Causes and Cures of Highway Congestion

Chao Chen, Zhanfeng Jia and Pravin Varaiya
{chaos,jia,varaiya}@eecs.berkeley.edu
Electrical Engineering & Computer Science
University of California, Berkeley

August 18, 2001

Introduction

People believe congestion occurs because demand exceeds capacity, so they support initiatives to build additional highway capacity or curtail highway travel demand. Politicians work to bring highway construction projects into their districts; environmentalists support proposals to make transit more attractive or automobile use more costly. This article argues that the facts do not support the belief that congestion occurs because demand exceeds capacity.

On the contrary, the major cause of congestion is the inefficient operation of highways during periods of high demand. Analysis of data shows that congestion reduces highway efficiency by 20 to 50 %; that is, vehicles take between 20 and 50 % more time to traverse sections that are congested than they would if congestion were prevented. Compensation of this efficiency loss through, say, a 20% capacity expansion is financially impossible; compensation through a 20% demand curtailment is practically impossible. The best way to combat congestion is through increases in operational efficiency. To increase efficiency, however, it is necessary to intelligently control access to highways through ramp metering.

We estimate that for Los Angeles, the annual congestion delay is 70 million vehicle-hours. If the highways were to be operated at 100 % efficiency, this delay would be reduced by 50 million vehicle-hours.

The paper is organized as follows. We first show that vehicles travel at 60 mph when there is maximum flow on a highway section (i.e., when the section is operating most efficiently). Thus congestion delay should be measured as the extra time vehicles spend on the highway traveling below 60 mph: a vehicle taking 20 minutes to travel 10 miles at 30 mph suffers a congestion delay of 10 minutes.

The efficiency η of any highway section may then be defined as

$$\eta = \frac{VMT/60}{VHT} \ ,$$

where *VMT* is the total number of vehicle-miles traveled and *VHT* is the total number of vehicle-hours traveled over the section over some time interval, such as the morning commute period. We will see that η overestimates efficiency; we use this definition anyway because it is easy to calculate from data.

We then present evidence to support the following model of traffic: if the occupancy on a highway section is kept below a certain critical level, its efficiency will be 100 %, congestion will not occur, and traffic will flow at 60 mph; attempts to increase occupancy above the critical level will cause congestion and a rapid drop in efficiency.

This model leads to an idealized ramp metering control policy (IMP), which holds vehicles back at the on-ramps so that the occupancy on each highway section is maintained at its critical level. The total travel time under IMP, VHT_{imp}, is the sum of highway travel time (at 100 % efficiency, 60 mph) and the delay at the ramps imposed by IMP:

$$VHT_{imp} = \frac{VMT}{60} + Delay_{ramp} \ .$$

Therefore, the travel time savings from IMP is

$$\begin{aligned} VHT_{saved} &= VHT - VHT_{imp} \\ &= VHT - \frac{VMT}{60} - Delay_{ramp}. \end{aligned}$$

As $VHT - VMT/60$ is, by definition, the congestion delay, this gives

$$Congestion\ delay = VHT_{saved} + Delay_{ramp}.$$

Observe that $Delay_{ramp}$ may be attributed to excess demand (i.e., demand that exceeds the maximum flow supported by the highway operating at 100 % efficiency). For Los Angeles, our estimates are $Congestion\ delay = 70$ million, $VHT_{saved} = 50$ million, and $Delay_{ramp} = 20$ million vehicle-hours per year.

In contrast to the belief that attributes *all* congestion delay to demand exceeding capacity, we find that the congestion delay consists of a (large) part that can be eliminated by IMP and a residual that can be reduced only by shifting demand during peak periods. Demand may be shifted to other modes, such as public transit, or over time to nonpeak periods.

The penultimate section compares the problems of highway congestion and strategies to relieve it by ramp metering with similar problems and proposed solutions in communication networks and power systems.

What Is Congestion?

Measures of congestion delay compare the actual travel time to some standard. There are two defensible standards: one is travel time under free flow conditions (nominally 60 mph), and the other is travel time under maximum flow. Drivers understand the first standard, transportation professionals approve the second. We analyze data to show that the two standards coincide at least for Los Angeles highways, where the maximum flow in most highway sections occurs near 60 mph.

California's Department of Transportation divides the state into 12 districts. The largest district, Los Angeles, comprises Los Angeles and Ventura counties. We obtained Los Angeles data from the PeMS (Performance Measurement System) database [1],[2].

PeMS receives real-time data from several districts. The data are produced by loop detectors buried in the pavement in each lane of the highway and spaced one-third to one-half mile apart. Every 30 s, the detectors report two numbers: flow and occupancy. *Flow* (often called count) is the number of vehicles that crossed the detector in the previous 30 s. We report flows in vehicles per hour, or *VPH*. *Occupancy* is the fraction of the previous 30 s that a vehicle was present over the detector.

A useful identity relates the three fundamental quantities of highway traffic:

$$Occupancy = \frac{Flow \times VehicleLength}{Speed} ,$$

where *VehicleLength* is the vehicle length (in miles) and *Speed* is the speed in mph. When occupancy exceeds a critical value, congestion sets in and speed drops, as is shown later. The critical occupancy varies with the section.

There are 4,199 detectors at 1,324 locations in Los Angeles highways. PeMS processes data from these detectors in real time and calculates 5-min averages of speed (mph) and flow (*VPH*). We analyze these averages for a 12-hr period beginning midnight of September 1, 2000, and bracketing the morning commute period. We limit the study to the 3,363 functioning detectors.

Detectors are located in all lanes. A section is a portion of a highway associated with a set of detectors (one per lane) and may contain one on- or off-ramp, as depicted in Fig. 1. We attribute a detector's data to the section in which it is located. So, for example, a recorded flow of 1,000 *VPH* for a half-mile section leads to a calculation of 500 *VMT* over that section during 1 hr.

For each detector, we find the 5-min interval in which it reported the maximum flow over the 12-hr study period. We then calculate the average speed reported by this detector over a 25-min interval surrounding this 5-min interval of maximum flow. That is, if the detector reported maximum flow in interval t, we calculate the average speed over the intervals $t-2, t-1, t, t+1, t+2$. This 25-min average is, therefore, a *sustained* speed. (The speed

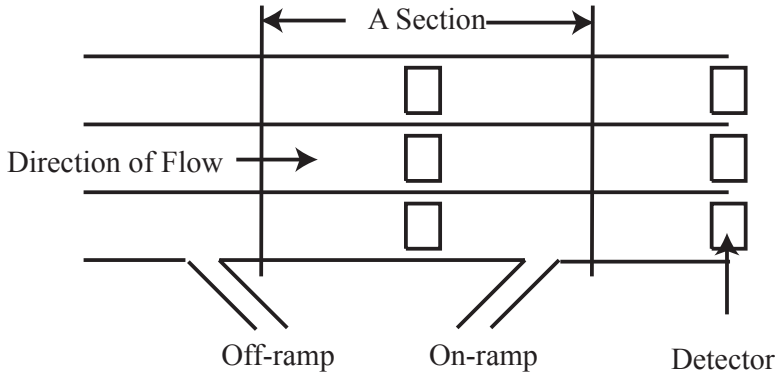

Figure 1: A section is a portion of highway at a detector location and may contain one on- or off-ramp.

at t is usually larger.) Fig. 2 gives the distribution of these speeds. The figure warrants the conclusion that the sustained speed at the time of maximum flow is 60 mph.

Fig. 3, which disaggregates by lane the data in Fig. 2, reinforces the conclusion, since the speed at maximum flow ranges from 65 mph in lane 1 (the innermost, fast lane) to 55 mph in lane 4 (the outermost, slow lane). Traffic on car-pool or HoV lanes is not included in the study.

Efficiency

We view a highway section as capital equipment that takes vehicle-hours traveled, *VHT*, as input and produces vehicle-miles traveled, *VMT*, as output. This is analogous to any other capital equipment that consumes certain variable inputs, such as labor, to produce some good or service.

According to this view, the output "produced" in one hour by a section of highway of length *SectionLength* miles is

$$VMT = Flow \times SectionLength.$$

The corresponding input is

$$VHT = \frac{VMT}{Speed}.$$

The ratio of output to input, *VMT*/*VHT*, is a measure of the *productivity* of this section (during this hour). Its unit is mph.

The maximum value of output produced is

$$MaxVMT = MaxFlow \times SectionLength,$$

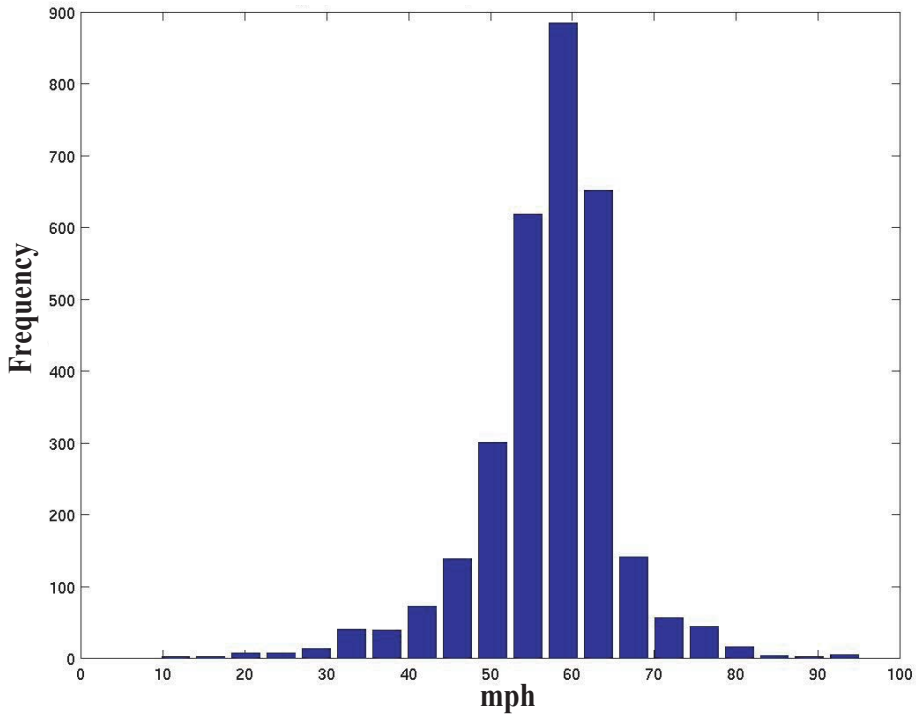

Figure 2: Distribution of average detector speed over a 25-min interval surrounding the time when the detector records maximum flow.

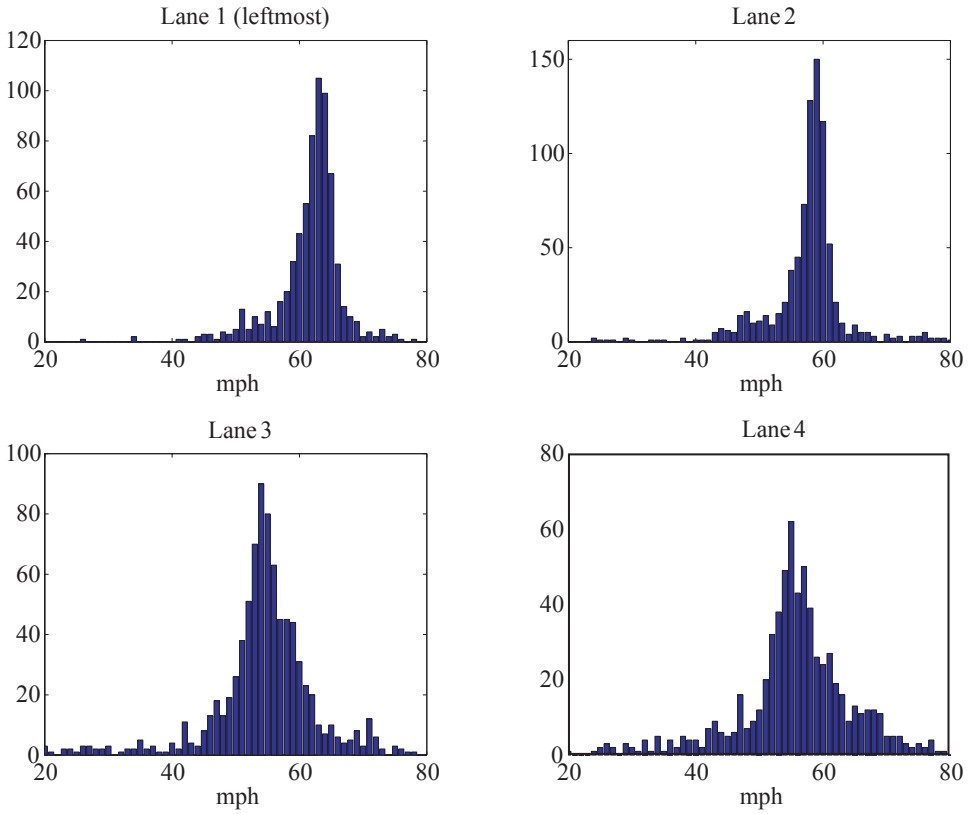

Figure 3: Distribution by lane of average detector speed over a 25-min interval surrounding the time when the detector records maximum flow.

where *MaxFlow* is the maximum flow that is observed on that section during the study period; the speed then is 60 mph. Thus, the maximum productivity is 60 mph. We define the *efficiency* index of a highway section as the ratio of actual to maximum productivity,

$$\eta = \frac{\mathit{VMT}/\mathit{VHT}}{60} \ . \tag{1}$$

This formula permits the following interpretation. Suppose the observed input over a section is $\mathit{VHT} = 10,000$ vehicle-hours and $\eta = 0.8$ or 80 %. Then for the *same* trips, the travel time would be reduced by 2,000 vehicle-hours were the section to operate at 100 % efficiency.

The formula also serves to calculate the efficiency not just for one section and one hour but for a highway network and any duration:

$$\eta_{\mathit{network}} = \frac{\sum_t \sum_k \mathit{VMT}_k(t) / \sum_t \sum_k \mathit{VHT}_k(t)}{60} = \frac{\mathit{VMT}_{\mathit{network}}/\mathit{VHT}_{\mathit{network}}}{60} \ , \tag{2}$$

where k ranges over all sections in the network and t ranges over the appropriate 5-min intervals; $\mathit{VMT}_{\mathit{network}}$, $\mathit{VHT}_{\mathit{network}}$ are simply the sum of the *VMT* and *VHT* over all the sections and time intervals.

We contrast our approach with the standard practice in traffic engineering. The observed *MaxFlow* depends upon physical characteristics, such as the section's grade and curvature, how it is connected to other sections, the location of on- and off-ramps, etc. It also depends on the pattern of traffic and how well the highway is operated. In the standard approach, the section is modeled in isolation to derive a theoretical maximum flow or *capacity* (which should be larger than the observed *MaxFlow*). The canonical model for deriving capacity is given in [3]. To prevent confusion with this notion of capacity, we do not use this term, and retain the empirically defined maximum throughput, *MaxFlow*. See [4] for a more detailed treatment of the distinction.

We estimate the congestion delay in Los Angeles using (2). PeMS provides 5-min averages of *VMT* and *VHT* for each section. For each day during the week of October 3-9, 2000, and the period midnight to noon, we calculate *VMT* and *VHT* for the network consisting of highways I-5, I-10, US 101, I-110, and I-405. For example, the calculation of *VMT* for I-5 for October 4 is

$$\mathit{VMT} = \sum_k \sum_t \mathit{VMT}_k(t),$$

where k ranges over all sections k in I-5 North and I-5 South and t ranges over all 5-min intervals on October 4 from midnight to noon. The results for the network are displayed in Table 1.

Table 1 deserves some comment. *VMT* is simply the sum of *VMT* over the five highways and the seven days in the week. *VHT* is obtained similarly. The % efficiency, η, is calculated using (2). The congestion delay, *VHT* - *VMT*/60, is the additional vehicle-hours spent

Table 1: Congestion delay and potential savings on five highways in Los Angeles during the week of October 3-9, 2000, midnight to noon.

Vehicle-miles traveled, *VMT*	= 86 million
Vehicle-hours traveled, *VHT*	= 1.85 million
Average efficiency, η	= 77 %
Vehicle-hours of congestion delay	= 404,000
Congestion delay saved by IMP	= 280,000
Delay due to excess demand	= 124,000

driving under 60 mph. The congestion delay saved by IMP is the potential reduction in congestion delay under the IMP ramp-metering policy, described later. Thus, the congestion delay is reduced by 280/404, or 70 %. The remaining delay is due to excess demand: it is the vehicle-hours spent behind ramps under IMP.

The period midnight to noon includes not only the morning congestion period but also periods when there is no congestion. The week of October 3-9 includes the weekend, when there is no congestion. The *VHT*, *VMT* include both highway directions, only one of which is congested during the morning commute. When there is no congestion, traffic is moving at 60 mph and efficiency is 100 %. The estimate η of 77 % average efficiency includes these non-congested periods and directions, so if we were to limit attention to the morning commute hours, the efficiency estimate would drop significantly. Fig. 4 shows the daily variation in congestion.

Evening traffic is more congested than morning traffic, so the congestion delay over the entire week is at least 808,000 vehicle-hours. For a 50-week year this amounts to 40 million vehicle-hours. The remaining highway network in Los Angeles is 75 % longer, so we estimate the annual congestion delay in all Los Angeles highways to be 70 million vehicle-hours. Assuming that 70 % of this delay can be saved by IMP, this amounts to 50 million vehicle-hours each year. Valuing the opportunity cost of time at $20 per vehicle-hour, this gives an annual savings of $1 billion.

From a more inclusive perspective, the efficiency index (1) is an underestimate, since it only accounts for changes in speed and not in flow. As a hypothetical example, consider a section with a maximum flow of 2,000 VPH at 60 mph, but which during congestion has a flow of 1,800 VPH at 30 mph. The efficiency according to (1) is 30/60 = 0.5, reflecting the drop in speed, but it does not reflect the 10 % reduction in flow. A better measure of the potential efficiency appears to be

$$\hat{\eta} = \frac{\textit{Flow} \times \textit{Speed}}{\textit{MaxFlow} \times SpeedAtMaxFlow}. \tag{3}$$

For the hypothetical example, $\hat{\eta} = 0.45$ instead of 0.5. (The product *Flow* $\times$ $Speed$ was proposed as a measure of performance in [5].)

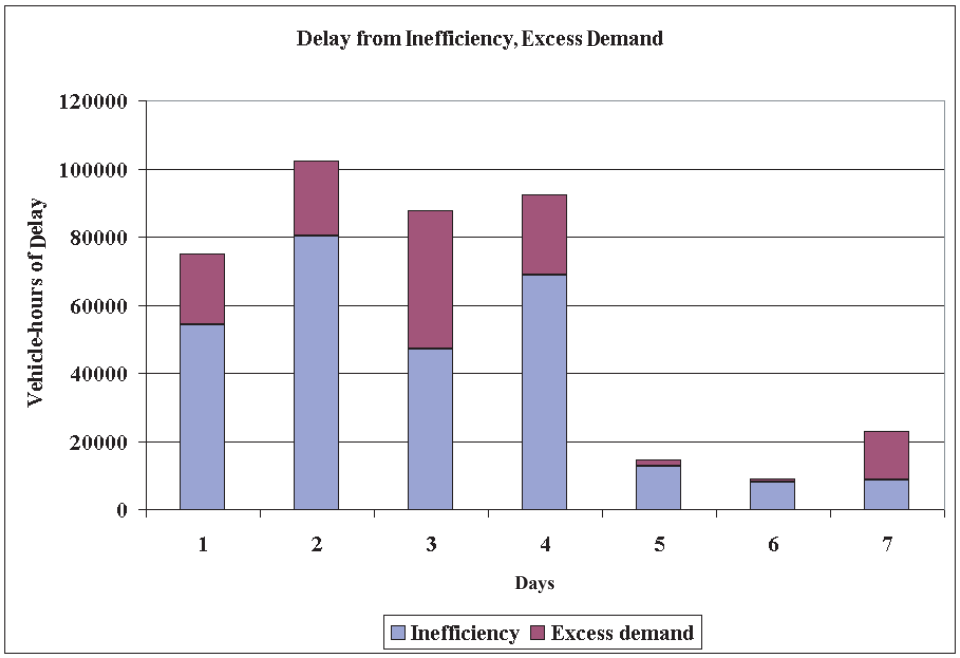

Figure 4: Variation in congestion during week of October 3-9, midnight to noon. Day 1, October 3, is Tuesday; Day 7, October 9, is Monday.

Viewing the highway section as a queuing system sheds light on the formula (3). The section provides a service to its customers (vehicles): the transport of a vehicle across the section. The service time of a vehicle is

$$\frac{SectionLength}{Speed}.$$

The system serves *Flow* vehicles in parallel, so its throughput is

$$\frac{Speed}{SectionLength} \times Flow.$$

The maximum throughput is

$$\frac{60}{SectionLength} \times MaxFlow.$$

$\hat{\eta}$ is the ratio of actual to maximum throughput.

We estimate $\hat{\eta}$ for all sections of I-10W during the morning congestion period on October 1, 2000, as follows. For each section we determine the 5-min interval between midnight and noon when its detector recorded the maximum occupancy. This is the time of worst congestion, and we find the speed and flow at that time. The efficiency during congestion for this section is

$$\hat{\eta} = \frac{FlowAtMaxOcc \times SpeedAtMaxOcc}{MaxFlow \times 60}.$$

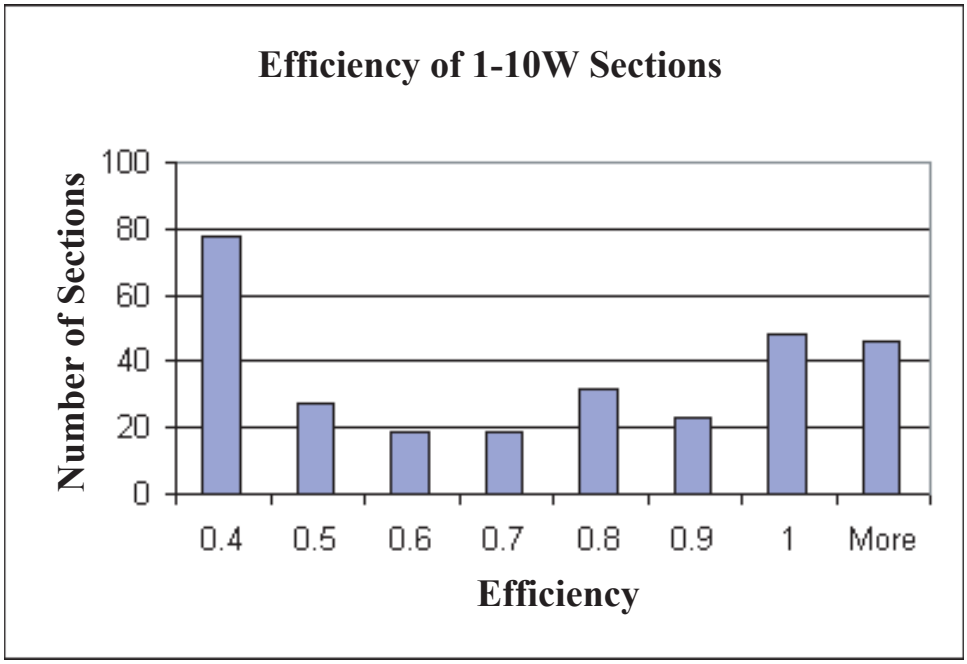

Figure 5: Variation in efficiency $\hat{\eta}$ during congestion along sections of I-10W, midnight to noon, October 1, 2000.

The distribution of $\hat{\eta}$ across 291 sections with functioning detectors on I-10W is shown in Fig. 5; 78 sections had an efficiency less that 40 %, 65 had an efficiency between 40 and 80 %, 71 had an efficiency between 80 and 100 %, and 46 had an efficiency above 100 % (recording speeds above 60 mph).

Behavior During Congestion

Fig. 6 is a plot of *Flow* vs *Occupancy* on one section of I-10W from midnight to noon on October 3, 2000. Each point is a 5-minute average of flow and occupancy. Successive points are connected by straight lines. Initially, vehicles travel at 60 mph and flow and occupancy increase in proportion. At 5:30 am, occupancy reaches a critical level, and flow reaches its maximum, 2400 VPH. Demand exceeds this maximum, congestion sets in, speed and flow decline while occupancy increases. At the depth of congestion, speed is 20 mph and flow has dropped to 1400 VPH. Demand then drops, and speed gradually recovers to 60 mph by 9:00 am. For this section, the critical occupancy level is 0.11.

This behavior suggests the model of congestion depicted in Fig. 7. Notice the three regimes in the "phase" portrait of the figure: free flow, then congestion, followed by recovery. The recovery phase is different from the congestion phase, reminiscent of hysteresis. Standard hydrodynamic models of fluid flow don't exhibit such hysterisis. It is a challenge to invent

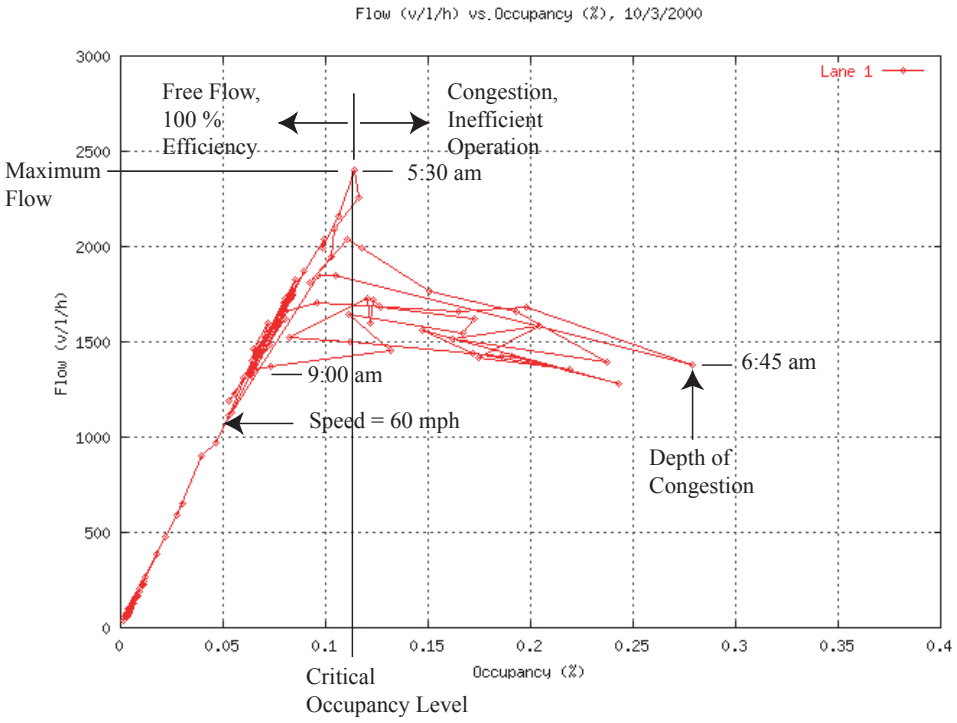

Figure 6: Flow vs. occupancy on a section at postmile 37.18 on I-10W, midnight to noon on October 3, 2000.

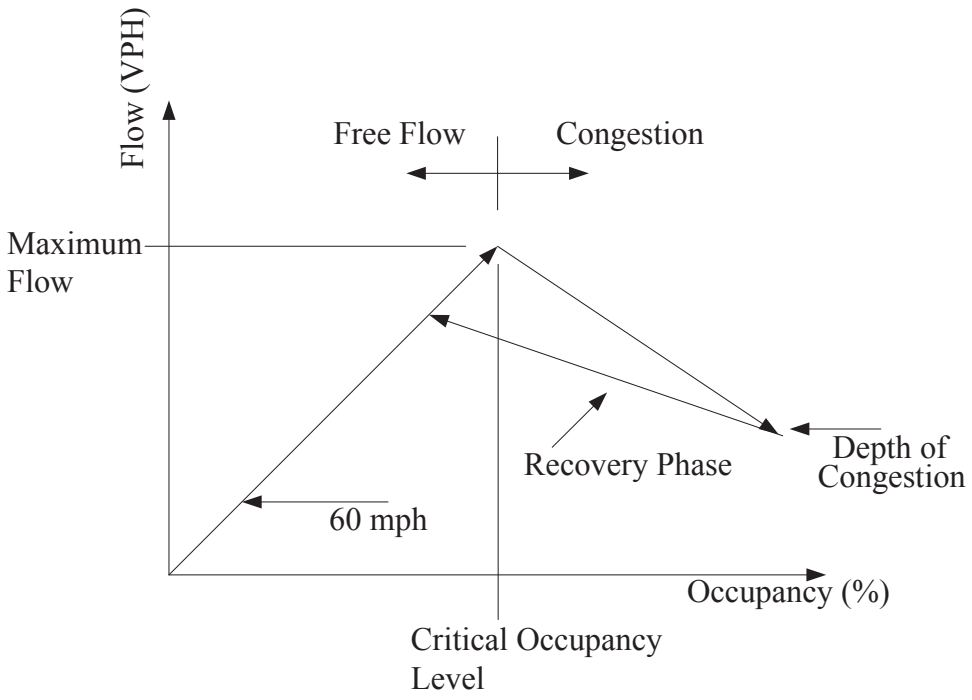

Figure 7: Model of congestion. If occupancy is maintained below critical level, section operates at 100 % efficiency and speed is at 60 mph.

a dynamic model that exhibits such transitions. Measurements of the recovery phase show erratic fluctuations as in Fig. 6. The model in turn supports the following hypothesis about traffic behavior:

> *If a metering policy keeps occupancy below its critical level in every section, efficiency will be 100 %, speed will be maintained at 60 mph, and highway congestion will be prevented. A consequence of the metering is that vehicles will be stopped at the ramps for some time.*

We call this the *ideal ramp metering (IMP) principle*. The IMP feedback strategy is to monitor the occupancy downstream of each on-ramp and to throttle the flow from the on-ramp whenever the occupancy exceeds its critical value. For a control-theoretic discussion of this policy, see [6].

The figures in Table 1 are computed as follows. For each highway section we calculate the critical occupancy level from PeMS data, as in Fig. 6. We assume that the pattern of demand is unchanged. We now simulate the traffic flow using the model of Fig. 7 and the IMP feedback strategy. In the simulation, vehicles will be held back at some ramps. We calculate the total time spent by the vehicles at the ramps. The ramp delay is 124,000 vehicle-hours. There is a net savings of 280,000 vehicle-hours.

Other Networks

As with highway networks, the movements of data and electric power are also organized in networks with high-capacity transmission links to take advantage of scale economies. Since they accommodate the demand from uncontrolled or unpredictable users, however, all three networks can experience congestion.

We explore similarities and differences in the congestion these three types of networks experience along the dimensions of demand, routing, and control.

The demand patterns in transportation and data networks are similar: users want to move commodities (vehicles, data packets) from some nodes to others. In power networks, users impose loads at some nodes that are supplied by power from other nodes. A power system is a single commodity network; the others are multicommodity networks.

Until recently, power system planners, like transportation planners, concerned themselves with finding effective ways to expand generation and transmission capacity to meet the forecast demand, assumed to be exogenous. Today there is a realization that curtailing demand for power is essential. One way to curtail demand is through real-time congestion pricing [7]-[9]. This is similar to the suggestion of transportation economists. There is also a growing literature exploring the use of pricing to shape demand in data networks [10]-[13]. Congestion pricing is not used in practice in data networks: it is used to an increasing extent in bulk power markets and to a very limited extent in highway transportation. It is practiced by airlines under "yield management." (We are ignoring the significant differences between spot prices for bulk power, time-of-day tolls in highways, and pricing of airplane seats for market segmentation. Only one component of these prices concerns congestion.)

In transportation networks, demand can be directly controlled by ramp metering. The counterpart in data networks is called admission control: at the entrance to the network certain flows or connections may not be admitted into the network. (Admission control is rare in data networks, but it is common in telephony: if the telephone network cannot route your call, it is blocked. Admission control can be based on the origin, destination, and other attributes of the data traffic. Ramp metering usually cannot distinguish between vehicles by their destination.) Power systems, too, employ admission control during emergencies by shutting down voluntary "interruptible" loads or by forced "rotating blackouts."

Routing of data packets is fully controlled by the routers located at the nodes. Power flows along routes determined by the laws of physics, given the patterns of sources (generation) and sinks (loads). Thus, the power flow routes cannot be directly controlled. (A limited amount of control can be exercised using expensive FACTS devices.) Similar to power, the routes chosen by drivers cannot be controlled. They prefer routes with shorter travel times, but the travel time on a congested highway section depends on the traffic flow, which in turn depends (through driver choice) on the travel time. Thus, routes and travel times are

jointly determined in a simultaneous equation system, similar to the load flow equations that determine power flows and phase angles.

The differences in demand and routing, and the controls that can be exercised, affect the nature of congestion in the three networks.

A transmission link in a power network is congested if the power flowing through it is close to its thermal capacity. Additional power through the link carries the risk of a line fault, endangering the transfer of power in other links. Since power flows cannot be controlled, the *only* way to reduce the flow through an overloaded link is to change the pattern of power generation and consumption.

In data networks, transmission links do *not* get congested. They transmit at a fixed line rate. (An exception is congestion due to contention for access in shared Ethernet and wireless links.) Instead, congestion occurs at a node or router when the rate at which data to be forwarded over a particular outgoing link exceeds that link's line rate. The router's buffer then overflows and the router is forced to drop the packet.

Conclusions

For many years, the increase in travel demand has outstripped additions to California's highway infrastructure. Congestion is worse each year. Rising housing costs in high-employment regions force people to live further away, lengthening commutes, and increasing congestion. The resulting low-density housing makes current transit options (rail and buses) costly and less effective.

We have argued that a large portion of highway congestion can be attributed to inefficient operation. The inefficiency is greatest when demand is greatest. Empirical analysis indicates potentially large gains in efficiency, with dramatic reductions in congestion. Intelligent ramp metering control strategies can realize these gains.

One reason these strategies are not implemented is the widely held belief that congestion is determined by demand, and ramp metering merely transfers delay that would occur on the highway to delay at the ramps. But our analysis concludes that intelligent ramp metering transfers only a *fraction* of the highway delay to the ramps; the rest of the delay is eliminated. Of course, further empirical studies that test this conclusion are needed.

Transportation economists have long recognized that congestion is a "negative externality" and proposed congestion tolls to limit highway access during periods of high demand [14], [15]. But equity and engineering considerations suggest that in most places, ramp metering is easier to deploy than congestion tolls.

Transportation, power, and data networks face congestion. Congestion in data networks has to date been contained by expanding capacity ahead of demand growth. Until recently,

that was also the strategy followed by transportation and power network operators; but that option today is frequently not available. The only option is to put in place efficiency-enhancing control strategies. But that poses challenges of control strategy design and the development and deployment of sensors and controller technologies to implement these strategies. Those challenges are just beginning to be addressed.

Acknowledgments

The paper builds on work of the PeMS Develpoment Group. Markos Papageorgiou and Alex Skabardonis helped us with their comments and criticisms. The PeMS project is supported by the State of California Department of Transportation, the EPRI/DoD Complex Interactive Networks Initiative under Contract WO8333-04, and the National Science Foundation under Grant CMS-0085739. The authors alone are responsible for the opinions expressed here and for any errors.

References

[1] transacct.eecs.berkeley.edu

[2] C. Chen, K. Petty, A. Skabardonis, P. Varaiya and Z. Jia, "Freeway performance measurement system: mining loop detector data," 80th Annual meeting of the Transportation Research Board, Washington, D.C., January 2001.

[3] Transportation Research Board. *Highway Capapcity Manual 2000*, Chapter 23. Washington, D.C., National Research Council, 1998.

[4] Z. Jia, P. Varaiya, K. Petty, and A. Skabardonis, "Congestion, excess demand, and effective capacity in California freeways," submitted to *Transportation Research*, December 2000.

[5] D.R. Drew and C.J. Keese, "Freeway level of service as influenced by volume and capacity characteristics," *Highway Research Record*, no. 99, pp. 1-47, 1965.

[6] M. Papageorgiou and A. Kotsialos, "Freeway ramp metering: an overview," *Proc. IEEE Intelligent Transportation Systems Conference*, Dearborn, MI, October 2000,

[7] C-W. Tan and P. Varaiya, "A model for pricing interruptible electric power service," in G.B. DiMasi, A. Gombani, and A.B. Kurzhanski (Eds.), *Modelling, Estimation and Control of Systems with Uncertainty*, pp. 423-444. Cambridge, MA: Birkhauser Boston, 1991.

[8] H-P. Chao, G. Huntington (Eds.), *Designing Competitive Electricity Markets*. Boston, MA: Kluwer, 1998.

[9] F.F. Wu and P. Varaiya, "Coordinated multilateral trades for electric power markets: theory and implmentation," *Electric Power and Energy Systems*, vol. 21, pp. 75-102, 1999.

[10] J. Walrand and P. Varaiya, *High-performance communication networks*, 2nd ed. San Francisco, CA: Morgan Kaufmann, 2000.

[11] R. Edell and P. Varaiya, "Providing Internet access: What we learn from INDEX," *IEEE Network*, vol. 13, no. 4, pp. 18-25, 1999.

[12] J.K. Mackie-Mason and H.R. Varian, "Pricing congestible network resources," *IEEE J. on Selected Areas in Communications*, vol. 13, no. 7, pp. 1141-9, 1995.

[13] S.H. Low and D.E. Lapsley, "Optimization flow control, I: basic algorithm and convergence," *ACM/IEEE Transactions on Networking*, vol. 7, no. 6, pp. 861-75, 1999.

[14] A. Waters, "Theory and measurement of private and social cost of highway congestion," *Econometrica*, vol. 29, pp. 676-99, 1961.

[15] T.E. Keeler, *The full costs of urban transport*, University of California, Inst. Urban and Regional Development, 1975.