the state-of-the-art and state-of-the-practice in congestion pricing, they still suffer from a number of limitations that make their applicability to large and complex urban areas questionable. For example, case studies on large urban networks were found scarce (Xu and Ben-Akiva, 2009). Although a few studies focused on non-flat (variable) pricing of an entire network, these studies relied mostly on hypothetical testing scenarios that lack any methodological justification. In addition, travellers' individual responses to pricing (e.g. choice of departure time, choice of mode, and choice of route) were not explicitly integrated in the overall pricing system design (Xu and Ben-Akiva, 2009). Also, most non-flat tolls lacked handling unpredicted disturbances (e.g., incidents) as prices were varied according to a fixed schedule rather than real-time traffic measurements.

These limitations pose a number of challenges to roadway agencies and highway authorities when they assess the viability of congestion pricing to solve their real-life traffic problems. These challenges are mostly related to how to concurrently address the multi-faceted nature of the congestion pricing problem into one framework including: impact

---

on congestion, user behavioural response, road network dynamics, potential of emerging technologies in enabling comprehensive pricing schemes, alternative mode choices, demand management strategies in relation to influencing mode choice and choice of departure time, supply management strategies in relation to choice of travel routes, and pricing structure in terms of temporal dynamics in response to congestion dynamics, and social welfare of pricing in relation to monopoly vs. social-welfare-maximizing pricing. In summary, these dimensions accentuate the need for a comprehensive approach and framework for an effective dynamic and socially conscious congestion pricing strategy.

In this report, therefore, we endeavour to answer these questions while considering two important and interrelated perspectives: economic and traffic engineering. The remainder of the chapter provides an in-depth review of the implications of congestion pricing strategies from economic and traffic engineering perspectives, followed by policy recommendations for selecting a balanced congestion pricing scheme that achieves various policy, social welfare and traffic control objectives. We conclude by summarizing the findings in the form of a practical win-win framework for dynamic congestion pricing strategies, with an eye on applicability in the GTHA.

### **Congestion? Microeconomic and Traffic Engineering Perspectives**

It is worth noting that economists and traffic engineers have different technical perspectives on congestion, which can have policy implications and create confusion. From an economic perspective, a system (e.g. road or transit lines) is congested if the performance of the system (e.g. travel time) starts to deteriorate with the intensity of use (e.g. traffic volume). The positively sloped portions of the curves in Figures 14-a and 14-b depict this situation. From a traffic engineering perspective, a road is considered congested when the traffic stream becomes unstable slipping into stop and go condition. More formally, when traffic density (vehicles per kilometre) exceeds the critical density (i.e. the density corresponding to capacity, Figure 14-b) resulting in traffic instability and breakdown. The negatively sloped portions of the curves in Figures 14-a and 14-b depict this situation.

This definitional difference can be a source of confusion in the congestion pricing circles (i.e., researchers, engineers, economists, and transportation practitioners working in the field of congestion pricing). What congestion means for traffic engineers is termed “hyper-congestion” for economists. Moreover, hyper-congestion causes a significant drop in capacity. This is notable at the critical density in Figure 14-b. Let’s consider, for example, a single freeway lane. When traffic density is very low (a few cars per kilometre), flow is also low in terms of vehicles served per hour and speeds are very high, say, 110–120 km/hr. These are known as free flow conditions. As density increases, flow increases, and speed remains relatively high but gradually drops (travel time increases),

---

until capacity (maximum flow) is reached. Further demand increases density beyond the tipping point, after which speeds rapidly drop and flow also drops. Capacity of a freeway lane is typically around 2,200-2,400 vehicles per hour. Corresponding speeds are roughly in the 80-90 km/hr range. Excessive demand can push the traffic stream to breakdown as drivers turbulently adjust their speed while surrounded by many other vehicles in closer proximity. If traffic is allowed to break down, the result can be significantly lower average speed (say ~40-60 km/hr) and reduced capacity (e.g. 1,800 vehicles per hour). In this example, the tolerable speed drop from 120 to 90 km/hr, is termed congestion by the economists' definition, but is of less significance for traffic engineers. The sudden breakdown that happens after is what is often perceived by drivers as congestion, and is termed by economists as hyper-congestion. The hyper part perhaps denotes that cost (travel time) increases, but service (flow) drops.

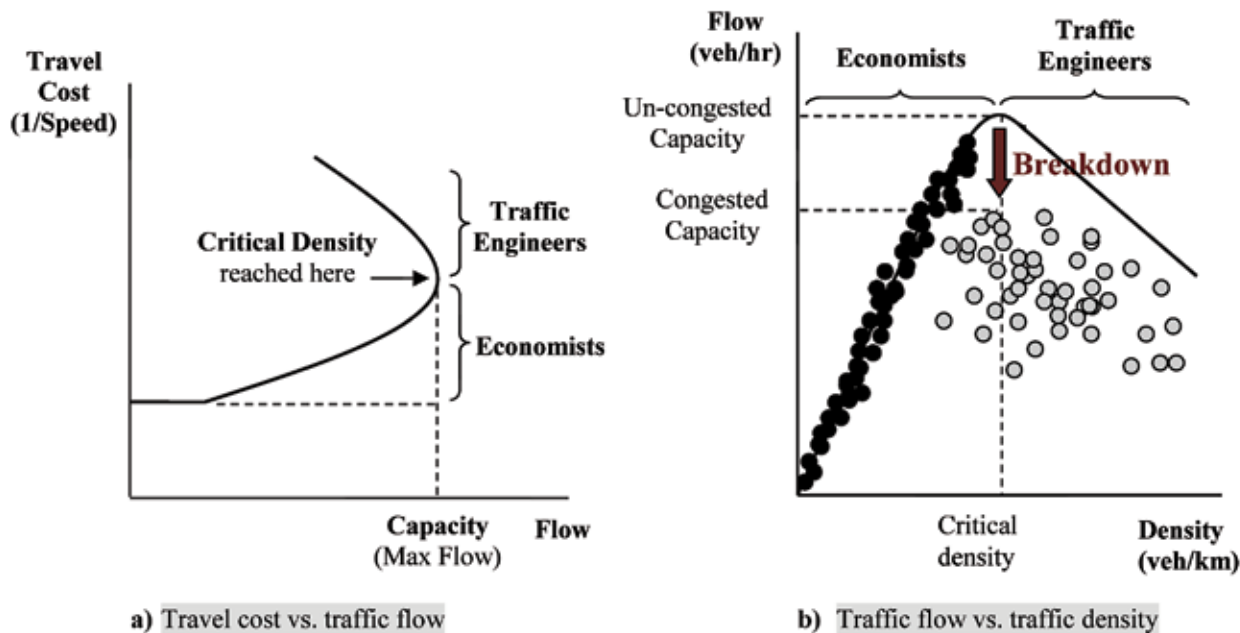
As shown in Figure 14-b, there are two capacity values at critical density; one on the uncongested (left) side of the curve and one on the hyper congested (right) side of the curve, which is 15-25% lower. Preventing hyper-congestion therefore helps the traffic stream to avoid flow breakdown and the related sudden drop of capacity (Small and Verhoef, 2007). Since the lower capacity value is the norm during rush hours, preventing hyper-congestion increases capacity by the 15-25%.

The implications of the seemingly minor difference in the definition of congestion are non-trivial from a practical perspective for the following reasons:

1. It makes sense to combat hyper-congestion in busy urban areas first before considering milder signs of congestion. Restoring free flow conditions (i.e. eliminating any sign of congestion or slow down) is neither practical nor desirable. It is not practical because it implies massive expansion of capacity or massive reduction in demand. It is undesirable because it implies underutilized road infrastructure during peak periods, let alone the off peak. Lastly, it would be much more irritating to the public to be charged the toll if they do not perceive a tangible problem in the first place.
2. As we will explain in detail later, targeting the elimination of hyper-congestion in large congested cities using pricing maybe a reasonable objective from traffic perspective, but is also socially optimal from economic perspective, i.e. a possibly very viable win-win approach.

Win-win means demand is rationalized (slightly reduced), capacity is increased significantly, and the price is very moderate relative to the performance gain, in addition to raising revenue for further improvements of the infrastructure (transit and roads). This way, stakeholders would be all satisfied, given their varying views and objectives as outlined earlier.

**Figure 14: Economic vs. Engineering Perspective of Congestion**



### Getting Technical with Tolls: Static vs. Dynamic, Revenue vs. Social Welfare Maximization, and Single Roads vs. a Network

There are two main methods for managing tolling time periods: static and dynamic tolling. In static methods, traffic demand is assumed to be constant and the resulting tolls are fixed over the period of time. For instance, users are charged \$5 per trip during peak demand periods (e.g. 3:30-6:30 PM) or during the whole working day (7:00 AM to 7:00 PM). In dynamic models, on the other hand, the variations of demand with time (how peak demand evolves during rush periods, then subsides after) are explicitly considered. Accordingly, they produce dynamic tolls that correspond to traffic variations. For instance, it would cost \$1 to travel during 3:30-4:00 PM, \$2 from 4:00-4:30 PM, \$3 from 4:30-5:00 PM, \$4 from 5:00-6:00 PM, \$3 from 6:00-6:30 PM, \$2 from 6:30-7:00 PM, \$1 from 7:00-7:30 and back to zero after 7:30 PM.

This toll can be in effect over a given full road, or per unit length, or over an area within a cordon. Using an everyday metaphor from home, static tolls are like an on-off light switch, while dynamic toll is like using a dimmer switch that controls the brightness of the room. Static tolls send some users away from the priced facility or time period. Dynamic tolls manage behaviour and choices with more precision to influence their choice of departure time, mode choice, and route choice, i.e. this system manages the spatio-temporal dynamics of tolling to match the dynamics of traffic. Such a system can more effectively produce the desired results.

The simplified details of static and dynamic congestion pricing methods are discussed in the following subsections.

### Static Pricing

Static tolling sets a price that is fixed over a certain period of time on a given facility (e.g. freeway). The amount of toll depends on the economic objective, i.e. maximize profit, or maximize social welfare. The difference between the two is very significant. In general, as shown in Figure 15 below, the Demand curve represents the quantity demanded as a function of price, whereas the Average Cost (AC) curve is the total production cost divided by the total quantity produced. The Marginal Cost (MC) is defined as the change in total cost required to increase the output by one unit, and the Marginal Revenue (MR) denotes the change in total revenue associated with an increase in output by one unit. In traffic, average cost can be, for instance, average travel time. For instance, 30 minutes on average to go from A to B on a given road. As you decide to enter that road, you not only incur the 30 minutes but you also increase everyone's travel time to, say, 30.01 min. This extra .01 minute is a price that you and 10,000 drivers pay, marginally increasing the total system cost by  $10,000 \times .01 = 100$  vehicle-minutes, a cost that is called the external cost of congestion. You don't pay this external cost, nor do you take it into account when making a travel decision.

Note that if the road is not priced (free), demand and cost equilibrate when the AC curve intersects the demand curve as shown in Figure 15. However, the marginal cost at this flow level is higher than average cost as the average cost does not consider the external cost of congestion, or the delay a traveller is imposing on all other travellers. This ignored external cost of congestion component is viewed as social subsidy, i.e. a cost borne by society (all travellers) that each individual traveller is not paying for.

If prices are set to maximize profits (defined as the difference between the total revenue and the total cost), we determine equilibrium in an unregulated environment resulting in what is known as monopoly price ( $P_m$ ), which is the price consistent with the output where the Marginal Revenue equals the Marginal Cost as follows:

$$\text{Profit} = \text{Total Revenue (TR)} - \text{Total Cost (TC)}$$

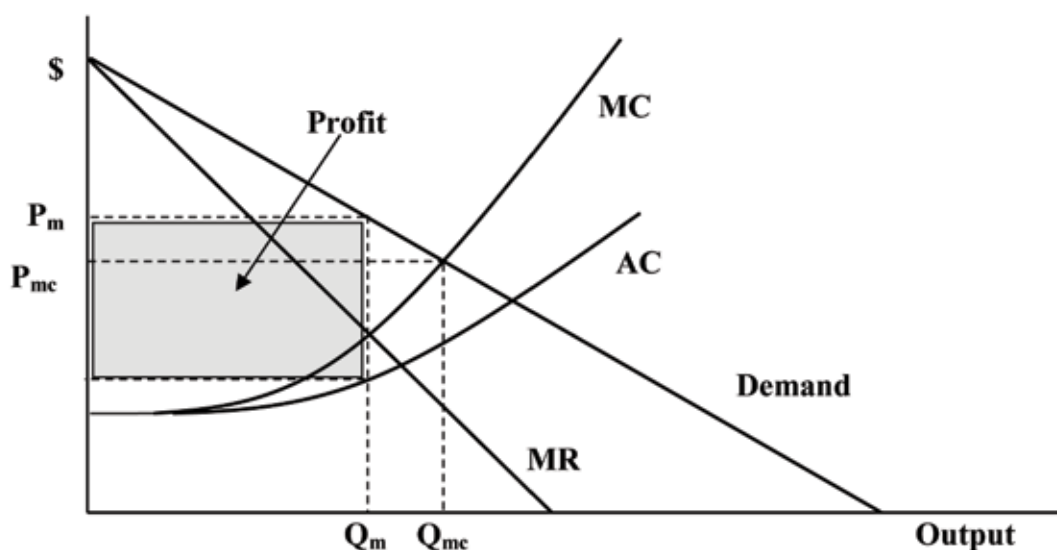
To maximize profit with respect to volume of production ( $Q$ ):

$$\frac{\Delta TR}{\Delta Q} = \frac{\Delta TC}{\Delta Q}$$

i.e., *Marginal Revenue (MR) = Marginal Cost (MC)*

But, if prices are set to maximize the social welfare (defined as the difference between the total benefits and the total costs), we determine a marginal-cost price ( $P_{mc}$ ), which is the price consistent with the output where the Marginal Cost meets the Demand curve. Figure 15 illustrates the difference between both pricing rules (monopoly vs. marginal-cost). In transportation, marginal-cost pricing means that each traveller faces a perceived full-cost price (i.e., the travel cost in addition to the road charges imposed) equal to his/her activity's social marginal cost (i.e., the monetary value of the travel time incurred by the traveller in addition to the extra time incurred by the existing travellers due to the entrance of that traveller to the system).

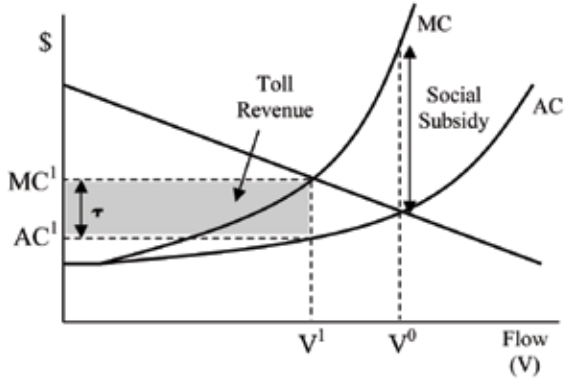
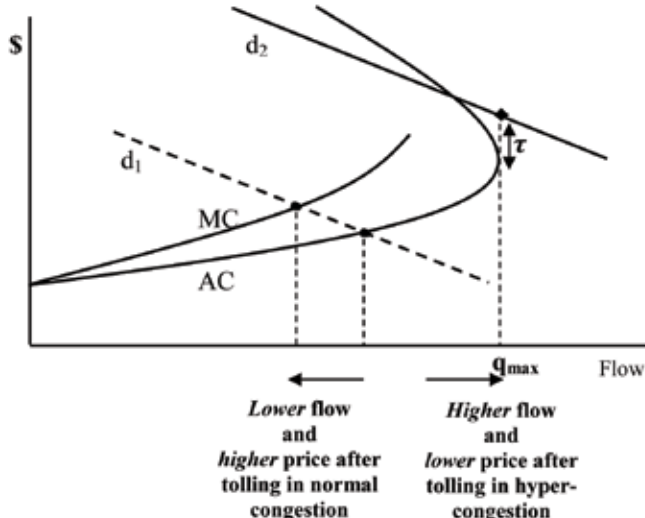
**Figure 15: Monopoly Price  $P_m$  vs. Marginal-Cost Price  $P_{mc}$**



Depending on the policies and constraints in place, social-welfare maximizing pricing may, in turn, be further broken down into two pricing schemes, namely, first-best pricing and second-best pricing. First-best pricing entails system-wide pricing in accordance with traffic levels everywhere. However, doing so in practice is often impossible. Various constraints on what prices can be charged must be considered (for example, the inability to price all links in a network and the political necessity of availing free options). We then enter the world of second-best pricing that involves optimizing social welfare given some constraints on policies. Table 5 reports the first-best pricing rules for three cases: single road at normal congestion (not hyper-congestion), single road at hyper-congestion, and an entire network at normal congestion.

Finally, it should be noted that static pricing approaches are appropriate only when traffic conditions do not change so quickly or when it is thought sufficient to focus on average traffic levels over extended periods, which is not the case in most large cities. Dynamic pricing can generally overcome such limitations, as will be illustrated in the following section.

**Table 5: First-Best Pricing Rules in Three Cases**

Facility Size	Congestion Status	Characteristics and Optimum Pricing
Single Road	Normal Congestion	 <ul style="list-style-type: none"> <li>• The demand curve intersects the cost curves in the <u>normal congestion</u> region.</li> <li>• The un-priced equilibrium occurs at the intersection of the demand and the <i>average cost</i> curves (involves a traffic flow <math>V^0</math>).</li> <li>• The optimal flow <math>V^1</math> occurs at the intersection of the demand and the <i>marginal cost</i> curves.</li> <li>• <math>V^1</math> can be achieved through an optimum toll <math>\tau</math> equal to the difference between the marginal cost and the average cost at <math>V^1</math>.</li> <li>• It is the excess congestion (difference between <math>V^0</math> and <math>V^1</math>) that should be the focus of policy makers and transportation planners. That is, higher tolls that would move the system towards free-flow travel conditions are not, generally, the socially optimal conditions.</li> </ul>
Single Road	Hyper-Congestion	 <ul style="list-style-type: none"> <li>• The demand curve is so high that it crosses the cost curves in the hyper-congestion region (where the traffic density exceeds the critical density, resulting in a traffic breakdown).</li> <li>• The marginal cost curve MC is infinite at a flow of <math>q_{max}</math>.</li> <li>• The optimal toll <math>\tau</math> (shown in figure above) <i>eliminates</i> hyper-congestion and maintains flow at <i>maximum capacity</i>, i.e. prevents typical capacity loss due to hyper-congestion and restoring some 20% more throughput. A toll <b>higher</b> than <math>\tau</math> will <i>unnecessarily</i> cut demand below capacity; some travellers whose <u>marginal benefit</u> (demand) exceeds their <u>marginal cost</u> (i.e. un-subsidized) will be <i>priced-off</i> the road. On the other hand, a toll <b>less</b> than <math>\tau</math> cannot fully eliminate hyper-congestion; it entails an equilibrium density above the critical value, resulting in a drop in capacity. The social welfares attained in both cases, over-tolling and under-tolling, are lower than that achieved at optimal tolling; this is due to the benefits forgone and the unpaid social externalities, respectively.</li> <li>• This case may be crafted as a <b>win-win</b> solution because it <i>increases</i> capacity (flow) but at the same time <i>reduces</i> congestion and raises funds to reinvest in sustainable transportation infrastructure.</li> </ul>
Network	Normal Congestion	<ul style="list-style-type: none"> <li>• This case highlights the correspondence between the <i>economic</i> perspective of pricing (maximizing social welfare) and the <i>traffic engineering</i> perspective (system optimal traffic conditions).</li> <li>• That is, first-best tolling on a complete network is proved to satisfy <u>system optimal</u> conditions (where the total travel time in the network is minimized) rather than <u>user equilibrium</u> conditions (where no one can improve his/her travel time by switching routes) (Small and Verhoef, 2007).</li> </ul>

---

## Dynamic Pricing

Dynamic pricing takes into consideration that congestion peaks over time then subsides. Therefore, in addition to (hyper) congestion-free travel time, there is a congestion (queuing) delay component that peaks with congestion as well and for which travellers need to account. To illustrate, let's assume that road users have a desired arrival time  $t^*$ , 8:00 AM arrival at work for instance. Deviations from this time imply early or late schedule delay costs. The summation of congestion delay and schedule delay is constant, i.e. if you try to avoid congestion, you arrive too early or too late. Travellers who must arrive on time during the peak, therefore, encounter the most congestion delay i.e. there is a tradeoff between incurring congestion delay and arriving too early or too late (schedule delay).

This conceptual model of dynamic congestion is known as the bottleneck model. It involves a single bottleneck that is either congested or not; i.e., for arrival rates of vehicles not exceeding the bottleneck capacity and in absence of a queue, the bottleneck's outflow is equal to its inflow and no congestion (delay) occurs. When a queue forms, vehicles exit the queue at a constant rate equal to the bottleneck capacity  $V_k$ . Figure 16-a illustrates this model and Figure 16-b shows the two components of the total cost  $c(t)$  in the unpriced equilibrium, namely, congestion delay cost  $c_T(t)$  and schedule delay cost  $c_S(t)$  (early and late arrival costs).

Note that the total number of travellers that enter the system ultimately exit the system after being queued for a while. The optimal toll in this case attempts to flatten the peak, i.e. to spread the demand evenly over the same time period. The price is set such that the inflow equals road capacity, which in turn equals the outflow. Note that pricing in this case spreads demand over the peak but does not extend the peak, which often causes misconception. To create this effect, we impose a triangular toll schedule, with two linear segments, that replicate the pattern of travel delay costs in the un-priced equilibrium. This toll is shown in Figure 16-b as  $\tau(t)$ . It results in the same pattern of schedule-delay cost as in the unpriced equilibrium, but it produces zero congestion delay cost, i.e. entirely eliminates hyper-congestion. Instead of queuing delay, travellers trade off the amount of toll to be paid vs. schedule delay such that a traveller who arrives right on time  $t^*$  pays the highest toll. It is worth noting that no congestion delay does not mean that people arrive on time  $t^*$ ; only those who pay max toll arrive at  $t^*$ . People who pay less incur some schedule delay. The triangular toll does not eliminate schedule delay costs; the peak is only flattened (but neither shrunk nor expanded). That is, queue exit times (i.e., work arrival times) have the same pattern around  $t^*$  as before tolling. However, the road entrance times (i.e., departure times from home) change after tolling in a way that makes everyone experience zero queuing delay by paying a toll that is directly proportional to the travel delays that (s)he used to experience and tolerate before tolling.

Accordingly, those people who used to experience the highest travel delay before tolling (and arrive to work on time) will now pay the max toll (and still arrive to work on time), whereas those people who used to experience zero congestion delays before tolling (but arrive to work too early or too late) will pay zero toll (but still arrive to work too early or too late).

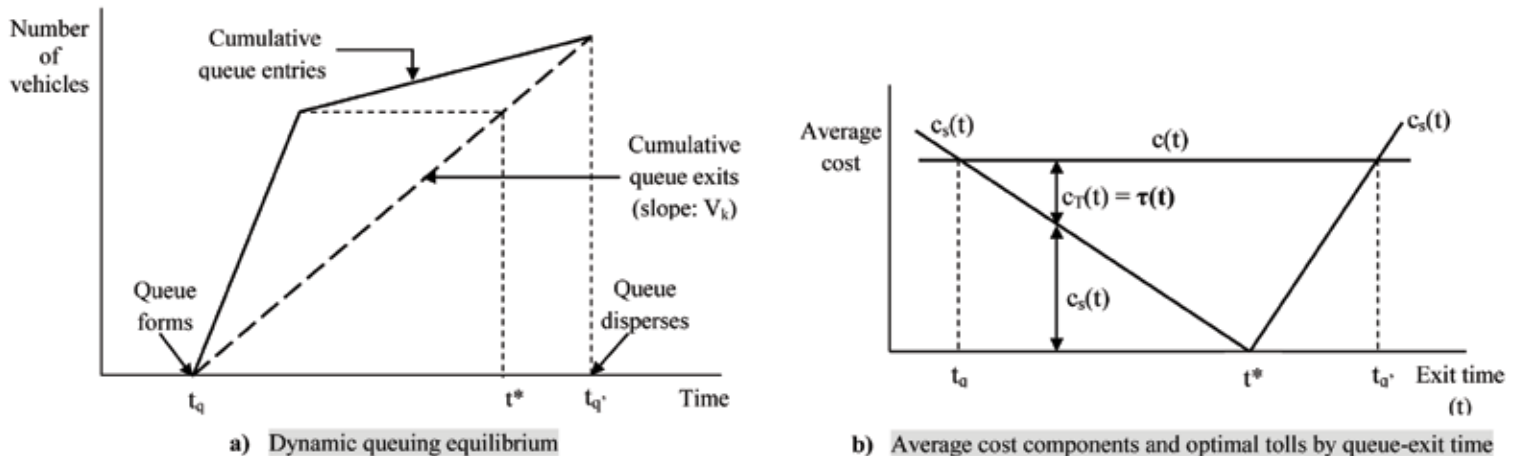
It is obvious that, in real life, not all travellers have the same desired arrival time  $t^*$ . However, elaborate demand data such as TTS data in Toronto would provide detailed time sliced demand that would be used as input to the system, i.e. not just a hypothetical  $t^*$ .

Further, this basic and simple triangular price structure based on the bottleneck model would work nicely only for a bridge-like case where people do not have routing options, i.e. their reaction to tolling can only be departure time variation. If travellers are free to choose their departure time, travel mode and route, a more comprehensive pricing system is needed as will be discussed later.

In conclusion, the main benefit of static marginal-cost congestion pricing is to achieve an optimum level of traffic flow by forcing travellers to pay the full cost of congestion externalities to society. Whereas, dynamic congestion models suggest that a main source of efficiency gains from optimal pricing would be the rescheduling of departure times (temporal distribution) from the trip origin. Later, we will show how to combine the two effects.

Based on the theoretical approaches of road pricing discussed in this section in addition to other practical schemes implemented in some major cities (e.g. London, Stockholm, and Singapore), Table 6 summarizes different pricing policies, their objectives and impacts and how they relate, if at all, to optimal pricing presented above (where black filled bullets denote a strong relation and so on). Table 6 therefore forms the first step toward brainstorming an effective congestion pricing strategy as it outlines the different policies to be followed based on the desired high-level objective.

**Figure 16: Equilibrium in the Basic Bottleneck Model (Small and Verhoef, 2007)**



**Table 6: Road Pricing - Objectives and Policies**

Policy options	Main objectives/impacts							Examples of each policy
	Reduce downtown traffic	Encourage carpooling	Maximize profits	Control traffic (temporal/spatial)	Reduce automobile use	Maximize social welfare (system optimal)	Alter departure-time choice	
Cordon/ Zonal tolls	●	●	○	●	●	○	○	<ul style="list-style-type: none"> <li>– London Congestion Zones</li> <li>– Stockholm Congestion Pricing</li> </ul>
HOT lanes	○	●	○	●	●	○	●	<ul style="list-style-type: none"> <li>– I-15 HOT Lanes, San-Diego, CA</li> <li>– I-394 in Minnesota</li> </ul>
Monopoly pricing	○	●	●	●	●	○	○	– ETR 407, ON, Canada
Variable tolls	●	●	●	●	●	●	●	– Singapore Electronic Road Pricing
Distance-based fees	●	●	○	○	●	○	○	
First-best pricing	●	●	○	●	●	●	●	----
Bottleneck pricing	○	●	○	●	○	●	●	----

### Turning Prices into Control Actions: How to Control Hyper-congestion using Dynamic Pricing

As mentioned earlier, the purpose of congestion pricing is to ensure a more rational use of road resources by charging fees for the use of certain roads in specific time periods in order to reduce traffic demand or distribute it more evenly across the network and over time. In other words, congestion pricing schemes can be viewed as traffic control strategies. Pricing decisions answer where to impose tolls, when to impose tolls, how much to charge.

---

Dynamic pricing structures can be determined offline then implemented in the field, which is referred to as an open-loop strategy. It is not based on actual real-time measurement of traffic. On the other hand, if the pricing structure is determined on the basis of continuous real-time measurements of traffic conditions, which are fed back in the price-determining mechanism, it is considered a closed-loop strategy. Another classification distinguishes congestion pricing based on the objective of pricing. Some pricing strategies seek the decisions yielding optimum process output (e.g. minimizing total travel time in the network), whereas other strategies only aim at maintaining/regulating the output at a certain desired level that is not necessarily optimal (e.g. maintaining roadways densities below critical density or any preset level).

These two classifications (open-loop vs. closed-loop and optimization vs. regulation) have a direct bearing on field implementation. For instance, open-loop strategies are less accurate, cannot respond to unanticipated events such as incidents, but are easier to implement because they do not need real-time measurements. We elaborate further in the following sub-sections.

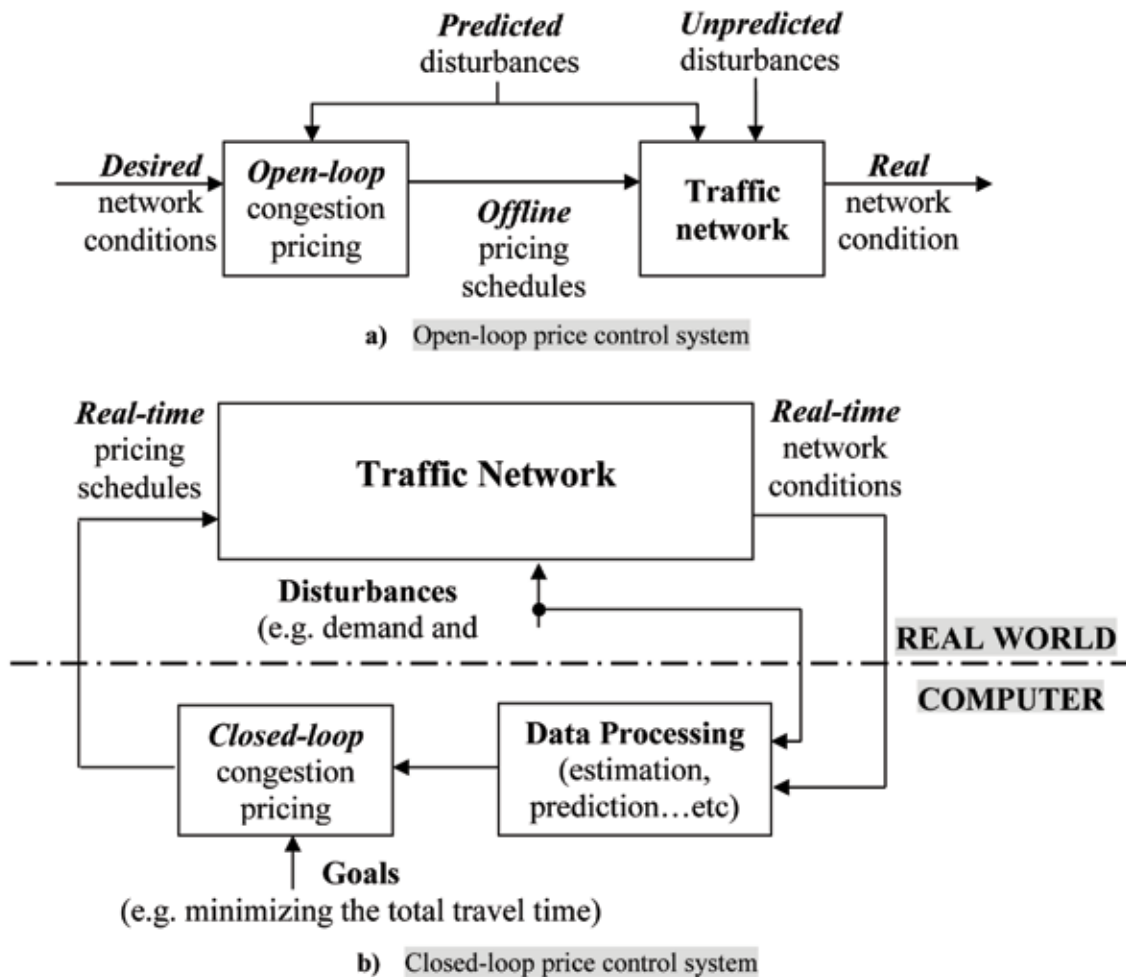
### **Open-Loop versus Closed-Loop Dynamic Congestion Pricing**

In open-loop dynamic pricing, no real-time measurements are needed; rather, the road charging schedules are determined offline based on the desired output (e.g., optimized or regulated occupancies or speeds) in addition to the predicted (known) network disturbances (e.g., known demand variations), as shown in Figure 17-a. That is, the pricing schedules are determined once and then deployed to the network; in other words they do not vary with any real-time network potential disturbances (e.g., unanticipated flows, incidents or lane closure).

In general, open-loop control has a quick response to known predicted system disturbances; i.e. there is no need to wait for the network known disturbances to affect the traffic and then to measure. Nevertheless, this type of control could lead to totally undesirable results in case of unexpected traffic disturbances (e.g., incidents). This means that the resulting control decision is optimal, when being deployed on the field, only if the traffic conditions do not change from what was anticipated, which is rarely the case. Therefore, feedback is necessary.

On the other hand, the role of a closed-loop pricing strategy is to specify the road charging schedules—based on the real-time traffic measurements, estimations, and predictions—so as to achieve the pre-specified goals regarding the network performance despite the influence of various unpredicted disturbances (e.g. accidents and severe weather conditions). This type, although being the most desirable, requires real-time computations and measurements. Figure 17-b illustrates the basic elements of a closed-loop congestion pricing control system. Closed-loop systems play an important role in dynamic congestion pricing where tolls are updated based on real-time observations as in the case of the HOT lane on I-15 in San Diego, I-394 in Minnesota, SR-167 in Seattle, and the fast lane to Tel-Aviv (Dong et al., 2007; and Bar-Gera and Gurion, 2012).

**Figure 17: Basic Elements of Open-Loop and Closed-Loop Congestion Pricing Control Systems (Abdulhai, 2011)**



### Traffic Network Optimization versus Regulation Using Pricing

An optimal controller determines the best pricing strategy that optimizes (minimizes or maximizes) a certain objective function (e.g. the total travel time). Whereas, a regulator tries to maintain the process output at a certain level that is exogenously set and that is not necessarily optimal, for example a HOT (high occupancy/toll) lane that is required to operate at a certain speed (Leonhardt et al., 2012). Furthermore, both regulator and optimal congestion pricing systems can be open-loop or closed-loop systems.

---

## **Revisiting Recent Advances in Congestion Pricing Methods**

Building on this information, the question becomes how to translate the economic and engineering fundamentals into practical but comprehensive road pricing practices and related policies that are applicable on a large scale in congested cities, while understanding and accounting for the users' heterogeneous behavioural responses. This question is a very active research area worldwide. This section of the report is therefore dedicated to reviewing recent advances internationally, in light of the previous sections. We divide the review into two related streams of thought or two sides of the congestion pricing coin. The first one involves studies related to developing general frameworks of congestion pricing, as a traffic control tool and policy model. On the other hand, the second category focuses on users' behavioural responses to congestion pricing.

### **General Congestion Pricing Framework**

In pursuit of a general pricing framework, some studies focused on facility pricing such as a freeway with HOT lane or an urban corridor (where one of two alternative routes is tolled and the other is left free). Other studies considered charging multiple links on a network such as area pricing and cordon tolls.

### **Facility Pricing**

Dong et al. (2007) developed an anticipatory state-dependent pricing for real-time freeway management. The tolling system imposes dynamic tolls with the objective of eliminating queuing on the tolled links. The system involves two components that operate in rolling horizon fashion: an anticipatory toll generator, and a prediction module. The anticipatory generator compares the predicted with the preset target link concentration (i.e., occupancy) values and adjusts the current link tolls accordingly, i.e. acts as a closed-loop regulator. The prediction module predicts future network states based on current states, past states, and previously predicted prices. However, the effect of tolling on the rest of the network is not taken into account while generating tolls.

In Bar-Gera and Gurion (2012), the authors presented a facility pricing project implemented in Tel-Aviv on a single left lane dedicated to public transport, high-occupancy vehicles, and toll payers. The system includes a dynamically responsive toll setting mechanism that guarantees a certain level of service (speed) on the fast lane as well as a sufficient utilization (flow). The toll is dynamically set, in a control centre, based on two components: a predictive component that estimates the demand and willingness to pay, and a feedback component that is used to automatically adjust the toll based on real-time measurements (Leonhardt et al., 2012). This system can also be regarded as a closed-loop regulator.

Yang et al. (2012) proposed a distance-based dynamic pricing algorithm that takes user responses to tolling into account. The authors applied a numerical approach to find the

---

optimum pair-wise tolls (between on-ramps and off-ramps of a hypothetical bridge) that maximize the total revenue. In this study, pair-wise demands are determined based on the associated tolls using a discrete choice model, and the algorithm is run every  $\Delta t$  time step, producing dynamic tolls. This system can be classified as an open-loop optimizer.

### **Network Pricing**

Verhoef (2002) developed an algorithm for finding second-best tolls where not all links of a congested transportation network can be tolled. Furthermore, a simulation model was used to study the performance of the algorithm for various archetype pricing schemes, e.g. a toll-cordon, pricing of a single major highway, and pay-lanes and free-lanes on major highways. Verhoef's algorithm can be regarded as an open-loop network optimizer.

In Kazem (2012), the author tested and compared many pricing scenarios (e.g. flat, distance-based, and peak tolls) in a study area in Southern California region. The pricing scenarios were obtained by consulting public groups along with transportation agencies, i.e., no theoretical rationale stands behind the pricing patterns presented in this work. The pricing scenarios tested in this study are classified as open-loop traffic regulators.

Xu and Ben-Akiva (2009) proposed a dynamic congestion pricing model in which traffic assignment accounts for travellers choice behaviour (route choice and departure time choice). The objective of this model is to find the optimum toll schedule (for specific links on the network) that minimizes the travel time of all network users. The authors, however, acknowledge that their model can be improved in several ways; e.g. by using: more robust optimization techniques, joint (instead of sequential) discrete choice models for departure-time choice and route choice, and elastic (rather than fixed) demand assumptions. The model proposed by Xu and Ben-Akiva follows the open-loop optimizer type of controllers.

### **User Responses to Congestion Pricing**

Given the above pursuits of how to price, the other side of the coin is focused on users' behaviour, i.e. how users respond to pricing and how they alter their travel behaviour. This class of studies does not focus on the determination of the pricing structure itself; rather, investigates the possible impacts of hypothetical (fixed or variable) pricing scenarios on the individual (disaggregate) traveller which give rise to the network (aggregate) performance. Within users' responses to pricing, route choice, and departure time choice attract the most attention in recent studies. Less attention, however, is given to mode choice.

For example, Mahmassani et al. (2005) presented algorithms to find time-dependent shortest paths (i.e. route choice) for use in dynamic traffic assignment applications to networks with hypothetical variable toll pricing and heterogeneous users (having different value of time preferences). The algorithms considered in this study assume

---

fixed departure times. However, in general, a trip maker facing a toll road with time-varying charges would not only change path but would also adjust departure time so as to minimize his/her total trip cost.

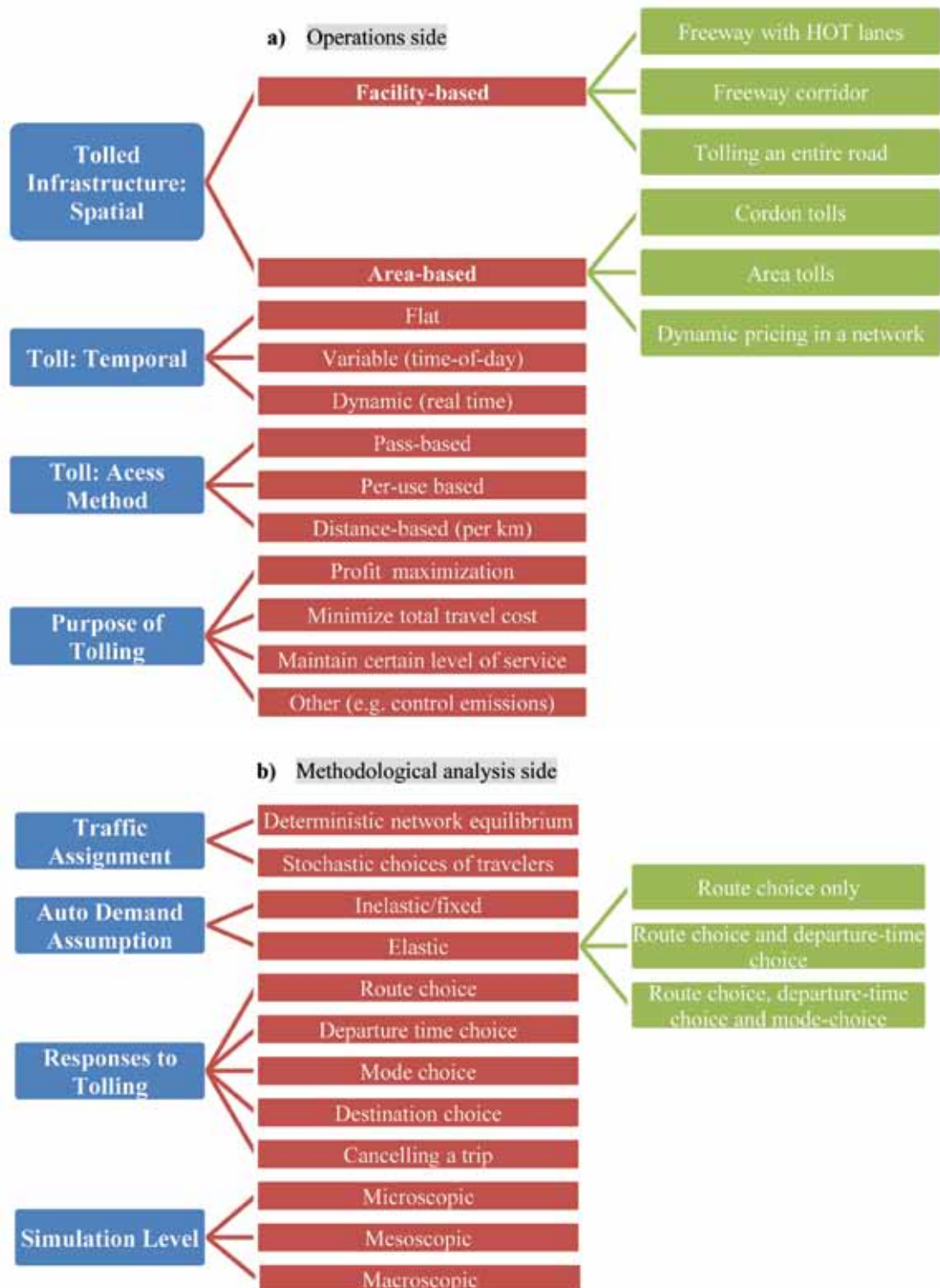
Another study at University Drive (Burnaby, BC) demonstrated a method for estimating SOV (Single-Occupant Vehicle) commuter mode choice responses to policies introducing financial disincentives for driving alone (road charges and parking charges) and improvements to alternative modes (Washbrook et al., 2006).

A recent study conducted at University of Toronto by Sasic and Habib (2012) focused on developing discrete choice models to describe mode-choice and departure-time choice in the GTHA. The empirical models were then used to evaluate mode and time switching behaviour in response to combined variable transit pricing with peak road pricing policies.

### **Putting the Pieces Together: Aspects to Consider in a Comprehensive Congestion Pricing Strategy**

As it is clear by now, there are numerous dimensions of the congestion pricing problem that need to be understood, considered, or at least reasonably assumed when planning a comprehensive non-trivial congestion pricing strategy. Stakeholders, including politicians, transportation authorities as well as travellers themselves, therefore, face significant challenges to grasp and integrate the dimensions of the problem, plan for a balanced congestion pricing strategy, and design a framework that simultaneously considers the multiple dimensions of the problem (Ed Pike, 2010; AECOM Consult Team, 2006; and Arnstein, 1969). Therefore, this section identifies the key dimensions/venues that should be considered early in the planning process in order to develop effective pricing strategies. Moreover, some of the limitations in recent dynamic congestion pricing research/projects will be highlighted. Figure 18 is developed to provide a multi-dimensional decision-making process for congestion pricing initiatives, from both the operations side and the methodological analysis side.

**Figure 18: Aspects of Congestion Pricing Decision Making Process**



The discussion demonstrated several dimensions, parameters, and assumptions of road pricing strategies from both operational and analytical perspectives. Implementing such strategies in practice, however, requires a rigorous dynamic congestion pricing approach. In fact, dynamic congestion pricing models and methods developed and/or implemented thus far suffer from several limitations as briefly summarized in Figure 19.

### Figure 19: Limitations in Current Dynamic Congestion Pricing Models

#### - Simplified Networks and Case Studies

- Case studies on large/complex networks are rare

#### - Link Toll Schedules Based on Hypothetical Scenarios

- No robust optimization approaches are generally employed to determine the toll patterns optimizing certain *network-wide* objective function. Instead, road charges are often determined based on trial and error approaches that aim at regulating specific traffic variable(s) (e.g. link speed) around certain target values (*that are not necessarily optimal*).

#### - Incomplete Dynamic Congestion Pricing Schemes

- Road charges are usually set based on traffic conditions of a *single road* instead of the *entire network*, which is inaccurate (the network is an inter-connected system that is wholly affected by the toll set on any individual link)

#### - Ignoring Traveler's Individual Responses to Pricing

- Traffic assignment is usually based on deterministic user equilibrium (aggregate models) rather than stochastic responses of travelers (disaggregate models of mode, route, and departure-time choice)
- This results in *unrealistic* modeling of travelers' true responses to pricing

#### - Inelastic (Fixed) Auto-Traffic Demand Assumption

- *Route choice* (and *sometimes* departure-time choice) is considered the only decision individuals make in *response to pricing*.
- *Mode-shift* impacts of pricing are not usually taken into account in the determination of link tolls (which results in unrealistic forecasting of charging impacts)

#### - Variable rather than Dynamic Tolling

- Studies usually come out with link tolls that are variable (vary according to a *fixed* predetermined schedule) but not dynamic (vary based on real-time traffic measurements).
- Unpredicted disturbances cannot be controlled.

---

## **The Roadmap: An Integrated Congestion Pricing Strategy for Large Cities and the GTHA**

In light of the foundational overview provided earlier and the congestion-pricing aspects and limitations of existing approaches summarized in Figure 18 and Figure 19, respectively, we provide a win-win multi-faceted framework for dynamic congestion pricing. The University of Toronto comprehensive congestion pricing framework is based on 1) users' discrete choices in response to pricing policies, and 2) control and optimization approaches for demand and supply management to maximize the transportation system performance and social welfare. Our strategy incorporates:

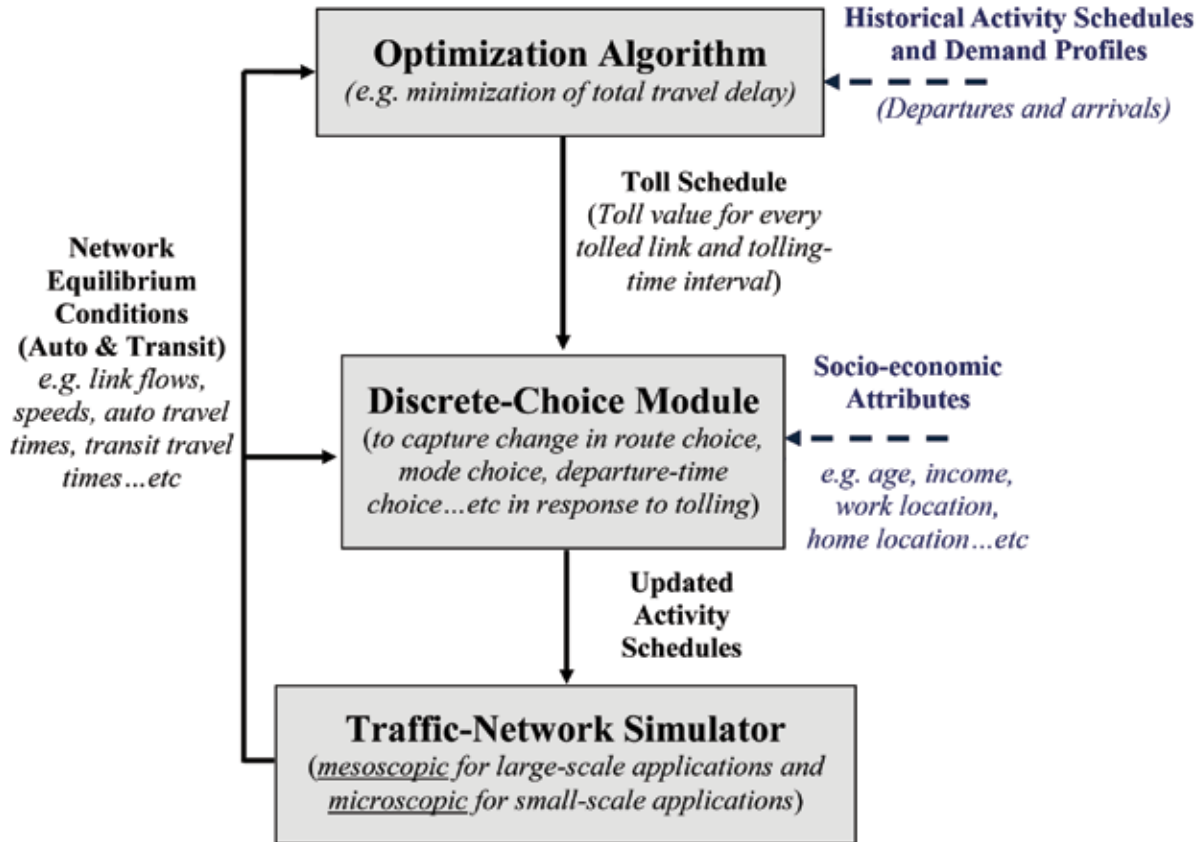
1. Robust optimizer to determine the optimum link toll schedules (on an entire network or sub-network rather than single facilities) under mild or hyper-congestion to produce the minimum total travel delay and maximum social welfare.
2. Users' stochastic response estimator embedding a discrete-choice module into the dynamic congestion pricing system. This behavioural model aggregates users' individual decisions (mode-choice, departure-time choice, and route choice) in response to dynamic pricing schedules.
3. Real-time toll regulator to fine tune the optimized pricing structure based on observed real-time traffic dynamics. The role of the regulator is to prevent pending traffic breakdowns due to unpredicted disturbances.

The offline toll optimizer and online toll regulator are shown in Figure 20 and Figure 21, respectively.

The optimizer first solves for the spatio-temporal toll structure (i.e. toll value on each tolled link for every time interval using actual GTHA travel patterns and inputs). The optimal toll schedule, on average, would not change from day to day and hence travellers adjust their mode choice and departure time in a manner that minimizes the total travel delay (and/or any other desired objective). The system takes as input the network topology, anticipated demand, user demographics and behaviour to form a hybrid dynamic traffic assignment and travel behaviour model. The optimizer seeks, for every facility of interest in the network (e.g. link, lane, road, corridor or area), a nonlinear (step-toll, e.g. Lindsey et al, 2012) version of the triangular price structure of the bottleneck model, i.e. that rises from zero to a maximum then fall back to zero when the demand subsides, as shown in Figure 22.

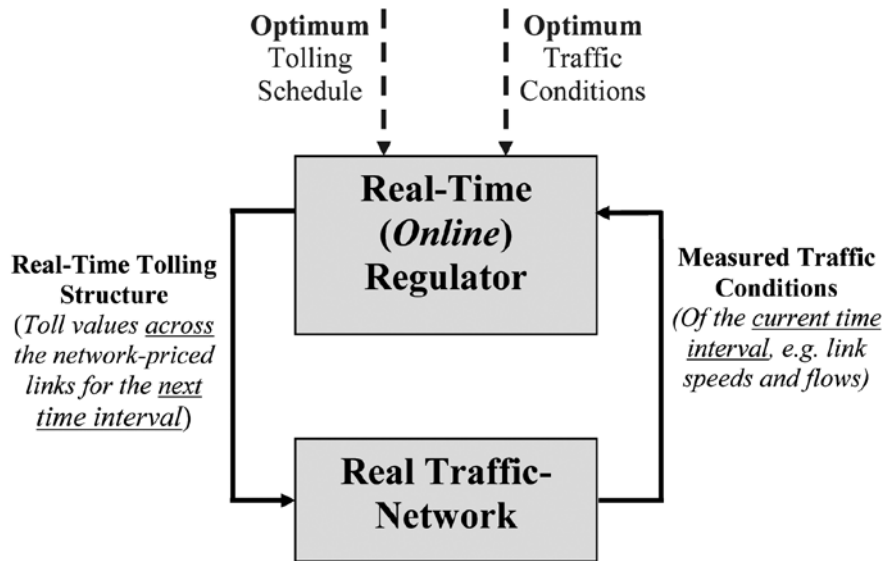
More specifically, as shown in Figure 20, the optimization is performed by iteratively evaluating the toll structure and minimizing (or maximizing) the objective function. At the end of each iteration, the discrete-choice module estimates the impact of the toll schedule and the most recent network conditions (obtained during that iteration) on travellers' individual choices (departure-time, mode and route choice). The updated choices are then fed into the dynamic assignment simulator, which, in turn, produces the new network conditions that are used to evaluate the objective function value. This process is repeated as many times as needed offline until certain convergence criterion is met and an optimal pricing structure is finalized.

**Figure 20: Open-Loop (Offline) Toll Optimizer Framework**

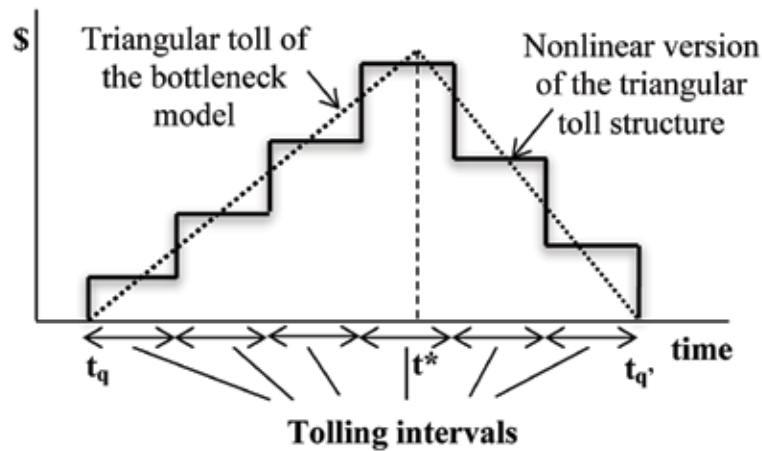


The optimizer described in Figure 20 is an open-loop congestion pricing controller that targets traffic network optimization. The resulting non-linear triangularly shaped price structures for all the priced facilities in the network constitute an optimal solution for the given demand level, network topology, and traveller characteristics. However, the inputs, being estimates, will always have inaccuracies and the traffic system may also experience possible unpredicted disturbances such as incidents. Therefore, the optimal price structure will need to be fine-tuned in real time based on actual measurement of traffic states. The optimization framework is hence extended by adding an online toll regulator (shown in Figure 21) through which real-time traffic measurements, in every time interval, are used to update the optimum link toll values for that interval based on the difference between the measured and the optimum traffic conditions. For instance, if the optimal toll on a freeway in the network in a given time interval is determined by the offline optimizer to be \$n, but traffic conditions are sensed to be deteriorating, an incremental increase in toll will be added to restore optimal conditions, and vice versa.

**Figure 21: Closed-Loop (Online) Toll Regulator**



**Figure 22: The Triangular (Nonlinear) Pricing Structure**



The open-loop optimal price system and the closed-loop regulator, combined, constitute a near-optimal closed-loop dynamic pricing controller that is applicable to any network.

In comparison to Figure 18, this pricing framework is dynamic, stochastic, applicable to area wide or specific facilities, sensitive to distance, location and time of the day, maximizes network performance in terms of minimum travel cost, maximizes social welfare by avoiding monopoly pricing and addressing the external cost of congestion, and explicitly accounts for demand elasticity and users' behavioural responses in terms of mode choice, departure time choice, and route choices. This framework is independent of the sensing and implementation method, i.e. whether using GPS, loop or other detection

---

methods, dynamic message signs or in-vehicle displays for information dissemination. Detection and dissemination technologies are not discussed in this report.

## **Summary and Future Work**

Congestion pricing is widely viewed among economists, transportation analysts, and policy makers as one of the promising control tools to tackle escalating traffic congestion. However, stakeholders need to capture the full dimensions of road pricing to create effective strategies to ensure successful implementation. This report presented several practical guidelines and recommended congestion pricing policies for practitioners, highway agencies, roadway authorities, and researchers to effectively decide upon the best congestion pricing strategy given a set of predefined objectives.

The report started with an in-depth understanding of the implications of congestion pricing strategies from economic and traffic engineering perspectives. Accordingly, and as a first step toward determining a practical framework for dynamic congestion pricing, the report highlighted some recommended pricing strategies to be followed to achieve certain high-level policy objectives. Finally, the report concluded with introducing a comprehensive framework (roadmap) that concurrently considers the key aspects of congestion pricing while attempting to address the main limitations of the state-of-the-art and state-of-the-practice.

The University of Toronto framework presented in this report is based on 1) rigorous control and optimization approaches for demand and supply management, 2) elaborate users' discrete choice assessment in response to pricing policies, and 3) a real-time price regulator to produce real-time dynamic tolls that can prevent potential traffic breakdowns. The framework is dynamic, stochastic, applicable to area wide or specific facilities, sensitive to distance, location and time of the day, maximizes network performance in terms of minimum travel cost, maximizes social welfare by avoiding monopoly pricing and addressing the external cost of congestion, and explicitly accounts for demand elasticity and users' behavioural responses in terms of mode choice, departure time choice, and route choices.

The next step is to apply the described system in the GTHA context via testing on multiple transportation networks in the GTHA, including:

1. Freeway only with HOT lanes, e.g. the Gardiner Expressway or the 401 Express lanes,
2. Freeway corridor, e.g. "Gardiner-LakeShore" where Gardiner would be tolled.
3. Cordoned network, e.g. downtown Toronto.

Toronto is currently one of the top ten most congested North American cities (TomTom International BV, 2012). Moreover, congestion costs commuters in the GTHA \$3.3 billion per year. Looking ahead to 2031, this cost is expected to rise to \$7.8 billion (GTTA, 2008). This, in fact, strengthens the need to explore, analyze, test, and deploy various traffic control policies, including dynamic pricing, in order to tackle the alarming congestion problems.

## Chapter III References

---

- Abdulhai, B. (2011). *CIV 1532 (ITS and Traffic Control) Lecture Notes*. University of Toronto.
- AECOM Consult Team. *International Urban Road Pricing*. Final Report, 2006.
- Arnstein, S. (1969). A Ladder of Citizen Participation, *AIP Journal*.
- Bar-Gera, H. and Gurion, B. (2012). Fast Lane to Tel-Aviv: High-Occupancy-Toll Project with Pareto Package. Transportation Research Board Annual Meeting 2012, Paper #12-0712.
- de Palma, A. and Lindsey, R. (2011). Traffic congestion pricing methodologies and technologies. *Transportation Research Part C*, 19(6), 1377-1399.
- Dong, J., Mahmassani, H. S., Erdogan, S., and Lu, C-C (2007). State-Dependent Pricing for Real-Time Freeway Management: Anticipatory versus Reactive Strategies. *Transportation Research Part C*, 19(4), 644-657.
- Duranton, G. and Turner, M. A. (2011). The Fundamental Law of Road Congestion: Evidence from US Cities. *American Economic Review*, 101(6): 2616-52.
- Ed Pike, P. E. (2010). *Congestion Charging: Challenges and Opportunities*. The International Council of Clean Transportation (icct).
- Gragera, A. and Sauri, S. (2012). Effects of Time-Varying Toll Pattern on Social Welfare: Case of Metropolitan Area of Barcelona, Spain. Transportation Research Board Annual Meeting 2012, Paper #12-4723.
- Greater Toronto Transportation Authority (GTTA, 2008.) *Costs of Road Congestion in the Greater Toronto and Hamilton Area: Impact and Cost Benefit Analysis of the Metrolinx Draft Regional Transportation Plan*. Final Report.
- Guo, X. and Yang, H. (2012). Pareto-Improving Congestion Pricing and Revenue Refunding with Elastic Demand. Transportation Research Board Annual Meeting 2012, Paper #12-6650.
- Kazem, O. (2012). Prototype Pricing Scenario Analysis for Southern California Association of Governments Travel Choices Study. Transportation Research Board Annual Meeting 2012, Paper #12-5318.
- Leonhardt, A., Sachse, T. M., and Busch, F. (2012). Dynamic Control of Toll Fees for Optimal High-Occupancy-Toll Lane Operation. Transportation Research Board Annual Meeting 2012, Paper #12-2699.
- Lightstone, Adrian (2011). *Congestion Charging in the City of Toronto: Distance Based Road Pricing on the Don Valley Parkway and Gardiner Expressway*. M.Sc. Thesis, Royal Institute of Technology, Stockholm, Sweden.
- Lindsey, R., Verhoef, E.T. and van den Berg, V. (2012). Step by step: Revisiting step tolling in the bottleneck model. *Journal of Urban Economics*, 72, 46-59.

---

Lindsey, R. (2008). Prospects for Urban Road Pricing in Canada.  
Brookings-Wharton Papers on Urban Affairs.

Mahmassani, H. S., Zhou, X., and Lu, C-C. (2005). Toll Pricing and Heterogeneous Users, Approximation Algorithms for Finding Bi-criterion Time-Dependent Efficient Paths in Large-Scale Traffic Networks. Transportation Research Record, 1923, 28-36

Morgul, E., F. and Ozbay, K (2010). *Simulation Based Evaluation of Dynamic Congestion Pricing*. M.Sc. Thesis, State University of New Jersey.

Sasic A. and Habib K. M. (2012). *Modelling Departure Time and Mode Choice for Commuting in the Greater Toronto and Hamilton Area (GTHA): Evaluation of Dynamic Travel Demand Management Policies*. M.Sc. Thesis, University of Toronto.

Small, Kenneth A. and Verhoef, Erik T. *The Economics of Urban Transportation*, Routledge, England, 2007.

*TomTom North America Congestion Index*. TomTom International BV, 2012.

Verhoef, E. T. (2002). Second-Best Congestion Pricing in General Networks, Heuristic Algorithms for Finding Second-Best Optimal Toll Levels and Toll Points. Transportation Research Part B, 36, 707-729.

Washbrook, K., Haider, W., and Jaccard, M. (2006). Estimating Commuter Mode Choice: A Discrete Choice Analysis of the Impact of Road Pricing and Parking Charges. Transportation, 33, 621-639.

Xu, S. and Ben-Akiva, M. E. (2009). *Development and Test of Dynamic Congestion Pricing Model*. M.Sc. Thesis, Massachusetts Institute of Technology.

Yang, L. Saigal, R., and Zhou, H. (2012). Distance-Based Dynamic Pricing Strategy for Managed Toll Lanes. Transportation Research Board Annual Meeting 2012, Paper #12-6639.

# Chapter IV: Open Transport Innovation

**Opening the Limit Beyond the Sky**

Abdulhai, B., El-Dariby, M., and Abdelgawad H.

## The Point

Technology is coming at us like a tsunami. Experts believe that cities that manage to embrace technology will simply rise and thrive. Governments (e.g. Department of Transportation at the federal, provincial, and municipal levels) that will put the energy into taming the technology beast to their advantage will serve people better, and equally importantly, they will stay relevant.

But how can governments tame the technology beast? How can they stay ahead of the technology curve on a shoestring budget? Technology is a never-ending race. By the time we develop a high-tech product or service and roll it out in the field, it is obsolete. Governments cannot afford such a mad race. It is not about technology for technology's sake. It is about managing cities.

One answer lies in open innovation. In this chapter, we explain our vision for open transport innovation and we offer enabling tools developed at the University of Toronto for open innovation in transportation services.



Surfing Mavericks

<http://photos.denverpost.com/2013/01/20/photos-mavericks-invitational-surf-contest-2013/#28>

## Introduction

Information and communication technologies (ICT) are emerging very rapidly offering a new paradigm for delivering services for smart cities, and in our context here, intelligent transportation systems (ITS) services. The demand for ITS services is therefore increasing with the evolution of these technologies. In addition, numerous information channels, data sources, and service providers are emerging daily with different abilities, offerings, quality and requirements. This poses challenges to transportation stakeholders if the methods/tools of delivering ITS services remain the same.

This chapter, therefore, discusses:

- I. The potential of a new open service innovation model for developing and sustaining ITS services in the knowledge-intensive tight-budget economy of the twenty-first century.
- II. The potential of methods/tools to ubiquitously collect traffic and travel information and deliver services at low cost (e.g., vehicles as probes, smartphone sensors, connected vehicle data, etc.), and
- III. The potential for sensory fusion of ubiquitous data and information.

---

These three themes are seen as complementary and cross-linked. The first theme establishes the open service innovation concept and introduces an enabling software platform for any future ATMS (Advanced Traffic Management Systems) or TIS (Traveller Information System) applications. The second theme presents a wide scale data collection via an array of sensors and communication channels to be strategically located and distributed across the transportation network. The third theme focuses on fusing the heterogeneous data collected from multiple sources (provided by Theme II) into a single estimate (e.g. travel time or speed information) for many ATMS and TIS applications openly sharable.

- Theme I: Open service innovation, co-creation and mass customization are the state-of-the-art models for delivering services in an era that is becoming highly driven by the customer needs rather than typical product-driven models. Major organizations (e.g., IBM (Pine J., 1993) are transforming the way they do business by adopting open service innovation that allows seamless flow of innovation from within to without the organization and vice versa. Open service innovation is the latest trend in open innovation that focuses on the provision of services to the customer (transportation stakeholders inclusive of the traveller side and the operator/service provider side). Mass customization recognizes that customers have different needs and empowers customers (travellers and stakeholders in general) with services that match their context-based needs such as subscribing to specific traveller information systems or incident/event notification system. This chapter will furnish the foundation for open transport service innovation and mass customization for the new generation of ITS by proposing an Open-ITS system.
- Theme II: One step towards open innovation is the open and ubiquitous collection of transport data and intelligence. Ubiquity will not only expand the spatial and temporal coverage of transport intelligence gathering but will also provide agile deployment in remote areas as needed and will relieve major cities of the burden of installing and maintaining expensive and traffic-disruptive traditional detection technologies. For example, connected vehicles data and the establishment of two-way communication with the car can enable the real-time transmission of rich smartphone as well as vehicle sensory and engine data to a central server. The system can upload data directly from the engine of the vehicle (through On-board Diagnosis Device- OBD) that will enable accurate measurement of speed, fuel consumption (and emissions), engine diagnostic data, to name a few. Moreover, smartphone sensor data can help determine travel activities and patterns, the collection of which is traditionally done using expensive and labour intensive transportation surveys.
- Theme III: With the richness of data tsunamis from all sources, transportation agencies and roadway authorities will always face the challenging question of which system monitoring technology is the best to use. In the authors' view, this should not be

---

the main question as technology rapidly changes. System performance monitoring should be technology agnostic. With this in mind, fusion of sensory information offers improved accuracy and retains healthy redundancy in the system in a technology agnostic fashion. This theme therefore discusses data fusion methods that could alleviate the burden of choice of a given technology and amalgamates all data sources into one single and most accurate estimate using rigorous mathematical data fusion techniques. It is our belief that no data or a single source of information should be wasted; however, the outcome will be more accurate than any of the individual data sources.

The remainder of the chapter highlights the motivation of the open innovation framework, discusses each theme in further detail, and concludes with three applications that illustrate the potential of the framework and the underlying processes.

## **Background and Motivation**

### **The Proactive Transportation Management Challenge**

The proactive management of transportation and freight infrastructure generally requires continuous and real-time performance monitoring, control and management decisions, and dissemination of information and guidance to motorists and trucks. Over the past two decades, numerous technologies and methods have been developed and deployed for real-time monitoring of these infrastructure systems. In general, emerging advanced solutions to congestion and delays (in urban areas or at border crossings) rely on the collection, dissemination and sharing of information among stakeholders including travellers and system operators. It is of benefit to all to equip vehicles with sensors and communication capabilities to talk to the infrastructure, and equip the infrastructure with economic means for gathering information from vehicles and relaying content and services back to vehicles. What is equally important to such ubiquitous sharing of data and information is the ubiquitous sharing of knowledge, applications, and services, which is still emerging. For instance, in a regional setting such as the GTHA, authorities would share online software resources and web services to move from localized and freeway management to integrated corridor management and emergency management.

### **The Sustainability of Transportation Innovation Challenge**

Traditionally, organizations, large and small, innovate internally to breed solutions to their self-identified technical and business challenges. For decades, agencies have relied on breeding innovation internally by identifying needs, procuring services and solutions, fully funding projects, fully bearing the operational and maintenance costs, and the full risk in the entire ideas-products-services cycle.

In the highly competitive business world and dwindling economy worldwide,

---

sustaining innovation using such closed business models is neither affordable nor sustainable. Government agencies in particular, as opposed to for-profit-organizations, feel the innovation challenge, which is exacerbated by a continuous trend of budget cuts, yet accompanied by the natural desire to stay in control and stay relevant to the affairs of the public.

Luckily, the trend of closed innovation is rapidly shifting towards service-orientation, open innovation, and mass customization. It is our vision that agencies should seize the opportunity and examine the potential of this paradigm shift to engage transportation stakeholders and customers in a sustained ecosystem of co-creators and open transportation service providers, and therefore the need for open service innovation. Open service innovation can reduce the cost of transportation innovation, quality service provision, customer satisfaction and help to share the risks and rewards of transportation innovation, and accelerate the time required to deliver innovative solutions to travellers, firms and the population at large. With this in mind, governments in particular are well-suited to orchestrate and balance the evolution of such eco-systems, for all the good obvious reasons.

### **Connected Vehicles and Aspirations for the Connected Traveller**

Cars and transit vehicles are increasingly outgrowing their traditional role as motorized cabins for transportation from A to B. The authors use the term “vehicles” here as opposed to “cars” because connectivity must not be limited to the private automobile but will rather encompass all moving vehicles, including public transit buses, light rail and rail vehicles, as well as commercial trucks. Vehicles are turning into social and technical (socio-technical) hubs that connect themselves to other machines (other vehicles and components of the infrastructure) and connect their driver/passengers to his/her social and business world, in a productive yet safe manner. Around 90% of new cars produced today (2013) have Bluetooth, which is just the dawn of vehicular connectivity. Vehicles will be more and more connected via mobile devices, embedded telematics, dedicated communication channels and broadcast services. Connections can enhance emergency services, security features, traffic, weather and navigation information and services, infotainment and, of course, business productivity.

Equally importantly, connectivity can be capitalized upon to draw ridership to public transit where travel time can become travel-time-well-spent if put to good use. Our daily transactions go through a complex intertwined mesh of “live” networks and the vehicle is just a mobile node navigating its passenger in time and space through those overlapping and intersecting networks. In essence, experts envision the vehicle as a mobile device that must safely connect drivers and passengers to the web of everything around them, while being particularly cognizant and aware of the context in which the traveller is. It is not just connected vehicles, however, that we aspire for. We aspire for connected travellers,

---

a concept that is not yet fully established. In such a rapidly changing world, traveller services (information provision is one of them) are becoming essential services that must seamlessly stream into the vehicle (car, bus, train, truck and even planes), into mobile devices and ultimately into our daily activity chains and activity scheduling of travellers.

### **The Tsunami of Data Opportunities and Challenges**

A key component of creative transportation service provision is data. The basis of any transportation improvement decision is a set of performance indicators that reflects how the transportation system has been and is operating. With the increased market penetration of modern communication technologies such as Bluetooth, WiFi, Zigbee, and the emergence of LTE (Long Term Evolution) communication, we face both an opportunity and a challenge. The opportunity is to tap into the wealth of information and availability of wide bands of communication that provide microscopic system information at the level of individual vehicles and individual travellers and on a second by second basis if desired.

The challenge, however, is how to pick and choose between the alternate surveillance technologies. For instance, a simple task such as gathering traffic speeds can be achieved using many technologies including conventional loop detectors, Bluetooth and WiFi data collection stations or gateways, or from GPS devices carried by the travellers or embedded into their cars. Different technologies offer different data accuracies and we therefore face the challenge of how to choose the best and most accurate technology. Luckily, modern data fusion techniques not only alleviate the burden of choosing among surveillance technologies but also offer a fused estimate that is better than any of the individual technologies, which makes data fusion a sensible choice.

### **Theme I: Open Transport Service Innovation Platform**

The concept presented herein is highly inspired from the work of renowned business experts around the world, including the work of Henry Chesbrough of the Haas School of Business at the University of California Berkeley (Chesbrough, 2011).

In the context of this report, transportation is viewed as a service; transportation agencies are viewed as agents of customer-oriented service providers eager to innovate to contribute to the well-being of the society, and ITS is viewed as a suite of knowledge-intensive services.

In the knowledge-intensive service economy of the twenty-first century, large and small organizations realize that they must open up and work with external partners to innovate and provide their customers with the services that they need.

The future of advanced businesses and government organizations and advanced economies is shifting from products to services and rethinking business models to innovate and build upon them. Services, in the context of this paper, refer to knowledge-intensive services that are becoming the engine of growth for the entire developed world. Today

services comprise roughly 80% of economic activity in the United States and 60% in other developed countries (Chesbrough, 2011). Isolated products are becoming a smaller and smaller share of the economic pie. Most of the growth in services is happening in the knowledge-intensive portion of the service sector (Hertog, 2000), particularly services through an open innovation platform that allows for a co-creation business model.

To clarify the concept more clearly, open service innovation is perhaps best explained via an example (Figure 23) that is intended to emphasize the need to confront and transcend the so called commodity trap (Chesbrough, 2011). The well-known Motorola once dominated the cell phone market when they released the Razr cell phone model. The Razr was the thinnest, slickest and coolest cell phone on the market that everyone wanted to buy. But, where is it today? Motorola has fallen in the commodity trap, where their innovative product was soon copied and improved upon by many competitors. Motorola is no longer near the top of cell phone suppliers in the world. Coming up with even better cell phone products is no longer



**Figure 23: iPhone Platform vs. Motorola**

enough. Cell phone manufacturers face fierce pressures from newer entrants like Apple, Google, and Microsoft who are working hard to continue to innovate new handsets, either by themselves or with partners. But each is doing far more than offering newer products (Chesbrough, 2011). They are building platforms that attract thousands of other companies to design applications and services that run on their portable platforms, i.e. they are crushing the competition via adopting an open service innovation business model and open platforms. The Apple iPhone for instance, introduced in 2007 also had a catchy sleek design, an elegant customizable user interface, and a novel touch screen (i.e. interesting product). However, the iPhone was not just a device; it was a platform that attracted many third-party applications and services to provide users with a wide range of experiences on a single device. They created an ecosystem of co-creators. More than 100,000 individuals and companies have created 'apps' that run on top of the iPhone,

---

and more than two billion apps have been downloaded by customers around the world.

Apple created an ecosystem of innovators that, in addition to innovating for their own sake, serve Apple's interests in the process. For other big companies such as Google and Microsoft to compete they are striving/struggling to create parallel ecosystems similar to Apple's. While it may be arguable which of the three dominant ecosystems (Apple, Google or Microsoft) is more "truly open" than the other, they all have in common an open business model and an open platform.

While transportation agencies are not producing cell phones and are not competing, in a sense, with other agencies, they provide transportation services to travellers and can benefit from a similar transportation-specific ecosystem to enhance the quality of serving the public and to leverage limited dollars. In the ITS domain in particular, transportation agencies face that challenge of how to fit amongst a myriad of new private sector entrants into this service market. Agencies also face the challenge of how to collaborate with the many organizations that hold a stake in this business.

Open innovation in services is a clear and sustainable way to sustain and grow the transportation agencies' services and influence. As in the Apple case, creating platforms that incorporate internal and external innovative services and surrounding these platforms with a variety of value-added services, agencies can invigorate transportation service innovation to ultimately offer travellers with state-of-the-art services, at low cost and low risk. This business model will lead the transportation sector by offering an entire constellation of innovative services created both internally (at the agency level) and externally by others and made available to their customers through an open single point of access or platform.

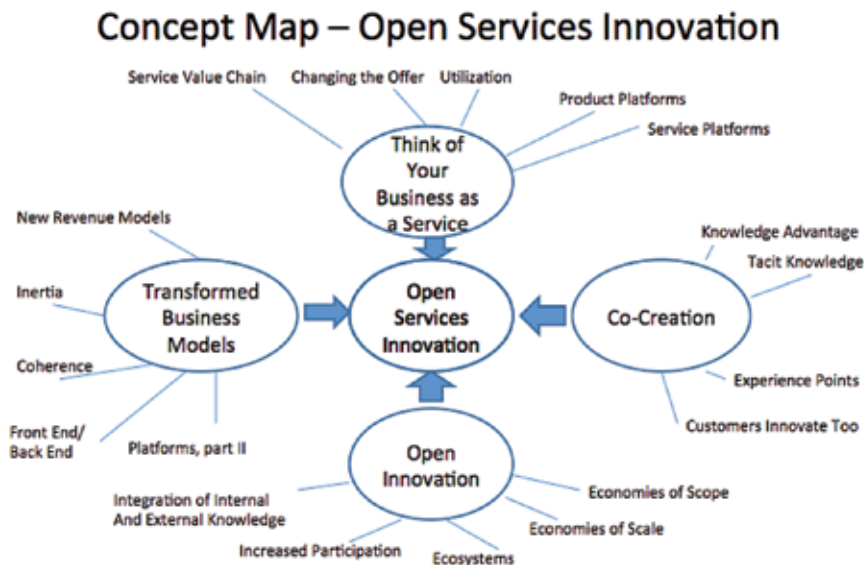
### **The Transport Open Service Innovation Framework**

Capturing the Chesbrough model (Figure 24), we summarize the four foundational concepts that must be established and that together create the driving framework of open ITS:

1. Adopt the mindset of transportation business (e.g. TIS, ATMS) as an open services business in order to create and sustain operations in the knowledge-intensive and limited budget economies of today and to create differentiation in a commodity-trap world (e.g. avoid creating and investing in a closed owned service such TIS that may be made obsolete in a few years).
2. Invite customers (or stakeholders in general such as travellers, municipalities, Emergency Management Services (EMS), universities, software providers, etc.) to co-create innovation in order to generate the experiences they will value and reward.
3. Use open innovation to accelerate and deepen services innovation, making innovation less costly, less risky, and faster. Use open innovation to create a platform for others to perpetually build on.
4. Transform the transportation services business model with open services innovation,

which will help to profit from internal innovation activities and from others' innovation activities as well.

**Figure 24: Open Service Innovation (Chesbrough H., 2011)**

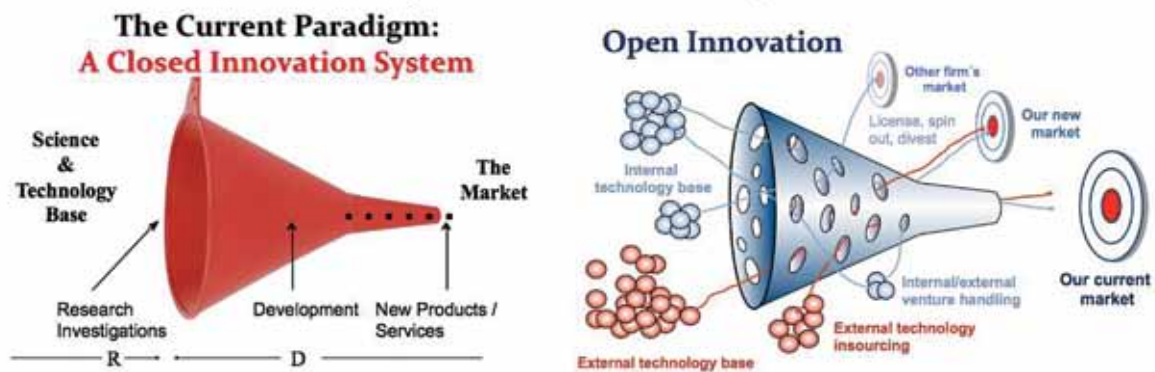


Open service innovation brings stakeholders and customers directly into the innovation process through a common platform as opposed to treating them as passive one-way recipients of whatever transportation agencies provide. An important aspect of this open innovation business model is to be able to capture stakeholders' tacit knowledge via adding social networking dimensions to the service innovation platform.

The key benefit from the open service innovation business model is to harness the power of participation of many more innovators from outside the transportation agencies. With the diffusion of knowledge, ideas, data and access to infrastructure, more organizations can experiment in parallel with possible ways of using and combining these resources in novel ways. No other single organization can hope to compete with this external explosion of potential offerings by relying exclusively on their own internal knowledge. In fact, external organization will not want to compete but rather collaborate. When the internal and external elements are combined, they produce a wealth of offerings for customers (travellers, operators etc.) while allowing the other providers to specialize in their own distinctive competences.

In summary, transportation agencies would purposively use external ideas as well as internal ideas, and internal and external paths to market, to create new services, new architectures, and new systems as illustrated in Figure 25 below.

**Figure 25: Closed vs. Open Service Innovation Business Models (Chesbrough H., 2011)**



## Theme II: A Sensing Platform and Gateway for Traffic, Transit and Freight Monitoring

This theme is motivated by the associated high costs as well as the inherent limitations (e.g. power consumption, telemetry, etc.) of existing traffic monitoring technologies. Infrastructure monitoring is typically conducted either on a short-term basis or only at the most critical locations or both. Continuous real time monitoring is typically deployed only at specific key locations (e.g. on major freeways in large cities). Furthermore, limiting the space and time of collected data may consequently limit the utility of collected data, particularly when real time applications are of interest for either efficiency and/or security purposes. This theme intends to encourage pervasive low cost technologies to enable ubiquitous/pervasive sensing and tracking of mobile units (cars, trucks, containers, cell phones, PDAs, etc.) and enable communication and connection between transportation infrastructure (roads, terminals, border crossings, etc.) and vehicles (aka connected vehicles) and/or the mobile units.

Generally, there are three main components for wireless monitoring of infrastructure systems. First, there are the sensing devices collecting the information from the environment. Second, there are the delivery components (i.e. devices, algorithms and protocols), delivering the sensed data to processing elements. Finally, there is the data management centre responsible for developing and/or orchestrating applications based on the collected data. In this chapter, we discuss an example of such integrated device or gateway as well as system components that facilitates sensing and delivery of information.

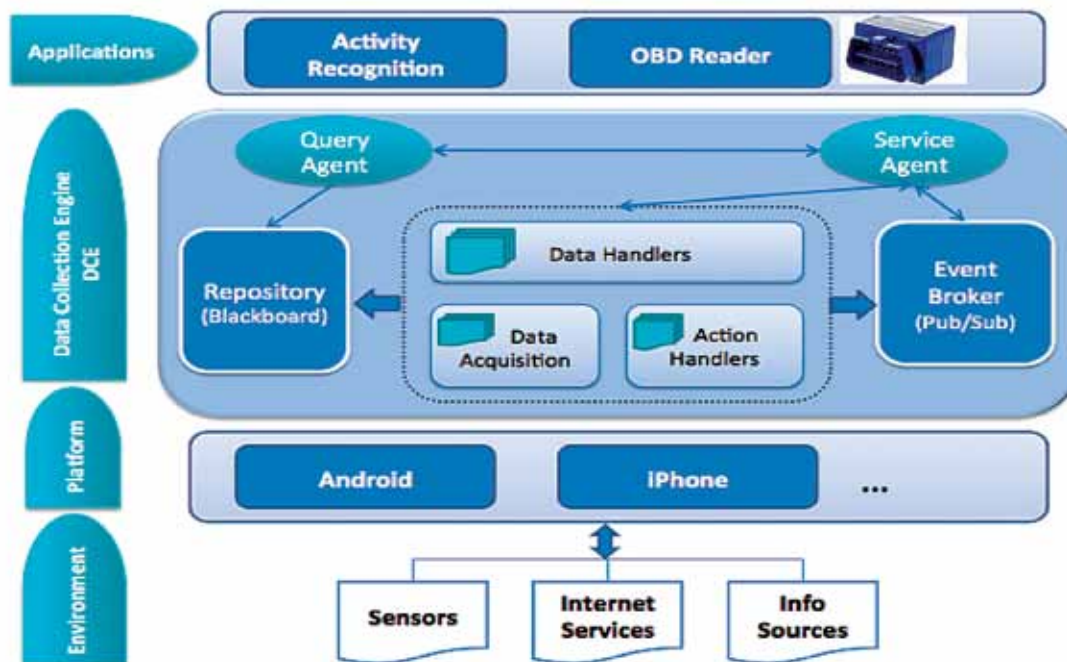
## A Gateway for Multi-Protocol Data Sensing and Delivery

This section highlights the design and integration of a sensing, delivery and management platform and gateway for monitoring traffic, transit, and freight information as follows:

1. As vehicles drive anywhere in the network, using On Board Diagnostic (OBD) scanner and mobile data sensing and telecommunication technology, and/or
2. As vehicles drive near a road-side mounted gateway.

In the first case, an OBD scanner sends vehicle sensor information to a mobile smartphone using Bluetooth and/or WiFi (depending on the user device preferences) or in the near future, connected vehicles equipped with built in Dedicated Short Range Communication (DSRC). The smartphone augments the OBD information with GPS geocoding and time stamp and sends the information via the Internet to a central server. The collected information is saved in a GIS database at the central location (e.g. in the ONE-ITS.net platform, which will be described later). Figure 26 shows the software architecture that runs on the smartphone with the OBD scanner.

**Figure 26: Smartphone Application Architecture**



---

In the second case, a roadside gateway gathers information for vehicles in the vicinity that have Bluetooth, WiFi or ZigBee devices, or DSRC in the near future, and then forward gathered information to the central server. Each gateway is configured to run an open-source Linux distribution for embedded devices. For the purposes of interfacing with different sensors, the gateway is augmented with dongles for Zigbee, WiFi and Bluetooth as well as a low-cost camera. With this setup, each gateway is controlled to scan the wireless medium for sensors in proximity. The gateway is capable of retrieving information reported by the sensors. The gathered information is then relayed to a centralized server on the Internet using TCP/IP.

It is to be emphasized that such gateway, components and overall platform offer the following desirable features:

- Infrastructure-light: as it relies on crowd-sourced data collection techniques;
- Open & Sustainable: as it can be extended to incorporate different data management algorithms and techniques and/or operate with other wireless technologies;
- Uses unlicensed wireless technologies (e.g. ZigBee, Bluetooth, WiFi), which are available as consumer electronics and with minimal requirements for usage licenses;
- Cost-effective: to allow for pervasive deployment;
- Scalable: uses hierarchy capability of increasing the size of the coverage area in a manner that does not adversely affect system performance; and
- Flexible: Both ZigBee and WiFi can self-configure themselves to build wireless mesh networks with virtually no deployment cost. This suits the dynamic environment inherent to travelling cars, buses, trucks and containers.

### **Theme III: Multi-Sensor Data Fusion**

Widespread technological development and deployment has created an abundance of data sources for traffic monitoring. Often, there are multiple independent measurements of the current traffic conditions for the same portions of the network. In these cases, a variety of data fusion techniques can be used to achieve better estimates while helping to overcome information overload. These data fusion techniques are applied to the data streams from the gateways described under Theme II above and fused together with any other data source such as loop detectors.

Data fusion is the process of combining data such that the fused estimate is better than those based on individual data sources (Mitchell, 2007; Hall, and Llinas, 2001). Fusing data from competitive sensors, those that provide independent measures of the same property (e.g., speed of traffic), can be challenging, as the quality of these data sources must be evaluated (Brooks, and Iyengar, 1998; Luo, and Kay, 1989).

---

Depending on the application at hand, data integration and fusion can realize a number of benefits including: Reliability/Robustness/Redundancy – as the system does not depend on a single source of input; Accuracy/Certainty – as combining several readings from the same sensor makes a system less sensitive to noise and temporary malfunctions; Completeness/Coverage/Complementarity – as more data sources will provide extended coverage of information on an observed object or state.

In summary, data integration is about bringing data together in one place. Data fusion is about using the data together such that there is some new or better inference to be had. Together, data integration and data fusion can create a variety of benefits.

### **Data Fusion Example: Emerging Technologies for Traffic Data Collection**

This section highlights the utility of data fusion using widely existing legacy technologies such as loop detectors and emerging technologies such as Bluetooth sensors. This is only one example, however. The techniques are generic and can be applied to any combination of technologies including Bluetooth, GPS, DSRC etc.

Loop detectors are the most widely, and conventionally, used sensors in freeway traffic management (MTO, 2013). The main function of loop detectors is to detect the presence and speed of vehicles on the freeway.

Probe vehicle data have been traditionally more difficult to obtain. But recently, there has been an interest in developing an anonymous probe vehicle monitoring system to measure travel times on highways and arterials based on wireless signals available from technologies such as Bluetooth. Other technologies also lend themselves well to probe vehicle data collection. For example, cellular telephone tracking is becoming another popular method for collecting probe vehicle traffic speeds, especially smartphone and OBD data collection as discussed in Theme II. There is a general shift from stationary (loop detectors, traffic cameras) to mobile (Bluetooth, GPS, OBD, Smartphones, etc.) sensors. Collecting probe vehicle data is desirable because the measurements can be more accurate, although of variable quality, and generally have good spatial coverage. In addition, the required infrastructure for probe vehicle tracking is very light and relatively inexpensive. However, while probe vehicles describe the state of traffic on the entire road segment, they are not exhaustive as only a small portion of the vehicles is tracked.

However, there are numerous issues to consider with each type of sensors. GPS data, cell phone data, and Bluetooth data require a substantial amount of refining (e.g., map matching) in order for the data to be useful and reliable. Also, there are limitations to using fleet vehicle data (including buses or taxis) because the operational characteristics of these vehicles are different from normal traffic. For example, when buses stop to serve passengers, their travel time will include the dwell time and deceleration and acceleration delay. Even when not stopping to serve passengers, buses have different performance

---

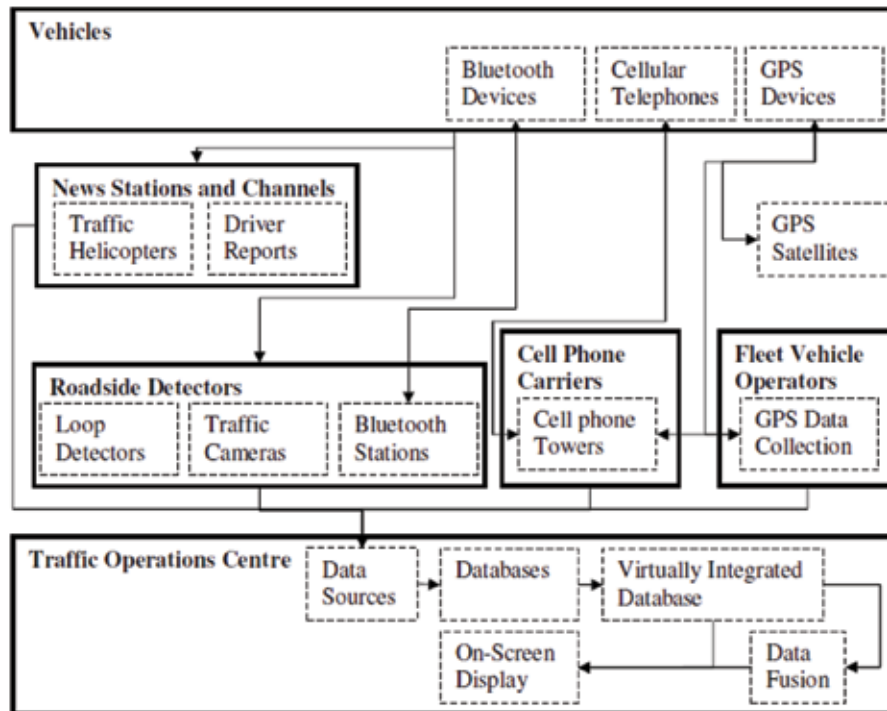
characteristics because of their large size, and they generally travel slower than standard passenger vehicles. Similar issues exist for truck and taxi fleets. The sample size of probes from these fleets is also an issue that should be considered in fusion efforts. Lastly, these data arrive in different formats and require various amounts of pre-processing before data fusion can be applied.

An integrated framework such as ONE-ITS ([one-its.net](http://one-its.net)), described in the next section, will maximize coverage of the network given the available data from different sources. For example, loop detectors and traffic cameras might monitor busy freeways, while GPS data from vehicle probes may cover urban streets. In this way, data integration alone is all that is necessary for successful traffic management and operations.

However, the same location might be covered by more than one competitive sensor. That is, the data will be providing independent measures of the same property at certain times in certain locations. For example, the traffic operations centre might receive loop detector data and Bluetooth device data describing the state of traffic on the same stretch of freeway at the same time. Note that a traffic operations centre would not benefit from several different technologies providing what should be the same information (e.g. travel time data). In fact, the abundance of data would only overwhelm decision makers and slow down their ability to make timely decisions. Rather, a single fused intelligent inference resulting from all of these data sources would be preferred. Therefore, data fusion is required to use the data most effectively and efficiently. Figure 27 summarizes the various technologies, collection methods, and stakeholders required for data integration and fusion.

The discussion in Theme III is built on the recent theoretical data fusion contribution conducted at the University of Toronto (Bachmann, Abdulhai, Roorda and Moshiri 2013), in which a fusion platform is presented based on the most promising methods for combining measurements from competitive sensor networks. After investigating data fusion methods from a wide variety of fields such as target tracking, artificial intelligence, multi-criteria decision-making, and other data fusion literature, it was found that all these methods share the ability to take multiple estimates, and fuse them to make one superior estimate.

**Figure 27: System Architecture for Traffic Monitoring (Bachmann, C., Abdulhai, B., Roorda, M.J. and B. Moshiri, 2013)**



The methods investigated at the University of Toronto include:

- Simple Convex Combination
- Bar-Campo/Shalom combination
- Measurement fusion Kalman filter
- Single Constraint At A Time (SCAAT) Kalman filter
- Ordered Weighted Averaging (OWA)
- Artificial Neural Networks

An application of the data fusion techniques is presented later in this chapter.

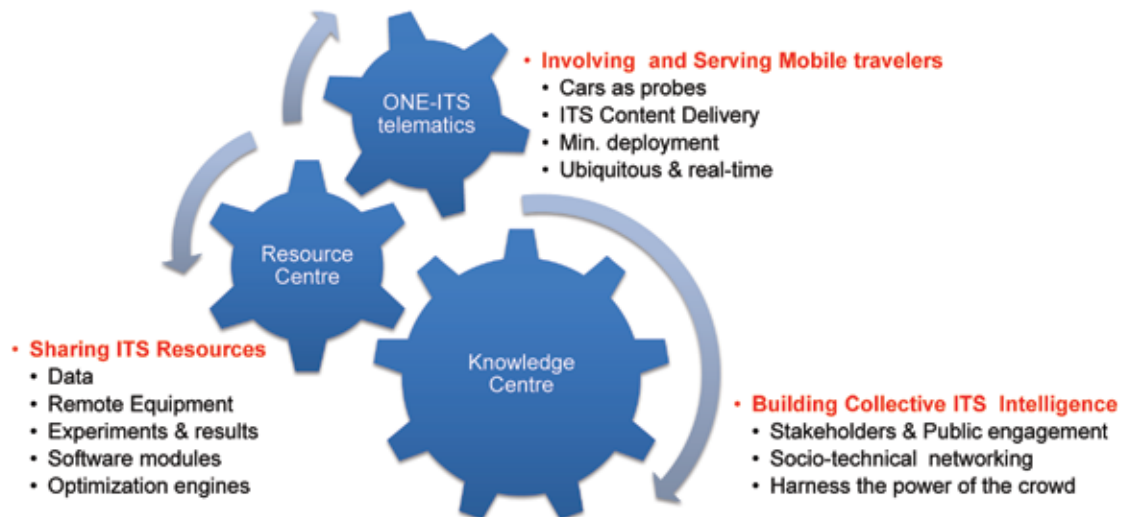
## Enabling Open Intelligent Transportation Systems: the Tools

### Online Network Enabled-Intelligent Transportation Systems (ONE-ITS)

ONE-ITS (one-its.net) is a pioneer multi-stakeholder initiative that exploits modern web, communications, and software technologies to enable collaborative innovation, research and development activities among widely dispersed stakeholders. Stakeholders have common objectives of easing traffic congestion, enhancing safety, reducing stress, reducing fuel consumption and pollution, protecting the environment, and promoting urban sustainability. ONE-ITS does this by allowing access to all participating parties to the applications, data and knowledge resources that have been acquired and accumulated in the system based on the concept of open innovation discussed in Theme I.

The power of the 'crowd' is what inspired the development of ONE-ITS to provide the means and tools to tap the potential of the many important, but fragmented ideas and views, experts and expertise, data sources, software solutions and applications, communication infrastructure, computing infrastructure, etc. This concept of harnessing the power of participants of many innovators enables ONE-ITS to integrate the following stakeholders: travellers (through ubiquitous and standard user interface), application developers (through client integration), service developers (through service integration and mashups), data providers (through data integration, security, scalability), system operators, and policy makers. This not only benefits collaborators, who contribute and benefit from the tacit technical knowledge, but also creates a social networking dimension to the service innovation platform. In essence, ONE-ITS is designed to establish a Socio-Technical network that transcends fragmentation via connectedness and inclusion of heterogeneous stakeholders. With this structure in place, ONE-ITS enables a three-prolonged approach via sharing resources, intelligence and services as shown in Figure 28 below.

**Figure 28: ONE-ITS: A Three-Prolonged Approach**

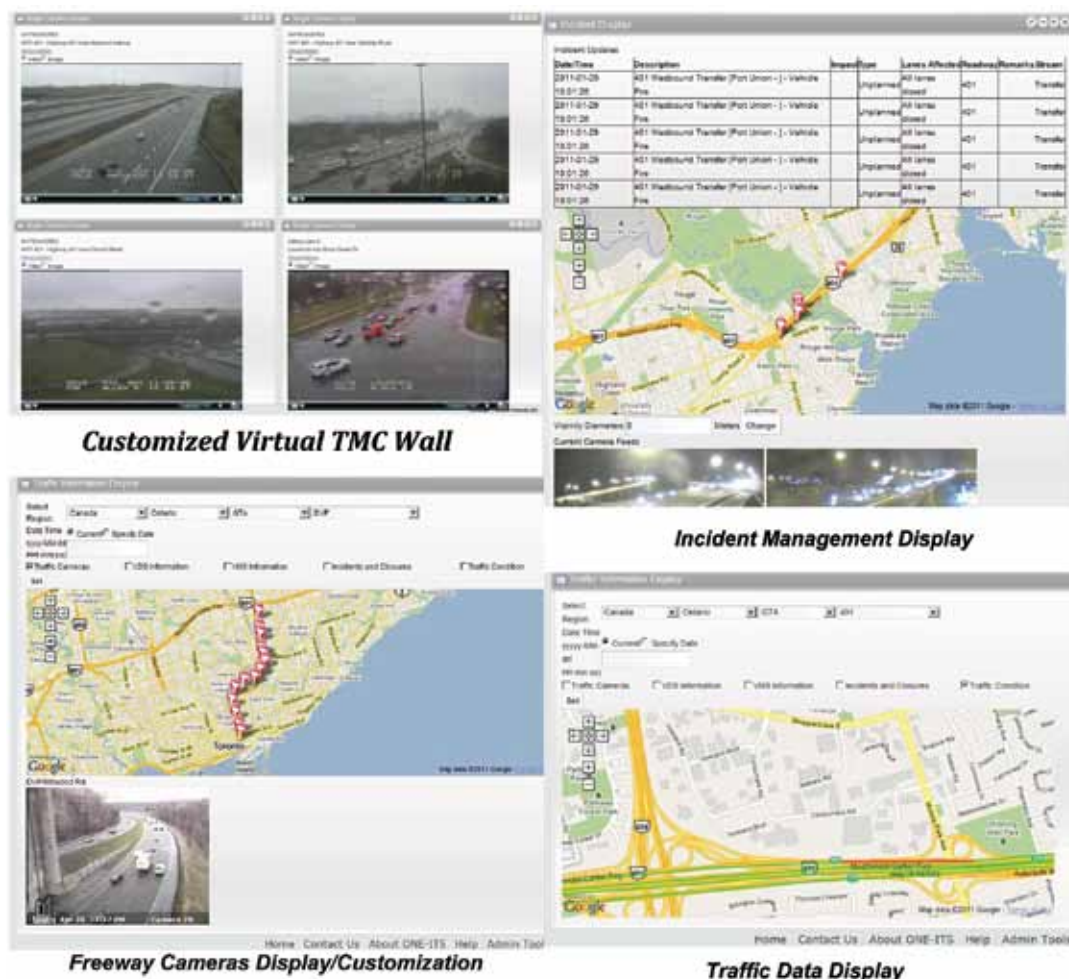


The following sections highlight some of the applications within the ONE-ITS framework and the benefits resulting in each of each sector (public, private, and researchers):

## Benefits to the Public Sector

Benefits to the public sector include multi-agency collaboration (e.g., City Toronto and Ministry of Transportation of Ontario). Practitioners and public sector staff can stay on top of the latest developments in the ITS research worldwide and voice their research needs and positive and negative experiences with various ITS products. In addition, it provides web services which means easy implementation with plenty of opportunity for “try-before-you-buy.” ONE-ITS examples for potential benefits to the public sector include: incident management (location and routing of emergency vehicles), virtual traffic management centre (customized wall for system operators and agencies staff), geospatial visualization (e.g., colour-coded traffic data, freeway cameras, etc.) all being integrated from different sources/services as shown in Figure 29.

**Figure 29: ONE-ITS Applications to Public Sector**

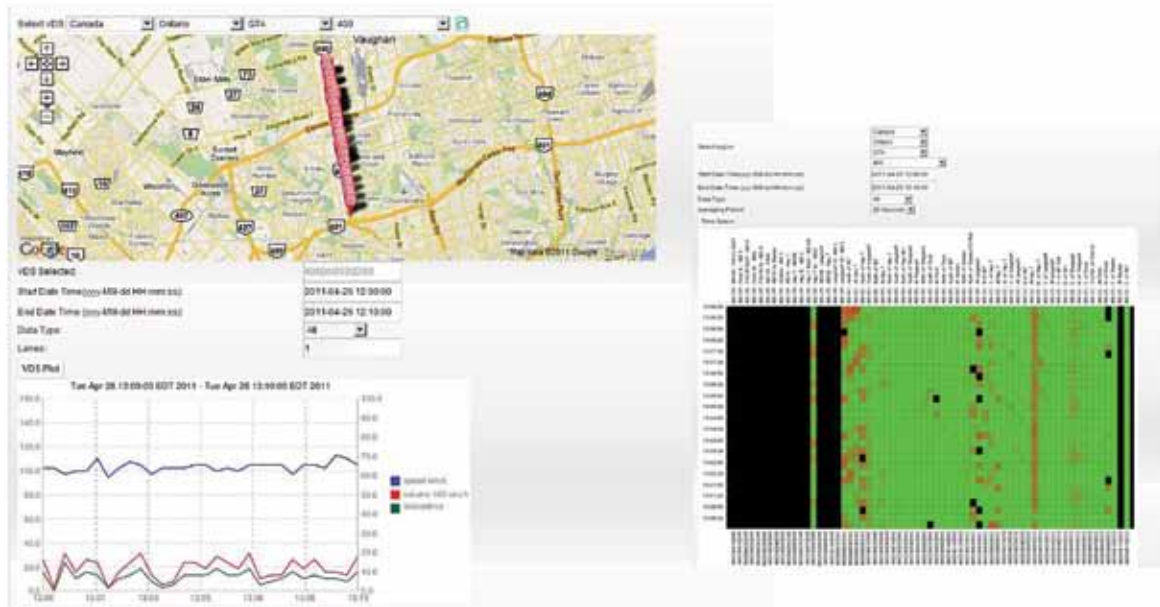


---

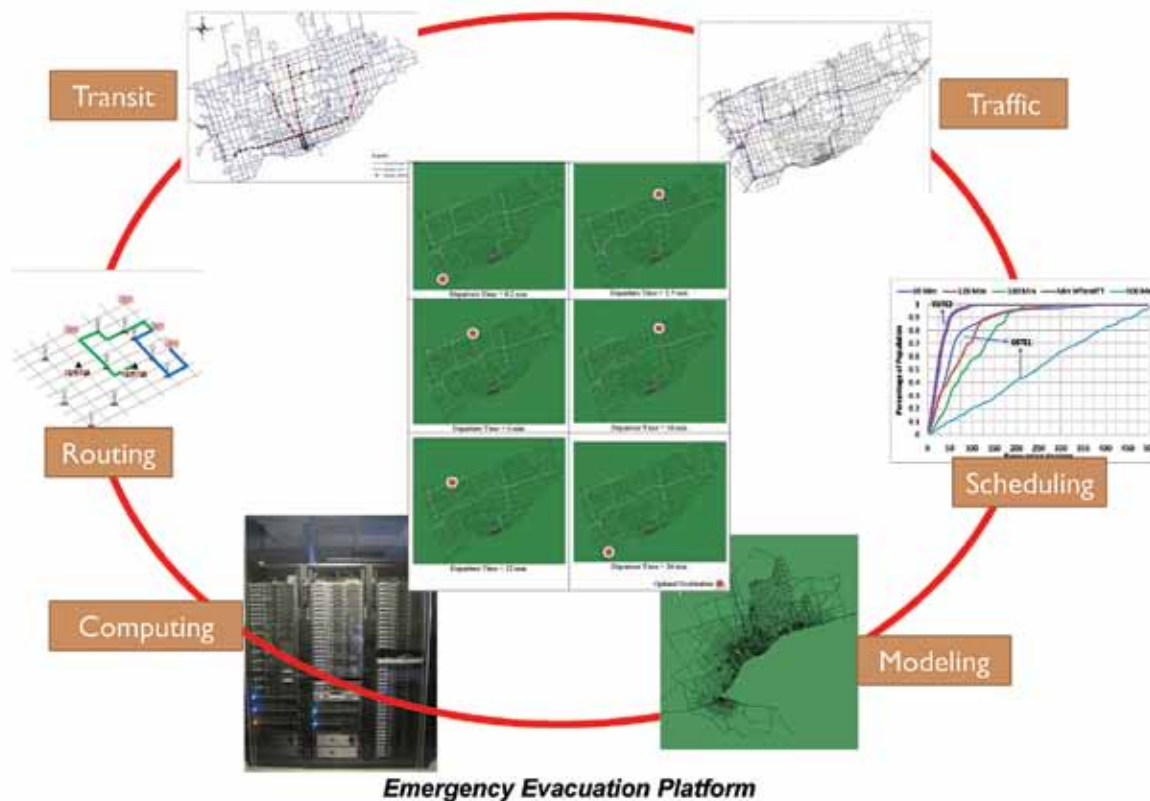
### **Benefits to Researchers**

ONE-ITS provide meaningful and rich data for researchers, a virtual online community of researchers who can share their ideas and findings online, web services that can be easily mashed up to develop more complex web services, and web services also means that applications are readily transferable and can easily be implemented in traffic management systems (e.g., multimodal emergency evacuation); all being mashed up using different software, engines and artificial intelligence techniques as shown in Figure 30. For example the emergency evacuation platform developed by (Abdelgawad and Abdulhai, 2010) is built on the following integral components of ONE-ITS: simulation models for traffic and transit, GIS, routing optimization algorithms, and Genetic Algorithm optimization and high performance computing facility. Each of these services could be integrated and used for other applications; such that researchers can harness the existing technologies and capabilities and use open innovation to create a platform for others to build on perpetually.

**Figure 30: ONE-ITS Applications to Researchers**



**Traffic Patterns Plots (Loop Detectors and VDS Plots)**



---

### **Benefits to Private Sector**

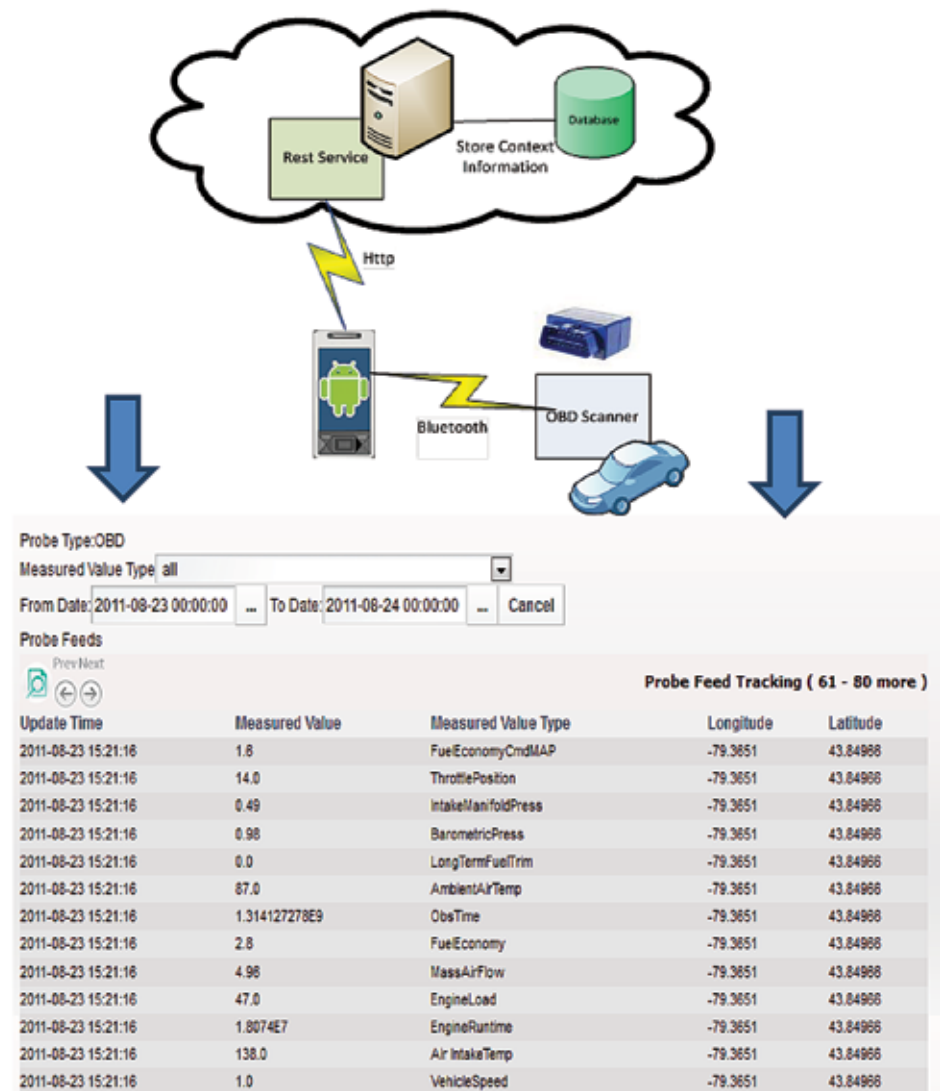
ONE-ITS also provides services and product exposure of private sector and enables working interactively with the research community to take promising research from concept to market. In addition, end-user applications can easily be developed using the base web services developed by the researchers as the heart of the system and eventually, structure and regroup these applications to respond to market needs.

### **A Sensing Platform using Smartphone and On Board Diagnostics Device Application**

As part of Theme II objectives, this application is developed to sense vehicles/travellers as they drive anywhere in the network, using OBD scanner and mobile data sensing and telecommunication technology.

In this application, two Android applications were developed and tested. The first application tracks users by periodically sending their GPS coordinates to a central server. The second application, as illustrated in Figure 31, collects the engine diagnostics of vehicles by interacting with an OBD scanner mounted on these vehicles.

**Figure 31: OBD Tracking Application and Sample Results**



---

This OBD information is then sent to the ONE-ITS server allowing the platform to keep track of many attributes such as vehicle speed, throttle position, ambient air temperature, etc. as well as user activity types such as work, shop, or travel by mode. With this data stored on the ONE-ITS open service platform, this rich data can then be fused with other data sources or can be used by researchers or agencies internally in their organization, and therefore harnessing the concept of open innovation.

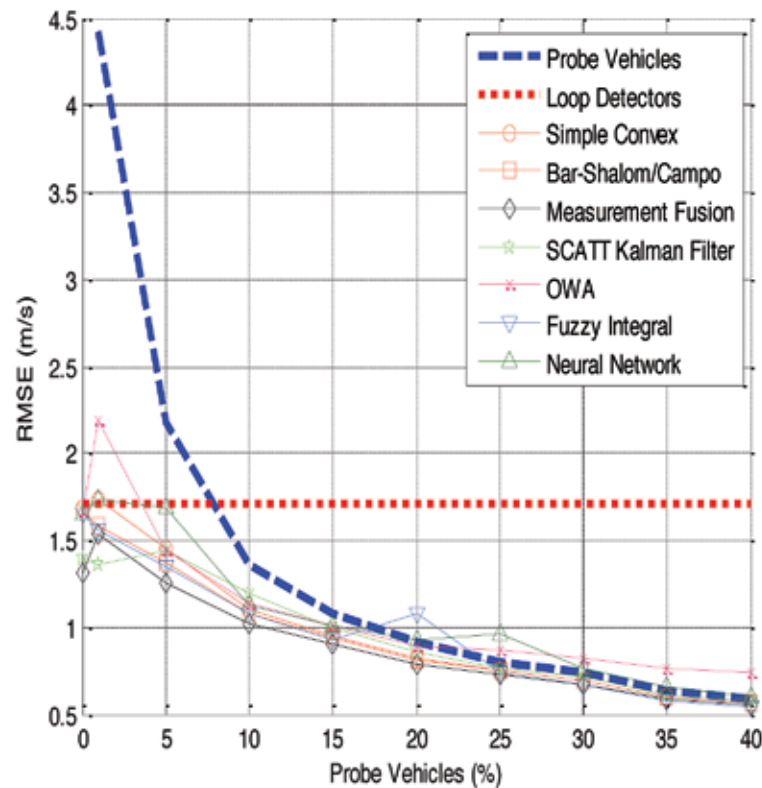
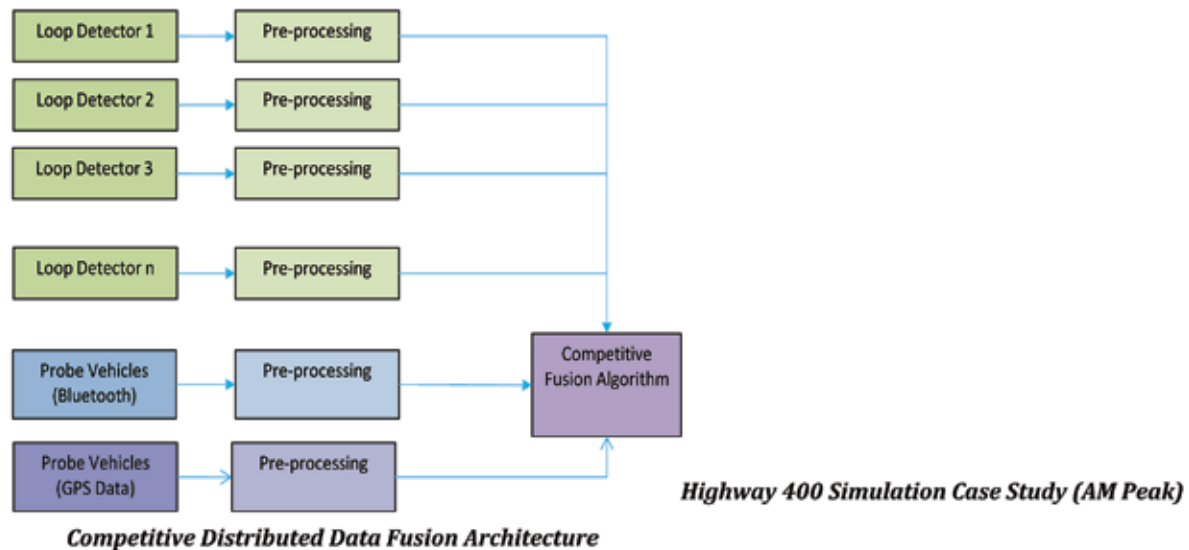
### **Data Fusion System Application for Speed and Travel Time Estimation**

As part of Theme III objectives, this application is conducted to evaluate the effectiveness of each of the data fusion methods discussed above for travel time/speed estimation. The test experiments were conducted on a simulation environment that fused probe vehicle data and loop detector data and compared the fused result to the ground truth traffic conditions (using all vehicle data) as shown in Figure 32. As shown in the figure, the application makes use of competitive distributed data including multiple loop detectors, probe vehicles and GPS data on a simulated freeway (Hwy 400) in the City of Toronto. The standard deviation and mean absolute error were used as measures of effectiveness and compared among different data fusion methods to analyze performance.

The following conclusions can be drawn from the experiment conducted on the simulation environment:

- All data fusion methods realize improvements that are statistically significant in some cases.
- Greatest improvement in accuracy with small numbers of probe vehicles.
- Simple convex combination, Bar-Shalom/Campo combination, and the measurement fusion Kalman filter perform best.
- No result is statistically worse than the best sensor used independently (i.e. never lose accuracy).

**Figure 32: Data Fusion Application (Bachmann, C., Abdulhai, B., Roorda, M.J. and B. Moshiri, 2013)**



**Sample Speed Estimation Error on a Highway Link using Different Data Fusion Methods**

---

## Summary

In this chapter we presented the concept of open service innovation for ITS services (Open – ITS) by integrating three intertwined Themes; namely: Open Transport Service Innovation Platform, Sensing Platform for Traffic Monitoring, and Multi-Sensor Data Fusion Framework. The open transport service innovation platform goal is to enable transportation agencies to purposively use external ideas as well as internal ideas, and internal and external paths to market, to create new services, new architectures, and new systems. The sensing platform for traffic monitoring builds on the concept of ubiquitously collecting data from multiple sources while harnessing the emerging technologies and smartphone sensory data. With these rich data and multiple sources of traffic information, the multi-sensor data fusion could not be more emphasized.

This chapter also presented three applications that demonstrate the three themes discussed; namely: Online Network Enabled- Intelligent Transportation Systems (ONE-ITS), A Sensing Platform using Smartphone and On Board Diagnostics Device Application, and Data Fusion System Application for Speed and Travel Time Estimation. To the best of the authors' knowledge, this is the first attempt to furnish the foundation for open transport service innovation concepts and mass customization for the new generation of ITS by proposing an Open-ITS system.

## Chapter IV References

---

Abdelgawad H. and Abdulhai, B. (2010). *Managing Large-Scale Multimodal Emergency Evacuations*. The Journal of Transportation Safety and Security, Volume 2, Issue 2 , pages 122 – 151.

Bachmann, C., Abdulhai, B., Roorda, M.J. and B. Moshiri. (2013). *A Comparative Assessment of Multi-Sensor Data Fusion Techniques for Freeway Traffic Speed Estimation*. Transportation Research Part C – Emerging Technologies. 26: 33-48.

Brooks, R.R., Iyengar, S.S. (1998). *Multi-sensor Fusion: Fundamentals and Applications with Software*. NJ.: Prentice Hall PTR, Upper Saddle River.

Chesbrough H. (2011). *Open Service Innovation: Rethinking your Business to Grow and Compete in a New Era*. Jossey - Bass A Wiley Imprint.

Hall, D.L., Llinas, J. (2001). *Handbook of Multisensor Data Fusion*. FL.: CRC Press, Boca Raton.

Hertog P. (2000). *Knowledge-Intensive Business Service as Co-Producers of Innovation*. International Journal of Innovation Management.

Luo, R.C., Kay, M.G. (1989). *Multisensor integration and fusion in intelligent systems*. IEEE Transactions on Systems 19 (5), 901-931.

Mitchell, H.B. (2007). *Multi-sensor Data Fusion: An Introduction*. New York, NY: Springer.

MTO. (2013). *Freeway Traffic Management Systems*. Retrieved Feb 28, 2013, from <http://www.mto.gov.on.ca/english/traveller/trip/compass-ftms.shtml>.

## Closing

---

### Looking Ahead: Autonomous-Vehicles – The Next Traffic Revolution

In closing, this report would not be complete without a look ahead. What is the next traffic revolution? Many experts believe that the next traffic revolution will be autonomous vehicles, or, less technically, driverless vehicles.

An autonomous car is a motor vehicle capable of fully automated driving and navigating entirely without human input, i.e. without a driver. Autonomous vehicles rely on advanced technologies, including arrays of sensors, software and electromechanical systems to navigate paths while avoiding obstacles in uncharted environments.

Putting aside the novelty factor, autonomous vehicles have significant potential to address congestion problems. A few examples of the anticipated advantages:

1. Increased road capacity and reduced congestion due to their abilities to operate at higher speeds, with shorter gaps and much faster reaction time compared to conventional driving. The difference can be dramatic. A conventional freeway lane carries 2,000-2,400 vehicles per hour. Full automation can double or triple these values.
2. Improved road safety and reduced accidents due to the elimination of driver error, which is the main contributor to traffic accidents. Eliminating accidents and particularly fatalities is priceless in terms of saving human lives. However, it also reduces congestion quite significantly because around half of the congestion we endure on a daily basis is attributable to traffic incidents (not just high demand).
3. Enhanced traveller convenience due to relieving the driver from the chores of driving.
4. Removal of the constraints on driver's state and whether the driver is well trained, physically fit to drive or mentally alert.
5. Reduced need for vehicle ownership if autonomous vehicles are shared amongst several users, in addition to reduced need for parking space as users need not park their car while at work for instance. The car would be called and used by others.
6. Many other benefits such as reduced fuel consumption, emissions and insurance cost.

Despite such potential, autonomous vehicles still face some challenges and/or unknowns, such as (Smith, 2012):

1. Autonomous driving may increase the amount of motor vehicle travel by groups who currently cannot legally drive themselves because of youth, age, disability, or incapacitation.
2. Autonomous driving may encourage suburban sprawl by increasing the acceptable commuting distance.
3. In the early stages, it is not clear how automated vehicles and human driven vehicles will mix together. For instance, for safety and liability reasons autonomous vehicles may be programmed to maintain generous time headways and to proceed overcautiously after stopping at intersections.

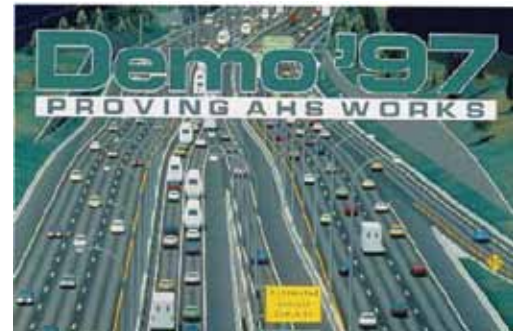
---

Nevertheless, autonomous vehicles came much closer to reality since the now famous Google Car, with a visually impaired person behind the steering wheel (I do not even know whether to call him passenger or driver!). Although Google deserves a big credit for “mainstreaming” the autonomous vehicle concept in North America, bigger credit is due to the science champions of the Automated Highway Systems Consortium (AHSC) and their groundbreaking Demo 1997 in California.

The core technologies behind autonomous vehicles date back to the 1950s and 1960s when, for instance, General Motors envisioned the “Highway of Tomorrow”. Since then, researchers have been busy inventing and tweaking bits and pieces of technology that would enable longitudinal and lateral control of vehicles. A myriad of technologies were developed and tuned including the use of wireless communication (e.g. DSRC or Dedicated Short Range Communication), radar, magnetic lane keep guides, vision and video-based obstacle avoidance, and of course GPS navigation and many other technologies. Some of these technologies are already in the market today as options and gadgets in higher end of vehicles, such as adaptive cruise control, collision avoidance systems, active driving and lane-keep assist, infra-red night vision and the like. Soon, however, these technologies will be cannot-do-without components or perhaps legislated the same way good brakes or functioning headlights are. It is worth noting that the State of Nevada issued the first license for self-driving car in 2012. Florida and California soon followed the emerging trend.

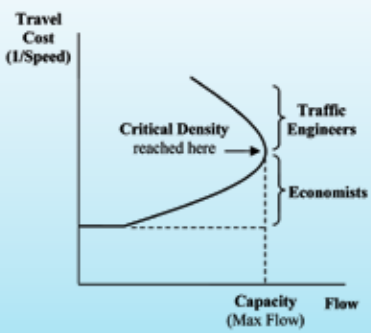
In our view, the AHSC Demo 97 in California broke new ground showing the world that the science and technology are there and driverless cars are ‘theoretically’ possible. As a side note, California often gets discredited for inventing congestion, but they get much less credit for inventing solutions, unduly so. After Demo 97, the challenges then were not so much related to the technology but were more so on the side of liability, interoperability, practicality and public acceptance of the possibility. Google, a decade later, gets significant credit bringing autonomous vehicles closer to the public minds, making lay people ponder, think and perhaps accept the possibility. Researchers and the auto industry are working around the clock to seize the ripe public minds to make autonomous vehicles a daily fact, much like what Apple did with the iPhone.

Stay tuned and look forward to the autonomous vehicle revolution on the roads.

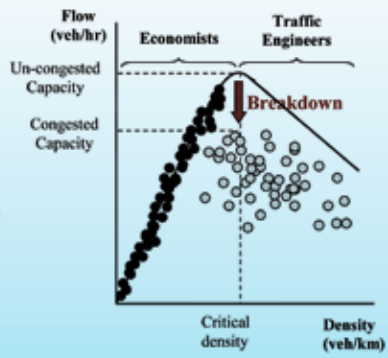


## References

Smith, B.W. (2012). Managing autonomous transportation demand. Santa Clara Law Review 52, 1405-1426



a) Travel cost vs. traffic flow



b) Traffic flow vs. traffic density



View this report and more at  
[rccao.com](http://rccao.com)