LOCALIZED SWITCHING RAMP-METERING
CONTROL WITH QUEUE LENGTH ESTIMATION AND
REGULATION AND MICROSCOPIC SIMULATION
RESULTS *

Xiaotian Sun 1 Roberto Horowitz 2

Department of Mechanical Engineering
University of California at Berkeley
Berkeley, CA 94720-1740, USA

Abstract: In this paper, we first review a localized ramp-metering strategy that achieves
the goal of reducing the congestion’s spatial and temporal span using locally available
information, with a switching mainline-traffic responsive ramp-metering controller that
adapts to the different traffic dynamics under different congestion conditions—free-flow
or congested, and a queue length regulator that improves the performance of the currently
used ad hoc “queue-override” scheme and prevents the queue from exceeding the ramp
storage capacity. Then a queue length estimator is designed to provide feedback to the
queue length regulator, using the queue-detector speed data that are available in the field.
Test results of this set of ramp-metering algorithms on a calibrated microscopic traffic
simulator demonstrate the performance and effectiveness of the switching ramp-metering
controller, the queue length estimator and regulator, and the control strategy. The Total
Vehicle and Passenger Delays are both reduced by 16%, while the Total Vehicle Time
and the Total Average Vehicle Speed are improved by 5.6% and 5.8%, respectively. As a
comparison, simulation results of ALINea are also presented. Copyright © 2005 IFAC

Keywords: Transportation Systems, Traffic Control, On-Ramp Metering, Switching
Control, Queue Length Estimation, Microscopic Simulation

1. INTRODUCTION

Freeway traffic congestion is a major problem in
today’s metropolitan areas. It occurs regularly during
commute hours. In addition, non-recurrent congestion
often takes place as a result of incidents, road work, or
public events. Congestion causes inefficient operation
of freeways, wasting of resources, increased pollution,
and intensified driver fatigue.

* This work has been supported by California Partners for Advanced
Transit and Highways (PATH) under Task Orders 4136 and 5503.
1 E-mail: sunx@me.berkeley.edu.
2 Corresponding author. E-mail: horowitz@me.berkeley.edu.

The 2004 Urban Mobility Report (Schrank and Lomax,
2004) finds: “Congestion has grown everywhere in
areas of all sizes. Congestion occurs during longer
portions of the day and delays more travelers and goods
than ever before.” In the report, it was calculated that
in 2002, congestion cost Americans 3.5 billion hours
of delay and 5.7 billion gallons of wasted fuel, with an
equivalent monetary cost of $63.2 billion.

On-ramp metering has been widely used as an effective
strategy to increase freeway operation efficiency. It
has been recommended to the U.S. Federal Highway
Administration as the No. 1 tool to address the con-
genestion problem, other than adding more capacity to
transportation infrastructures (Cambridge Systematics,
Inc. and Texas Transportation Institute, 2004). It has
been reported that ramp metering was able to reduce delay by 101 million person-hours in 2002, approximately 5% of the congestion delay on freeways where ramp-metering is in effect (Schrank and Lomax, 2004).

In this paper, we will first briefly review a set of localized ramp-metering algorithms that we have developed (Sun and Horowitz, 2005), including a switching mainline-traffic responsive ramp-metering controller that employs different feedback structures, depending on whether the freeway is in a free-flow or congested mode; a queue length regulator that keeps the queue below the ramp storage capacity limit; and a localized control strategy to achieve the goal of reducing the spatial and temporal span of the congestion.

In addition, we will design and validate a queue length estimator using the speed data measured by the queue detector. This estimator provides the feedback that is needed by the queue length regulator.

Test results of this set of ramp-metering algorithms, as well as those of ALINEA, on a calibrated microscopic traffic simulator will also be presented.

2. PERFORMANCE MEASURES

In this section, some measures are defined for quantitative evaluation of a given freeway segment. All the quantities are defined for the time period $T$ and the freeway segment $L$.

$D_{V \text{tot}}$ Total Vehicle Distance, which is defined as the sum of the distances traveled by all the vehicles in $L$ within $T$.

$T_{V \text{tot}}$ Total Vehicle Time, which is the sum of the time that is spent by all vehicles in $L$ within $T$. It includes the time spent while vehicles are waiting in the on-ramp queues.

$DL_{V \text{tot}}$ Total Vehicle Delay, which is the difference between the Total Vehicle Time and the time that would be spent by all the vehicles if there were no congestion. $DL_{V \text{tot}} = T_{V \text{tot}} - D_{V \text{tot}}$, where $D_{V \text{tot}}$ is the nominal free-flow speed.

$\bar{v}_{V \text{tot}}$ Average Total Vehicle Speed $\bar{v}_{V \text{tot}} = D_{V \text{tot}}/T_{V \text{tot}}$.

$\bar{v}_{V \text{ml}}$ Average Mainline Vehicle Speed, which is similar to $\bar{v}_{V \text{tot}}$ but calculated using the data from the mainline portion of the freeway only.

Another set of passenger-weighted performance measures can be defined by first collecting the traffic quantities separately for the low- or high-occupancy vehicle classes, and then weighting these quantities by the average passenger number in each vehicle class when calculating the performance measures. This set of passenger-weighted performance measures include Total Passenger Distance $D_{P \text{tot}}$, Total Passenger Time $T_{P \text{tot}}$, Total Passenger Delay $DL_{P \text{tot}}$, Average Total Passenger Speed $\bar{v}_{P \text{tot}}$, and Average Mainline Passenger Speed $\bar{v}_{P \text{ml}}$.

3. TEST SITE

A segment of Interstate 210 Westbound (I-210W) in Pasadena, California, has been selected as the test bed for new ramp-metering algorithms. It is approximately 14 miles long, from Vernon Avenue (Mile Post 39.159) to Fair Oaks Avenue (Mile Post 25.4), with 20 metered on-ramps, 1 uncontrolled freeway-to-freeway connector (I-605), and 18 off-ramps. In our previous effort, microscopic (Gomes et al., 2004) and macroscopic (Muñoz et al., 2004) traffic simulation tools has been calibrated to this test site.

4. THE SWITCHING MAINLINE-TRAFFIC RESPONSIVE RAMP-METERING CONTROLLER, QUEUE LENGTH REGULATOR, AND LOCALIZED CONTROL STRATEGY

In this section, we will briefly review the switching mainline-traffic responsive ramp-metering controller, the queue length regulator and the localized control strategy that we have developed (Sun and Horowitz, 2005).

4.1 The Switching-Mode Model and Ramp-Metering Controller

Based on the observation that traffic dynamics are different under different congestion conditions—free-flow or congested, we have piecewise linearized the cell transmission model (Daganzo, 1994, 1995) and derived a switching-mode traffic model (Muñoz et al., 2003). The properties of this switching-mode model can be summarized as follows:

1. In free-flow mode, traffic moves freely at drivers’ desired speeds without restriction. Therefore, the information travels from the upstream to the downstream, and the downstream vehicle densities are affected by the upstream densities. As a consequence, the vehicle density at one location is observable using a downstream measurement, and it can be controlled by metering an upstream on-ramp.

2. In congested mode, traffic moves slowly and is restricted by the downstream available spaces, which means the information travels from the downstream back to the upstream. Therefore, the upstream densities are affected by the downstream densities. The observability and controllability are opposite to those of free-flow mode: the vehicle density at one location can be observed by an upstream measurement and can be controlled by a downstream ramp.

It is therefore natural to employ different feedback structures for the mainline-traffic responsive ramp-metering controller, depending on different congestion modes, as shown in Fig. 1. The mixture Kalman filter (MKF) based traffic state estimator that we have developed (Sun et al., 2003, 2004) is used to estimate, in real time, the most probable congestion mode and
Fig. 1. Different control structures for different congestion modes.

the cell vehicle densities in a freeway section. The estimated congestion mode is used to determine the appropriate control structure, and the estimated vehicle densities are used as feedback.

To compensate for disturbances and to accommodate the difference between the model sampling time and the metering-rate update interval, a multirate linear quadratic control with integral action (multirate LQR) approach (Sun and Horowitz, 2005) was used to synthesize the ramp-metering controller for both of the congestion modes.

In either mode, the desired metering rate is first calculated using

\[
    r_\epsilon(t) = r(t) - K(t) \rho(t),
\]

and then saturated

\[
    r_\epsilon(t) = \min[\max[r_{\text{max}}, r_{\text{min}}, r_\epsilon(t)]] \tag{2}
\]

where \( r(t) \) is the actual ramp flow measured by the entrance loop-detector, \( \rho(t) \) is the mainline density error, \( z(t) = \rho(t) - \rho(t-1) \), and \( r_{\text{max}} \) and \( r_{\text{min}} \) are the established maximum and minimum metering rates. There is an anti-windup scheme in (1) that is similar to what is used in ALINEA (Papageorgiou et al., 1991) to address the metering-rate saturation problem.

In (1),

\[
    K(t) = \begin{cases} 
    K_p, & \text{when } t = np \text{ for some } n \in \mathbb{Z}, \\
    0, & \text{when } t \neq np \text{ for any } n \in \mathbb{Z},
    \end{cases} \tag{3}
\]

where \( p \) is the ratio between the metering-rate update interval and the model sampling time, and \( K_p \) is determined by solving a periodic Riccati equation. See (Sun and Horowitz, 2005) for detail.

4.2 Queue Length Regulator

A typical configuration of loop detectors and signals on an on-ramp on California freeways is shown in Fig. 2.

Fig. 2. A typical configuration of loop detectors and signals on an on-ramp.

To prevent the on-ramp queue from spilling over into surface streets and interfering with the street traffic, the queue length must be regulated. The “queue-override” scheme currently used on California freeways steadily increases the metering rate (e.g., 120 vehicles per hour per lane every 30 seconds) whenever the end of the queue reaches the queue detector, until the metering rate saturates to the maximum value. After the queue dissipates and becomes shorter than where the queue detector is, the metering rate is reset to the value determined by the mainline-traffic responsive metering controller. This scheme is equivalent to an integral control with a saturated integrating rate and resetting, which can be easily shown that is not asymptotically stable, given that the queue length dynamics is a simple integrator. It has been noted (Gordon, 1996; Smaragidis and Papageorgiou, 2004) that this queue-override scheme leads to oscillatory behavior and under-utilization of on-ramp storage capacities. Gordon (1996) attempted to improve the performance of the queue-override scheme by filtering the occupancy signal and reducing the sampling time interval. Smaragidis and Papageorgiou (2004) proposed a proportional controller that relies on the on-ramp vehicle demands, of which the measurements are not available in the field.

If the queue length \( \tilde{h}(t) \) could be measured, an asymptotically stable PI-controller

\[
    r_\epsilon(z) = \left( k_p + k_1 \right) \tilde{h}(z) \tag{4}
\]

would be able to regulate the queue length precisely at a specified value. This controller can be designed by choosing proper gains \( k_p \) and \( k_1 \), using the root-locus method on the closed-loop sensitivity function from the disturbance to the error that is given by

\[
    \frac{\tilde{h}(z)}{d(z)} = \frac{T_s(z - 1)}{(z - 1)^2 - k_p T_s z + (k_l - k_p) T_s},
\]

where \( \tilde{h}(t) \) is the queue length error, and \( d(t) \) is the vehicle arriving rate (the demand), which is regarded as a disturbance.

The anti-windup and saturating mechanisms in (1) and (2) need to be implemented in this queue length regulator too.

4.3 The Localized Control Strategy

Localized strategies are desired for reasons including reduced algorithmic complexity, lower computational requirements, and higher robustness to changing traffic conditions such as unpredicted demands. We have
Fig. 3. A schematics for on-ramp queue length estimation.

proposed a localized metering strategy and tested it on a calibrated macroscopic traffic model (Sun and Horowitz, 2005). It is described as follows:

1. The set-point for the switching mainline-traffic responsive ramp-metering controller is chosen to be the critical density, i.e., the density at which congestion is about to form. This is adopted to slow down congestion shock-wave propagating toward the upstream and to speed up the congestion shock-wave moving downstream.

2. The set-point of the queue length regulator is the maximum allowed queue length. This value is chosen to fully utilize the available storage capacity on the ramp and to deter short-trip travelers from using the freeway, saving the freeway capacity for longer-distance travelers.

3. The higher of the two rates determined by the mainline-traffic responsive metering controller and the queue length regulator is chosen to be the actual metering rate that is sent to the signal control box. This rule is designed to properly resolve the conflict between the objectives of these two controllers. Smaragdis and Papageorgiou (2004) proposed this same formula for this purpose.

5. AN ON-RAMP QUEUE LENGTH ESTIMATOR

Though it has a more stable response than the queue-override scheme, the PI regulator described in Section 4.2 needs the current queue length as its feedback, which unfortunately is not available in the field. A suitable estimator has to be designed using available information, such as the vehicle speed measured by the queue detector.

We assume the following simplified driving behavior model for a vehicle approaching the end of the queue: The vehicle decelerates at a constant rate, $-a$, from its cruising speed to a target speed $v_0$ at the position where the distance from the end of the queue is $s$. We also assume a uniform effective vehicle length $g$. Let $l_0$ be the number of vehicle spaces from the stop line to the queue detector and $v(t)$ be the vehicle speed measured by the queue detector. See Fig. 3.

A straightforward kinetic calculation yields

$$ g(l_0 - l(t)) - s = \frac{v(t)^2 - v_0^2}{2a}, \tag{6} $$

where $l(t)$ is the current queue length, in number of vehicles. From (6), we obtain

$$ g(t) = g_0 - s + \frac{v_0^2}{2a} - \frac{v(t)^2}{2a} = c_0 - c_2v(t)^2. \tag{7} $$

To determine the coefficients $c_0$ and $c_2$ in (7), a curve fitting was performed on the $g(t)$ and $v(t)$ data collected using the VISSIM (PTV AG, 2004) microscopic traffic simulator. Fig. 4 shows a typical scatter plot of the queue lengths versus the speeds. A few points need to be noted:

1. When the queue is shorter than a certain length, the approaching vehicles pass the queue detector at the drivers’ desired cruising speeds, which are independent to the queue length. This phenomenon corresponds to the data points at the lower-right corner of the scatter plot.

2. When the queue is longer than $g_0$, i.e., the queue has extended beyond the queue detector, the measured speed is also a constant, which is related to the queue discharging rate and the vehicle lengths, and is also independent to the queue length. This phenomenon corresponds to the data points at the upper-left corner of the scatter plot.

3. There are many outliers among the data points. Therefore, the usual least-squares curve fitting method, which is biased toward outliers, is not suitable.

For these reasons, we neglected the data points whose speeds are below $v_{\min}$ or above $v_{\max}$ and those whose queue lengths are below $l_{\min}$ or above $l_{\max}$ in the curve fitting. These values were determined by visual inspection of the scatter plots.

To be more robust to the outliers, we used the least median-of-squares (Rousseeuw, 1984) curve fitting method, instead of the usual least (sum-of-)squares. The fitted curve is also shown in Fig. 4.

After the $l-v$ curve is fitted for each on-ramp, the difference between the actual and desired queue length, which is used as the feedback to the regulator (4), is estimated as

$$ l(t) = \begin{cases} 
(c_0 - g_0 - c_2v(t)^2)/g, & \text{if } v(t) \geq v_{\min}, \\
-kc_2(v(t)^2 - v_{\min}^2)/g, & \text{if } v(t) < v_{\min}. 
\end{cases} \tag{8} $$

Fig. 4. A scatter plot of the queue lengths vs. the queue detector speeds and the least median-of-squares curve fitting for one of the on-ramps.
where $k$ is a tuning parameter.

When $v(t) < v_{\text{min}}$, the end of the queue is very close to or beyond the queue detector, and the measured speed $v(t)$ by the queue detector is a constant, which is roughly $gr_e$. Therefore, (8) can be thought of as saturating $l$ to $-kc_2(t^2 - v_{\text{min}}^2)/g$, which is larger when the metering rate $r_e$ is lower. This has a desirable effect on the regulator: The metering rate $r_e$ will be increased more aggressively when there is more room for this increase, and more slowly when $r_e$ is close to its maximum value. In addition, this saturation value can be further tuned by changing the value of $k$.

It also worths to point out that the coefficients $c_0$ and $c_3$ identified by the least median-of-squares fitting are very close to the nominal values predicted by using the actual distance between the stop-line and the queue detector and a nominal vehicle deceleration of $2.5$ m/s$^2$. Therefore, when the queue length measurements are unavailable through any means to perform a curve fitting, these nominal values can be used in the queue length estimation.

### 6. RESULTS

The switching mainline-traffic responsive metering controller and the queue length regulator were implemented and interfaced with the VISSIM microscopic traffic simulation model that has been calibrated to the I-210W test segment by Gomes et al. (2004). The localized control strategy described in Section 4.3 was used. Fig. 5 shows the congestion patterns, as determined by the MKF traffic state estimator (Sun et al., 2003, 2004), before and after ramp metering. In the plots, red indicates congested mode and blue free-flow. The vertical axis is the time, from 5:30 to 11:00 in the morning. The horizontal axis is the mile post along the freeway, and the traffic travels from the left to the right. It can be seen that the localized ramp metering strategy was able to reduce the congestion, in terms of both the spatial span and the time duration.

We also implemented ALINEA (Papageorgiou et al., 1991) and combined it with the queue length estimator and regulator that we have developed in Sections 5 and 4.2. Different ramp-metering algorithms, including 1) switching LQI plus queue regulation, 2) switching LQI only, without queue regulation, and 3) ALINEA plus queue regulation, were tested with the VISSIM I-210W model. Under each scenario, 8 simulation runs were carried out, with 8 different VISSIM random seeds. To ensure the randomness of these random seeds, they were chosen to be the second of the computer clock when the random seed was changed.

Some of the performance measures for this freeway segment, as defined in Section 2, are listed in Table 1. The listed numbers are the averages from the 8 simulation runs for each scenario. In calculating these quantities, the average number of passengers per one low- and high-occupancy vehicle is assumed to be 1.2 and 2.5, respectively, and the nominal free-flow speed $v_0$ is $63$ miles per hour.

Under all the scenarios, the freeway segment served almost the same amount of demand, as measured by

![Fig. 5. Congestion modes for the I-210W test segment under different metering scenarios, blue: free-flow; and red: congested.](image-url)
the Total Vehicle Distance $D_{V,\text{tot}}$ or Total Passenger Distance $D_{P,\text{tot}}$. Ramp-metering was able to reduce the congestion under all the metered scenarios. For example, with the switching LQI mainline control and queue length regulation, the Total Vehicle Delay $D_{V,\text{tot}}$ was reduced by 16%, while with the switching LQI mainline control only, $D_{V,\text{tot}}$ was reduced by 20%.

When only the switching mainline-traffic responsive metering was used, without activating the queue length regulator, on-ramp queues can be accumulated to arbitrary lengths, sometimes hundreds of vehicles. In this case, almost all the congestion on the mainline was eliminated, as evidenced by the average mainline vehicle speed $v_{V,\text{ml}}$, which is 55.8 mph. Another interesting phenomenon in this case is that the relative improvements in terms of passenger-weighted performance measures are greater than those in terms of vehicle performance measures. This is because many of the metered on-ramps on this freeway segment have designated lanes for HOVs to bypass the long queues.

It can also be seen from the numbers in Table 1 that the switching control algorithm outperforms ALINEA, when both algorithms are combined with the queue length estimator and regulator.

7. CONCLUSIONS

In this paper, we first reviewed a localized ramp-metering strategy that achieves the control goal of reducing the spatial and temporal extent of the congestion using the locally available information. This control strategy works with a switching mainline-traffic responsive ramp-metering controller that adapts to the different traffic dynamics under different congestion conditions, and a PI queue length regulator that improves the performance of the currently used “queue-override” scheme and keeps the queue under the ramp storage capacity limit. In addition, a queue length estimator was designed to provide feedback to the queue length regulator, using the field-available queue-detector speed data.

Test results of this set of ramp-metering algorithms on the calibrated VISSIM I-210W microscopic model demonstrated the performance and effectiveness of the switching ramp-metering controller, the queue length estimator and regulator, and the control strategy. The Total Vehicle and Passenger Delays were both reduced by 16%, while the Total Vehicle Time and the Total Average Vehicle Speed were improved by 5.6% and 5.8%. As a comparison, simulation results of ALINEA were also presented. The switching mainline-traffic responsive control was able to outperform ALINEA, when both algorithms were combined with the same queue length estimator and regulator.

REFERENCES


