Improving City Mobility through Gridlock Control: an Approach and Some Ideas

Carlos F. Daganzo

WORKING PAPER
UCB-ITS-VWP-2005-1

UC Berkeley Center for Future Urban Transport
A VOLVO Center of Excellence

July 2005

IMPROVING CITY MOBILITY THROUGH GRIDLOCK CONTROL: AN APPROACH AND SOME IDEAS

Carlos F. Daganzo
Berkeley Center of Excellence on Future Urban Transport
Working Paper
Institute of Transportation Studies
University of California, Berkeley, CA 94720

July 5, 2005

ABSTRACT

This paper examines the effect of gridlock on urban mobility. It defines gridlock and shows how it can be modeled, monitored and controlled with parsimonious models that do not rely on detailed forecasts. The proposed approach to gridlock management should be most effective when based on real-time observation of relevant spatially aggregated measures of traffic performance. This is discussed in detail. The ideas in this paper suggest numerous avenues for research at the empirical and theoretical levels. An appendix summarizes some of these.

1. INTRODUCTION

The current paradigm for the development and evaluation of transportation policies in cities all over the world relies heavily on forecasting models. Government agencies often stipulate by law the outputs that evaluation models must produce before a policy can be rolled out—even the kind of model in some countries (Muñoz, 2004). But the objective of these laws may not be achieved if the models and data used to produce the outputs are unreliable. Unfortunately, unreliability is the order of the day for the reasons explained below.

The level of detail and complexity of available urban transportation models have steadily increased over decades: from the static and largely aggregate "four-step" models of the 1950 and 60's; to the "disaggregate demand" and "network equilibrium" extensions of the 1970's and 80's; and now the "multi-modal" and "dynamic" models of the 1990's and 2000's. Given correct inputs, the most recent computer models have the capacity to predict almost anything on a multi-modal transportation network in minute detail. But for this to be a reality, some inputs must be first obtained. They include: (i) highly disaggregated, and time-dependent, origin-destination data; and if the model is traffic responsive (ii) a psychological model of driver information acquisition and reaction to existing and anticipated route congestion conditions.

Input-set (i) is customarily constructed with econometric models that use data from a variety of sources (census; home interviews, etc.) Unfortunately, the estimation problem is formidable for the level of resolution required by a detailed model. A model with reasonable spatio-temporal resolution should have no more than 10^4 people per zone, and a time slice many times smaller than the average trip time. According to these criteria a medium-size metropolitan area with 1 million people should be modeled with 100 zones and a 5-minute time slice—small compared with a 30-minute commute. The study of a two-hour period including the rush would then require an origin-destination table with 240,000 entries; a large number. For a large metropolitan area with 10 million people, the same considerations show that the number would be 24 million;[1] *more than the city's population*! Estimating all these numerical values is obviously problematic.

The difficulty is compounded by requirement (ii) because people do not just decide at one point in time when, how, where and whether to travel, but they re-evaluate their route choice (how to travel) in real-time, and may change routes as conditions change. Nobody really knows how they make these continuing decisions, but the possibilities are troublesome. As pointed out in Heydecker and Addison (1996), to minimize travel time, "rational" drivers would try to anticipate the congestion level along their possible paths at the relevant (future) times when they would traverse them. But to predict these conditions, they would need to know the decisions of "rational" drivers from other origins (e.g., traveling in the opposite direction) who may not have yet left their origins but could arrive at the location in question before them. Conversely, these new drivers may face the same conundrum, in reverse, trying to guess the decisions of drivers from the first origin. In essence, drivers from the two origins would be engaged in an unpredictable game of "poker."

It should be clear from the above that, although sophisticated models for mechanical prediction exist, reliable data to support them cannot be obtained. Yet, current practice for policy development and evaluation continues the tradition of strong reliance on model predictions. We can continue the tradition, hoping that model predictions match reality in the aggregate despite the imperfection of the model inputs, but we should then candidly admit that our policy evaluations are hope-based. Unfortunately, this hope is not soundly based because highly congested networks can exhibit chaotic behavior—where slight changes to the input O-D table, or perturbations to drivers' route choices, can drastically change the aggregate outputs; see Daganzo (1998) for a discussion. Thus, model predictions for large systems cannot always be trusted. This suggests that the model-based forecasting approach for mobility assessment should be complemented with more reliable approaches.

[1] Our estimate assumes that the number of time slices remains the same. This is reasonable. A study period including the rush would now have to be longer, but we could also use wider time slices since the typical commute should also be longer.

Fortunately, one such approach is close at hand. Cities can benefit significantly if we succeed in understanding and applying this approach. The idea is that implementation priority should go to (robust) policies, whose benefits and disbenefits can be measured directly in the field, without questionable data inputs. The advent of new technologies for sensing, vehicle-tracking and web-based data processing expands the kinds of things that can be measured. Improved measurements in turn expand the feasible set of (robust) policies that can be accurately tested.

As a first exploratory step of this alternative paradigm, we show below that two key determinants of city mobility are the aggregate vehicular accumulations and cumulative flows by district and time-of-day. These indicators are ideal policy beacons because they can be measured directly (without modeling) if sufficient sensors are deployed, and because they correlate extremely well with measures of interest to the public such as the aggregate number of vehicle-hrs (VHT), vehicle-km (VMT), emissions or noise.[2] We also show that parsimonious models can be constructed to predict how some (robust) policies affect these beacons. But this is less important. If a city is properly instrumented, the effects of the (robust) policies can be measured by the detection system immediately after implementation. The policies can then be fine-tuned over a period of weeks with real data feedback, to achieve real benefits. In essence, the instruments constantly take a city's pulse and can replace the model.

Section 2 below discusses what our two determinants reveal, and how they relate to each other. Section 3 describes the physics of gridlock in terms of these determinants, and how to control it in an idealized scenario. Section 4 shows how the gridlock ideas can be applied on a city-wide scale, and proposes policies to improve mobility. An appendix discusses qualitatively other research ideas related to the new paradigm.

[2] The indicators also yield important people-based measures such as the people-hrs and people-km of travel—since the contribution of automobile trips toward the measures can be obtained by factoring into our indicators passenger occupancy data, and the contribution of transit trips is independently available from ridership and vehicle timetable data.

2. ACUMULATION (VHT) AND FLOW-SUMS (VMT)

Consider a city and let A be a set of *directed* links, i, describing its street network; $i \in A$. The city may be partitioned into sub-regions, r, with network links A^r, where every link belongs to only to one sub-region. Define $n_i(t)$ as the number of vehicles traveling on link i at time t. (This excludes parked vehicles – with no occupants and engines off.) Also define $A_i(t)$ and $L_i(t)$, respectively, as the cumulative number of vehicles to have *arrived* and *left* link i.by time t; and initialized in such way that $n_i(t) = A_i(t) - L_i(t)$.

Link arrivals and departures are either exogenous (from/to other links) or endogenous (from/to the origins/destinations in the link). The endogenous portions of A_i and L_i will be respectively denoted $O_i(t)$ and $E_i(t)$, representing the trips *originated* and *ended* within i. The exogenous portions will be denoted $U_i(t)$ and $D_i(t)$, respectively representing *upstream* arrivals and *downstream* departures. By definition, $A_i(t) = O_i(t) + U_i(t)$, and $L_i(t) = E_i(t) + D_i(t)$.

Before examining our two determinants, note that the total number of trips (TT) starting and ending in a region r is: $O^r(t) =\sum_i O_i(t)$ and $E^r(t) = \sum_i E_i(t)$, where the sums are evaluated for $i \in A^r$. We claim that $O^r(t)/E^r(t) \cong 1$, regardless of when we start counting, for neighborhoods r with many people if t = 1 week. The basis for this assertion is that most people repeat their routines on a weekly basis, thus, the collective travel pattern of a large group should have a weekly cycle. Or putting it another way, if we were to inspect the number of vehicles in a neighborhood r on a weekly interval for a single season, e.g., each Tuesday at 3:00 AM during winter, we would likely find little variation in these numbers—and this of course implies balance between the number of trips originated and ended in r per week. The approximation $O^r(t)/E^r(t) \cong 1$ may also be reasonable for the 24 hour cycle and for shorter time periods during the middle of the day. We will soon find that if this approximation holds the formulae relating our determinants to VHT and VMT simplify considerably.

Aggregate accumulations and VHT: The total VHT in link i during a short time interval (dt) when nobody enters or exits the link is simply $n_i(t)dt$. If we partition a long time interval into short time slices with this property, the total number of vehicle-hrs is the sum of the VHT for each slice, $VHT_i = \sum_t n_i(t)dt$, where the sum is evaluated over the relevant slices. In the limit of vanishing dt this is the integral: $VHT_i = \int_t n_i(t)dt$. The total VHT in a sub-region of the city VHT^r is the sum of the VHT_i over the links in the region. In practice, an approximation for VHT^r can be obtained by sampling $n_i(t)$ every Δt time units and evaluating $\sum_t \sum_i n_i(t)\Delta t$ for $i \in A^r$. Estimates can be substituted for the actual $n_i(t)$'s.

For links without endogenous flows such as freeway segments between ramps the $n_i(t)$'s can be obtained from detectors that continually measure U_i and D_i , since in this case $n_i(t) = U_i(t) - D_i(t)$. (Methods that compensate for detector measurement errors have been developed.)

City streets, however, are a different matter. Street links are rarely instrumented with detectors to measure their exogenous flows, let alone their endogenous flows. No reliable method seems to exist for determining accumulation on city streets. But if as we argue below accumulation is not just an informative measure of performance but also an important (and controllable) driver of congestion, this needs to be rectified.

<u>Flow-sums and VMT</u>: Let us now see how the number of vehicle-miles (VMT) in a link i of length l_i is related to the endogenous and exogenous curves of cumulative counts. Assume for now that the link is empty at both ends of a time interval of interest (0, t); hence, $A_i(t) = L_i(t)$. We neglect (reasonably) the number of trips that both begin and end in the link, without crossing either of its ends. Then, the total number of endogenous link visits (i.e., with at least one end rooted in the link) by time t is $O_i(t) + E_i(t)$. Since the total number of link visits is $A_i(t) \equiv O_i(t) + U_i(t) = L_i(t) \equiv E_i(t) + D_i(t)$, the number of through visits is: $U_i(t) - E_i(t) = D_i(t) - O_i(t)$

Each through visit contributes l_i distance units to VMT. If the average distance contribution of each endogenous link visit is $l_i/2$ (a reasonable assumption), the VMT for link i is then: $l_i(U_i(t) + ½[O_i(t) - E_i(t)]) \equiv l_i(D_i(t) + ½[E_i(t)- O_i(t)])$. For major arterials and collector streets, the number of endogenous link visits should be much smaller than the number of exogenous visits. In this case VMT is roughly given by either $l_iD_i(t)$ or

$l_iU_i(t)$. But this approximation is not good for local streets—certainly not for cul-de-sacs, and may or may not hold for a neighborhood including many local streets.

To estimate VMT for a region r we have to add $l_i(U_i(t) + ½[O_i(t) - E_i(t)])$ for all i in the region with the result: $VMT^r = \sum_i l_iU_i(t) + \sum_i ½\, l_iO_i(t) - \sum_i ½\, l_iE_i(t)$. If all links have unit length the above reduces to: $VMT^r = \sum_i U_i(t) + ½\sum_i O_i(t) - ½\sum_i E_i(t)$. This assumption is not outlandish; it applies if the city blocks are uniform in size, and also if we imagine artificial nodes uniformly spaced every unit distance along all the streets. (We can choose the unit of distance to be as big or small as convenient.) With this convention, the first term of VMT^r is the sum of all the cumulative flows at the sampled locations; i.e., the "flow-sum." The second term is 50% of the trips generated in the region by time t, and the last term 50% of the trips attracted. We argued earlier, however, that the net trip generation $O^r(t) - E^r(t)$ could be neglected for neighborhoods of sufficient size when the period of observation was either 1 day or 1 week.[3] Thus, the flow-sum, $\sum_i U_i(t)$, is an accurate estimate of VMT. This should also be the case for shorter periods of time if the city region includes freeway portions, arterials and spans many blocks, e.g., with a diameter just a few times smaller than a typical trip, because in these cases, the flow-sum should be large compared with $\sum_i O_i(t)$ and $\sum_i E_i(t)$, even if $\sum_i O_i(t) \neq \sum_i E_i(t)$. Thus, VMT in large regions can be estimated with flow sums, without following vehicles around.

Interpretation: Flow-sums at evenly spaced locations give an indication of VMT, which in turn is an indicator of system productivity. Accumulation-sums, on the other hand, give VHT. The ratio of VMT and VHT is an indicator of system performance. We call it the "generalized" average speed or "space-mean" speed for the network and time period under consideration, in agreement with Edie's definition of space-mean speed for single links (Edie, 1963). The excess vehicle-hours over the "nominal" time that would be consumed by a given VMT under smooth flowing conditions (e.g., at night time) are the VHT of delay. VHT delay is a waste to society.

[3] This is useful because endogenous trips are difficult to observe.

These aggregate measures of travel, and the total number of trips started (TT = $\sum_i O_i(t) \cong \sum_i E_i(t)$), are important because they also capture other resources wasted to society such as energy consumed and emissions generated. We know for example (Evans et al, 1976) that the total energy consumed by vehicles in a city can be expressed as a linear combination of cold starts (TT), VMT and VHT.

The definitions in this section can be extended to the movement of any objects that are conserved. Thus, truck and bus trips, miles and hours; and even people trips, miles and hours, can be treated in the same way. We now show how these ideas can be used.

3. THE PHYSICS OF GRIDLOCK

3.1 Definition

We discuss here urban gridlock and policies to manage it. The term "gridlock" has a specific meaning in this paper. We apply it to "input/output" systems that can be modeled as reservoirs, or sets of interconnected reservoirs, where specific items flow into the system, spend some time in it and then flow out. There are many examples: checkout stands at supermarkets, water reservoirs, baggage carousels at airports, highway roundabouts, your desktop, restaurants, the Internet and, of course, a city street network. In some cases, such as supermarket checkout stands, the output rate is relatively independent of the number of items in the system. In most cases, though, including all the remaining examples in the list, there is some dependence between outflow and accumulation. For example, the outflow from a water reservoir with an open spillway increases monotonically with accumulation. In all other examples, though, the output rate can decline with accumulation. The reason is that in these cases items must complete a task before leaving, and crowded conditions reduce the efficiency with which the tasks can be completed. Severe inefficiencies obviously reduce the system's output rate.

To recognize this effect more formally, we say that a single-reservoir system is subject to gridlock for a range of accumulations if *its steady-state outflow*, g = dL(t)/dt, *declines with the steady-state accumulation*, n, *when* n *is within the range*. (We imagine that the accumulation level is regulated by controlling the inflow, f.) The definition can be extended to sets of reservoirs. In this case, n is a (regulated) vector, and we say that gridlock exists if the combined steady-state system outflow declines with some component of n.

When conditions change with time, the outflow g(t) could theoretically depend on the history of inflows {f(t)} and accumulations {n(t)}, and not just on n(t), but in many cases the dependence on history is weak. We can then write

$$dL(t)/dt \equiv g(t) \cong G(n(t)), \tag{1}$$

where G is a function of the accumulation alone. In some application contexts we may choose to track the accumulation of multiple item types; e.g., passengers and bags at carousels, or cars on different links on road networks. The accumulation n is then treated as a vector. The existence and character of G has been examined for baggage carousels (Ghobrial et. al, 1982) and closed-loop roads (Daganzo 1995). In both cases the function G is unimodal, increasing from the origin to a maximum and then declining toward zero. (We call the value at which it reaches zero the "jam accumulation," the maximum outflow the "capacity.") We now examine the gridlock mechanism for unimodal G, and then show how to control it.

3.2 Mechanism and control

Assume that prior to t = 0 the system is in a steady state with $A(t) = n_o + f_o t$, $L(t) = g_o t$, and $f_o = g_o = G(n_o)$. Consider now the evolution of n(t) from t = 0 onward when the arrival rate, f := dA/dt , is allowed to vary in a narrow range around f_o; i.e., $f(t) \cong f_o = G(n_o)$. Since n = A – L, we can write dn/dt = f – g. But g = G(n). Thus, dn/dt = f(t) – G(n), which equals 0 for t < 0 but not thereafter. Thus, n(t) will in general differ from n(0) for t > 0. If n(0) is in the declining range of G and n(t) > n(0) for some t, then $dn(t)/dt = f(t) - G(n) \cong G(n(0)) - G(n) > 0$. The last inequality holds because G declines for $n \geq n(0)$. This sets in motion a "positive feedback" cycle where accumulation increases at a rate that itself grows with accumulation. (Growth is exponential when G is linear.) When the system reaches its jam accumulation it ceases to emit flow. Accumulation could continue to grow, but at some point would prevent further inflows.[4] The system would then have reached a static state of *gridlock*.

Similar considerations show that the vicious cycle leading to gridlock becomes "virtuous" if we start with $n(t) < n(0)$. In this case, the self-sustaining mechanism involves outflow increases and accumulation reductions until capacity is reached. Consideration also shows that vicious and virtuous cycles do not arise if n_o is on the rising branch of G. The system is then stable.

We see from the above that if a system is in an unstable equilibrium state (i.e., on the declining portion of G and hence susceptible to gridlock) then minor perturbations on opposite sides of its equilibrium input flows will have vastly different effects. This strongly suggests that systems susceptible to gridlock can be improved significantly by gently controlling their inputs. In particular, as the previous footnote suggests, they can be improved by releasing inputs at a "metered" rate, holding items outside the system if necessary. This can be easily understood with a queuing diagram; see Fig. 1a. We assume that A(t) and n(0) are given and that L is given by (1). The function G is as we stipulated earlier: unimodal with maximum outflow g_m, declining between n_m and the jam accumulation n_g, and equal to 0 for $n \geq n_g$; see Fig. 1b. Note from Fig. 1a how the slope of L becomes shallower as n increases, and eventually flattens when n reaches the jam value. The metering approach uses a curve A* of released arrivals into the system such that $A^*(0) = A(0)$ and $A^*(t) \leq A(t)$ for all t. Items between curves A and A* would be held outside the system. The idea is to generate a "higher" departure curve, L*, such that $L^*(t) \geq L(t)$ for all t. This is would be good since the items between L* and L at any time

[4] Input restrictions have been observed in empirical data of baggage carousels (Ghobrial et al, 1982); for these systems, f(n) is bounded from above by a function of n, F(n), which declines toward zero but exceeds G(n) for large n. For traffic links obeying the kinematic wave theory of Lighthill and Whitham (1955) and Richards (1956) a bound of the form $F(n) \cong G(n)$ can be justified when n is on the declining branch of the fundamental diagram (FD). The results in Newell (1993) justify its use for triangular FDs, and those in Daganzo (1993) for general FDs under slow-varying conditions.

would have been served with the new control scenario but not the old. The shaded area between L* and L would indicate the total number of item-hours saved.

Figure 1a illustrates the following "bang-bang" strategy: Choose dA* = 0 if $n > n_m$; and otherwise, set dA* as large as possible without exceeding the optimum accumulation or curve A. The idea is to avoid accumulation growth when $n > n_m$. The recursion for the points of A* is then:

$$\begin{array}{rcll} A^*(t+dt) & = & A^*(t) & \text{if } n(t) > n_m \quad (2a) \\ & = & L^*(t) + n(t) & \text{if } n(t) = n_m \quad (2b) \\ & = & A(t) & \text{if } n(t) < n_m. \quad (2c) \end{array}$$

Curve L* continues to be given by (1); i.e., $L^*(t+dt) = L^*(t) + G(n)dt$.

Note from the dashed curves how A* = A(0) until the optimum accumulation is reached, as per (2a). From this point until the outside queue vanishes A* and L* increase at the maximum rate, g_m, with $A^* - L^* = n_m$, as per (2b). In the final dissipation regime, $n < n_m$ and A = A*, as per (2c). The benefit of the bang-bang strategy is very large for this example, since the area between L and L* is unbounded.

Although, benefits could be smaller in other cases, it is possible to show—but we don't do it here—that the bang-bang strategy A* is uniformly optimal in the following sense. Let A' be another strategy and L' its exit curve according to (1). Then, we have:

Theorem 1. *If $A' \neq A^*$, then $L'(t) \leq L^*(t)$ for all t.* □

Theorem 1 shows that in a FIFO system the proposed policy gives *every* item the most advanced departure time possible. Thus, it distributes benefits widely. The proposed policy is also powerful because, as we can see from (1) and (2), it does not rely on forecasts; only on current values of A*(t), A(t) and n(t), which can be readily measured. Thus, it is *robust*.

Performance with spillovers: In practical applications reservoirs have finite capacity, and *spillbacks* can prevent flows from entering the system. If spillbacks are a possibility, we should make sure that the proposed control strategy accounts for this feature. We saw in footnote 4 that the inflow f(t) into a single traffic link (without endogenous flow) is approximately bounded by $F(n) \cong G(n)$ if $n > n_m$. But the constraint $\{f(t) \leq G(n)$ if $n > n_m\}$ never comes into play with our bang-bang strategy, since rule (2a) automatically sets $f(t) \equiv dA/dt = 0$ for $n > n_m$. Thus, the bang-bang policy automatically avoids spillovers when applied to a single link. Since the policy solves optimally the (less constrained) problem without spillovers, it must be optimal too if spillovers are allowed. Obviously, this is also true for any $F(n) \geq 0$ that does not restrict inflows for $n \leq n_m$. Actually, it can be more generally shown that the bang-bang policy is optimal for any $F(n) \geq 0$—although its restrictions for $n \leq n_m$ may generate spillovers. Thus, the bang-bang policy is indeed quite robust.

4. URBAN MOBILITY AND GRIDLOCK CONTROL

4.1 Complications for cities

We are interested in applying the ideas of Sec. 3 to cities by decomposing them into districts that can be modeled as interconnected reservoirs obeying the physics underlying Theorem 1. Information from the real world can then be used to control the flows in and across reservoirs to improve mobility. Control can be exercised with street closures, pricing, signing, metering, signal timing and suitable combinations of these measures.

The idea is not so far fetched. The Swiss city of Zurich already does something along the lines of Theorem 1 (Cervero, 2004). The average speed of public buses in Zurich's central area is used to regulate how much traffic is released into it. Ostensibly, this system ensures that the performance of large occupancy public vehicles is not adversely affected by interference from automobiles, but Theorem 1 also suggests that the restrictions could benefit automobile users. To assess this accurately would require information that is not being collected. The city of London has also implemented a congestion-pricing scheme for its central area, but as in the case of Zurich its policies have been implemented with incomplete feedback. More complete knowledge of $O^r(t)$,

$E^r(t)$ and $n^r(t)$ could open the door for policies that could target more precisely specific problems in time and space.

These zonal-level data would also allow us to verify our theories. But more importantly, even if the theories fail, real-time zonal data would allow policy developers to test their ideas immediately, accurately and with a fair amount of detail.

Problems with multiple reservoirs can sometimes be treated on an aggregate level as if they consisted of a single reservoir. As we show below, the spillover constraint may need modification but the optimization framework remains the same.

4.2. Rings, symmetry and aggregation

Consider a closed-loop freeway system of length C with a symmetric O-D table and closely spaced ramps. Let the distribution of trip lengths be exponential with mean $c << C$. We divide the freeway into sections from on-ramp to on-ramp. Each section can be viewed as a reservoir that is fed by its upstream on-ramp and upstream neighboring section. The reservoir discharges into off-ramps and a downstream section. Because the system is symmetric we can study it as a single reservoir, recognizing that the inter-reservoir transfers cancel out by symmetry.

The total demand along the perimeter of the loop is assumed to be $A(t) = MH(t)$, where M is the user population size (assumed to be very large) and H(t) is the Heaviside unit step function. We also assume that if the loop contains $n < n_g$ vehicles uniformly distributed along its perimeter, then the speed of traffic along the loop is approximately uniform: v(n). This function declines with n and is such that $v(n_g) = 0$. The VMT per unit time is therefore, v(n)n. Since trip distances are exponential, the exit rate per unit time is: $G(n) := nv(n)/c$. Experiments show that v(n) is relatively close to a constant v_f (called the free-flow speed) if $n \leq n_m$. Thus, $G(n) = nv_f/c$ if $n < n_m$. Experiments also show that G(n) is roughly linear for $n \geq n_m$ and approximately triangular overall; i.e. $G(n) = nv_f/c$ if $n < n_m$, and $G(n) = (n_g - n)/\beta$ otherwise. (The constant β is proportional to c.) The merge model in Daganzo (1996) indicates that the rate at which the system can admit new vehicles is bounded by $F(n) = (n_g - n)/\alpha$ if $n > n_m$, where α is a constant that increases with the number of lanes.

Assume that the system includes n_m vehicles at $t = 0$. The freeway is so narrow and trips so long that $\beta > \alpha$. Thus, $F(n) > G(n)$. Since the demand is heavy, more vehicles would enter than exit the system if left uncontrolled. In this case, $dn/dt = \gamma(n_g - n)$, where $\gamma = 1/\alpha - 1/\beta$. The solution of this ordinary differential equation for n(t) is: $(n_g - n)/(n_g - n_m) = \exp(-\gamma t)$. The cumulative number of entrances is the integral of $F(n) = (n_g - n)/\alpha = [(n_g - n_m)/\alpha] \exp(-\gamma t)$ from 0 to t. For $t \to \infty$ the result is: $[(n_g - n_m)/\alpha\gamma]$. Thus, there is an upper bound to the number of cars that the system can serve. If the population of users is so large that $M > [(n_g - n_m)/\alpha\gamma]$, the queue would never dissipate.

On the other hand, if the system is controlled with strategy (2), the inflow would equal $A^*(t) = L(t) = (n_g - n)/\beta$ as long as $n = n_m$. This state of affairs would persist until the last person enters the system; i.e. until $t = M\beta/(n_g - n_m) < \infty$. Therefore, with strategy (2) the queue would dissipate even if M is arbitrarily large.

Simulations (and analytical considerations) show that this result is qualitatively valid even if the symmetry assumptions of the example are only approximately true. But, cases with strong asymmetry exhibit complex features that can be exploited and should be studied as such. This is done next.

4.3 Time-independent behavior of asymmetric rings

We now examine the performance of a closed loop with an asymmetric O-D travel demand pattern. We assume that the demand arises from an "intervening opportunities model" with a homogeneous driver population. Drivers look for "opportunities" at the various exits, and take the first exit that satisfies their need. Opportunities are distributed along the road with density $\lambda(x)$. All opportunities are equally likely to satisfy a driver's need (with probability $p << 1$) independently of where s/he comes from. We assume that that the total number of opportunities, N, is so large that $Np >> 1$.

If we denote by N(x) the cumulative number of opportunities along the road starting from a reference point x = 0 (i.e., N(x) is the definite integral of $\lambda(\cdot)$ from 0 to x) then the probability that a driver finds its opportunity in the interval (x, y) given that it did not find it before x is a function of the intervening opportunities, N(y) −N(x): $p(x, y) = 1-(1-p)^{[N(y)-N(x)]} = 1-\exp(-p[N(y)-N(x)])$. If x > y, i.e. the reference point is between x and y, then the intervening opportunities are N(y)−N(x)+N instead of N(y)−N(x) but the formula is the same. If (x, y) is the segment containing opportunities for exit j , p(x, y) will approximate the fraction of vehicles taking the exit, p_j. Given our assumptions, this probability is independent of the driver's origin and can be viewed as a property of the road.

Let k(x) be the density at x, Q(k(x), x) be the flow at x as predicted by the kinematic wave (KW) theory of Lighthill and Whitham (1955) and Richards (1956). Assume that Q is concave in k, and that the system is in a steady state with balanced inflows and outflows. The function Q is called the *fundamental diagram* (FD) in traffic lingo. Off-ramps have significant capacity to discharge the desired outflows. Define too the density of opportunities at x, p(x), with p(x)dx = p(x, x+dx). The desired outflow in (x, x+dx) is then Q(k(x), x)p(x)dx, and the total desired outflow in an interval (y, y') the definite integral

$$g(y, y') = \int Q(k(x), x)p(x)dx, \tag{3a}$$

evaluated over the interval. If the interval in question is long compared with the separation between ramps, this will also be the actual outflow. If integral (3a) is taken for the whole length of the road we obtain the total outflow.

We now ask: Given an accumulation, n, what is the distribution of cars that maximizes total outflow? The answer is the function k(x) that maximizes (3a) (evaluated over the whole road), subject to the accumulation constraint (also evaluated over the whole road):

$$n = \int k(x)dx. \tag{3b}$$

Going through Lagrange multipliers we find that the solution of this optimization problem, k*(x), must satisfy:

$$p(x)w(k(x), x) = \text{constant} . \qquad (4)$$

where w(k, x) is the partial derivative of Q(k, x) with respect to k. This is the kinematic "wave-speed". The constant on the right side of (4) is the value for which k*(x) satisfies (3b); the constant can be positive or negative, depending on n.

If the constant is positive, then w must be positive, indicating that the road is uncongested for all x; if the constant is negative w is negative and the road is congested everywhere. If the constant is zero the road is at capacity everywhere. This leads to the following insight:

> ROBUST PROPERTY 1: *If conditions are not changing rapidly with time a road should not have both, congested and uncongested portions.*

We say that the property is robust because it holds independently of the distribution of opportunities and the O/D table.

We also see from (4) that locations with the largest p (the most popular) should have the smallest "w" in absolute value; i.e., the flows closest to capacity. In the limit of a triangular Q (a good approximation for reality) these flows would be capacity flows. Thus, (4) is saying that accumulation should be managed so as to ensure that flow is maximum on the stretches of road that contain the maximum number of desired destinations. We encapsulate this as follows:

> ROBUST PROPERTY 2: *If conditions do not change rapidly with time, system output is maximized when flow is at capacity only along road stretches with the greatest density of destinations.*

Metering strategies should aim for the goals of Properties 1 and 2, which suggest practical strategies for queue storage in congested roads. Essentially, we want to have pockets of congestion where exit rates are low, and maximize flow where exit rates are high. The idea is simple, but current traffic management schemes do not make it explicit. We believe that strategies explicitly built on these principles should be very practical because they can be monitored from readily observable data: accumulations and outflows. We now examine whether they can be extrapolated to complex networks.

4.4 Practical impacts for cities.

The ideas of Sec. 4.3 are quite generic because we never used in our derivation the geometry of the road; we simply parameterized position by "x". We just assumed that each opportunity was associated with a unique point on the road. But as long as this continues to be roughly true for a new (more complex) geometry our conclusions should continue to apply. They should apply to very large networks; e.g., corresponding to a whole city. Thus, we state:

> NETWORK PROPERTY 3: *If conditions do not change rapidly with time, the rate at which trips are served in a metropolitan area is maximized when capacity flows are observed in the neighborhoods with the greatest density of destinations, and elsewhere the network is either congested or uncongested.*

This suggests that in a congested city—as in an asymmetric ring—we should strive for having pockets of congestion where exit rates are low and high flows where exit rates are high. Therefore, there could be some merit in treating a city as a system of large, neighborhood-sized, interconnected reservoirs containing origins, destinations and travelers, and then controlling the flows across reservoirs so as to approach the ideal of property 3. A two-reservoir approach to improve mobility is currently being used in London and Zurich. London restricts travel into the city center through a time-dependent pricing mechanism; and Zurich through a state-dependent metering mechanism informed by the real-time speeds of its bus fleet.

Perhaps we could improve on the experience of these cities by monitoring and controlling the performance of more reservoirs, which could be better adapted to city structure. This generalization could be particularly fruitful if models existed to create managemnt schemes for parking, signal timing, bus flow and pricing in a multi-reservoir context. The development of such models should receive some priority. Unfortunately, this may not be easy if it turns out that management measures change accumulations in complicated ways. We conjecture that this is not always the case, and that some dynamic models can succeed at the aggregate reservoir level (at least for some measures) if the number of reservoirs is small.

To gain some insight into the level of aggregation we can get away with, we now quantify the error introduced when a complex network, perhaps representing a whole neighborhood, is reduced to a single reservoir. If the errors are small for large neighborhoods, then models with few reservoirs may succeed. If this is not the case, a reservoir-based control method may still improve on the status-quo—since a system with few reservoirs could be monitored easily in real time and would have few degrees of freedom.

Let L be the total length of the network, $L = \sum_i l_i$, and P(x) denote the least upper bound to the number of destinations found in all portions of the network with total length x ($x \leq L$). By construction, this function is increasing, concave and satisfies: $P(x) \geq N(x)$, and P(L) = N(L). Note that the minimum number of destinations across all network portions with length x ($x \leq L$) is P(L-x), and that $P(x) \geq P(L-x)$. The difference $\varepsilon(x) = P(x)-P(L-x)$, and its maximum across x, ε, measure uniformity. If the density of destinations is uniform, then P(x) = N(x) = N(L)(x/L), and $\varepsilon = 0$. The following insight pertains to homogeneous networks with uniformly distributed destinations.

> AGGREGATION INSIGHT 4: *If a network is homogeneous and the density of destinations along its links does not change much over space (i.e., ε is small) then the total outflow is roughly given by the number of vehicles on the network, independently of where they are.*

It is in fact possible to show that for triangular FD's, conditional on the accumulation, n, the actual outflow must be in the interval $[Q(n/L) \pm \varepsilon]\ pN(L)$ for any distribution of density. Recall that p is the probability that a trip chooses a specific single opportunity. For non-triangular FDs the distribution of traffic should matter even less. Thus, equation

(1) with $G(n) := Q(n/L)pN(L)$ is a good approximation in our case if ε is small. Note that the aggregate FD (G) is geometrically similar to the FD (Q).

A version of Insight 4 is also true for an important class of inhomogeneous networks. We say that a network is "self-similar" if $Q(k, x)$ can be expressed as the product of the number of lanes $\beta(x)$ and a homogeneous FD with the normalized density per lane as its argument; i.e., if: $Q(k, x) = \beta(x)Q_o(k/\beta(x))$. Then we have the following:

> AGGREGATION INSIGHT 5: *If a network is self-similar and the density of destinations normalized by the number of lanes is space-independent, then* (1) *holds approximately with* $G(n) := (pN(L)/L)\int Q(n/L,x)dx$..

Now, the relation between G and Q is no longer one of similarity, but exit rates can still be predicted independent of the density distribution. It is not a great leap of faith to assume that inhomogeneous road networks should exhibit a fundamental aggregate relation G(n) when uniformly congested.

These aggregation insights should also apply to time-dependent networks (e.g., controlled by traffic signals), if conditions are monitored on a coarse scale of observation where the time-dependence is averaged out (e.g., every several minutes). On such a scale, a single link and signal have a reproducible steady-state accumulation-flow relationship. This relationship can be shown to be of the form $g_i = \beta_i Q(n_i/l_i\beta_i)$ and to depend on the length of the link, the number of lanes and the timing of the signal. Our aggregation insights can and should be tested empirically. If accurate, they would provide a strong basis for the reservoir-based modeling approach.

4.5 Discussion

The results of Secs. 3 and 4 suggest that the inner reservoir of a two-reservoir problem (as in Zurich or London) can be managed with the bang-bang approach of Theorem 1, where flows into the inner core are freely permitted when its accumulation is sub-critical and restricted just enough to maintain a steady critical accumulation when/if the critical level is reached. By monitoring accumulation closely over time, the times when metering starts and ends--and the metering-rate itself--could be set more precisely. This could refine the results currently achieved by these cities.

The two-reservoir approach essentially prevents queues from forming in the inner core by holding them outside. According to Network Property 3 it should be most effective if the density of destinations is much higher in the inner core than outside. But, the approach can be improved if destinations are not so neatly clustered.

If destinations are highly distributed over a metropolitan area, one should partition it into quasi-homogeneous neighborhoods with similar destination densities--treating them as storage "cells" in an aggregate storage network--and still apply Network Property 3. The idea now is to control traffic most closely in the neighborhoods with the highest density of destinations—trying to maintain "optimal" accumulations there as long as possible, without exceeding the critical levels, by controlling the transfer of flows from abutting neighborhoods. One has now more degrees of freedom than in the two-reservoir problem--since one has a choice of locations for storing queues--but the decision process should be manageable if the number of neighborhoods is small. Decision support tools should be developed to deal with aggregate reservoir networks. Since such networks should be quite simple and observable, and since the decisions to be made are quite basic, we are hopeful that the support tools envisioned will avoid the route choice and O-D conundrums of traditional models, even in large-scale application contexts. Preliminary simulations with just a few reservoirs suggest that mobility can be enhanced quite significantly with this approach. The appendix presents eight mobility-enhancing ideas with varying degrees of connection to the macroscopic modeling and management paradigm just discussed.

ACKNOWLEDGEMENT

The research was supported by Berkeley's International Center of Excellence on Future Urban Transport supported by the Volvo Foundations.

REFERENCES

Cervero, R. (2004) Private communication. Professor, DCRP, U. California, Berkeley.

Daganzo, C.F. (1993), "A finite difference approximation for the kinematic wave model", Institute of Transportation Studies, Research Report, UCB-ITS-RR-93-11, Univ. of California, Berkeley, CA; abridged in Trans. Res. 29B(4) 261-276.

Daganzo, C.F. (1995), "The nature of freeway gridlock and how to prevent it", Institute of Transportation Studies, Research Report UCB-ITS-RR-95-1, Univ. of California, Berkeley, CA; and in Transportation and Traffic Theory, pp. 629-646, J.B. Lesort, editor, Pergamon-Elsevier, New York, N.Y.)

Daganzo, C.F. (1998), "Queue spillovers in transportation networks with a route choice," *Transportation Science* 32, 3-11.

Daganzo, C.F., Laval, J.A. and Muñoz, J.C. (2002) "Some ideas for freeway congestion mitigation with advanced technologies" *Traffic Engineering and Control* 43, 397-403.

Edie, L.C. (1963) "Discussion of traffic stream measurements and definitions," *Proceedings 2nd Int. Symposium on the Theory of Traffic Flow*, (J. Almond, editor), pp. 139-154, OECD, Paris, France.

Eichler, M. and Daganzo, C.F. (2005) "BLIP lanes: operation, benefits and domain of application. (Draft working paper; Berkeley International center of Excellence for Future Urban Transport.).

Evans, L. Herman, R. and Lam, T.N. (1976) "Gasoline consumption in urban traffic." General Motors Research Laboratories Publication GMR-1949, Warren, Michigan.

Ghobrial, A., Daganzo, C.F. and Kazimi, T. (1982) "Baggage Claim Area Congestion at Airports: An Empirical Model of Mechanized Claim Device Performance," *Transportation Science* 16, 246-260.

Heydecker, B.G. and Addison, J.D. (1996) "An exact expression of dynamic equilibrium" in Transportation and Traffic Theory (J.B. Lesort, editor) pp. 359-384, Pergamon-Elsevier, New York, N.Y.

Kuwahara, M. (2005) Private communication. Professor, Institute of Industrial Science, University of Tokyo, Tokyo, Japan.

Lago, A. (2003) "Spatial models of morning commute consistent with traffic flow," Ph.D. thesis, University of California at Berkeley, Berkeley, CA.

Lighthill, M.J. and G.B. Whitham (1955), "On kinematic waves. I flow movement in long rives. II A theory of traffic flow on long crowded roads," Proc. Roy. Soc., A. 229, 281-345.

Menendez, M. (2005) PhD Thesis proposal (draft). Department of Civil and environmental Engineering, University of California Berkeley.

Muñoz, J.C. (2002) "Driver –shift design for single-hub transit systems under uncertainty" PhD Thesis. Department of Civil and environmental Engineering, University of California Berkeley.

Muñoz, J.C. (2004) Private communication. Professor, Civil Engineering Department, Pontificia Universidad Catolica de Chile, Santiago, Chile.

Newell, G.F. (1993), "A simplified theory of kinematic waves in highway traffic, I general theory, II queuing at freeway bottlenecks, III multi-destination flows", Trans. Res., 27B, 281-313.

Richards, P.I. (1956), "Shockwaves on the highway," Opns. Res., 4, 42-51.

Robuste, F. (2004) Private communication. Universitat Politecnica de Catalunya, Barcelona, Spain.

Viegas, J. and B. Lu (2001) "Widening the scope for bus priority with intermittent bus lanes", *Transportation Planning and Technology,* vol. 24, no. 2, p. 87-110.

APPENDIX: SOME MOBILITY-ENHANCING IDEAS

This appendix discusses eight ideas to improve mobility that should receive some attention in the near future. A common thread is their non-reliance on detailed forecasts. Some require new uses of sensing and communication technology. The ideas are:

1. Gridlock management
2. Smart parking
3. Green logistics
4. Self-organizing bus systems
5. Intelligent special lanes
6. Best transit system for a given city
7. Flexible staffing
8. Land use and congestion pricing

A.1. Gridlock management

Problem: Conventional forecasts are unreliable and so detailed they do not provide useful insights to guide policies for congestion abatement.

Solution: Focus on robust macroscopic policies (a la London/Zurich) using aggregate models based on observable measures (regional accumulation and outflow). Benefits: Immediate feedback, robustness and transparency.

Needed activity: (a) Mathematical theory of city dynamics; (b) measurement/observation procedures; and (c) field work, verification and fine-tuning of the basic ideas.

Rationale: The ideas in sections 1-3 suggest that two key determinants of mobility in cities are the “aggregate vehicular accumulation” and “length-weighted-cumulative flow” for each city-district and time-of-day. These indicators can be measured directly (without modeling) if sufficient sensors are deployed. The indicators turn out to be ideal beacons for policy guidance because they correlate well with measures of interest to the public such as the aggregate number of trips ended and started, total vehicle-hrs, vehicle-kms, emissions and noise. Key to the success of this activity is an ability to collect, process, store, communicate and display relevant data in real-time.

Two-way communication gadgetry has achieved a penetration bordering on 10% of the automobile fleet in Japan. This makes Japanese cities (Nagoya, being the prime example) an ideal laboratory to determine how vehicle probe data can enrich the information that would otherwise only be available from roadside sensors. With our Japanese partners (Kuwahara, 2005), we are designing data-fusion algorithms that combine data from different sources (including probe vehicles) and will test their predictions against reality in settings where "reality" is known. This research will tell us which type of data should be collected and how it should be processed in cities that are not so well instrumented.

A.2 Smart parking

Problem: increased congestion, delay and energy consumption caused by (i) vehicles looking for parking, and (ii) parked vehicles.

Solution: (i) parking meters reserved remotely; and (ii) dynamic allocation of on street parking spaces, regulated according to traffic conditions.

Needed activity: (a) Implementation would require development of appropriate hardware (smart meters; car computers; sensing and communication devices.) and passage of ordinances for enforcement; (b) evaluation of these systems would have to consider, both, user (parking performance) and non-user (congestion reduction) issues; and the distribution of their impacts (positive and negative) across societal segments.

Rationale: In many cities all over the world, vehicles looking for on-street parking are a source of traffic congestion and accidents. Better parking management can therefore improve efficiency, safety and at the same time mitigate environmental impacts. Parking reservation systems could support better pre-trip plans, reducing unnecessary vehicle travel to find parking places. Some parking garages are already doing this, but the real benefit could come from on-street parking. Our partners in Barcelona (Robuste, 2004) and Tokyo (Kuwahara, 2005) have already begun to explore the idea. The number of ordinary meters and those available for reservation (and their rates) could be changed spatially and temporally according to traffic conditions to achieve a desired target of efficiency and equity. Congestion improvements could be quantified and monitored with the methods of research topic 4.1.

A.3 Green logistics

Problem: Freight distribution in cities is inefficient and insufficiently coordinated.

Solution: Coordination strategies can reduce VMT and VHT's. In addition, "green" distribution vehicles can be given priority; e.g., to work in restricted areas at the most desired times.

Needed activity: (a) a survey of current practices; (b) understand groupings of goods that can be consolidated for joint-distribution; (c) develop logistics planning models to predict truck VMT's and VMH's as a function of the location of their depots; (d) integrate the LPM predictions with gridlock models to better choose a green-logistics structure.

Rationale: This is similar to the rationale of idea 5.2. Excess delivery truck miles (like excess miles by parking hunters) contribute to VMT and increase congestion. Therefore, a rationalization of this process has a double benefit. The benefit of proposed solutions for specific cities should be assessed from a life-cycle perspective with existing methods of logistics systems analysis, and traffic modeling. We plan to build on the preliminary work of our partner in Barcelona (Robuste, 2004).

A.4. Self-organizing bus systems

Problem: Conventional approach to eliminate bus pairing does not work with high frequency systems.

Solution: Distributed control. "Car-following" concepts can be used to design algorithms in which buses would "talk" to their neighbors, and possibly to traffic signals.

Needed activity: (a) Concept development of robust policies for complex networks; (b) mathematical modeling; (c) simulation and testing.

Rationale: Recent work on supply chains and car-following algorithms shows that these systems can be stabilized with distributed control even when under the influence of random exogenous disturbances. Work on distributed communication networks suggests that a centralized level of performance (avoiding clock-drift) can be achieved without a central communication node. The merging of these two ideas opens the door for control approaches to transit systems that are scalable and reliable. The transit system of Santiago (Chile) could be an excellent test-bed for this idea because it is extremely complex, large and has been proven difficult to control. Our Chilean partners are interested in pursuing this line of research (Muñoz, 2004).

A.5 Intelligent special lanes

Problem: Lanes allocated to special vehicles are often underused or mismanaged.

Solution: Match time-space lane designations (active/inactive) with desired usage. Current communication, sensing and control technology make this a possibility. Two application contexts seem possible: (a) HOV lanes; (b) bus lanes.

Needed activity: (a) survey current practices and ideas; (b) develop new control concepts; (c) understand the physics, evaluate with models and classify proper application contexts; (d) demonstrate with field tests.

Rationale: HOV lanes that pass through a bottleneck without carrying saturation flows create unnecessary congestion. Strategies to increase bottleneck flow without penalizing HOV users should be used. Currently, pricing schemes are in vogue, but dynamic control strategies that turn them on and off for a small length upstream of the bottleneck can achieve the same effect (Daganzo et al, 2002). HOV lanes can also create new bottlenecks if they induce lane changes (Menendez, 2005). Strategies that would minimize lane changes should be studied. When bus flows are low, dedicated bus lanes are inefficient. Intermittent bus lanes have also been proposed to alleviate this problem (Viegas and Lu, 2001); different ways to flush traffic out of the bus lane ahead of a bus arrival (with minimal disruption to other traffic) should also be explored. Eichler and Daganzo (2005) evaluate these systems.

The last three ideas are in the planning realm. They do not require special technologies but are also forecast-free.

A.6. Best transit system for a given city

Question: How should a city with given critical statistics such as size (population and area); motorization; number of trips; trip length; trip orientation; and peaking be served by transit. What level of service should one expect for a given level of investment?

Answer approach: Describe all transit systems as a single multi-parameter family in terms of as few relevant variables as possible. Develop a formula to predict transit performance for a city with any set of critical variables. Use formula to predict what is possible. Modeling tools adapted from the field of logistics systems analysis can be used to develop "best designs" for bus and metro systems.

A.7. Flexible staffing

Problem: Transit systems serve two peaks separated by the length of a workday. Staffing is difficult if workers work a workday.

Solution: J.C. Munoz showed in his PhD thesis (Muñoz, 2002) that a menu of work schedules solves this problem. We are refining and applying these ideas in Santiago (Chile).

A.8. Land use and congestion pricing

Problem: Commuting distances are long because people want to live separated from their neighbors. But the price of a lot does not include its true cost to society.

Solution: A. Lago showed in his thesis (Lago, 2003) the differential advantage that people living close to the city center of a mono-centric city enjoyed during the morning commute. Their advantage depended on location and density (sprawl), and could be quantified. Lago’s model allows us to evaluate residential tax policies. This should be explored systematically.

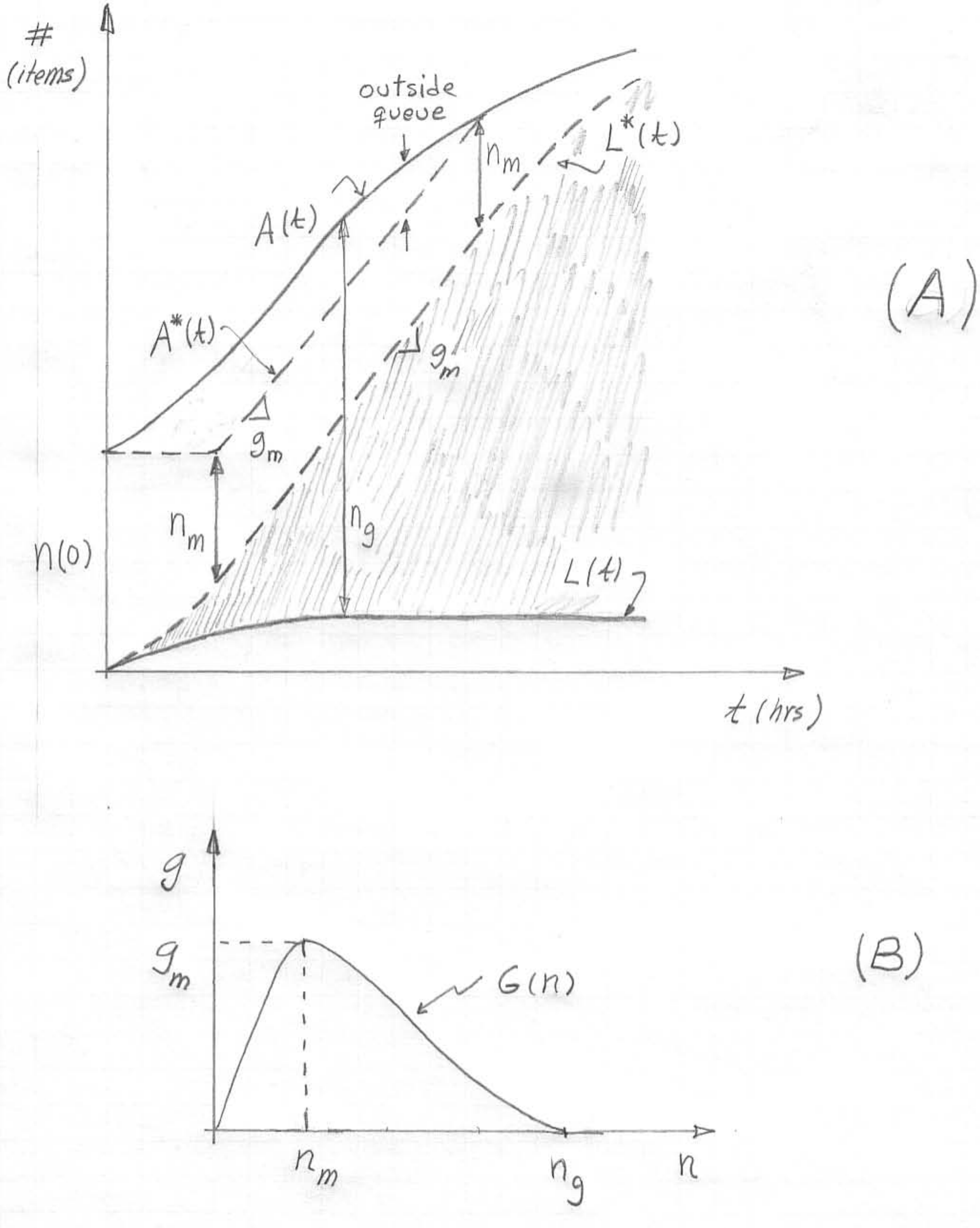

FIGURE 1. GRIDLOCK DEVELOPMENT AND CONTROL: (A) QUEUING DIAGRAM; (B) EXIT FLOW VS. ACCUMULATION DIAGRAM