Chapter 7

Urban Traffic Signal Optimisation

Gaetano Fusco, Guido Gentile, Pietro Meschini*

DICEA, University of Rome "La Sapienza"

ABSTRACT

This chapter aims to introduce the reader to the fundamentals of signal based traffic regulation, and trace the progress made so far in the attempt to render signalisation of urban areas a truly proactive feature of the infrastructure, capable of adapting to — and in some cases positively influencing — the road user demand.

The main features that characterise a signalisation plan are first formalised, so as to introduce the basic principles for the optimal management of conflicting manoeuvres at a single intersection in relation to specific traffic conditions.

An overview is then given of the most prominent strategies used to further optimise signalisation plans in sight of complex network-wide objectives, or to enact specific traffic management policies, following a distinction between offline Signal Setting processes used to design pre-timed plans, and online Signal Control methods relying on real-time data to adapt dynamically to the immediate circumstances.

Keywords: Signal Setting, Signal Control, Adaptive Regulation, Intersection Coordination, Traffic Gating, Green Bandwidth Maximisation.

INTRODUCTION

The fundamental role of *traffic signals* is to equitably and efficiently administer the right of way amongst conflicting streams of road users.

Since the first sporadic appearances around the turn of the 20th century, traffic lights have become a ubiquitous feature in the everyday life of all road users, regardless of their preferred mode of transportation: whether they sit behind the wheel of their own car, walk, or let the public service carry them about their business, there will be traffic lights regulating their movements and those of others around them (most noticeably those in front).

It is therefore natural that traffic signals should garner so much attention: on one hand they are perceived (only sometimes unfairly) as a major source of delay and frustration to drivers; on the other, the tantalising idea of an intelligent traffic control system often comes to identify, in the general public's fantasy, with the very notion of ITS.

In fact, a history of case studies shows that wherever public money has been invested into the development and maintenance of a signalisation system tailored to the transportation needs of a community, the returns have invariably surpassed expenditures by far [16].

* Corresponding Author Email: pietro.meschini@uniroma1.it

Carefully planned signalisation allows a more efficient use of the existing road infrastructure, minimising the stress suffered by drivers as well as the risk of accidents, favouring public transport and improving air quality, with a positive impact on virtually every aspect of life in a modern city.

This chapter outlines the main features that characterise the traffic signalisation system of an urban network, and proceeds to expand upon some of the most successful design and optimisation techniques used for Signal Setting and Signal Control, the distinction between which may seem to be a fine line — often blurred even in technical literature — and is worth clarifying right away, as will be understood throughout the chapter.

The term *Signal Setting* refers to the pre-determination of *timings* (green durations, cycle and offsets) and *schedules* (the daily program sequence at each intersection), which may be done by well-known pen-and-paper techniques or using advanced simulation-based heuristics: there is no theoretical limit to the accuracy with which a *pre-timed* signal plan can meet the daily demand, other than the unavoidable uncertainty in flow prediction.

The concept of *Signal Control* on the other hand encompasses all strategies and rules applied to enable *actuation*, i.e. dynamic adaptivity of signal timings to the real-time traffic conditions. As such, it always involves some level of user demand detection and the need to enforce road administration policies. Only real-time signal control has the potential to mould to the intrinsically unpredictable nature of traffic demand.

It should be noted that this is in no way a distinction based on complexity or functionality: advanced signal setting techniques may enact traffic regulation policies that aim to influence the drivers' long term route choices, relying on algorithms that take days to solve, while a simple pedestrian crossing push-button to request a red phase at a signal that would otherwise remain green fully qualifies as actuated signal control. Nevertheless, both the most promising advancements in technology and the greatest challenges in modelling pertain to the field of Signal Control.

Main Notation

A quick glossary of the relevant variables is provided below, alongside the units of each dimensional quantity. For a leaner presentation of the model, subscripts referring to topological elements may be dropped to simplify notation.

Network Topology

$i, j \in \mathrm{N}$	nodes (junctions)
$a, b \in \mathrm{A} \subseteq \mathrm{N} \times \mathrm{N}$	arcs (lane groups)
$a = \left(\mathrm{N}_a^-, \mathrm{N}_a^+\right)$	tail and head nodes of arc a
$\mathrm{A}_i^+ = \left\{a \in \mathrm{A} : \mathrm{N}_a^- = i\right\}$	forward star of node i
$\mathrm{A}_i^- = \left\{a \in \mathrm{A} : \mathrm{N}_a^+ = i\right\}$	backward star of node i
$y, z \in \mathrm{Y}$	manoeuvres

Phases

$p,q \in \mathrm{P}_j$		phases of junction *j*
$\mathrm{A}_p \subseteq \mathrm{A}_j^-$		arcs of phase $p \in \mathrm{P}_j$

Signal Timings

t_j^C	s	cycle time at junction *j*
t_p	s	nominal duration of phase *p*
g_a	s	effective green duration of arc *a*
$\gamma_a = \frac{g_a}{t_j^C}$		effective green share of arc *a*
t_j^L	s	time lost of junction *j*
t_j^O	s	offset of junction *j*

Supply and Demand

q_a	veh/s	demand flow on arc *a*
$\hat{q}_a$	veh/s	saturation flow of arc *a*
$\phi_a = \frac{q_a}{\hat{q}_a}$		flow ratio of arc *a*
$\chi_a = \frac{\phi_a}{\gamma_a}$		saturation level of arc *a*

Performance

t_a^Q	s	queue clearance time of arc *a*
ω_a^{stop}		share of stopped vehicles of arc *a*
ω_a^d	s	average delay of arc *a*

Anatomy of a Signal Plan

The following section briefly illustrates the main features of a *signal plan* devised for urban traffic regulation. This term encompasses all timings and schedules behind the delicate clockwork of traffic signals, from the elements that constitute a single signal program at one of the many junctions of the network, to the succession of network-wide program changes designed to meet the daily evolution of traffic demand and the propagation of vehicle flows.

The features presented in this section fully define what is commonly called a pre-timed plan, and as such do not describe any real-time actuation or decision making logic.

They are themselves, however, the decision variables of most optimisation methods and adaptive strategies, and it is crucial to understand their significance in order to appreciate the diversity of setting and control approaches illustrated in more detail throughout this chapter.

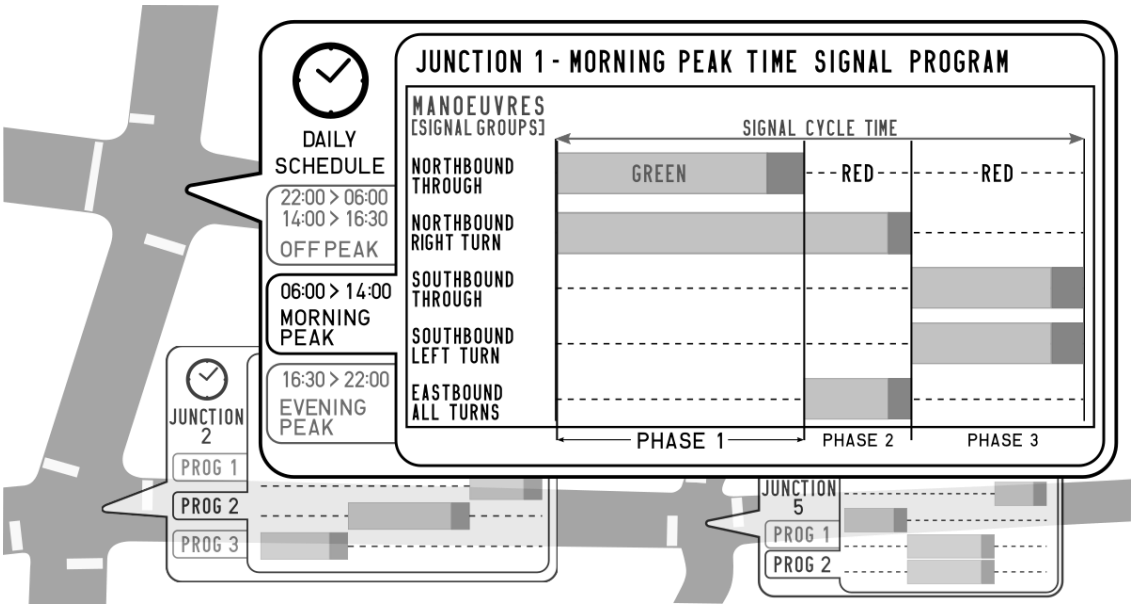

Figure 1 – Elements of a network-wide signal plan: a daily schedule specifies the signal programs running at each intersection. The sequence and duration of signal phases repeats over the course of every signal cycle as specified by the different signal programs, administering junction capacity amongst the expected traffic flows. During each phase, a set of compatible manoeuvres is allowed through while the others remain closed.

Signal Phases

Traffic signals exist mainly to separate conflicting traffic flows competing for the right of way at a road intersection. The natural way of doing so is to bundle compatible (e.g. non-secant) manoeuvres which may be safely performed simultaneously into *signal phases*, so that the corresponding flows may be allowed through the junction in turn.

Phases are the fundamental blocks of a signal program, and are usually repeated in the same order at every signal cycle, although some signalisation systems provide phase skipping, usually as part of their public transport prioritisation strategy.

Manoeuvres may pertain to different modes of transport, meaning that cars, trams and pedestrians are taken into joint consideration and can be given the right of way during the same signal phase.

With reference to Figure 2, consider a *junction*, i.e. a network node $j \in \mathrm{N}$ where it is possible to perform a given set of *manoeuvres* Y_j . The generic manoeuvre $y \in \mathrm{Y}_j$ may be a turn, from an arc $a \in \mathrm{A}_j^-$ of the node backward star, to a forward star arc $b \in \mathrm{A}_j^+$, or a pedestrian crossing affecting one or more arcs either entering or leaving the junction.

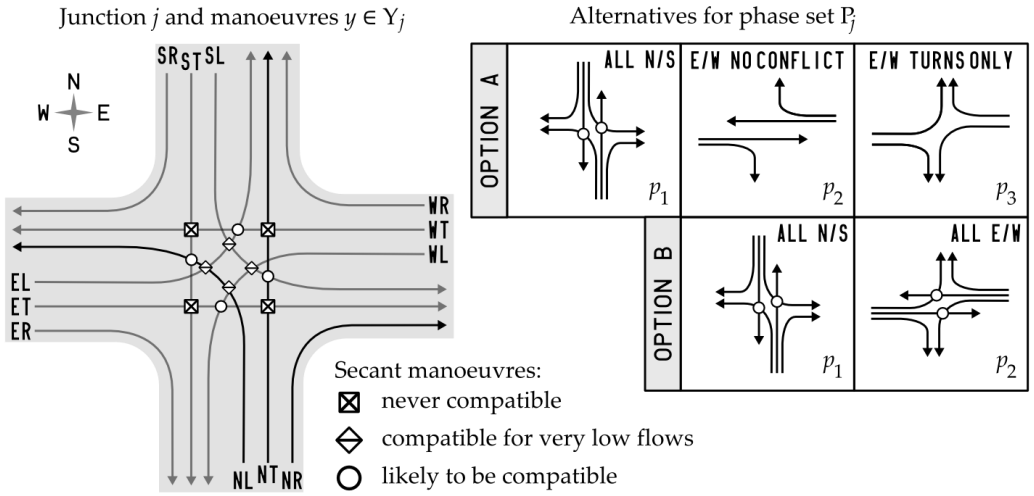

Figure 2 – Manoeuvres at an intersection, conflict areas and possible phasing options: option A avoids direct conflicts between Eastbound (E-) and Westbound (W-) manoeuvres, as would be desirable if high volumes were expected along that direction; option B favours a lower number of phase changes (less lost time) assuming flows to be such that left turning vehicles have space to wait at the middle of the intersection, and time to clear after the opposite through flow has decreased sufficiently to let them safely complete the manoeuvre.

From this point on, however, in order to present a more straightforward correlation between manoeuvres, junction geometrical layout and signalisation, the focus will be on private transport vehicles only. It shall be clear that the principles illustrated may be easily extended to handle more heterogeneous combinations of users.

Under given flow conditions and for a given junction layout, manoeuvres may or may not be safe to perform simultaneously: this information is easily represented by a square Boolean matrix where rows and columns correspond to each manoeuvre and elements comply with the following rule:

$$\delta_{yz} = \begin{cases} 1 & \text{if } y \text{ and } z \text{ are compatible} \\ 0 & \text{otherwise} \end{cases} \qquad \forall y, z \in \mathrm{Y}_j \tag{1}$$

Each signal phase $p \subseteq \mathrm{Y}_j$ identifies with a subset of compatible manoeuvres. The set of phases P_j selected for the junction must therefore belong to the space of feasible signal phases, i.e. all possible combinations of manoeuvres contained in the power set $\wp(\mathrm{Y}_j)$ such

that no two are incompatible. The union of all phases must also include every available manoeuvre at least once. Formally, P_j must therefore comply with the following properties:

$$P_j \subseteq \left\{ p \in \wp(Y_j) : \prod_{y \in p} \prod_{z \in p} \delta_{yz} = 1 \right\} \quad , \quad \bigcup_{p \in P_j} p = Y_j \, . \tag{2}$$

Clearly, the power set $\wp(Y_j)$ contains sets of manoeuvres that, although compatible and technically feasible, make little practical sense. The selection of an optimal set of phases P_j satisfying relation (2) with respect to a specific objective (e.g. minimum total delay), for given demand flows, is a combinatorial bi-level problem whose lower level is Signal Setting. The solution is usually obtained through a *what-if* approach, in which the selection of a good set of phases remains largely a traffic engineer's task.

Conceptually, the determination of signal phases is thus driven by the interactions between manoeuvres. From a practical point of view, however, administration of the right of way by means of traffic signals cannot transcend the junction layout: it is only possible to separate different manoeuvres coming from the same approach to a junction if each group of manoeuvres has a dedicated lane, allowing those vehicles to queue without hindering the others; conversely, as everyday experience testifies, traffic signals do not allow or prohibit manoeuvres directly, but rather regulate vehicle egress from lanes (or lane groups) dedicated to specific sets of manoeuvres.

Each lane or group of adjacent lanes sharing the same manoeuvre set $Y_a \subseteq Y_j$ can be conceptually assimilated into a *lane group:* a single independent arc $a \in A_j^-$ of the node backward star, whose flows, queues and physical characteristics are derived from a combination of demand and supply properties for each individual lane.

Let A_p be the set of lane groups which are given the green light during signal phase *p*, and Y_p the corresponding set of manoeuvres:

$$Y_p = \bigcup_{a \in A_p} Y_a \, . \tag{3}$$

The set of manoeuvres Y_a specific to each lane group *a* is relevant for the determination of the arc effective outflow capacity, which may be affected by partial conflicts with other manoeuvres allowed during the same phase. The HCM (2010) manual presents practical methods for quantifying such effects.

Henceforth, the discussion will only concern manoeuvres implicitly, in the aggregate form of lane groups referred to as arcs: a signal phase is simply a time period during which certain arcs are open, and the others are closed.

Signal Program

A signal program contains the state switching times for all signals at a given junction.

For signal planning and optimisation, it is practical to view the program as a succession of signal phases with specific durations, as portrayed in Figure 1, rather than the sequence of switching times of individual lights opening and closing each lane group: the present chapter adopts this phase-based signal program definition.

A *signal program* is a cyclic set of instructions spanning a period called *cycle time.* For a given phase set, it specifies the start and end of each signal phase, with respect to the beginning of the cycle. Transitions between subsequent phases are generally pre-timed sequences of light state changes that handle the closure of a set of lane groups before opening the next e.g. pre-red and pre-green amber signals.

Daily Schedule

It is common practice to tailor several signal programs to the traffic conditions normally observed at different times of the day, in order to meet each scenario with the best possible allocation of resources. The daily schedule defines the sequence of programs that each junction will run over the course of the day.

Cycle Time

The cycle time t_j^C is the period of the signal program, i.e. the time lapse between two occurrences of the same signal phase at a given junction. It affects the average delay and the level of saturation at which the intersection may operate. In general, longer cycle times imply larger average delays, but increase the total throughput, which may be necessary to deal with high demand flows by attenuating the effects of the time lost in signal phase changes.

Effective Green Shares

The nominal duration of each phase t_p is seldom exploited by demand flows at the full capacity of the corresponding arcs: even assuming that vehicles are not held back by downstream congestion, it is necessary to account for some transient phenomena affecting the performance of a junction.

As the signals turn green at the beginning of each phase, some time is lost before the queuing vehicles start moving and until the flow through the stop line reaches the arc capacity; conversely, vehicles may keep crossing the junction during the amber light, but a fraction of its duration must be allowed for vehicles to clear the junction.

After taking into account these delays and extensions, the portion of cycle time during which a given lane group a may allow traffic onto the junction at full capacity is referred to as its *effective green share.* The absolute and relative durations of effective green experienced by each lane group during phase p are denoted respectively as:

$$g_{ap} \in \left[0, t_p\right] \qquad \text{and} \qquad \gamma_{ap} = \frac{g_{ap}}{t_j^C}. \tag{4}$$

It is not uncommon to have a lane group open during more than one phase: typically, an approach experiencing high traffic volumes is given the right of way over two or more consecutive phases without incurring further lost time in the phase change.

The effective green of each arc a is then calculated from the total effective green time it gathers over all relevant phases:

$$g_a = \sum_{p \in \mathrm{P}_j} g_{ap} \quad \text{with} \begin{cases} 0 < g_{ap} \le t_p & \text{if } a \in \mathrm{A}_p \\ g_{ap} = 0 & \text{otherwise} \end{cases} \qquad \text{and} \qquad \gamma_a = \frac{g_a}{t_j^C} \tag{5}$$

Offset

When multiple intersections are considered, synchronisation issues are addressed by defining a global time reference, with all junctions sharing the same cycle time, or integer fractions thereof. Each junction may then have all of its phase switching times anticipated or delayed in order to operate in concert with the neighbouring ones.

The amount of time t_j^O, by which the beginning of a cycle at one junction j lags or leads the global reference instant, is referred to as a positive or negative *offset*, respectively.

OFFLINE SIGNAL SETTING

Long before microprocessors and sensors made adaptive real-time traffic control an everyday reality, the notion of signal plan optimisation identified with a range of techniques for designing good signal plans based on historical demand flows.

Such methods are not only still used for offline planning, but lie at the core of several online optimisation approaches: given a signal setting policy, it makes little difference from the methodological point of view whether the input variables are determined from historical data or fed in real-time by sensors.

At any rate, this does not mean that the dynamic interaction between signal setting and driver behaviour can be disregarded altogether: for example, the assumption often made that route choices are fixed and unaffected by signal settings has warranted the formulation of planning strategies which have proven quite patently inadequate in the real world, as first discussed in Dickson (1981).

The main subjects of optimisation in network signalisation design, including cycle time, offset and green share allocation, are illustrated in the following sections, alongside methods commonly used to tackle each specific problem.

Traffic Performance at Isolated Signalised Intersections

Traffic performances at signalised junctions depend on the time at which the different vehicles arrive at the junction and on the signal regulation that determines their departure times. Several flow models were introduced in the scientific literature to reproduce arrival and departure phenomena. Here, traffic flow is assimilated to a fluid stream, according to the macroscopic paradigm, which differ substantially from the microscopic approach where the trajectory of each single vehicle is explicitly considered.

More specifically, vehicle departures are modelled as a uniform flow. If the arrival flows are high enough but lower than capacity, their inherent random component can be neglected and they are also considered deterministic. Conversely, if stochasticity of arrival flows is significant, as it occurs when flows approach capacity or are very low, a random component is added to the simple deterministic model as in Webster, (1956).

In the following, the basic relationships between signal timing variables and traffic performances are presented with reference to the simple deterministic model.

Queue clearance

Consider a single arc (lane group) $a \in A_j^-$ entering a signalised junction $j \in N$, with a constant *demand flow* of vehicles q_a arriving over the entire cycle. Instead, the queue can be discharged into the junction during the effective green only, at the constant *saturation flow*

rate $\hat{q_a}$. The latter is given by the arc capacity, possibly degraded due to conflicts with other arc flows. The *flow ratio* between demand and saturation is denoted as:

$$\phi_a = \frac{q_a}{\hat{q}_a}. \tag{6}$$

The departure rate is zero during the rest of the cycle. So $\hat{q_a}$ must be sufficient to discharge the queue accumulated over a red of duration $t_j^C - g_a$, plus the flow of vehicles that keep arriving, during the effective green g_a.

Let t_a^Q the time required to discharge the queue. The total number of stopped vehicles (see Figure 3) can be expressed alternatively as:

$$q_a \cdot \left(t_j^C - g_a + t_a^Q\right) = \hat{q}_a \cdot t_a^Q. \tag{7}$$

The *queue clearance* time is then given by:

$$t_a^Q = \frac{q_a \cdot \left(t_j^C - g_a\right)}{\hat{q}_a - q_a} = \phi_a \cdot t_j^C \cdot \frac{1-\gamma_a}{1-\phi_a}. \tag{8}$$

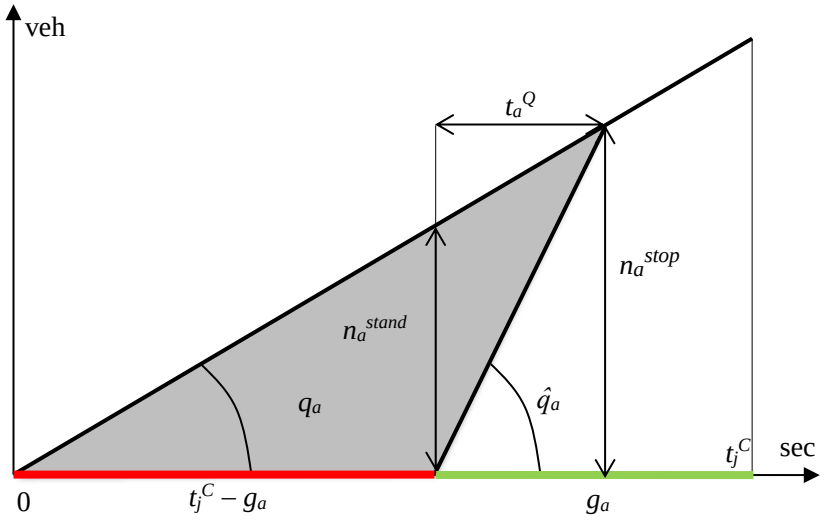

Figure 3 – Geometric determination of stopped vehicles and queue clearance for one approach, given cycle and green time, as well as demand and saturation flow. The grey triangle between the arrival cumulative, the departure cumulative and the horizontal axis covers the number of vehicles queuing at any given moment. Notice that the number of standing vehicles n_a^{stand} at the beginning of the effective green is smaller than the total number of stopped vehicles n_a^{stop}.

Number of Vehicle Stops

Based on equation (6), the fraction ω_a^{stop} of stopped vehicles is:

$$\omega_a^{stop} = \frac{\hat{q}_a \cdot t_a^Q}{q_a \cdot t_j^C} = \frac{1-\gamma_a}{1-\phi_a}, \tag{9}$$

which is proportional to the red share of the cycle time and increases as the arrival rate approaches the discharge capacity. For values of $\phi_a \geq 1$ or $\gamma_a < \phi_a$ queues cannot be fully discharged and all vehicles end up stopping.

Average Delay

Assuming constant arrival and departure rates, the total delay experienced at each cycle by all users from a given approach *a* corresponds to the integral over time of the queue size (the area of the blue triangle). Whence the average delay $\omega_a{}^d$ per vehicle is:

$$\omega_a^d = \frac{\left(t_j^C - g_a\right) \cdot \left(\hat{q}_a \cdot t_a^Q\right)}{2 \cdot \left(q_a \cdot t_j^C\right)} = \frac{\left(t_j^C - g_a\right)^2}{2 \cdot \left(1-\phi_a\right) \cdot t_j^C}. \tag{10}$$

Clearly, the above equation (10) assumes no standing queues at the end of a cycle.

More complex delay functions can be obtained by considering stochastic fluctuations of arrival flows (Webster, 1958). Flows exceeding the arc capacity require the introduction of either simulation models or empirical adaptations of analytical models, such as the coordinate transformation method introduced by Kimber and Hollis (1979) and later adopted by the popular HCM traffic manual (2010).

Critical Flow Ratio and Saturation

The saturation flow characterising each lane group depends on various factors, such as road width, visibility, or the presence of dedicated turn bays for the relevant manoeuvres which may alleviate conflicts with other manoeuvres served during the same phase. This is particularly relevant to left turns, or turns encroaching a pedestrian crossing: scrupulous phase planning can minimise the number and entity of such conflicts.

The flow ratio ϕ_a quantifies the expected demand on a given lane group *a* in relation to its saturation capacity. For each lane group *a* the *saturation level* χ_a is determined by the ratio of demand flow to its *outflow capacity*, which is further limited by the signal, inasmuch as each arc can only be open for a limited share of the available green time:

$$\chi_a = \frac{q_a}{\gamma_a \cdot \hat{q}_a} = \frac{\phi_a}{\gamma_a}. \tag{11}$$

For values of $\gamma_a < \phi_a$ the saturation level is greater than 1 and queues cannot fully discharge.

When multiple lane groups are to be open simultaneously during phase *p*, the *critical flow ratio* ϕ_p is given by the approach which is relying most heavily on the phase in question.

The concept is formalised by scaling the flow ratio of each approach according to the portion of flow that must be dealt with during phase *p*, and selecting the maximum value:

$$\phi_p = \max\left(\phi_a \cdot \frac{\gamma_{ap}}{\gamma_a} : a \in \mathrm{A}_p\right), \tag{12}$$

whence the *critical lane group* of phase *p* is also identified:

$$\mathrm{A}_p^* = \left\{ a \in \mathrm{A}_p : \phi_p = \phi_a \cdot \frac{\gamma_{ap}}{\gamma_a} \right\} . \tag{13}$$

The *critical saturation* χ_p of signal phase *p* is obtained from applying (11) to its critical lane group A_p*:

$$\chi_p = \frac{\phi_p}{\gamma_{\mathrm{A}_p^* p}} . \tag{14}$$

In the particular case where each lane group is only open during a single phase, the critical flow ratio and critical saturation definitions are reduced to the identification of the critical approach, which maximises both indicators**Error! Reference source not found.**.

The total *junction flow ratio* is simply the sum of the critical flow ratios of all phases:

$$\phi_j = \sum_{p \in \mathrm{P}_j} \phi_p . \tag{15}$$

Although the flow ratio is an effective gauge of the intersection capability to meet user demand, the junction performance depends on the saturation levels of the critical approaches as expressed by (11).

Finally, considering Equation (14), the effective green of phase *p* is defined as the effective green experienced during the same phase by the corresponding *critical lane group*:

$$g_p = g_{\mathrm{A}_p^* p} . \tag{16}$$

Lost Time

Driver reactions are not instantaneous, and vehicles take a finite amount of time to accelerate and clear the junction. As already mentioned: at every phase start, a few seconds pass before vehicles can flow at full capacity; at every phase end, sufficient time must be provided to allow vehicles to clear the junction before others may safely attempt a conflicting manoeuvre. This implies that a non-negligible share of the signal cycle goes wasted, since demand is not served efficiently during the phase transitions.

The *start-up loss* may be reduced by helping drivers to react more promptly, e.g. using a pre-green amber light or red count-down timers, which also seem to alleviate the stress of being stuck in a queue. The *clearance loss* may only be mitigated by an optimal choice of phases and phase sequence for given traffic conditions or, wherever possible, by appropriate modification of the junction layout, e.g. implementation of protected turn bays.

The total *lost time* t_j^L then depends on phase design and sequence, which in turn should be tailored to the geometry of junction *j* in relation to the expected traffic conditions.

Each phase contributes its own time losses t_p^L to the total lost time, which may be considered in the following relation with the effective phase green g_p:

$$t_p^L = t_p - g_p . \tag{17}$$

The total effective green of all phases and the total lost time account for the whole cycle:

$$t_j^L + \sum_{p \in \mathrm{P}_j} g_p = t_j^C \,. \tag{18}$$

Formulation of the signal optimisation problem

Conflicting sets of manoeuvres compete for the right of way at road intersections, and the main purpose of signalization is to distribute the junction capacity amongst them. It follows naturally that the allocation of green time to signal phases is the single most important step in signal setting: green time must be allotted according to the relative distribution of demand, lest the junction capacity go wasted and queues will form on critical approaches.

As far as fixed timing is concerned, optimal allocation of green time is a straightforward process, yet it can be undertaken according to a number of different principles: early studies aimed to develop analytical equations, while modern simulation based methods rely on heuristics to shape the signal setting around a cost function that formalises the chosen signal setting policy. In the following we provide a general formulation of the optimal signal setting problem.

Lagrangian Formulation

The Signal Setting of junction *j* can be formulated as an optimisation problem, i.e. to find effective green durations for each phase and cycle time that minimise an objective function ω while complying with a set of constraints. The objective function is usually a measure of the average delay at the intersection, given by the sum of the average delay $\omega_a^d(g_a, t_j^C)$ for each lane group $a \in \mathrm{A}_j^-$ weighted by the corresponding flow q_a. The average delay on each lane group can in turn be expressed through equation (10) as a function of the effective green of the arc g_a and of the signal cycle t_j^C.

The constraints are the consistency relations **Error! Reference source not found.** and (5), as well as the maximum saturation level of each phase:

$$\min_{\mathbf{g}, t_j^C} \quad \omega\left(\mathbf{g}, t_j^C\right) = \sum_{a \in \mathrm{A}_j^-} q_a \cdot \omega_a^d\left(g_a, t_j^C\right) \quad \text{subject to:} \tag{19}$$

$$t_j^L + \sum_{p \in \mathrm{P}_j} g_p = t_j^C \tag{20}$$

$$g_p \geq \phi_p \cdot t_j^C \quad \forall p \in \mathrm{P}_j \tag{21}$$

The set of constraints then amounts to the same number of Signal Setting variables; but because the saturation levels are inequalities, the problem has as many degrees of freedom as the number of phases.

By introducing the auxiliary positive variables η_p^2 every inequality can be transformed into an equation:

$$g_p - \phi_p \cdot t_j^C - \eta_p^2 = 0 \tag{22}$$

Thus, the optimisation problem can be solved by applying the Lagrange method, that consists in finding the stationary points of the so called Lagrangian function £, that is the linear combination of the objective function and the equality constraints. The coefficients λ_p and μ of the linear combination (Lagrangian multipliers) are additional auxiliary variables.

$$£ = \omega\left(\mathbf{g}, t_j^C\right) + \sum_{p \in \mathrm{P}_j} \lambda_p \cdot \left(g_p - \phi_p \cdot t_j^C - \eta_p{}^2\right) + \mu \cdot \left(\sum_{p \in \mathrm{P}_j} g_p + t_j^L - t_j^C\right) \tag{23}$$

By definition, at the stationary points of the Lagrangian function the partial derivatives with respect to all control and auxiliary variables shall be zero.

$$\begin{aligned}
&\frac{\partial £}{\partial t_j^C} = \frac{\partial \omega\left(\mathbf{g}, t_j^C\right)}{\partial t_j^C} - \sum_{p \in \mathrm{P}_j} \lambda_p \cdot \phi_p - \mu = 0 && \text{(a)}\\
&\frac{\partial £}{\partial g_p} = \frac{\partial \omega\left(\mathbf{g}, t_j^C\right)}{\partial g_p} + \lambda_p + \mu = 0 \quad \forall p \in \mathrm{P}_j && \text{(b)}\\
&\frac{\partial £}{\partial \lambda_p} = g_p - \phi_p \cdot t_j^C - \eta_p^2 = 0 \quad \forall p \in \mathrm{P}_j && \text{(c)}\\
&\frac{\partial £}{\partial \mu} = \sum_{p \in \mathrm{P}_j} g_p + t_j^L - t_j^C = 0 && \text{(d)}\\
&\frac{\partial £}{\partial \eta_p} = \eta_p \cdot \lambda_p = 0 \quad \forall p \in \mathrm{P}_j && \text{(e)}
\end{aligned} \tag{24}$$

For sake of simplicity, the discussion of this non-linear system of equations is illustrated by referring to the simplest case of a single junction *j* with two phases *p* and *q*, which enables a more immediate interpretation of the constraints.

In this case, the complementarity conditions $\eta_p \cdot \lambda_p = 0$ and $\eta_q \cdot \lambda_q = 0$ have 4 possible solutions. The optimal solution must be identified by comparing the objective functions for all candidate stationary points.

Case a) $\lambda_p \neq 0$ and $\eta_p = 0$; $\lambda_q \neq 0$ and $\eta_q = 0$

Capacity constraints (24).c are active for both phases. Lagrangian multipliers become inconsequential and the problem degenerates into the system of equations (24).c and (24).d that are sufficient to determine effective greens and cycle time:

$$\begin{aligned}
&g_p = \phi_p \cdot t_j^C \quad \forall p \in \mathrm{P}_j\\
&\sum_{p \in \mathrm{P}_j} g_p + t_j^L - t_j^C = 0 \quad \cdot
\end{aligned} \tag{25}$$

The system has a unique solution corresponding to the minimum green times that exactly match demand on the critical lane groups of each phase. The cycle length, found by substituting for g_p from the first equation into the second equation of (25), is the shortest possible cycle that allows to serve the demand at the junction despite the unavoidable time losses associated with phase changes.

Error! Reference source not found. It is straightforward to conclude that each phase should get at the very least a green share of the total cycle time equal to its critical flow rate. Based on (15) the minimum cycle for the junction is then found:

$$t_j^C = t_j^{Cmin} = \frac{t_j^L}{1-\phi_j}. \tag{26}$$

Under these conditions, every critical lane group operates at full saturation $\chi_a = 1$, meaning that each phase gets *just* enough green time to meet demand.

Case b) $\lambda_p \neq 0$ and $\eta_p = 0$; $\lambda_q = 0$ and $\eta_q \neq 0$

In this case the capacity constraint is active only during phase *p*. The other Lagrangian multiplier λ_q is zero and the corresponding derivative can be removed:

$$\begin{aligned}
&\frac{\partial £}{\partial t_j^C} = \frac{\partial \omega\left(g_p, g_q, t_j^C\right)}{\partial t_j^C} - \lambda_p \cdot \phi_p - \mu = 0 \\
&\frac{\partial £}{\partial g_p} = \frac{\partial \omega\left(g_p, g_q, t_j^C\right)}{\partial g_p} + \lambda_p + \mu = 0 \\
&\frac{\partial £}{\partial g_q} = \frac{\partial \omega\left(g_p, g_q, t_j^C\right)}{\partial g_q} + \mu = 0 \\
&\frac{\partial £}{\partial \lambda_p} = g_p - \phi_p \cdot t_j^C = 0 \\
&\frac{\partial £}{\partial \mu} = g_p + g_q - t_j^C + t_j^L = 0
\end{aligned} \tag{27}$$

By solving the equation corresponding to the active constraint, the green time g_p is set at its minimum value, while the other green time g_q can be determined by using the last equation that defines the cycle length:

$$\begin{aligned}
&g_p = \phi_p \cdot t_j^C \\
&g_q = t_j^C - t_j^L - g_p = \left(1-\phi_p\right) \cdot t_j^C - t_j^L
\end{aligned}. \tag{28}$$

The green times are both expressed in terms of the cycle length: by substitution, the objective function can then become a function of a single variable. In the case of deterministic constant arrivals, minimisation of delays yields a unique stationary point in terms of cycle length, which can be found in closed form:

$$t_j^C = t_j^{Cdet} = \sqrt{\frac{\rho_q \cdot \left(t_j^L\right)^2}{\rho_p \cdot \left(1-\phi_p\right)^2 + 2 \cdot \rho_q \cdot \phi_p^2}}, \tag{29}$$

where the coefficient ρ_p (and analogously ρ_q) is obtained as follows:

$$\rho_p = \frac{1}{2} \cdot \frac{\sum_{a \in A_p} \frac{q_a}{1-\phi_a}}{\sum_{a \in A_j^-} q_a} \tag{30}$$

This solution assigns minimum green to phase *p* and is reasonable if the major demand flow is handled during phase *q*, while the opposite scenario covers the specular case.

Case c) $\lambda_p = 0$ and $\eta_p \neq 0$

In this case no capacity constraints are active. Thus, the problem consists in searching for the solution of the following system of three nonlinear equations in the three unknowns g_p , g_q and t^C:

$$\begin{aligned} &\frac{\partial \omega\left(g_p, g_q, t_j^C\right)}{\partial g_p} + \frac{\partial \omega\left(g_p, g_q, t_j^C\right)}{\partial t_j^C} = 0 \\ &\frac{\partial \omega\left(g_p, g_q, t_j^C\right)}{\partial g_q} + \frac{\partial \omega\left(g_p, g_q, t_j^C\right)}{\partial t_j^C} = 0 \\ &g_p + g_q - t_j^C + t_j^L = 0 \end{aligned} \quad . \tag{31}$$

In this case the analytical solution is more difficult than in the previous cases. However, it can be easily found numerically, since the delay function is bounded and convex under usual realistic assumptions on vehicle arrivals.

The explicit formulation of this method rests upon the assumption that the *capacity condition* be respected, i.e. the capacity of the junction be sufficient to serve demand. This may be relaxed in practice for heuristic optimisation, but little changes about the fundamental fact that no green share allocation will ever enable a junction to operate *above capacity* without delay.

Webster optimal solution

The first and foremost formulation of optimal signal settings under probabilistic assumptions of vehicle arrivals is due to Webster (1958), who tackled the problem by simulating a queueing system with Poissonian arrivals and a constant service rate equal to the capacity of the signalised lane group. The average delay (10) was extended to obtain a more complete delay function for random arrivals, with an additional empirical term needed to improve the fit with experimental observations.

To simplify the optimisation problem, a reasonable green share allocation policy (widely known as *Equisaturation Policy*) was chosen. This revolves around the idea that an equitable distribution of green share is obtained when all critical manoeuvres operate at the same saturation level: the higher the demand for a manoeuvre *with respect to the capacity* of the relevant infrastructure, the higher the green share allocated to the corresponding signal phase.

Furthermore, Webster worked under the assumption that *demand flows are stable* i.e. no oversaturation occurs and path choices made by road users are in no way a consequence of the signal setting. This assumption was removed by later scholars who tackled the global optimisation signal setting who route choice problem (Smith, 1984; Cipriani and Fusco, 2008).

Under the *equisaturation* policy, all phase saturation levels at a given junction are equal by definition. The *available green time* can simply be allocated proportionally to the critical flow ratio of each phase:

$$\gamma_p = \frac{\phi_p}{\phi_j} \cdot \frac{t_j^C - t_j^L}{t_j^C} \quad \forall p \in \mathrm{P}_j \tag{32}$$

which yields meaningful results provided that the junction total flow ratio does not exceed its maximum value of 1 and the cycle time is sufficiently long to amortise the lost time.

The approach can be extended to design for specific (not necessarily even) saturation values for each phase by rearranging Equation **Error! Reference source not found.** and solving for the green share.

This may have practical sense in order to design a higher tolerance to high arrival rates into a given phase e.g. if it is strategically more important to keep queues at a minimum on a certain set of lanes than it is elsewhere.

After the criterion to set the green share has been determined, the problem of minimising the average delay is reduced to a single variable function of the cycle length.

The resulting expression of the cycle time that minimises *average delay* under probabilistic arrivals is rather complex and was approximated it through an *empirical* formula, widely known as the Webster optimum cycle time:

$$t_j^{Webster} = \frac{3/2 \cdot t_j^L + 5}{1 - \phi_j} . \tag{33}$$

Notice from equations (26) and (33) how the cycle time invariably grows with the total flow ratio of the junction. It is also possible to extend (26) to get a target saturation level χ_j for the junction:

$$t_j^C\left(\chi_j\right) = \frac{t_j^L}{1 - \phi_j / \chi_j} , \tag{34}$$

or a specific critical saturation level for each phase χ_p :

$$t_j^C\left(\boldsymbol{\chi}\right) = \frac{t_j^L}{1 - \sum_{p \in \mathrm{P}_j} \phi_p / \chi_p} \tag{35}$$

It should be evident that saturation values greater than 1 correspond to *oversaturated* conditions, under which the demand flows are not met with sufficient capacity and queue buildup is inevitable: such traffic conditions require radically different timing approaches.

The rule of thumb mentioned in HCM (2008) and generally followed in practice is that approaches operating at saturation levels below 0.85 may be expected to deal with reasonable traffic fluctuations efficiently and discharge any queues within a few signal cycles.

P_0 Policy

This is an alternative green allocation paradigm first presented in Smith (1984). The author deemed some assumptions made in previous formulations responsible for excessive capacity reductions of signalised networks. In particular, the new approach sought to:

- account for user response to signalisation: path choices cannot be considered fixed and their dynamic evolution must be considered as a consequence of signal settings;

- build upon a paradigm of dynamic user/signal equilibrium: given the green shares, flows tend to shift onto the cheapest routes – given the flows, green shares are adapted according to the signal setting policy until a stable configuration is reached;
- maximise the *network capacity:* a measure of the maximum total demand that the network can serve without becoming oversaturated, given the O-D demand flows that describe a certain scenario.

With the above in mind, the work led to the demonstration that the policy to *guarantee maximum network capacity*, given the origin-destination flows and infrastructure capacity, is one that at each junction $j \in \mathrm{N}$ equalises the product of saturation flow and average delay across the *critical lane groups* of each phase.

Consider two different signal phases *p* and *q* of the same junction *j*, and the respective critical lane groups *a* and *b* as per (13). The P_0 policy may then be expressed as follows:

$$\hat{q}_a \cdot \omega_a^d = \hat{q}_b \cdot \omega_b^d \quad \text{with} \quad \begin{cases} \mathrm{A}_p^* = \{a\} \\ \mathrm{A}_q^* = \{b\} \end{cases} \quad \forall p, q \in \mathrm{P}_j \,. \tag{36}$$

Delay on lane group *a* served during phase *p* may be modelled using (10) or any monotone decreasing function $\omega_a{}^d(\bullet)$ of the corresponding *residual capacity*, which accounts for the difference between the maximum flow compatible with the current green share and the actual demand flow, such that

$$\begin{cases} \omega_a^d\left(\gamma_a \cdot \hat{q}_a - q_a\right) = \omega_a^d \geq 0 \\ \omega_a'^d\left(\gamma_a \cdot \hat{q}_a - q_a\right) \leq 0 \end{cases} \quad \forall \gamma_a, q_a, \hat{q}_a : \gamma_a \cdot \hat{q}_a - q_a \geq 0 \tag{37}$$

This can be used to prove the existence of a stable equilibrium point between signal setting and route choices, where the network capacity is maximised — with considerable margin over the equisaturation method — whence ideal fixed timings can also be determined. Furthermore, it is straightforward to find a Lyapunov function of the dynamic user/signal system governed by this policy, which can be used to prove that over time it naturally evolves towards the optimal equilibrium point, as detailed in several papers by the same author.

The policy expressed by (36) was proven to be an effective improvement over the equisaturation principle in numerous studies, both as a tool to develop fixed time plans (Smith 2010) and as dynamic plan update policy (Smith 2011) extendable to large networks.

Signal Synchronisation

Traffic light synchronisation between adjacent junctions is an essential aspect of an optimal signalisation plan, with disposition of *green waves* as its most notable and popular feature. Traffic mostly travels along a limited number of main corridors, commonly referred to as *arteries* carrying *arterial traffic.* It has long been accepted as a reasonable compromise to minimise discomfort along those, rather than taking on the much more intricate problem of reducing the total network delay.

Being able to drive through a streak of green signals already goes a long way towards improving the quality of a trip from the user point of view, but signal synchronisation chiefly serves the purpose of ensuring an efficient use of the available infrastructure.

It is in fact of the utmost importance to avoid unnecessary signal-induced delays and stops which could rapidly lead traffic to a grinding halt, even under mild traffic conditions which the road network could otherwise easily cope with.

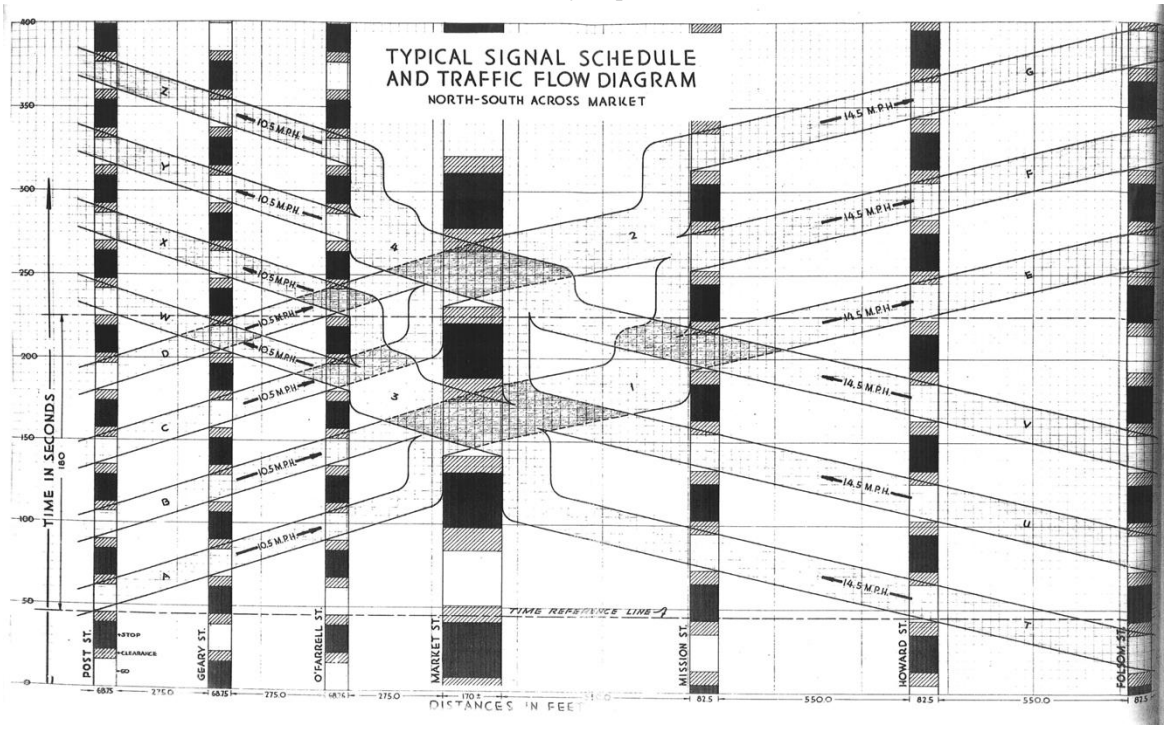

Figure 4 – Early signal synchronisation along a San Francisco arterial road, circa 1929. Bands A through T represent vehicle platoons.

The search for a synchronisation solution that maximises usability of urban arteries under specific traffic conditions is still mostly carried out offline — as it was for the first attempts at smart arterial signalization, such as the pen-and-paper method portrayed in Figure 4.

To this end, a wide variety of methods have been the object of intensive research since the early 1980s, ranging from simple analytical approaches to heuristics. Analytical methods have brought about a number of popular applications which are still in use despite the fact that they mainly apply to low congestion scenarios; more complex methods, which account for demand flows and their propagation along the arterial, can deal with heavy congestion related phenomena, but invariably require a more detailed network model and rely on computationally demanding simulations rather than a closed-form problem formulation. An overview of the most prominent approaches to the synchronisation problem is given in the following sections.

Bandwidth Maximisation

In relation to arterial traffic, the concept of *progression bandwidth* emerges as a measure of the quality of a green wave setup, and according to its most recent formulations can be defined as *the duration of the time window through which a vehicle may enter the artery and travel its entire length without encountering neither red lights nor standing queues.*

By reducing delays and number of stops along the most critical paths, bandwidth maximisation is a relatively straightforward but effective way to help the system meet user expectations about traffic fluidity, mitigating the stress associated with driving in a congested urban environment. Moreover, this type of signal coordination has proven highly beneficial in reducing the chance of rear end collisions and red signal violations (Li, Tarko 2010) as well as pollution levels associated with the hiccupping stop-and-go driving often experienced under poorly synchronised signalisation.

Bandwidth maximisation has been formulated as a Linear Optimisation problem since Little et al. (1981) which led to development of the MAXBAND/MULTIBAND series of software solutions. These considered the offsets between junctions as the only decision variables, but provided a computationally viable method for one-way and two-way bandwidth maximisation relying solely on the target travel times between junctions and predetermined signal cycle length and green times. A more efficient solution method was introduced by Papola and Fusco (1988a) who exploited periodicity properties of synchronised signal systems to developed a new algorithm based on ideal equivalent systems.

However, relevant discrepancies — dubbed *bandwidth degradation* — were observed between the expected outcome and the real-world performance of the signal plans generated by these early methods. It is now universally accepted that, as Tsay and Lin (1988) amongst many others pointed out, the underlying models were oversimplified and no account was taken of side flows and platoon dispersion. Extensions of the original method that factored in queue and side flow clearance times were proposed, further providing the capability to handle phase lead/lag and left turn phases to produce a more realistic bandwidth model. The analytical relationship between maximal bandwidth and minimum delay problems were finally fromulated by Papola and Fusco (2000), who devised an analytical model that provides delays of a function of the maximal bandwidth and other variables.

Currently, any arterial progression optimisation that do not rely on *theoretical* bandwidth maximisation also fall outside the scope of *offline* signal setting, as they rely on real time detection to synchronise signal phases with the passage of major vehicle platoons, as illustrated in the relevant section of this chapter.

A Mixed Integer Linear Programming Approach

Recently, De Nunzio (2015) extended the mixed integer linear programming approach to include *Variable Speed Limits* (VSL) as control variables, allowing for a wider range of high bandwidth solutions. The concept of VSL has been applied to motorway traffic for quite some time, to enhance traffic fluidity in response to congestion, accidents or adverse weather, but its application to urban traffic presents new challenges, not least the need for effective means of introducing it and getting it across to the drivers: in pilot projects this is quite effectively achieved by variable led panels, mimicking an ordinary speed limit sign, showing the target synchronisation speed. Were such measures to gain popularity, the already promising degree of driver compliance can only be expected to improve.

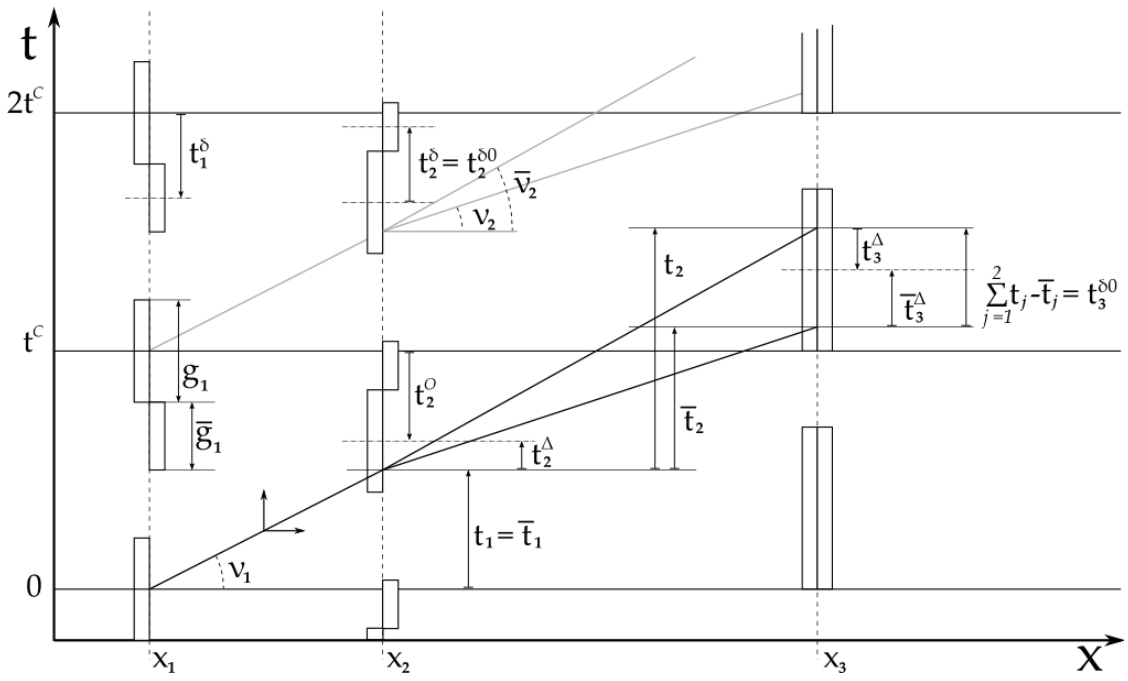

Figure 5 – Bandwidth Problem Formulation: the signal coordination parameters are portrayed on a time-distance graph. Temporal references are given by integer multiples of the cycle time and by the synchronisation frame of reference, moving along the diagonal trajectories at speeds v_j. The green phases in the main direction are drawn on the left of each junction's temporal line, that of the inverse direction is to its right. Notice the offsets measured between the phase midpoints and the time of arrival of the moving FoR.

The simple MILP approach to two-way bandwidth maximisation can be summarised by considering a major traffic artery as a path $k \in \mathrm{K}$ running through an ordered set of $|k|$ intersections $\mathrm{N}_k = \{i_1^k, i_2^k, \dots, i_{|k|}^k\}$ along its *main* driving direction, while the opposite, possibly lower priority direction traverses the same nodes in reverse order.

From defining positive travel speeds v_n and $\bar{v}_n$ between nodes i_n^k and i_{n+1}^k in the *main* and *other* direction respectively, and a generic spatial coordinate x which increases along with $n = 1, \dots, |k|$ it follows that travel times should be:

$$\begin{cases} t_n = \dfrac{\overrightarrow{i_n\, i_{n+1}}}{v_n} = \dfrac{x_{n+1} - x_n}{v_n} > 0 \\ \bar{t}_n = \dfrac{\overrightarrow{i_{n+1}\, i_n}}{\bar{v}_n} = \dfrac{x_n - x_{n+1}}{\bar{v}_n} < 0 \end{cases} \qquad \forall n \in \left[1, \left|\mathrm{k}\right| - 1\right] \tag{38}$$

Assuming a common cycle time t^C, consider at each node and for both directions:

- *effective green duration* of the arterial *through* movement phases g_i and $\bar{g}_i$
- *absolute offset* as the distance between the half-duration of a green phase and the closest multiple of the cycle time:

$$t_i^O, \bar{t}_i^O \in \left]-t^C/2, t^C/2\right]$$

A nonstandard modulo operation $\|\bullet\|_t$ can be defined for brevity at this point, such that

$$\left\| t^* \right\|_t \in \left]-t^C/2, t^C/2\right] \tag{39}$$

returns the distance from t^* to the nearest multiple of t^C.

It is used to define the *internal offset* as $t_i^d = \left\| \bar{t}_i^O - t_i^O \right\|_{t^C}$ and the *relative offset* t_i^Δ.

The latter represents the time coordinate of the mid-green instant of the relevant phase with respect to a moving frame of reference travelling along the *main* driving direction, starting in i_1^k at the zero instant and moving with the specified speeds v_n between nodes.

It follows that the relative offset at each node after the first can be computed easily from the offsets at upstream nodes:

$$t_i^O - t_i^D = t_{i-1}^O - t_{i-1}^D + t_{i-1} \Rightarrow t_i^D = \left\| t_{i-1}^D + t_i^O - t_{i-1}^O - t_{i-1}^D \right\|_{t^C}$$

$$\Rightarrow \begin{cases} t_i^D = \left\| t_1^D - t_1^O + t_i^O - \sum_{j=1}^{i-1} t_j \right\|_{t^C} \\ t_i^O = \left\| t_1^O - t_1^D + t_i^D + \sum_{j=1}^{i-1} t_j \right\|_{t^C} \end{cases} \tag{40}$$

In order to express the bandwith in both directions in terms of the relative offsets, it is also beneficial to map all t_i^δ to the time reference of the first junction using

$$t_i^{\delta 0} = \left\| t_i^\delta + \sum_{j=1}^{i-1} t_j - \bar{t}_j \right\|_{t^C} \tag{41}$$

considering that the t_i^δ are given by the signal program timing at each intersection, which leads to the vector equation linking the offsets in the two directions

$$\overline{\mathbf{t}^\Delta} = \mathbf{t}^\Delta - \mathbf{t}^\delta \quad \text{with} \begin{cases} \mathbf{t}^\Delta = \left(t_1^\Delta, t_2^\Delta, \ldots, t_{|\mathrm{N}_k|}^\Delta \right) \\ \mathbf{t}^\delta = \left(t_1^{\delta 0}, t_1^{\delta 0} - t_2^{\delta 0}, \ldots, t_1^{\delta 0} - t_{|\mathrm{N}_k|}^{\delta 0} \right) \end{cases} \tag{42}$$

Finally, it is possible to approach the maximisation of bandwidth as a function of travel times and offsets. According to the definition given at the start of this section and considering Figure 5, it is the intersection of all green windows as seen in the moving frame of reference:

$$\bigcap_{i \in \mathrm{N}_k} \left[t_i^\Delta - g_i/2, t_i^\Delta + g_i/2 \right] \tag{43}$$

The bandwidth value in the *main* direction is then calculated from the decision variables as

$$\mathrm{b}\left(t_1^\Delta, \ldots, t_{|\mathrm{k}|}^\Delta \right) = \max\left(0, \min\left(t_i^\Delta - t_j^\Delta + g_{ij} \right) \right) \forall i, j \in \mathrm{N}_k \quad \text{with } g_{ij} = \frac{g_i + g_j}{2} \tag{44}$$

while the equivalent in the other direction is found using the relevant green times and (42).

The sum of the bandwidths in the two directions can then be the objective of the linear optimiser — bounded by appropriate constraints such as maximum speed values — in

conjunction with any function of the decision variables used to favour a certain type of solution: for example, the optimisation presented in De Nunzio (2015) is driven by an extended utility function aiming to favour low travel times and minimise the speed indication variance across segments so as to ease drivers into complying with apparently arbitrary limits.

Real world statistics are beginning to back up the simulation results that originally validated these studies, proving the following interesting points about modern bandwidth maximisation techniques:

- the best combinations of optimal offsets and VSL drastically reduce the number of stops and energy consumption
- lower and smoother speed limits reduce energy consumption *at no disadvantage to the total arterial travel time*
- VSL brings about larger bandwidth and faster solution of the LP

It must be noted however that despite the practically negligible computation times associated with the LP methods just mentioned, these remain conceptually unfit for *real time* signal optimisation since they take in no account the flow and speed of the actual traffic, nor they apply outside the safe boundaries of *capacity conditions* (whereby green time is always assumed sufficient to deal with demand).

Traffic Performance Optimisation

The simple signal setting problems presented so far are quasi-convex, but more realistic traffic models that include and quantify global performance indicators such as total delay introduce an inherent non convexity, better addressed with the aid of heuristic methods.

With the increase in computing power availability, metaheuristics have seen a substantial rise in popularity as means to overcome the inherent limitations of analytical formulations: heuristic approaches to this class of problems involve the generation of a large — yet manageable, compared to the dimensions of the search space — number of candidate timing solutions, the effects of which are then simulated to evaluate their fitness. At each iteration, a variety of methods ranging from Genetic Algorithms to Simulated Annealing and Particle Swarm Optimisation can then be used to modify and combine the most successful solutions into a new set of candidates.

Such methods are particularly suited for solving obscure problems as they require no attempt to establish an explicit correlation between the control variables and the desired outcome. Rather, they rely on the assumption that if any relevant phenomena can be modelled with sufficient accuracy and a performance index can describe the degree of achievement of the optimization objectives, then the system can be made to *naturally* evolve towards an optimal solution.

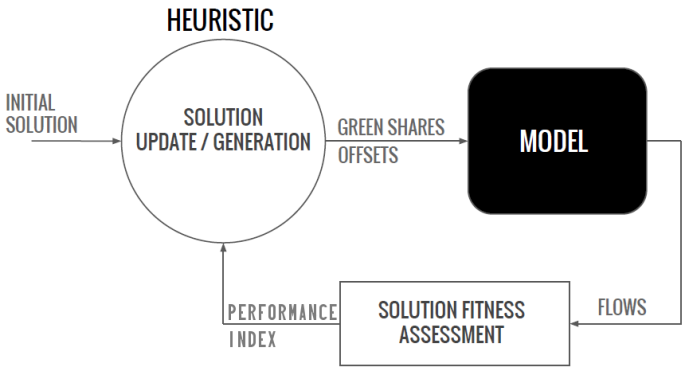

Figure 6 – Conceptual information flow in a heuristic approach to signal optimisation

It is therefore obvious that the model used to assess the fitness of candidate solutions should represent a sensible trade-off between speed and completeness: the real-world performance will inevitably be disappointing if the optimisation could not account for relevant traffic phenomena that were simplified out of the solution assessment, while on the other hand the need to evaluate huge numbers of candidate solutions calls for a lean and fast method to predict the outcome of a given timing plan. Heuristics that depend heavily on the choice of initial conditions often use maximum bandwidth solutions as starting point in the search for minimum total delay, to shave off convergence time and increase the quality and applicability of solutions.

This approach has been taken most notably by the Transport Research Laboratory, the UK based institution that since Robertson (1969) has been developing the *TRAffic Network StudY Tool*, which was born as a software tool to minimise stops along arterial roads while accounting for reasonably realistic vehicle behaviour, and was gradually extended to model ever more complex phenomena. Today, TRANSYT can handle pedestrian flows, optimise green shares as well as junction offsets and include actuated signals, all the while monitoring a custom set of network-wide performance indicators that can implement whatever policy the traffic administration desires. The optimisation relies on the availability of a complete transportation network model, possibly including detailed junction geometry. A range of search algorithms can be used to explore complex timing solutions, which are then evaluated using either micro– or macrosimulation models.

Earliest version of Transyt implemented a simple hill climbing algorithm that explored the non convex performance function by executing a predetermined set of short and long steps to vary each control variable in both directions alternatively. At each step, the changed value of the control variable is kept if it improved the performance index. Park et al. (1999) introduced a traffic signal optimization program for oversaturated intersections consisting of two modules: a genetic algorithm optimizer and mesoscopic simulator. Colombaroni et al. (2009) devised a solution procedure that first applies a genetic algorithm and then a hill climbing algorithm for local adjustments. The fitness function is evaluated by means of a traffic model that computes platoon progression along the links, their combination and possible queuing at nodes through analytical delay formulations. Colombaroni et al. (2010) extended the model to design optimal signal settings of a synchronised artery with predetermined rules for dynamic bus priority. A visual example is given in Figure 7 of the genetic algorithm representation of signal settings as a chromosome population.

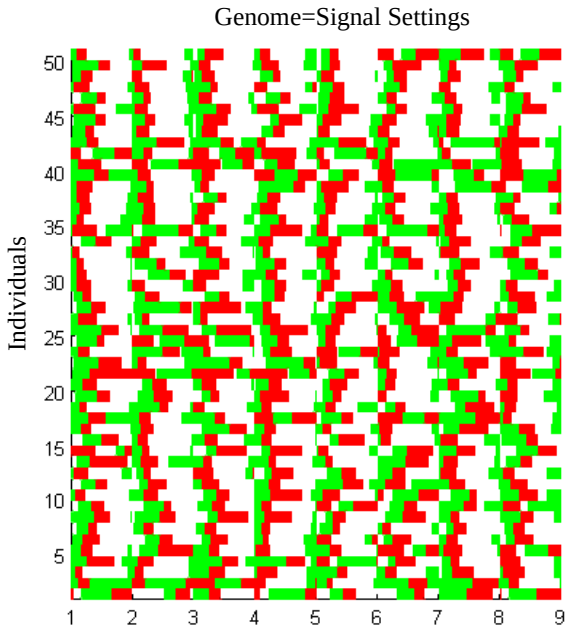

Figure 7 – Example of solution population generated by a genetic algorithm at a generic iteration: each individual of the population represents a synchronisation solution for a 9-intersection road artery whose green, red and offsets are represented by the genome of one individual.

Metaheuristics often see applications in traffic signal engineering that reach beyond ordinary signal planning, and have more than once played an important role in *research* by aiding the formalisation of less intuitive correlations between signal settings and traffic behaviour. Gentile and Tiddi (2009) use a Genetic Algorithm to venture out into the yet uncharted territory of arterial synchronisation under heavy congestion and queue spillback. To predict the outcome of candidate signal plans, the heuristic method relies on the General Link Transmission Model (Gentile 2015), which implements the Kinematic Wave Theory to allow accurate simulation of traffic dynamics and model physical blockage of links, while requiring sufficiently short computation times to deal with the very large number of solutions to be evaluated.

The optimisation reveals a crucial difference between subcritical and supercritical flow conditions: while in the former case the optimal green wave is led as usual by the flow velocity, the same approach proves completely *ineffective* under supercritical conditions, which oppositely demand that the *backwards propagating jam wave speed* should set the pace of upstream signals, to ensure that the residual capacity of saturated links is fully exploited.

It must be noted that the level of detail taken into account when using metaheuristics comes at a heavy cost in terms of computation speed, which restricts the functionality of this type of software to that of advanced yet *offline* development tools. As long as ordinarily accessible computing power remains insufficient for true real time functionality, advanced optimization suites are staying on top of the game by attempting to streamline the interactions between the development environment and the street-level equipment, e.g. providing offline optimisation based on real time readings and quick and simple deployment of new plans.

Such efforts are driving this type of software towards a sort of mid-term, day-to-day real time adaptivity which, possibly in combination with a true real-time program selection logic may well prove to be an effective approach to slowly and steadily improve the signalization of large urban networks.

ONLINE SIGNAL CONTROL

Over the years, many attempts have been made to render the signalisation system of urban networks capable of *reacting* autonomously to the traffic conditions, to address the mutable nature of demand. In this context, the term *optimisation* is used in its broader sense of *choice of the best option*, whether this is picked out of a set of previously planned solutions, tailored on-the-fly onto the current traffic conditions, or simply the result of a sequence of best possible actions evaluated individually: the main features of each class of very different approaches will be illustrated in the following sections.

The one thing that all responsive traffic control systems have is the need to *perceive* the traffic state on the network by means of *detectors*. The type and quantity of information required for different optimisation approaches may vary, but in the end it always boils down to one, or a combination, of the following quantities:

- flow: the number of vehicles crossing a road section in a given amount of time
- occupancy: the percentage of time the sensor spends occupied by a vehicle
- velocity: the average speed of the vehicles through a road section

Detectors, which may also serve monitoring and simple *counting* purposes, are described extensively in Chapter 2.

Optimal Plan Selection

The most straightforward traffic actuated control strategy is represented by *plan selection* systems, such as the *Urban Traffic Control System* developed by the Federal Highway Administration. Their aim is to ensure, based on real time information about the traffic conditions, that the most suitable amongst a set of predetermined signal plans is enacted.

Automatic plan selection developed as an enhancement for both isolated traffic lights and centralised traffic control systems, which previously relied only on accurate signal schedule setting to take advantage of multiple plans, each tailored to specific traffic conditions either expected or observed on the network.

Plan selection is typically performed by comparing real time detector readings with the conditions for which each plan was designed. Readings may be validated using historical data and otherwise filtered to protect the stability of the system against measurement errors and faults. The pre-processed input is then fed into an *objective function* that computes the degree of suitability for each plan.

Consider for example a bank of signalisation plans $s \in S$, each representing a *solution* designed around a given traffic scenario — the generalisation applies at the network level just as well as for a single intersection, where the concepts of *plan* and *program* are equivalent.

Each scenario is represented by a *snapshot* of the traffic conditions: assume this to come in the form of flow and occupation values measured on a subset $A^{\oplus} \subseteq A$ of detector equipped arcs of the network.

The core objective function of a plan selection method quantifies the degree of coincidence between the flow and occupancy values $\bar{q}_{a,s}$ and $\bar{o}_{a,s}$ associated with each of the pre-timed solutions with those measured on the corresponding network arcs in real time.

A possible form for such a function is e.g.

$$\omega_S = \sum_{a \in A^{\oplus}} \alpha_a \cdot \left(\beta^q \cdot \left(q_a - \bar{q}_{as} \right)^2 + \beta^o \cdot \left(o_a - \bar{o}_{as} \right)^2 \right) \tag{45}$$

where the current flow and occupation values q and o refer to each individual arc a, as do the location weights α (some locations may be strategically more important than others) and the measurement weights β which reflect the relevance (or accuracy) of each reading at the given location. Equation (45) can easily be extended to account for additional reading types.

The most suitable plan is the one that minimises the performance index ω_s, representing the divergence of the current traffic conditions from its signature traffic snapshot.

The system may further require the best candidate solution to beat the currently running plan by more than a predefined threshold before confirming a plan change: a cautionary measure called *Anti-hunting*, and taken to avoid continuous switching between similar plans, particularly in applications where a large number of plans are used to closely follow the evolution of demand throughout the day.

Switching between different plans may momentarily disrupt corridor progression, therefore in some cases a hybrid transition cycle is synthesised from the outgoing and incoming plans.

The above principles equally apply to single junctions, areas or entire networks, and require a relatively low number of strategically placed detectors, making automated plan selection a viable and cost-effective option for many applications.

Traffic Actuated Control

The class of traffic control methods referred to as *actuated* generally don't rely much (if at all) on an underlying network model, and seldom deal with the very concept of signal program as anything beyond a predetermined sequence of phases.

In its simplest form, an actuated controller put in charge of a junction does the job that once was a traffic officer's: by applying a set of rules it attempts to give as much right-of-way as possible to congested approaches — for as long as it's needed — while keeping other flows in check to avoid having any queue standing for *too* long. Just like their human counterparts, actuated signals are extremely effective at maximising the throughput of their own junction, thanks to the direct gauge of traffic on every approach and very fast reaction times, but may prove disastrous at the wider network level since a poorly designed control strategy may introduce self-induced oscillations in the traffic flows, rendering the whole system unstable — particularly when flows approach critical values.

Signal actuation depends on real time data acquired at the junction by short range sensors that monitor individual approaches: to this end, cameras have recently started replacing street level inductive loops, as a single device is often capable of monitoring several approaches.

The first traffic actuated intersection was tested in the USA in 1930. The controller relied on microphones to detect vehicles waiting on the lesser approaches, and drivers had to honk

to signal their presence. Since then, the available technologies have improved, but the simple fact remains that with cheap electronics and very simple logic (analog friendly, if necessary) an actuated controller has long been capable of looking after a junction better than any pre-timed plan ever will, no matter how well the timings are optimised to fit *expected* flows.

Different levels of automation are generally classified into two categories:

- *semi-actuated:* the controller monitors the low flow approaches to allocate them green time as required, and otherwise only serves the main approaches;
- *actuated:* the controller monitors all approaches.

Principles of Operation

Every few seconds, an actuated controller must answer the question: "*should the transition to the next phase start right now?"* or some variant thereof (e.g. to include the possibility to skip a phase). Actuated junction control is now commonplace around the world: any given implementation may rely on different types of sensor and data (e.g. cameras or pressure plates, simple counts or occupancy), but could likely be reduced to the basic principles (based on common induction loop readings) presented in this section.

Consider a signal phase serving a single approach a to a junction.

The incoming lane is equipped with an induction loop a short way back from the stop line (just enough to remain upstream of the back of the queue for most of the time), through which the signal controller measures the time interval t_a^h between subsequent vehicle detections, commonly referred to as *headway*.

After the phase has started, the signal controller determines its duration on-the-fly, based on sensor readings in the context of a few fundamental parameters such as minimum and maximum green durations $g_p^{\min}$ and $g_p^{\max}$, and maximum headway $\hat{t}_a^h$: the latter could be seen as a minimum flow rate required to extend the phase duration and does not necessarily concern a single lane group. In the scope of this example, since only one phase and one lane group are considered, the subscript will be dropped.

The actuation parameters can be fixed, or determined in real time by taking into account the traffic flow on the relevant manoeuvres, the standing queues before the phase start d_a, the number of vehicles queuing for lane groups served by *other* phases $d_{b\neq a}$, or all of the above.
Between the minimum and maximum green duration values, the junction signal controller continuously checks whether the time elapsed since the last vehicle passage has exceeded the maximum headway value for the current phase: if so, the transition to the next phase begins:

$$\text{Initiate phase transition } p \rightarrow p+1 \text{ if } \begin{cases} t \geq \mathrm{g}_p^{\min} \\ t^h \geq \hat{\mathrm{t}}^h \quad \vee \quad t > \mathrm{g}_p^{\max} \end{cases} \tag{46}$$

where the minimum green value can be obtained similarly to (8) in order to at least ensure discharge of the standing queue, possibly using an estimate of the incoming flow, and the maximum may depend on the minimum green of other manoeuvres and residual cycle time. After the initial minimum green, the headway threshold can be a dynamic function of flows, queues and time since phase start t, as could be formalised e.g. in the following

$$\hat{t}^h(\tau) = \hat{t}^{h0}\left(1 - \tau^{\beta\left(d_{q\neq p}\right)}\right) \quad \text{with} \quad \tau = \frac{t - g_p^{\min}}{g_p^{\max} - g_p^{\min}} \tag{47}$$

whereby the maximum headway decays from its initial value $\hat{t}^{h0}$ over the time span between the minimum and maximum green at a rate determined by any positive nondecreasing function β of the queues accumulated on other approaches $q_b \; \forall \; b \notin A_p$.
Note that the independent variable $\tau \in [0,1]$, which means that in case of very high or very low queues on other approaches the headway threshold drops to zero right after the minimum green or not until the end of the maximum green, respectively.

Furthermore, even as queues on other approaches grow between $g_p^{\min}$ and $g_p^{\max}$ the maximum threshold remains monotonic nonincreasing as long as β is a sensible function of the (nondecreasing) queues as seen in Figure 9.

The basic principles expressed so far in (46) and (47) can be shaped into any of the most common categories of actuated control:

Volume actuation: in the simplest case, the green signal duration is bound between fixed minimum and maximum design values. It can be extended beyond the minimum value only as long as vehicles keep reaching the junction at sufficiently short intervals, as seen in Figure 8. Each vehicle arrival starts or resets a timer, and the next phase is initiated as soon as the gap between subsequent vehicles surpasses the headway threshold $\hat{t}^h$, which is also constant.

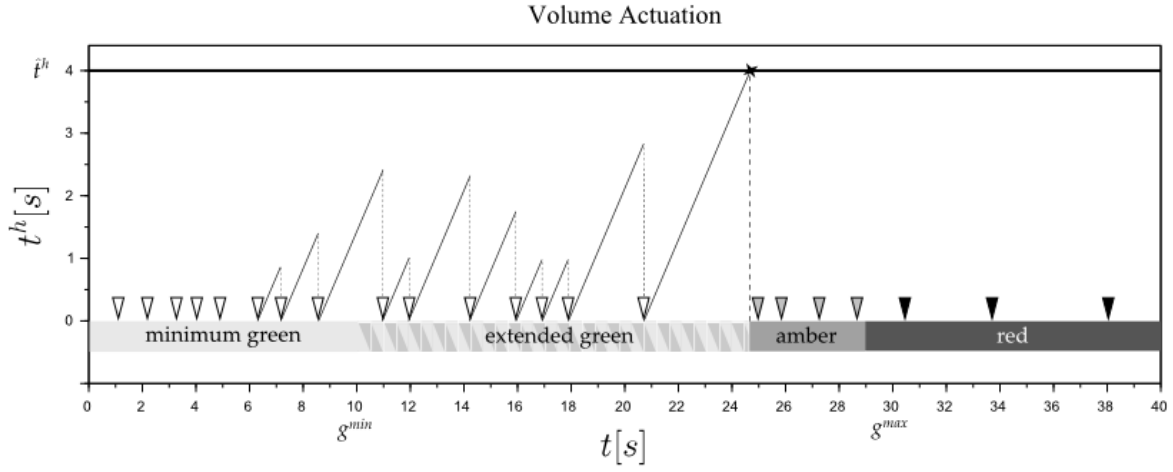

Figure 8 – Volume actuation: on the horizontal axis, time since the start of the current phase; on the vertical axis, time elapsed after each vehicle detection (triangular markers). Each vehicle reaching the sensor before the headway threshold resets the timer. Shaded and black markers respectively represent vehicles reaching the sensor after the maximum headway time has been exceeded, and vehicles that must stop at the red light.

Volume-density actuation: follows the same principles of volume actuation but the minimum green time is determined by the amount of vehicles initially queuing at the stop line. The maximum headway allowed to extend the current phase becomes more and more

restrictive as the maximum green duration is approached, as portrayed by any of the lightly-shaded curves in Figure 9, each corresponding to a different fixed value of β.

Density actuation: the headway threshold decay rate is governed by the number of vehicles detected on the other approaches through the exponent β, so that at high saturation levels a drop in arrival rate, which denotes the end of a queue or the rear of a dense vehicle platoon, may trigger the transition to the next phase.

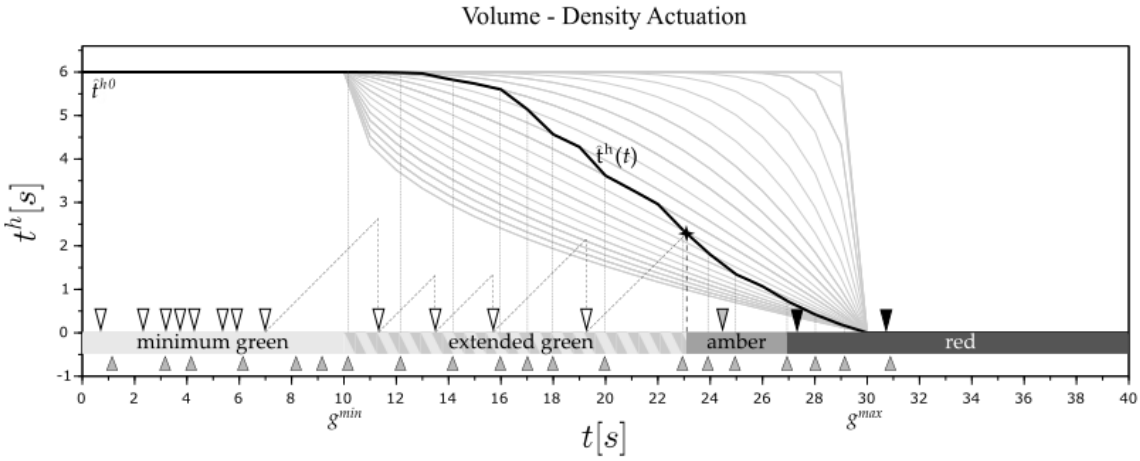

Figure 9 – Density actuation: symbols and quantities as in Figure 8. On the vertical axis, the headway threshold is also shown declining to zero over the green extension period following the shaded lines in the background, which correspond to polynomial curves as in (47) with fixed values of β. The maximum headway curve latches onto increasingly rapid decay curves with each arrival detected on other approaches, marked by the small triangles.

Although actuated controllers are mostly regarded as *autonomous* entities, it should be evident that phase duration limits and threshold function parameters associated with each approach can be finely tuned by a centralised system to deal with specific traffic scenarios.

Isolated actuated controllers are relatively undemanding from the infrastructural point of view, but the considerable drawback is that without junction coordination the flexibility in phase duration may come at a heavy cost in terms of arterial progression disruption.

Real Time Plan Generation

Real time optimisers that perform plan generation are a class of *proactive* signal control systems that, based on current traffic conditions, seek to develop an optimal plan to apply in the immediate future, either from first principles or by continuous update of an existing pre-timed plan. While each plan is played out, the system gathers information to make the next.

This mode of operation is often referred to as *rolling horizon*, and in order for the system to respond effectively (i.e. to capture and react to rapid changes in traffic conditions) the rolling horizon time step should be reasonably short, which imposes austere constraints on the optimisation methods. Some real-time optimisers with a very short rolling horizon step update the signalisation plan *at every cycle*, so that their behaviour may appear indistinguishable from that of an actuated controller. It is important however to understand the clear conceptual difference between the two: actuated controllers perform second-by-second decisions about

the best action to perform instantly, while the systems considered in this section *plan ahead,* producing fully featured signal plans made of cycle times, offsets and green shares deemed optimal for dealing with the traffic conditions observed.

Incremental Analytical Optimisation

The most prominent member of this category is the *Split Cycle and Offset Optimisation Technique* developed for research purposes in Glasgow, and first applied there in 1975 under the acronym SCOOT by which it is now popular all over the world, counting over a hundred active installations.

It revolves around a centralised control unit which generates plans based on a real-time traffic snapshot gathered from detectors. The signalisation plans are continuously updated, with a frequency in the order of one to three cycle times, and may concern the entire network or *regions* thereof which are expected to feature homogeneous traffic conditions.

One of the main advantages is that optimisation requires very little information about the network. All the system needs, for *each approach* to a controlled junction, is the following:

- distance from each detector (at least one is needed) to the stop line
- saturation flow of the detector lane at the junction
- total vehicle storage capacity
- initial lost time and clearance time for the corresponding signal phase

The SCOOT optimisation method described in Robertson (1986) is based on Cyclic Flow Profiles: for each approach to a controlled junction these represent the continuously updated flow profile covering the span of a signal cycle with a resolution of 4s, obtained from the readings gathered by sensors. The centralised control unit integrates CFPs to estimate the number of vehicles arriving at the stop line while the signal is red, which combined with saturation flows yields the queue sizes and clearance times as pictured in Figure 10.

The system is therefore all the more effective if detectors are placed far from the stop line — possibly just downstream of the previous junction — to give as early a warning as possible of changes in the expected flow pattern. This also allows the system to detect significant spillback situations, triggering different operation modes aimed at gridlock avoidance.

It may be advisable to trade accuracy for early detection by overlooking minor side streets which may alter the flow rate and progression between two major intersections. However, best results are obtained with more sensors spaced out along each inbound arc.
With the flow conditions described by this simple traffic model, the optimiser proceeds to calculate cycle times, offsets and green shares based on explicit mathematical formulations. The fitness of the solution found is quantified by a global cost function built on a linear combination of delays and number of stops. The method runs as follows:

1. *Cycle Optimisation:* each region of the network shares a single cycle time. Its ideal value is determined by an empirical formula similar to Equation (33) based on the saturation conditions at the *critical* (i.e. most saturated) intersection, see **Error! Reference source not found.**.

2. *Green Share Optimisation:* once the cycle time is determined, green shares are updated at every intersection. As soon as it possesses relevant flow information, the optimiser decides whether to anticipate/delay each phase change by up to 4s, depending on which alternative scores best according to an objective function aiming to reduce the saturation level of the most saturated approach to the junction.

3. *Offset Optimisation:* at every cycle, the central unit may shift the pre-timed offsets by up to 4s in either direction, if this leads to an improvement in an explicit objective function which accounts for the degree of synchronisation with the *adjacent* junctions, possibly accounting for updated travel times of the relevant arcs.

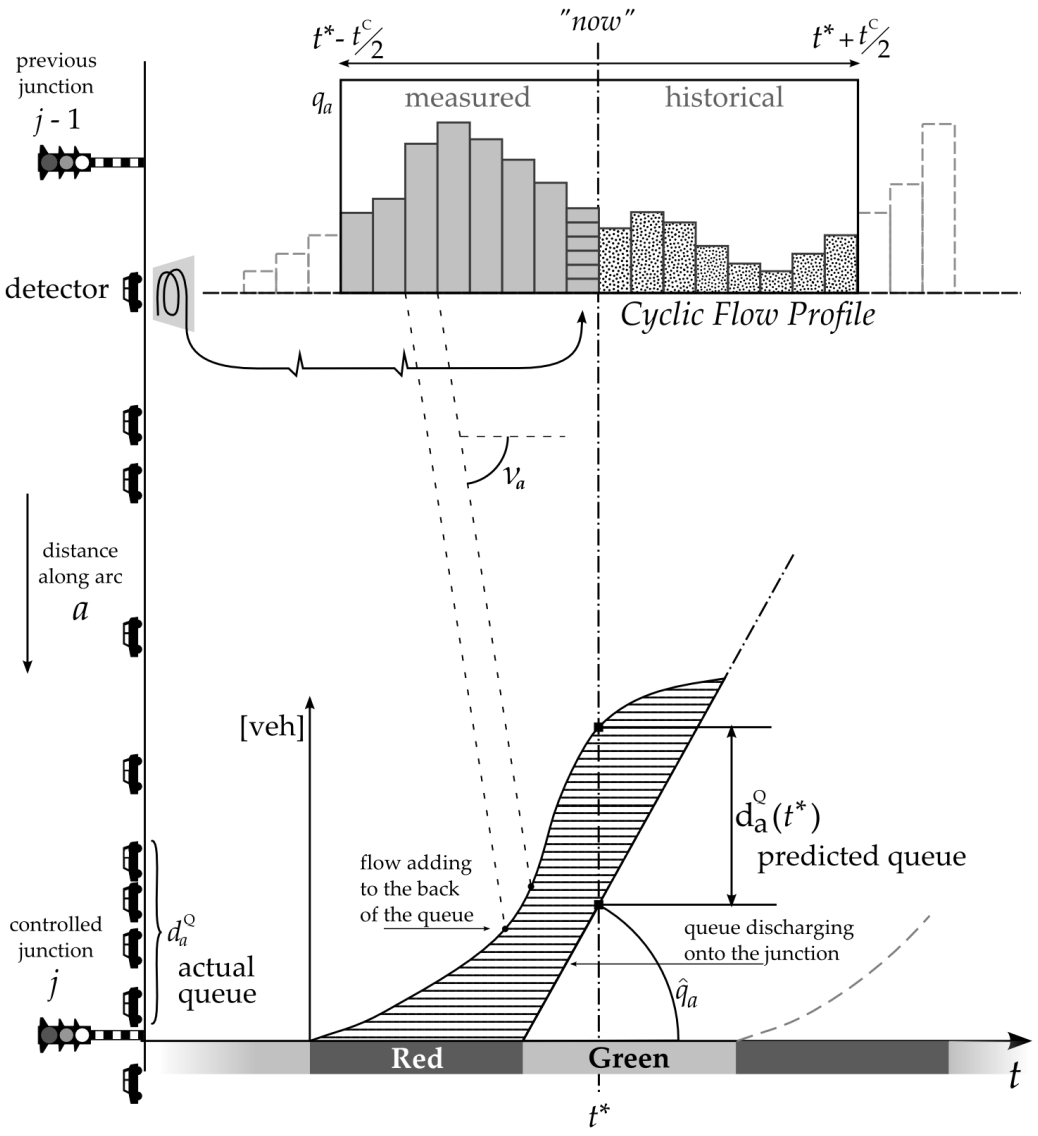

Figure 10 – SCOOT Cyclic Flow Profiles and queue prediction: detector readings are used to update the flow profile, which is integrated to predict the queue forming at the downstream junction during the red

phase. The information may prompt the system to anticipate or delay a phase change in order to accommodate the measured demand.

This type of optimiser has the advantage of low modelling requirements and very fast computation times, combined with the ability to operate quite close to complete saturation — allegedly up to 90% critical junction saturation.

Even with modest prediction capabilities and no full network model, it has proven capable of dealing reasonably well with moderate flow pattern alterations and unusual route choices such as might be caused by accidents or road works.

It does however rely heavily on the accuracy of detectors, which if insufficient may cause the performance of the system to decline rapidly: the modified timings are in fact set to degrade back to the pre-timed plan if sensor faults are detected.

The small adjustment step sizes are also chosen to increase the robustness of the system to detection faults: unfortunately, this goes to the detriment of its responsiveness, which has been pointed out as the main weakness of SCOOT.

Linear Quadratic Optimal Control

The *Traffic Urban Control* system commonly referred to as TUC was developed in the scope of TABASCO (Telematics Applications in BAvaria SCotland and Others), a late 90s European project aimed at demonstrating the applicability of advanced transport telematics as innovative solutions for traffic management. Initially conceived for green split optimisation, it was extended to deal with cycle and offsets as well, and later enabled to perform on-the-fly Public Transport prioritisation.

It therefore constitutes a direct alternative to the SCOOT system mentioned in the previous section, and was designed to build upon the latter's ease of applicability while addressing its main issues: most notably its slow response to rapid traffic variations (due to the incremental correction approach), and scarce effectiveness under high saturation conditions. The former was made unnecessary by the verified robustness and stability of the system, while a stronger interdependence of measurements and signal settings across the entire network helped counteract the tendency shown by more localised control policies to accelerate the onset of saturation by blindly favouring high flows.

The system inputs are the average numbers of vehicles on network links (which may be estimated from occupancy readings if video detection is not possible) and public transport information, at least accurate enough to detect the *presence* of public vehicles on a given link. Cycle and offset optimisation are carried out independently and in much the same way as it was described in the previous section.

What characterises the TUC strategy however is its approach to *green split optimisation,* based on a Store-and-Forward traffic model (Aboudolas 2009) and simple control theory.

These are combined to formulate the control problem as a Linear Quadratic optimisation, as illustrated in detail in Diakaki (2002) and summarised here.

The instantaneous network state is represented solely by the number of vehicles on each link. The discrete-time evolution rule of the network dynamic system encapsulates its

dependency on the decision variables and on previous instantaneous states, and in matrix form may be written simply as:

$$\mathbf{d_a}(t+1) = \mathbf{A}\,\mathbf{d_a}(t) + \mathbf{B}\,\Delta\boldsymbol{\gamma}_\mathbf{p}(t) \tag{48}$$

where $\mathbf{d_a}$ is the vector of states containing the number of vehicles on each link and $\Delta\boldsymbol{\gamma}_\mathbf{p}$ is the vector of variations in green share applied to each signal phase, with respect to a baseline signal plan assumed to lead to steady state queues under non-saturating conditions.

A and **B** are the state and input matrices: they respectively encapsulate the network topology and the expected impact of signalisation (based on signal staging, turning rates, saturation flows) on the movements of traffic volumes across time intervals.

It should be noted that the *expected* demand is taken into no account: this is reasonable as TUC aims to react to the *manifest* impact of disturbances on the controlled network rather than to their forecast consequences.

At the core of this approach lies a simple gain matrix — introduced in Equation (50) — rather than accurate modelling of physical phenomena and constraints: its calculation and calibration however are most computationally demanding processes.

In order to minimise the risk of oversaturation and queue spillback on all network links, the chosen strategy is to attempt to balance the link relative occupancies (with respect to each link's jam storage capacity), as expressed by the following quadratic criterion:

$$\omega^{TUC} = \frac{1}{2}\sum_{t=0}^{\infty} \left\| \mathbf{Q}\,\mathbf{d_a}(t) \right\|^2 + \left\| \mathbf{R}\,\Delta\boldsymbol{\gamma}_\mathbf{y}(t) \right\|^2 \tag{49}$$

where **Q** and **R** are non-negative definite *diagonal* matrices of weights, so that the cost function is compatible with the standard form of a Linear Quadratic Cost. Matrix **Q** contains the inverse storage capacities of links, so that the first term of the sum drives the relative occupancy balancing, while the second term favours smooth changes in the control variables, influencing the magnitude of control reactions through appropriate scaling factors contained in matrix **R**. The infinite time horizon of the sum reflects the necessity to obtain a time-invariant feedback control law in accordance with the LQ optimisation theory.

The LQ feedback control law is then obtained by minimisation of the performance criterion (49) subject to (48): calculation of the control matrix **L** is straightforward, but can only be performed offline by solving the infinite-horizon *Discrete-time Algebraic Riccati Equation* from the network topology and objective function weights described by the matrices **A**, **B**, **Q** and **R**. These must be computed and calibrated individually for the specific network topology, capacities, signal staging etc, by simulation or other optimisation methods: this is a lengthy and demanding task to be performed as part of the system setup.

However, after finding the stabilising solution $\mathbf{L}$ to the dynamical system expressed by the DARE, things get much simpler, with the control law taking the standard form

$$\boldsymbol{\gamma}_\mathbf{p}(t) = \bar{\boldsymbol{\gamma}}_\mathbf{p} - \mathbf{L}\,\mathbf{d_a}(t) \tag{50}$$

where $\bar{\gamma}_p$ is the vector of baseline green shares. Optimal modifications to the green times are linearly dependent on the current network state vector of link occupancy measurements through the matrix **L**, which provides both the discharge and gating functionalities: intuitively, as the occupancy of a link increases, so does the green share that favours its outflow, while upstream arcs experience a reduction in green time to avoid its oversaturation.

These effects can be accentuated or mitigated by weighing elements of the state vector according to specific rules, e.g. to prioritise desaturation of certain links as they approach critical saturation levels.

A simple form of *public transport prioritisation* can similarly be integrated into the application of the control equation (50), by further weighting link occupancy values in function of the number of public transport vehicles detected on them.

Since control constraints such as green time upper and lower bounds cannot be directly accounted for by the LQ methodology, the green shares output by the regulator are further processed on the fly by a simple optimisation algorithm that, in linear time, finds the set of feasible green times that least deviate from the optimum.

Although several software packages are available for solving the DARE for this standard LQ control problem using a variety of well documented methods, the calculation of an effective control matrix remains a time consuming task, particularly for large networks, and must be performed anew every time the controlled network is modified or extended.

This lack of flexibility represents the main drawback of the approach just presented, although it has been proven that reasonable variations of traffic parameters such as turning rates and link saturation flows have little effect on the control matrix.

On the other hand, the real-time operation of the TUC control strategy only consists of the solution of the simple matrix equation (50) followed by the application of green time constraints, which are both extremely fast and undemanding operations, making the quadratic regulator a particularly suitable approach for real time applications. Furthermore, the feedback controller is perfectly capable of responding appropriately to very specific traffic anomalies such as accidents or roadworks, as confirmed by both simulation and empirical data gathered from real world installations.

Since the first TUC installation in Glasgow, further applications of optimum control theory to the signal setting problem, including open-loop Quadratic Programming and Nonlinear Optimum Control based on the same store-and-forward traffic paradigm, have been developed and investigated. These aim to improve upon the performance of the simple feedback controller by accounting for more detailed network dynamics, factoring in time varying demand, or allowing for a larger and more effective set of decision variables: encouraging results presented in Diakaki (2009) suggest that despite an increased real-time computation complexity, these may be considered strong competitors and potential successors to the Linear Quadratic TUC approach.

Traffic Gating

Feedback Traffic Gating (Ekbatani 2012) is a form of actuated signal control aiming to prevent oversaturation of critical portions of the network by holding back the incoming traffic flows — using deliberately exaggerated red phases — rather than attempting to deal with the flows already trapped in a congested area. In these respects, it constitutes a simple yet innovative method to induce more efficient utilisation of the existing infrastructure, and an answer to the patent performance degradation that currently feasible real-time optimisation solutions face under saturated conditions.

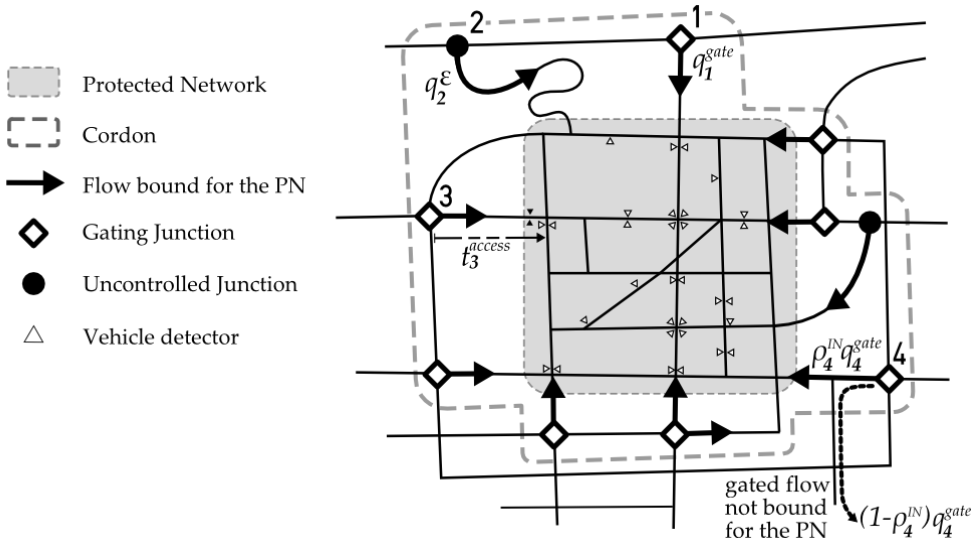

Figure 11 – Traffic Gating: a cordon of gating junctions holds back traffic attempting to access the *Protected Network*. If it is not possible to implement gating at one or more cordon junctions (see j = 2), these may allow some disturbance flows to sneak past the feedback controller. Conversely, part of the gated flow from a cordon junction may not in fact be bound for the PN (see j = 4). The system must account for the delay between control application and effect (due to the physical distance between the gating junctions and the PN, see j = 3) and for incomplete or uneven detector placement in the protected network.

Based on the general principle that even from the users' point of view there's no advantage to getting *close* to one's destination sooner, only to be stuck in traffic for longer, the system delays incoming vehicles in order to keep the controlled network near to but *below* its saturation occupancy level, which can be monitored effectively even with a small number of detectors (Ekbatani 2013): this was proven the most effective strategy to maximise *network throughput*, which constitutes a good measure of how efficiently the network is being used.

A feedback controller is used to ensure that the only vehicles pre-emptively delayed are those which not only would, on average, be delayed anyway further down their path, but would *critically* increase congestion — causing themselves greater delays as well as others — were they to access the critical region.

Network Fundamental Diagram Formulation

Feedback Traffic Gating revolves around the concept of Network Fundamental Diagram introduced in Ekbatani (2012), profiling *throughput* as a function of *occupancy*, as seen in Figure 12 where the axes of the sample NFD correspond to *Total Time Spent* (in *vehicle hours per hour*) and *Total Travelled Distance* (in *vehicle kilometres per hour*) cumulatively by all users hourly. Such relationship may be obtained empirically from observation of the area of interest, and allows to identify with certainty the optimal operation point for the feedback controller to suit the behaviour of a specific network.

An operational NFD is derived from real or simulated occupancy measurements o_a taken on a set of detector equipped arcs $A^{\oplus} \subseteq A$ at discrete time intervals *t*, corresponding to signal cycles. Occupancy is converted into an estimate $d_a(t)$ of the number of vehicles on each arc during the t^{th} signal cycle

$$d_a(t) = \frac{l_a \cdot n_a^{lane}}{100 \cdot \ell} \cdot o_a(t) \tag{51}$$

where n_a^{lane} is the number of lanes of link *a*, l_a its length, and ℓ the average vehicle length. Hence, the relevant quantities are obtained by summing over the measurement arcs:

$$\omega^{TTS}(t) = \sum_{a \in A^{\oplus}} \frac{d_a(t) \cdot t^C}{t^C} = \sum_{a \in A^{\oplus}} d_a(t) \tag{52}$$

$$\omega^{TTD}(t) = \sum_{a \in A^{\oplus}} \frac{q_a(t) \cdot l_a \cdot t^C}{t^C} = \sum_{a \in A^{\oplus}} q_a(t) \cdot l_a \tag{53}$$

The values thus obtained are sufficiently precise for the purpose of traffic gating, especially if detectors are located around the arc midpoints. Although a high number of detector links (ideally $A^{\oplus} = A$) yields a more accurate NFL, Ekbatani (2013) proves that fully functional results can be obtained also from a *reduced* NFL in more likely scenarios where only a cost-effective subset of links has detection capabilities, such as would be sufficient for ordinary traffic monitoring, plan-selection schemes, or actuated signal control applications, as portrayed e.g. in Figure 11.

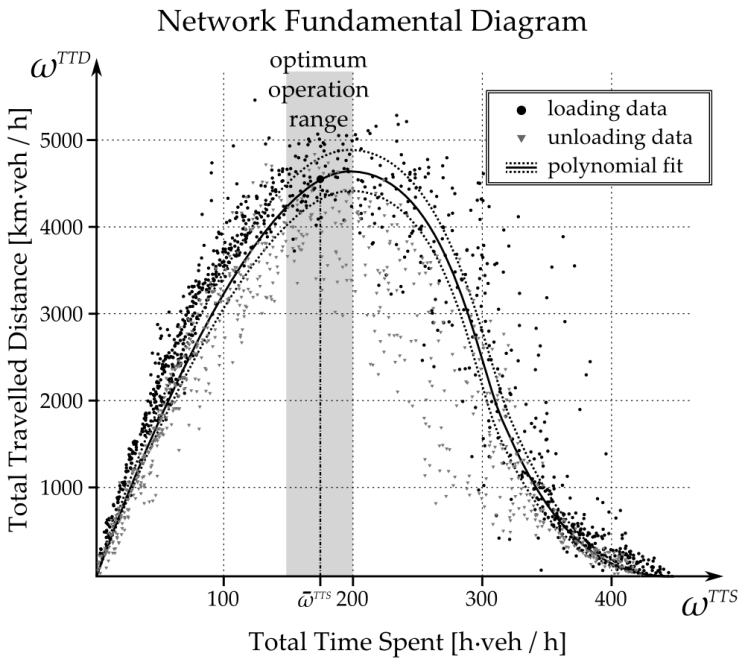

Figure 12 – An experimental Network Fundamental Diagram: a polynomial fit of the TTD curve is obtained from flow and occupancy measurements, identifying an optimum operation point on the TTS axis. As long as ω^{TTS} is kept within the optimum operation range, the number of vehicles in the Protected Network is expected to maximise the infrastructure efficiency, resulting in shorter travel times for all users. The dynamics of the system during network loading and unloading can be expected to differ slightly as flows tend to be slower during the relaxation of a more congested state.

Feedback Controller Design

The gating control problem is to regulate the *TTS* in the *Protected Network* via appropriate manipulation of gated inflows, so as to maintain the *TTD* around its optimal maximum value. The task can be accomplished as summarised in Figure 13 based solely on real-time measurements via a simple and robust feedback regulator, taking advantage of the basic system dynamics described by the NFD.

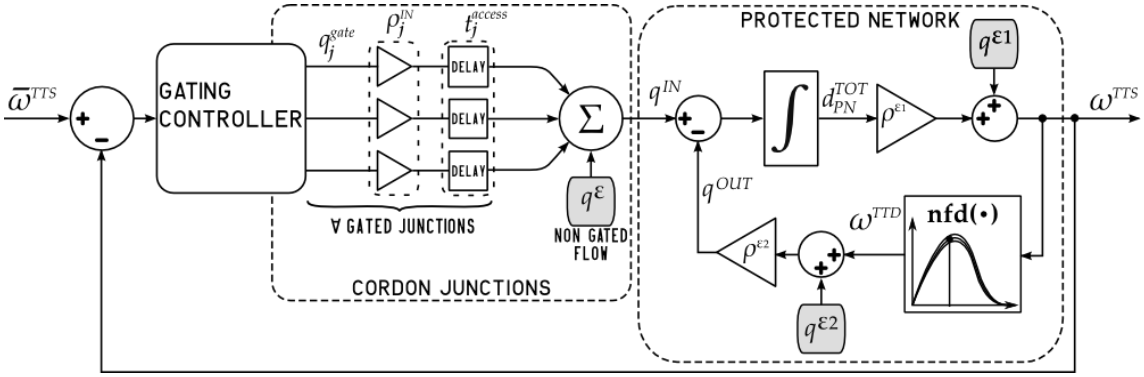

Figure 13 – Gating Feedback Controller and Protected Network Dynamic Plant.

The controlled input to the PN system is the gated flow q^{gate}, and the main disturbance the uncontrolled inflow q^{ε}. Referring to Figure 13, the inflow q^{IN} in continuous time is

$$q^{IN}(t) = \rho^{IN} \cdot q^{gate}\left(t - t^{access}\right) \tag{54}$$

where ρ^{IN} is the portion of gated flow entering the PN, and t^{access} the time it takes for vehicles to reach the PN from gating junctions not directly located on its boundary.

Consider the total number d_{PN}^{TOT} of vehicles in the PN: its rate of change is determined form vehicle conservation, which reads

$$\dot{d}_{PN}^{TOT}(t) = q^{IN}(t) + q^{\varepsilon}(t) - q^{OUT}(t) \tag{55}$$

However, $\omega^{TTS} = d_{PN}^{TOT}$ only if all PN links are monitored, which is not generally the case.

Realistically, the *TTS* is smaller than the true number of vehicles by some factor $\rho^{\varepsilon 1} \leq 1$. Allowing for an additional measurement error $q^{\varepsilon 1}$ the *TTS* value to be used in the NFD is

$$\omega^{TTS}(t) = \rho^{\varepsilon 1} \cdot d_{PN}^{TOT}(t) + q^{\varepsilon 1}(t) \tag{56}$$

and finally, if nfd(•) is a nonlinear best fit of the NFD data (see Figure 12) and $q^{\varepsilon 2}$ the error due to the data scatter, the resulting *TTD* is

$$\omega^{TTD}(t) = \text{nfd}\left(\omega^{TTS}(t)\right) + q^{\varepsilon 2} \tag{57}$$

which as seen in Figure 13 is proportional to the network outflow q^{OUT} aside for a scaling factor $\rho^{\varepsilon 2}$ analogous to $\rho^{\varepsilon 1}$ yielding a time delayed nonlinear first-order model between the initial q^{gate} and the resulting *TTS*, which can be linearised around the optimum steady state.

The following proportional-integral controller is then well suited to handle the gated flows:

$$q^{IN}(t) = q^{IN}(t-1) - K^{P}\left(\omega^{TTS}(t-1) - \omega^{TTS}(t-2)\right) + K^{I}\left(\bar{\omega}^{TTS} - \omega^{TTS}(t-1)\right) \tag{58}$$

where K^I and K^P are the integral and proportional gains to be fine tuned. The flow values thus determined have to be shared amongst all gated junctions, after accounting for monitored or estimated disturbance flows, and subjected to minimum/maximum green time constraints.

The resulting system is largely robust to measurement errors, low signal timing resolution, and fluctuations in demand. It may be activated at specific times or as the traffic conditions approach a critical state, and requires virtually no additional infrastructure with respect to an ordinary plan-selection centralised signal setting system. Provided that appropriate gating locations can be found, where gate-delayed flows do not risk compromising the mobility of vehicles not bound for the PN, the principles just illustrated will undoubtedly form the core of future sustainable approaches to relieve urban congestion by delaying or avoiding the extreme traffic conditions that frustrate most currently available signal optimisation techniques.

References

[1]. Aboudolas, K., Papageorgiou, M., & Kosmatopoulos, E. (2009). Store-and-forward based methods for the signal control problem in large-scale congested urban road networks. *Transportation Research Part C: Emerging Technologies*, *17*(2), 163-174.

[2]. Cipriani, E., & Fusco, G. (2004). Combined signal setting design and traffic assignment problem. European Journal of Operational Research, 155(3), 569-583.

[3]. Colombaroni, C., Fusco, G. & Gemma, A. (2009). Optimization of Traffic Signals on Urban Arteries through a Platoon-Based Simulation Model, In Proceedings of the 11th WSEAS International Conference on Automatic Control, Modeling and Simulation, Istanbul, Turkey, May 30–June 1, 2009.

[4]. Colombaroni, C., Fusco, G., Gemma, A. (2010). A Model and an Algorithm for Signal Synchronization and Bus Priority on Urban Arteries. Int. Conf. Models and Technologies for Intelligent Transportation Systems, Aracne. ISBN: 978-88-548-3025-7.

[5]. De Nunzio, G., Gomes, G., de Wit, C. C., Horowitz, R., & Moulin, P. (2015, December). Arterial bandwidth maximisation via signal offsets and variable speed limits control. In *2015 54th IEEE Conference on Decision and Control (CDC)* (pp. 5142-5148). IEEE.

[6]. Diakaki, C., Dinopoulou, V., Aboudolas, K., Papageorgiou, M., Ben-Shabat, E., Seider, E., & Leibov, A. (2003). Extensions and new applications of the traffic-responsive urban control strategy: Coordinated signal control for urban networks. *Transportation Research Record: Journal of the Transportation Research Board*, 1856, 202-211.

[7]. Diakaki, C., Papageorgiou, M., & Aboudolas, K. (2002). A multivariable regulator approach to traffic-responsive network-wide signal control. *Control Engineering Practice*, *10*(2), 183-195.

[8]. Dickson, T. J. (1981) A note on traffic assignment and signal timings in a signal-controlled road network. *Transportation Research B,* 267 – 271.

[9]. Ekbatani, M. K., Kouvelas, A., Papamichail, I., & Papageorgiou, M. (2012). Congestion control in urban networks via feedback gating. *Procedia-Social and Behavioral Sciences*, *48*, 1599-1610.

[10]. Fusco, G., Gentile, G., Meschini, L., Bielli, M., Felici, G., Cipriani, E., Nigro, M. (2006). Strategies for signal settings and dynamic traffic modelling. *XI EURO Working group meeting and EXTRA EURO Conference, (str. 851--864)*.

[11]. Gentile, G. (2015). Using the general link transmission model in a dynamic traffic assignment to simulate congestion on urban networks. *Transportation Research Procedia*, 5, 66-81.

[12]. Gentile, G., Tiddi, D. (2009). Synchronisation of traffic signals through a heuristic-modified genetic algorithm with GLTM. In *Proceedings of the XIII Meeting of the Euro Working Group on Transportation*.

[13]. Gomes, G. (2015). Bandwidth maximisation using vehicle arrival functions. *IEEE Transactions on Intelligent Transportation Systems*, 16(4), 1977-1988.

[14]. Keyvan-Ekbatani, M., Papageorgiou, M., & Papamichail, I. (2013). Urban congestion gating control based on reduced operational network fundamental diagrams. *Transportation Research Part C: Emerging Technologies*, *33*, 74-87.

[15]. Kimber, R. M., & Hollis, E. M. (1979). Traffic queues and delays at road junctions. LR909 Monograph, Transport and Road Research Laboratory.

[16]. Lee, R., Lee, K., & Quayle, S. (2008). Traffic signal timing manual, federal highway administration. *Turner-Fairbank Highway Research Center*, 6.

[17]. Li, W., & Tarko, A. P. (2010, March). Safety Consideration in Signal Coordination and Road Design on Urban Street. In *4th International Symposium on Highway Geometric Design*.

[18]. J. Little J., M. Kelson M., and N. Gartner N. (1981). MAXBAND. A Versatile Program for Setting Signals on Arteries and Triangular Networks. *Transportation Research Record*.

[19]. Papola, N. (1988). Teoria del deflusso veicolare e suo impiego nella progettazione e nella regolazione. *Atti di Trasporti.* Università degli Studi di Roma "La Sapienza".

[20]. Papola, N., & Fusco, G. (1998). Maximal bandwidth problems: a new algorithm based on the properties of periodicity of the system. Transportation Research Part B: Methodological, 32(4), 277-288.

[21]. Papola, N., & Fusco, G. (2000). A new analytical model for traffic signal synchronization. In Proceedings of 2nd ICTTS Conference, Beijing, China, Vol. 31.

[22]. Park B., Messer C.J. & Urbanik II T. (1999). Traffic Signal Optimization for Oversaturated Conditions: Genetic Algorithm Approach. Transpn. Res. Rec. 1683, 133-142.

[23]. Robertson, D. I. (1969). TRANSYT: a traffic network study tool.

[24]. Robertson, D. I. (1986). Research on the TRANSYT and SCOOT Methods of Signal Coordination. *ITE journal*, *56*(1), 36-40.

[25]. Smith, M. J. (1984). The stability of a dynamic model of traffic assignment-an application of a method of Lyapunov. *Transportation Science*, 18(3), 245-252.

[26]. Smith, M. J. (2010). Intelligent Network Control: Using an Assignment Control Model to Design Fixed Time Signal Timings. Chapters.

[27]. Smith, M. J. (2011). Dynamics of route choice and signal control in capacitated networks. Journal of Choice Modelling. 4(3), 30-51.

[28]. Tsay, H. S., & Lin, L. T. (1988). New Algorithm for Solving the Maximum Progression Bandwidth (With Discussion and Closure). *Transportation Research Record No. 1194.*

[29]. Webster, F. V. (1958). Traffic signal settings.